

## The searching for agents for Alzheimer's disease treatment via the system of self-consistent models

Andrey A. Toropov, Alla P. Toropova, P. Ganga Raju Achary, Maria Raškova & Ivan Raška

To cite this article: Andrey A. Toropov, Alla P. Toropova, P. Ganga Raju Achary, Maria Raškova & Ivan Raška (2022): The searching for agents for Alzheimer's disease treatment via the system of self-consistent models, Toxicology Mechanisms and Methods, DOI: [10.1080/15376516.2022.2053918](https://doi.org/10.1080/15376516.2022.2053918)

To link to this article: <https://doi.org/10.1080/15376516.2022.2053918>



View supplementary material [↗](#)



Published online: 07 Apr 2022.



Submit your article to this journal [↗](#)



Article views: 9



View related articles [↗](#)






View Crossmark data [↗](#)

RESEARCH ARTICLE



# The searching for agents for Alzheimer's disease treatment via the system of self-consistent models

Andrey A. Toropov<sup>a</sup> , Alla P. Toropova<sup>a</sup> , P. Ganga Raju Achary<sup>b</sup> , Maria Raškova<sup>c</sup> and Ivan Raška<sup>c</sup>

<sup>a</sup>Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano, Italy; <sup>b</sup>Department of Chemistry, Institute of Technical Education and Research (ITER), Siksha 'O'Anusandhan University, Bhubaneswar, India; <sup>c</sup>Medical Department, Faculty of Medicine, Charles University in Prague, Prague, Czech Republic

## ABSTRACT

Robust quantitative structure-activity relationships (QSARs) for hBACE-1 inhibitors (pIC<sub>50</sub>) for a large database ( $n = 1706$ ) are established. New statistical criteria of the predictive potential of models are suggested and tested. These criteria are the index of ideality of correlation ( $IIC$ ) and the correlation intensity index ( $CII$ ). The system of self-consistent models is a new approach to validate the predictive potential of QSAR-models. The statistical quality of models obtained using the CORAL software (<http://www.insilico.eu/coral>) for the validation sets is characterized by the average determination coefficient  $R^2_v = 0.923$ , and  $RMSE = 0.345$ . Three new promising molecular structures which can become inhibitors hBACE-1 are suggested.

## ARTICLE HISTORY

Received 11 January 2022  
Revised 19 February 2022  
Accepted 11 March 2022

## KEYWORDS

QSAR; hBACE-1; Alzheimer's disease; CORAL software; index of ideality of correlation; correlation intensity index

## Introduction

Alzheimer's disease is a chronic neurodegenerative disease affecting more than 30 million people worldwide. The development of small-molecule inhibitors of human  $\beta$ -secretase1 (hBACE-1) which are potential therapeutic agents for the treatment of Alzheimer's disease is the well-known task of theoretical chemistry and medicinal chemistry (Subramanian and Poda 2018).

hBACE-1 is the target for the prevention and treatment of Alzheimer's disease (Ribaud et al. 2019). However, the main cause of Alzheimer's is not known yet but some abnormalities in the brain are the hallmark of this disease. These include the accumulation of insoluble deposits of beta amyloid plaques outside the neurons and intracellular neurofibrillary tangles (Dhanjal et al. 2014). During the last decade, numerous research groups have synthesized inhibitors of  $\beta$ -amyloid cleaving enzyme-1 (BACE-1) in the hope of developing a therapy to halt or even reverse the progression of Alzheimer's disease. The amyloid hypothesis has long been the central dogma in drug discovery for Alzheimer's disease (Prati et al. 2018). Many big pharma companies were expending great resources in the search for BACE-1 inhibitors (Hunt and Turner 2009). The inhibition of  $\beta$ -secretase activity in vivo remains one of the most attractive strategies for the treatment of Alzheimer's disease, although progress in getting inhibitors to the clinic as rule is slow (Zhang et al. 2014). Despite experimental and clinical efforts, there is a high unmet medical need for the effective treatment of Alzheimer's disease, one of the most prevalent and

debilitating neurological diseases (Neumann et al. 2018). The inhibitors of BACE-1 can be also used for traumatic brain injuries as a therapeutical agent (Blasko et al. 2004). It should be noted, besides beta-secretase human BACE-1 as potential agents for the treatment of Alzheimer's disease, also, inhibitors of human acetylcholinesterase (hAChE) and human butyrylcholinesterase (hBuChE) are studied (Panek et al. 2018; Tripathi et al. 2019; Choubey et al. 2021) as well as other endpoints (Toropov et al. 2018). Computation models for the above biochemical phenomena are suitable tools to reduce the space of potential effective substances.

Quantitative structure-property/activity relationships (QSPRs/QSARs) are the source of models in general (Bhargava et al. 2017; Ničković et al. 2020; Kumar and Kumar 2021; Toropova and Toropov 2022) and for hBACE-1 inhibitors activity, in particular (Cruz and Castilho 2014; Panek et al. 2018; Tripathi et al. 2019; Choubey et al. 2021).

There are various QSPR/QSAR approaches (Berhanu et al. 2012) such as linear regression (Ghasemi et al. 2007), partial least squares, neural network, support vector machine (Cao et al. 2010), k-nearest neighbors (Stoyanova-Slavova et al. 2014), and Monte Carlo technique (Toropova and Toropov 2014, 2022; Ahmadi and Akbari 2018; Toropov and Toropova 2021). The Monte Carlo approach was applied to the development of therapeutic agents for the treatment of Alzheimer's disease (Toropova et al. 2015; 2018; Kumar et al. 2021).

The aim of the present study is the assessment of the Monte Carlo technique as a tool to build up models and the

check the predictive potential of these models using three the following recently suggested innovations (i) so-called system of self-consistent models (Toropov and Toropova 2021); (ii) the index of ideality of correlation (Kumar et al. 2019; Toropova and Toropov 2019; Toropov and Toropova 2019; Ahmadi 2020; Bagri et al. 2020; Javidfar and Ahmadi 2020; Nimbhal et al. 2020; Ghiasi et al. 2021; Kumar and Kumar 2021); and (iii) the correlation intensity index (Toropov, Sizochenko et al. 2020; Toropov, Toropova et al. 2020; Toropova and Toropov 2020; Toropov and Toropova 2020; Jafari et al. 2022).

## Method

### Data

A collection of experimental  $pIC_{50}$  values for 1706 compounds was taken from the literature (Subramanian and Poda 2018). These compounds were randomly split 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 1 shows that these splits are far to be identical.

Each above-mentioned set has a special task: (i) Active training set is a creator of a model, i.e., compounds of this set are used to build up a predictive model; (ii) Passive training set is an inspector of the model, i.e., compounds of this set are used to check up whether the model is satisfactory for substances that are absent in the active training set; (iii) Task of the calibration set is to detect the start of the over-training; (iv) The validation set is used for the final validation of the predictive potential of the model.

### Optimal SMILES-based descriptor

The optimal SMILES-based descriptor  $DCW(T, N)$  is applied for a predictive model of the endpoint using the following regression equation:

$$pIC_{50} = C_0 + C_1 \times DCW(T, N) \quad (1)$$

$$DCW(T, N) = \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (2)$$

where  $T$  is an integer to separate SMILES attributes into rare and non-rare, i.e. the non-rare SMILES are applied to build up the model. The rare SMILES are not applied to build up the model;  $N$  is the number of epochs of the optimization of the correlation weights;  $S_k$  is a SMILES atom, i.e., one symbol of SMILES line (e.g., '=', 'O') or a group of symbols which cannot be examined separately (e.g., 'Cu', '%11');  $SS_k$  and  $SSS_k$

are fragments of SMILES including two and three SMILES-atoms, respectively; and  $CW(S_k)$ ,  $CW(SS_k)$ , and  $CW(SSS_k)$  are the correlation weights of the above SMILES attributes, i.e. the numerical values of correlation weights are calculated using the Monte Carlo method.

### The Monte Carlo optimization

Equation (2) needs the numerical data on the above correlation weights. The Monte Carlo optimization is a tool to calculate those correlation weights. Here three target functions for the Monte Carlo optimization are examined:

$$TF_1 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (3)$$

$$TF_2 = TF_1 + IIC_C \times 0.5 \quad (4)$$

$$TF_3 = TF_2 + CII_C \times 0.5 \quad (5)$$

where  $r_{AT}$  and  $r_{PT}$  are correlation coefficients between observed and predicted endpoint for the active training and passive training sets, respectively.

The  $IIC_C$  is the index of ideality of correlation (Toropova and Toropov 2019). The  $IIC_C$  is calculated with data on the calibration set as follows:

$$IIC_C = r_C \frac{\min(-MAE_C, +MAE_C)}{\max(-MAE_C, +MAE_C)} \quad (6)$$

$$\min(x, y) = \begin{cases} x, & \text{if } x < y \\ y, & \text{otherwise} \end{cases} \quad (7)$$

$$\max(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases} \quad (8)$$

$$-MAE_C = \frac{1}{-N} \sum |\Delta_k|, \quad -N \text{ is the number of } \Delta_k < 0 \quad (9)$$

$$+MAE_C = \frac{1}{+N} \sum |\Delta_k|, \quad +N \text{ is the number of } \Delta_k \geq 0 \quad (10)$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k \quad (11)$$

The observed and calculated are corresponding values of the endpoint.

The correlation intensity index ( $CII$ ) (Toropov and Toropova 2020) similarly to the above  $IIC$ , was developed as a tool to improve the quality of the Monte Carlo optimization aimed to build up QSPR/QSAR models.

The  $CII$  is calculated as follows (Toropov and Toropova 2020):

$$CII_C = 1 - \sum Protest_k \quad (12)$$

$$Protest_k = \begin{cases} R_k^2 - R^2, & \text{if } R_k^2 - R^2 > 0 \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where  $R^2$  is the correlation coefficient for a set that contains  $n$  substances and  $R_k^2$  is the correlation coefficient for  $n-1$  substances of a set, after removing of  $k$ th substance. Hence, if the  $(R_k^2 - R^2)$  is larger than zero, the  $k$ th substance is an "opponentist" for the correlation between experimental and predicted values of the set. A small sum of "protests" means a more "intensive" correlation.

Table 1. Matrix of identity (%) for five random splits examined here.

	split1	split2	split3	split4	split5
split1	100	41.8	44.2	44.4	43.5
split2	46.4	100	42.5	44.8	41.5
split3	43.8	44.3	100	39.9	40.9
split4	46.2	45.3	45.5	100	42.8
split5	48.7	44.8	43.1	45.7	100

Matrix Element[i,j],  $i > j$  the identity for the active training sets.  
Matrix Element[i,j],  $i < j$  the identity for the validation sets.

### The system of self-consistent models

Each  $i$ th model has a  $i$ th validation set. As it is demonstrated (Table 1), the validation sets are far from to be identical. It is a question of whether the arbitrary model can be used for an arbitrary validation set? Naturally, compounds from the validation set which are involved in creating the model should be then excluded from the corresponding assessment of the predictive potential. If the answer is 'yes', these different models should be considered as self-consistent ones.

The measure of self-consistency (Toropova and Toropov 2021) is the average and dispersion of the correlation coefficient on different validation sets. The corresponding computational experiments are represented by this matrix:

$$MV = \begin{bmatrix} (M_1 : V_1 \rightarrow Rv_{11}^2) & \cdots & (M_5 : V_1 \rightarrow Rv_{51}^2) \\ \vdots & & \vdots \\ (M_1 : V_5 \rightarrow Rv_{15}^2) & \cdots & (M_5 : V_5 \rightarrow Rv_{55}^2) \end{bmatrix} \quad (14)$$

where  $M_i$  is an  $i$ th model, the  $V_j$  is the list of compounds applied as the validation set in the case of  $j$ th split, and the  $Rv_{ij}^2$  is the correlation coefficient observed for the  $j$ th validation set if applied  $i$ th model.

## Results and discussion

### Histories of the monte carlo optimization with different target functions

Figure 1 shows the graphical representations of the Monte Carlo optimization with  $TF_1$ ,  $TF_2$ , and  $TF_3$  for split 1. It is to be noted that the histories of the other splits are the same as indicated in Figure 1. One can see, the  $TF_1$  gives the overtraining starting from the third epoch of the Monte Carlo optimization. The  $TF_2$  and  $TF_3$  do not give the overtraining,

but the statistical quality of the model for the validation set in the case of  $TF_3$  is better than in the case of  $TF_2$ . The analysis of Figure 1 indicates that the preferable optimal descriptors for  $TF_1$  are  $DCW(1,2)$ , but for the cases of  $TF_2$  and  $TF_3$  descriptor  $DCW(1,15)$  is most adequate.

### The comparison of the Monte Carlo optimizations based on $TF_1$ , $TF_2$ , and $TF_3$

Table 2 shows the statistical characteristics of the models obtained for the five different splits into the active training, passive training, calibration, and validation sets. The statistical quality of obtained models indicates the advantage of models obtained with  $TF_3$  (Equation 5).

### The weirdness of correlations on the training sets

Models obtained by the Monte Carlo optimization with target function  $TF_3$  demonstrate the pair of correlations on the active and passive training sets. The common determination coefficient is poor, but separated correlations on two clusters are quite good. Figure 2 shows a graphical representation of the situation for split #1.

### Domain of applicability

In general, the domain of applicability should be defined before building up a model. However, even if a model is built up for a congeneric set of compounds (e.g., alkanes, benzene derivatives, etc.) some compounds will be characterized by a poor prediction for an endpoint. In the case of CORAL software, the statistical defect of SMILES is used as an indicator of compounds for which one should expect a poor prediction of the endpoint.

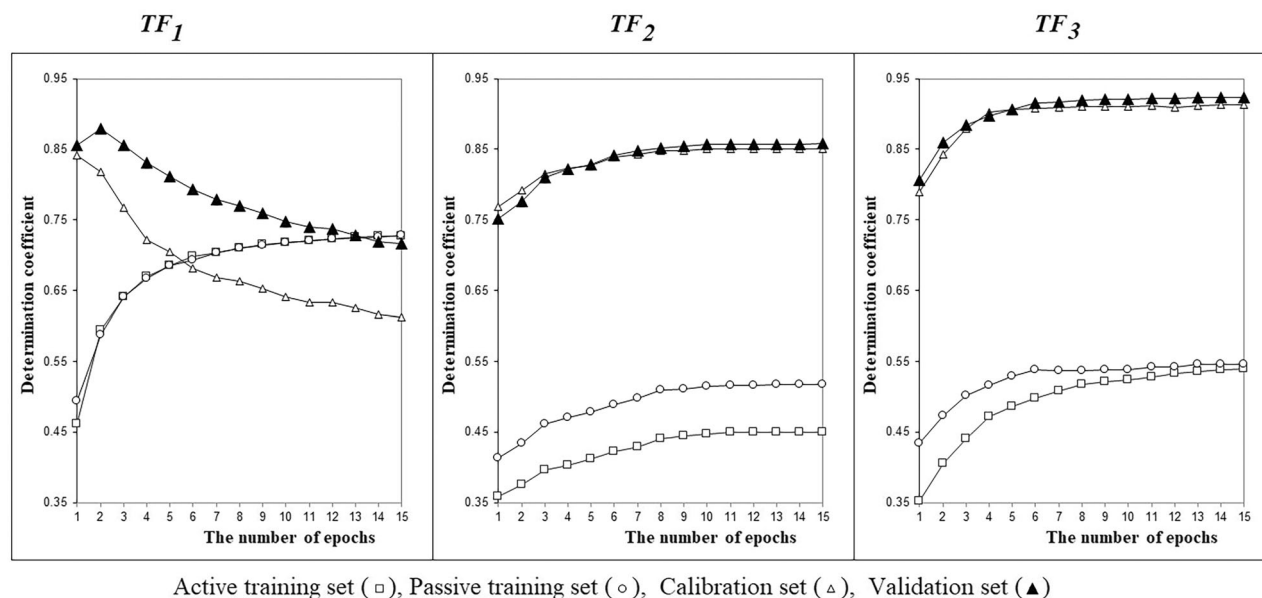


Figure 1. Histories of the Monte Carlo optimization with  $TF_1$ ,  $TF_2$ , and  $TF_3$  in the case of the split 1.

Table 2. Statistical characteristics of models for five random splits were obtained using different target functions.

Target function	Split	Set*	n	$R^2$	CCC	IIC	CII	$Q^2$	RMSE	MAE
$TF_1$	1	A	423	0.6088	0.7569	0.6867	0.7824	0.6056	0.915	0.768
		P	420	0.6101	0.7705	0.7004	0.7944	0.6067	0.943	0.800
		C	428	0.7962	0.8869	0.7297	0.8757	0.7942	0.590	0.458
		V	435	0.8453	0.9170	0.8990	0.9061	0.8439	0.521	0.403
	2	A	420	0.6010	0.7508	0.7322	0.7908	0.5975	0.933	0.809
		P	438	0.5951	0.7481	0.7358	0.7760	0.5916	0.931	0.774
		C	417	0.8082	0.8954	0.7926	0.8877	0.8064	0.568	0.446
		V	431	0.8461	0.9164	0.7626	0.9078	0.8446	0.541	0.426
	3	A	418	0.6007	0.7505	0.7318	0.7771	0.5971	0.950	0.806
		P	441	0.6089	0.7530	0.7153	0.7862	0.6055	0.950	0.832
		C	429	0.8288	0.9043	0.7769	0.9010	0.8271	0.523	0.414
		V	418	0.8739	0.9323	0.8268	0.9238	0.8726	0.461	0.367
	4	A	410	0.5931	0.7446	0.6917	0.7846	0.5896	0.934	0.812
		P	428	0.6469	0.7646	0.6820	0.8004	0.6440	0.942	0.815
		C	433	0.8101	0.8974	0.8647	0.8789	0.8085	0.541	0.432
		V	435	0.8361	0.9133	0.7930	0.8943	0.8347	0.497	0.391
	5	A	413	0.5918	0.7436	0.7016	0.7930	0.5881	0.968	0.844
		P	439	0.5028	0.6987	0.6718	0.7647	0.4984	1.03	0.899
		C	414	0.8618	0.9261	0.7921	0.9160	0.8605	0.460	0.369
		V	440	0.8821	0.9373	0.8175	0.9283	0.8810	0.457	0.364
$TF_2$	1	A	423	0.5204	0.6845	0.6348	0.7579	0.5161	1.01	0.894
		P	420	0.5255	0.7012	0.6362	0.7797	0.5211	1.04	0.922
		C	428	0.8893	0.9413	0.9430	0.9338	0.8882	0.404	0.327
		V	435	0.9120	0.9544	0.8826	0.9470	0.9112	0.370	0.290
	2	A	420	0.5210	0.6851	0.6316	0.7692	0.5165	1.02	0.910
		P	438	0.5252	0.6857	0.6914	0.7635	0.5209	1.00	0.883
		C	417	0.8766	0.9352	0.9363	0.9287	0.8754	0.429	0.349
		V	431	0.8978	0.9469	0.9147	0.9363	0.8969	0.410	0.327
	3	A	418	0.5163	0.6810	0.6467	0.7656	0.5119	1.05	0.936
		P	441	0.5151	0.6683	0.6335	0.7591	0.5109	1.06	0.965
		C	429	0.8709	0.9323	0.9332	0.9196	0.8698	0.412	0.329
		V	418	0.8904	0.9428	0.9411	0.9297	0.8894	0.404	0.323
	4	A	410	0.5069	0.6728	0.6271	0.7662	0.5023	1.03	0.928
		P	428	0.5655	0.7003	0.6408	0.7738	0.5617	1.04	0.932
		C	433	0.8893	0.9418	0.9430	0.9300	0.8882	0.395	0.314
		V	435	0.8934	0.9439	0.8862	0.9338	0.8924	0.385	0.309
	5	A	413	0.5438	0.7045	0.6857	0.7746	0.5396	1.02	0.904
		P	439	0.4694	0.6716	0.6511	0.7557	0.4648	1.06	0.929
		C	414	0.8682	0.9285	0.9317	0.9202	0.8669	0.439	0.354
		V	440	0.8888	0.9412	0.8600	0.9283	0.8879	0.432	0.342
$TF_3$	1	A	423	0.4660	0.6358	0.6122	0.7534	0.4610	1.07	0.968
		P	420	0.5036	0.6768	0.6171	0.7647	0.4993	1.06	0.948
		C	428	0.9014	0.9463	0.9494	0.9427	0.9004	0.379	0.315
		V	435	0.9184	0.9546	0.9355	0.9493	0.9177	0.359	0.294
	2	A	420	0.4799	0.6485	0.6004	0.7564	0.4750	1.07	0.950
		P	438	0.4827	0.6524	0.6362	0.7511	0.4781	1.05	0.934
		C	417	0.8747	0.9307	0.9352	0.9262	0.8735	0.434	0.354
		V	431	0.8932	0.9431	0.8208	0.9333	0.8923	0.417	0.338
	3	A	418	0.5455	0.7059	0.6711	0.7709	0.5415	1.01	0.900
		P	441	0.5437	0.6965	0.6514	0.7648	0.5398	1.03	0.913
		C	429	0.9121	0.9543	0.9550	0.9450	0.9113	0.344	0.283
		V	418	0.9224	0.9602	0.8599	0.9536	0.9216	0.345	0.277
	4	A	410	0.5034	0.6696	0.6188	0.7676	0.4988	1.03	0.932
		P	428	0.5448	0.6873	0.6116	0.7674	0.5409	1.07	0.956
		C	433	0.9031	0.9489	0.9416	0.9423	0.9022	0.367	0.304
		V	435	0.8888	0.9418	0.8924	0.9320	0.8878	0.391	0.315
	5	A	413	0.5592	0.7173	0.7021	0.7868	0.5551	1.01	0.902
		P	439	0.4601	0.6685	0.6147	0.7583	0.4554	1.08	0.945
		C	414	0.9007	0.9475	0.9490	0.9430	0.8997	0.381	0.312
		V	440	0.9089	0.9519	0.7898	0.9423	0.9081	0.396	0.314

\*A: active training set; P: passive training set; C: calibration set; V: validation set; n: the number of compounds in a set;  $R^2$ : determination coefficient; CCC: concordance correlation coefficient; IIC: index of ideality of correlation; CII: correlation intensity index;  $Q^2$ : cross-validated (the leave-one out)  $R^2$ ; RMSE: root mean square error; and MAE: mean absolute error.

Defects of SMILES attribute  $A_k$  are calculated as follows:

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N'(A_k)} \quad (15)$$

where  $P(A_k)$  and  $P'(A_k)$  are the probability of  $A_k$  in the active training set and passive training set, respectively;  $N(A_k)$  and  $N'(A_k)$  are frequencies of  $A_k$  in the active training set and

passive training sets, respectively. The statistical defects of SMILES ( $D_j$ ) are calculated as follows:

$$D_j = \sum_{k=1}^{NA} d_k \quad (16)$$

where NA is the number of non-blocked SMILES attributes in the SMILES.

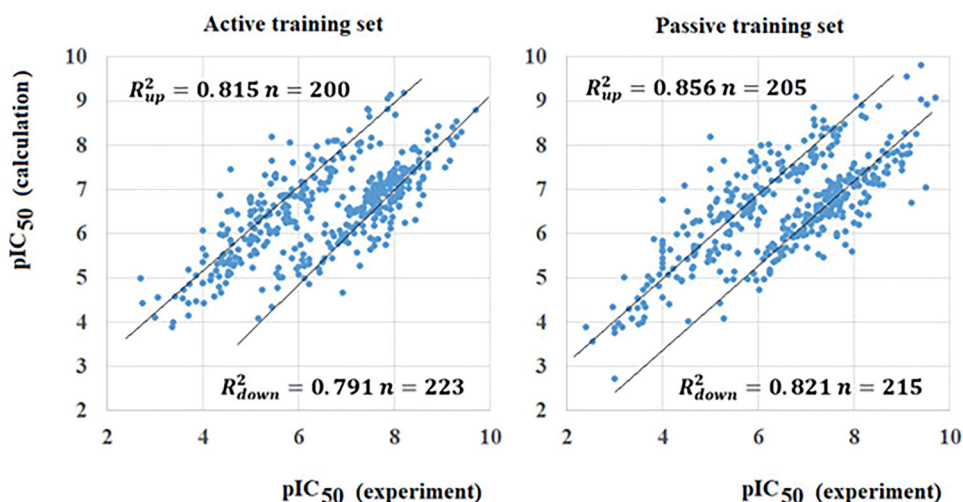


Figure 2. Correlation clusters which obtained for the active training set and the passive training set of the model for split #1 ( $n$  is the number of participants of a correlation cluster).

Table 3. Determination coefficients for the validation sets of models are based on the  $TF_1$ ,  $TF_2$ , and  $TF_3$ .

$TF_1$	Validation set, $V_1$	Validation set, $V_2$	Validation set, $V_3$	Validation set, $V_4$	Validation set, $V_5$
Model, $M_1$		0.8600 (201)	0.8804 (187)	0.8550 (201)	0.8717 (213)
Model, $M_2$	0.8718 (201)		0.8901 (188)	0.8644 (196)	0.8905 (195)
Model, $M_3$	0.9054 (187)	0.9170 (188)		0.9026 (194)	0.8924 (185)
Model, $M_4$	0.8569 (201)	0.8617 (196)	0.8548 (194)		0.8430 (200)
Model, $M_5$	0.8970 (213)	0.9182 (195)	0.8810 (185)	0.8854 (200)	
$TF_2$	Validation set, $V_1$	Validation set, $V_2$	Validation set, $V_3$	Validation set, $V_4$	Validation set, $V_5$
Model, $M_1$		0.9190 (201)	0.9304 (187)	0.9120 (201)	0.9258 (213)
Model, $M_2$	0.9180 (201)		0.9268 (188)	0.9142 (196)	0.9166 (195)
Model, $M_3$	0.9090 (187)	0.9194 (188)		0.9155 (194)	0.8969 (185)
Model, $M_4$	0.9333 (201)	0.9304 (196)	0.9371 (194)		0.9229 (200)
Model, $M_5$	0.8972 (213)	0.9065 (195)	0.8984 (185)	0.8864 (200)	
$TF_3$	Validation set, $V_1$	Validation set, $V_2$	Validation set, $V_3$	Validation set, $V_4$	Validation set, $V_5$
Model, $M_1$		0.9383 (201)	0.9343 (187)	0.9191 (201)	0.9379 (213)
Model, $M_2$	0.9101 (201)		0.9147 (188)	0.8978 (196)	0.9161 (195)
Model, $M_3$	0.9418 (187)	0.9376 (188)		0.9411 (194)	0.9380 (185)
Model, $M_4$	0.9168 (201)	0.9295 (196)	0.9283 (194)		0.9155 (200)
Model, $M_5$	0.9102 (213)	0.9342 (195)	0.9123 (185)	0.9156 (200)	

$\bar{R}_v^2 = 0.880 \pm 0.018$  ( $TF_1$ ).

$\bar{R}_v^2 = 0.916 \pm 0.011$  ( $TF_2$ ).

$\bar{R}_v^2 = 0.923 \pm 0.012$  ( $TF_3$ ).

The quantities of compounds that are not involved in building up corresponding models are indicated in brackets. The diagonal elements for the MV matrix (Equation 14) are available from Table 2 as the corresponding determination coefficient for validation sets.

Table 4. Comparison of the statistical characteristics of the described model obtained for split #1 with models suggested in Subramanian and Poda (2018) for similar distribution into the training sub-set and validation subset.

Method	$n$	Training* set		$n$	Validation set	
		$R^2$	RMSE		$R^2$	RMSE
PLS**	1365	0.629	0.827	341	0.577	0.883
RF	1365	0.904	0.421	341	0.657	0.795
CORAL	1271	0.590	0.894	435	0.918	0.359

\*Training in the case of CORAL: Active training set + Passive training set + Calibration set;  $n$ : the number of compounds in a set;  $R^2$ : correlation coefficient; RMSE: root mean square error; \*\*PLS: partial least squares; RF: random forest.

A SMILES falls in the domain of applicability if

$$D_j < 2\bar{D} \quad (17)$$

where  $\bar{D}$  is the average statistical defect of SMILES.

### The system of self-consistent models

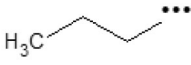
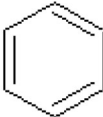
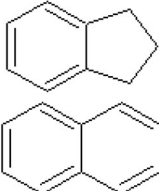
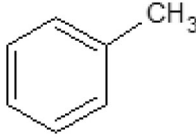
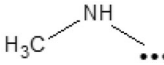
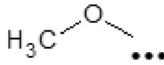
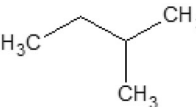
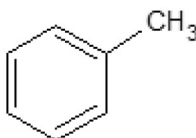

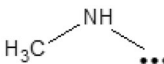

Table 3 shows the correlation coefficient between experimental (Subramanian and Poda 2018) and predicted pIC50 observed for five random splits examined here. One can see that most self-consistency of described models is observed in the case of Monte Carlo optimization with target function  $TF_3$ .

### Comparison with models from the literature

The comparison of the statistical quality of models should be based on analogical distributions into the sub-system of the training and the sub-system of the validation. Table 4 shows the comparison of the statistical characteristics of the model



**Table 5.** Molecular features extracted from SMILES which are promoters of *pIC50* increase.

Molecular feature	Graphical comments	<i>CWs run 1</i>	<i>CWs run 2</i>	<i>CWs run 3</i>	<i>N1*</i>	<i>N2</i>	<i>N3</i>
C ... ..		0.23727	0.05564	0.19789	423	419	427
c ... ..		0.11269	0.01867	0.08093	420	413	425
2 ... ..		0.57707	0.60379	0.24666	414	412	423
c ... ( ... ..		0.12197	0.07471	0.66682	414	411	414
N ... ( ... ..		0.61909	0.05188	0.11007	410	407	413
O ... ( ... ..		0.10734	0.33108	0.03332	386	376	359
( ... C ... ( ...		0.01442	0.00882	0.65119	361	359	353
c ... ( ... C ...		0.37849	0.29414	0.39177	322	317	307
C ... C ... ..		0.39868	0.06869	0.32817	284	273	304
N ... ( ... C ...		0.55295	0.55176	0.86751	279	275	279
C ... ( ... C ...		0.01780	0.29643	0.54863	277	287	275

\**N1*, *N2*, and *N3* are frequencies of a molecular feature in the active training set, the passive training set, and the calibration set, respectively.

obtained for split #1 and the model for the same database described taken in work (Subramanian and Poda 2018).

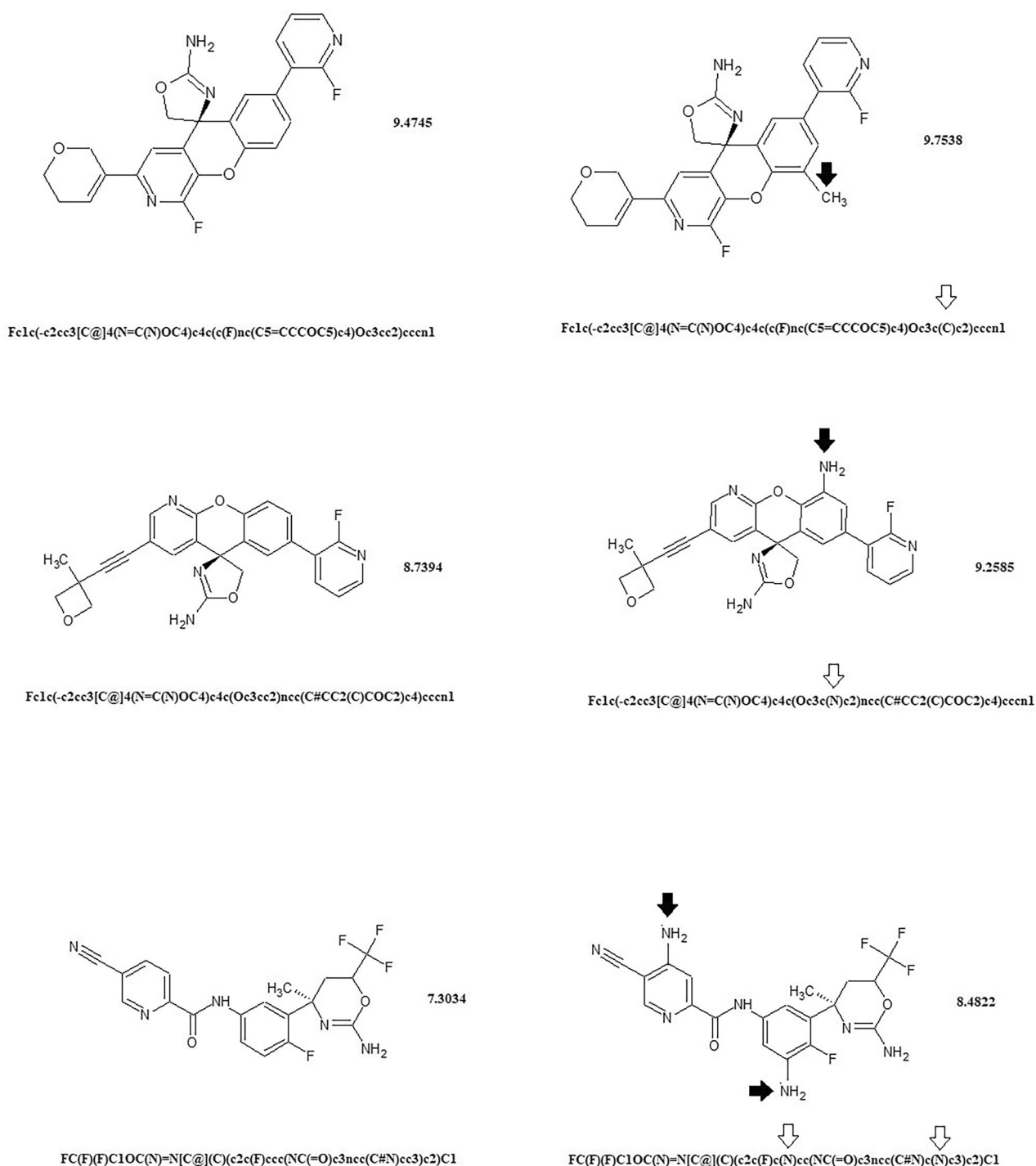
values of correlation weights. Table 5 shows a collection of promoters of *pIC50* increase.

### Mechanistic interpretation

Mechanistic interpretation of the CORAL models is the list of molecular features extracted from SMILES which have in several runs of the Monte Carlo optimization solely either positive (these are the promoters of endpoint increase) or negative (these are the promoters of endpoint decrease)

### The list of perspective hBACE-1 inhibitors

Using the list of promoters of *pIC50* increase it is possible to compare the different versions of molecular architecture in order to extract in the aspect of the biological activity more perspective compounds. Figure 3 shows examples of the perspective hBACE-1 inhibitors.



**Figure 3.** Modifications (for compounds at left), which can increase pIC<sub>50</sub> according to the model. Arrows show modifications that can increase the inhibitor potential according to promoters of increase for pIC<sub>50</sub> (Table 5): black arrows show modifications in the molecular structure, white arrows show modifications in SMILES.

Thus, the approach described here is presented that has its advantages and disadvantages. The statistical quality of the approach is comparable or even better than the statistical quality of the models described in the literature (Table 4). It can be noted that the approach does not involve additional physicochemical data, 3D molecular structure, and descriptors of quantum mechanics.

It is to be noted that the reliability of the system of self-consistent models can be checked up with a larger number of models. Most probably, in this case, owing to the increase

in the number of situations (models) considered the statistical reliability of this analysis will be improved.

The paradoxical effect of using the index of ideality of correlation is that the corresponding models obtain improve statistical quality for the calibration set and the external validation set but to the detriment of the training set (both the active and the passive sub-sets). In addition, the effect is accompanied by the division of dots in coordinates experimental vs. calculated values pIC<sub>50</sub> in two clusters (Figure 2). Consequently, on one hand, improving the factual predictive



potential (attractive statistical quality for the validation set), but, on other hand, a decrease of the statistical quality is observed for the training set. From a practical point of view, the above situation is rather attractive than a poor one.

However, if the index of ideality of correlation has been checked in building up models for different endpoints (Kumar et al. 2019; Toropova and Toropov 2019; Ahmadi 2020; Bagri et al. 2020; Javidfar and Ahmadi 2020; Nimbhal et al. 2020; Ghiasi et al. 2021; Kumar and Kumar 2021), the correlation intensity index gives a practical improvement of models, as the rule, incorporation with the index of ideality of correlation, only (Toropov, Toropova et al. 2020; Toropova and Toropov 2020; Toropov and Toropova 2020; Jafari et al. 2022).

Unfortunately, the distribution (splits) into the active training set, the passive training set, the calibration set, and the validation set can affect the statistical quality of a model: different distributions provide different models with different predictive potential. However, namely, the principle of comparison of several splits provides the possibility to assess the reliability (reproducing) of results.

**Supplementary materials** section contains the technical details of the described computational experiments for split 1. Similar information for all splits is available on request. The division of correlations into layers for active and passive training sets is presented in the section of the **Supplementary material** for all five random splits examined here.

## Conclusions

The described approach gives quite good models for hBACE-1 inhibitor activity ( $IC_{50}$ ). The statistical quality of these models was confirmed by computational experiments with five different distributions into the active training set, passive training set, calibration set, and validation set. The statistical characteristics of models for the validation set are examined as the measure of the predictive potential of the models. The best choice of the target function for the Monte Carlo optimization produced by these models is the use of  $TF_3$  calculated with Equation (5). Despite the stochastic character of described approach, the suggested system of self-consistent models provides the measure for both the predictive potential of the applied approach (selected model) together with the level of the reproducibility of the results.

## Acknowledgments

AAT and APT are grateful to the EC project LIFE-CONCERT contract [LIFE17 GIE/IT/000461] for the financial support.

## Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## ORCID

Andrey A. Toropov  <http://orcid.org/0000-0001-6864-6340>

Alla P. Toropova  <http://orcid.org/0000-0002-4194-9963>

P. Ganga Raju Achary  <http://orcid.org/0000-0002-9631-3356>

## Data availability statement

The data used in this work and developed models are freely available **Supplementary materials** section.

## References

- Ahmadi S. 2020. Mathematical modeling of cytotoxicity of metal oxide nanoparticles using the index of ideality correlation criteria. *Chemosphere*. 242:125192.
- Ahmadi S, Akbari A. 2018. Prediction of the adsorption coefficients of some aromatic compounds on multi-wall carbon nanotubes by the Monte Carlo method. *SAR QSAR Environ Res*. 29(11):895–909.
- Bagri K, Kumar A, Nimbhal M, Kumar P. 2020. Index of ideality of correlation and correlation contradiction index: a confluent perusal on acetylcholinesterase inhibitors. *Mol Simul*. 46(10):777–786.
- Berhanu WM, Pillai GG, Oliferenko AA, Katritzky AR. 2012. Quantitative structure-activity/property relationships: The ubiquitous links between cause and effect. *ChemPlusChem*. 77(7):507–517.
- Bhargava S, Adhikari N, Amin SA, Das K, Gayen S, Jha T. 2017. Hydroxyethylamine derivatives as HIV-1 protease inhibitors: a predictive QSAR modelling study based on Monte Carlo optimization. *SAR QSAR Environ Res*. 28(12):973–990.
- Blasko I, Beer R, Bigl M, Apelt J, Franz G, Rudzki D, Ransmayr G, Kampfl A, Schliebs R. 2004. Experimental traumatic brain injury in rats stimulates the expression, production and activity of Alzheimer's disease  $\beta$ -secretase (BACE-1). *J Neural Transm*. 111(4):523–536.
- Cao D-S, Xu Q-S, Liang Y-Z, Chen X, Li H-D. 2010. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J Chemom*. 24 (9):584–595.
- Choubey PK, Tripathi A, Tripathi MK, Seth A, Shrivastava SK. 2021. Design, synthesis, and evaluation of N-benzylpyrrolidine and 1,3,4-oxadiazole as multitargeted hybrids for the treatment of Alzheimer's disease. *Bioorg Chem*. 111:104922
- Cruz DS, Castilho MS. 2014. 2D QSAR studies on series of human beta-secretase (BACE-1) inhibitors. *Med Chem*. 10(2):162–173.
- Dhanjal JK, Goyal S, Sharma S, Hamid R, Grover A. 2014. Mechanistic insights into mode of action of potent natural antagonists of BACE-1 for checking Alzheimer's plaque pathology. *Biochem Biophys Res Commun*. 443(3):1054–1059.
- Ghasemi J, Saaidpour S, Brown SD. 2007. QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis. *J Mol Struct THEOCHEM*. 805(1-3): 27–32.
- Ghiasi T, Ahmadi S, Ahmadi E, Talei Bavil Olyai MR, Khodadadi Z. 2021. The index of ideality of correlation: QSAR studies of hepatitis C virus NS3/4A protease inhibitors using SMILES descriptors. *SAR QSAR Environ Res*. 32(6):495–520.
- Hunt CE, Turner AJ. 2009. Cell biology, regulation and inhibition of beta-secretase (BACE-1). *Febs J*. 276(7):1845–1859.
- Jafari K, Fatemi MH, Toropova AP, Toropov AA. 2022. The development of nano-QSPR models for viscosity of nanofluids using the index of ideality of correlation and the correlation intensity index. *Chemometr Intell Lab Syst*. 222:104500.
- Javidfar M, Ahmadi S. 2020. QSAR modelling of larvicidal phytocompounds against *Aedes aegypti* using index of ideality of correlation. *SAR QSAR Environ Res*. 31(10):717–739.
- Kumar A, Bagri K, Nimbhal M, Kumar P. 2021. In silico exploration of the fingerprints triggering modulation of glutamyl cyclase inhibition for the treatment of Alzheimer's disease using SMILES based attributes in Monte Carlo optimization. *J Biomol Struct Dyn*. 39(18):7181–7193.
- Kumar P, Kumar A. 2021. Correlation intensity index (CII) as a benchmark of predictive potential: Construction of quantitative structure activity

- relationship models for anti-influenza single-stranded DNA aptamers using Monte Carlo optimization. *J Mol Struct.* 1246:131205.
- Kumar P, Kumar A, Sindhu J. 2019. 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.* 30 (2):63–80.
- Neumann U, Ufer M, Jacobson LH, Rouzade-Dominguez M-L, Huledal G, Kolly C, Löönd RM, Machauer R, Veenstra SJ, Hurth K, et al. 2018. The BACE-1 inhibitor CNP520 for prevention trials in Alzheimer's disease. *EMBO Mol Med.* 10(11):e9316.
- Ničković VP, Mitić NR, Krdžić BD, Krdžić JD, Nikolić GR, Vasić MZ, Ranković G, Babović P, Sokolović D, Veselinović AM. 2020. Design and development of novel therapeutics for brucellosis treatment based on carbonic anhydrase inhibition. *J Biomol Struct Dyn.* 38(6):1848–1857.
- Nimbhal M, Bagri K, Kumar P, Kumar A. 2020. The index of ideality of correlation: A statistical yardstick for better QSAR modeling of glucokinase activators. *Struct Chem.* 31(2):831–839.
- Panek D, Więckowska A, Pasięka A, Godyń J, Jończyk J, Bajda M, Knez D, Gobec S, Malawska B. 2018. Design, synthesis, and biological evaluation of 2-(benzylamino-2-hydroxyalkyl)isoindoline-1,3-diones derivatives as potential disease-modifying multifunctional anti-Alzheimer agents. *Molecules.* 23(2):347.
- Prati F, Bottegoni G, Bolognesi ML, Cavalli A. 2018. BACE-1 Inhibitors: From Recent Single-Target Molecules to Multitarget Compounds for Alzheimer's Disease. *J Med Chem.* 61(3):619–637.
- Ribaudo G, Coghi P, Zanforlin E, Law BYK, Wu YYJ, Han Y, Qiu AC, Qu YQ, Zagotto G, Wong VKW. 2019. Semi-synthetic isoflavones as BACE-1 inhibitors against Alzheimer's disease. *Bioorg Chem.* 87:474–483.
- Stoyanova-Slavova IB, Slavov SH, Pearce B, Buzatu DA, Beger RD, Wilkes JG. 2014. Partial least square and k-nearest neighbor algorithms for improved 3D quantitative spectral data-activity relationship consensus modeling of acute toxicity. *Environ Toxicol Chem.* 33(6):1271–1282.
- Subramanian G, Poda G. 2018. In silico ligand-based modeling of hBACE-1 inhibitors. *Chem Biol Drug Des.* 91(3):817–827.
- Toropov AA, Sizochenko N, Toropova AP, Leszczynska D, Leszczynski J. 2020. Advancement of predictive modeling of zeta potentials ( $\zeta$ ) in metal oxide nanoparticles with correlation intensity index (CII). *J Mol Liq.* 317:113929.
- Toropov AA, Toropova AP. 2019. Use of the index of ideality of correlation to improve predictive potential for biochemical endpoints. *Toxicol Mech Methods.* 29(1):43–52.
- Toropov AA, Toropova AP. 2020. Correlation intensity index: Building up models for mutagenicity of silver nanoparticles. *Sci Total Environ.* 737: 139720.
- Toropov AA, Toropova AP. 2021. The system of self-consistent models for the uptake of nanoparticles in PaCa2 cancer cells. *Nanotoxicology.* 15(7):995–1004.
- Toropov AA, Toropova AP, Benfenati E. 2020. 'Ideal correlations' for the predictive toxicity to *Tetrahymena pyriformis*. *Toxicol Mech Methods.* 30(8):605–610.
- Toropov AA, Toropova AP, Benfenati E, Salmons M. 2018. Mutagenicity, anticancer activity and blood brain barrier: similarity and dissimilarity of molecular alerts. *Toxicol Mech Methods.* 28(5):321–327.
- 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.
- Toropova AP, Toropov AA. 2019. Does the Index of Ideality of Correlation Detect the Better Model Correctly? *Mol Inf.* 38(8-9): 1800157.
- Toropova AP, Toropov AA. 2020. Fullerenes C60 and C70: a model for solubility by applying the correlation intensity index. *Fuller Nanotub Carbon Nanostructures.* 28(11):900–906.
- Toropova AP, Toropov AA. 2021. The System of Self-Consistent of Models: A New Approach to Build Up and Validation of Predictive Models of the Octanol/Water Partition Coefficient for Gold Nanoparticles. *Int J Environ Res.* 15(4):709–722.
- Toropova AP, Toropov AA. 2022. Nanomaterials: quasi-SMILES as a flexible basis for regulation and environmental risk assessment. *Sci Total Environ.* 823:153747.
- Toropova AP, Toropov AA, Begum S, Achary PGR. 2018. Blood brain barrier and Alzheimer's Disease: Similarity and Dissimilarity of Molecular Alerts. *Curr Neuropharmacol.* 16(6):769–785.
- Toropova MA, Toropov AA, Raška I, Rašková M. 2015. Searching therapeutic agents for treatment of Alzheimer disease using the Monte Carlo method. *Comput Biol Med.* 64:148–154.
- Tripathi A, Choubey PK, Sharma P, Seth A, Tripathi PN, Tripathi MK, Prajapati SK, Krishnamurthy S, Shrivastava SK. 2019. Design and development of molecular hybrids of 2-pyridylpiperazine and 5-phenyl-1,3,4-oxadiazoles as potential multifunctional agents to treat Alzheimer's disease. *Eur J Med Chem.* 183:111707.
- Zhang L, Shen C, Chu J, Zhang R, Li Y, Li L. 2014. Icaritin decreases the expression of APP and BACE-1 and reduces the  $\beta$ -amyloid burden in an APP transgenic mouse model of Alzheimer's disease. *Int J Biol Sci.* 10(2):181–191.