Journal of Molecular Modeling (2023) 29:218
https://doi.org/10.1007/s00894-023-05632-2

ORIGINAL PAPER

The validation of predictive potential via the system of self-consistent models: the simulation of blood–brain barrier permeation of organic compounds

Alla P. Toropova[1] · Andrey A. Toropov[1] · Alessandra Roncaglioni[1] · Emilio Benfenati[1] · Danuta Leszczynska[2] · Jerzy Leszczynski[3]

Received: 3 April 2023 / Accepted: 22 June 2023 / Published online: 29 June 2023

Abstract

Context To apply the quantitative relationships "structure-endpoint" approach, the reliability of prediction is necessary but sometimes challenging to achieve. In this work, an attempt is made to accomplish the reliability of forecasts by creating a set of random partitions of data into training and validation sets, followed by constructing random models. A system of random models for a helpful approach should be self-consistent, giving a similar or at least comparable statistical quality of the predictions for models obtained using different splits of available data into training and validation sets.

Method The carried out computer experiments aimed at obtaining blood–brain barrier permeation models showed that, in principle, can be used such an approach (the Monte Carlo optimization of the correlation weights for different molecular features) for the above purpose taking advantage of specific algorithms to optimize the modelling steps with applying of new statistical criteria such as the index of ideality of correlation (*IIC*) and the correlation intensity index (*CII*). The results so obtained are good and better than what was reported previously. The suggested approach to validation of models is non-identic to traditionally applied manners of the checking up models. The concept of validation can be used for arbitrary models (not only for models of the blood–brain barrier).

Keywords Blood–brain barrier permeation · QSAR · System of self-consistent models · Mathematical modelling · Monte Carlo method · CORAL software

Introduction

The extremely high levels of biocide products determined in different wastewater treatment plants may be due to the COVID-19 pandemic situation. During a recent pandemic, a much more significant amount of cleaning and hygiene products have been used [1]. Facade renders containing biocides are often used on external thermal insulation composite systems to avoid the growth of algae, fungi, and bacteria. Different factors such as soil type, percolation rate, and soil organic carbon sorption coefficient affect the time needed for the biocide substances to leave the soil and consequently keep a reasonable level of ecologic safety [2]. Ceramics for tableware (vases, tea sets, flower pots, plates, etc.) and sanitary ware also are affected by biocide substances [3]. The environmental fate of different biocides is relatively poorly understood, with nearly all data derived from the assessment reports on biodegradation [4]. In addition, the role of biocides in the search and design of drugs is quite noticeable [5–11]. Biocidal products are used in homes for hygienic purposes or in industries to control the growth of bacteria [12]. Their use should not provoke adverse effects for the environment and human health.

We address here the role that these substances may have in affecting the blood–brain barrier (BBB), which has been

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

[1] Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 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, MS 39217, USA

studied in medicine and pharmaceuticals [13–16] but not for home related products.

In particular, this work aims to develop quantitative structure–activity relationships (QSARs) for logBBB of various biocides [12] based solely on the analysis of molecular features extracted from SMILES [17]. This approach gives the uncial possibility of extracting the similarity of endpoints in terms of structural fragments extracted from SMILES [18, 19]. The availability of these simple tools based on the Monte Carlo method and representation of the molecular structure by SMILES can find applications that do not require any additional calculations.

Computational details

Data

The experimental data on logBBB ($n = 231$) for biocides was taken from the literature [12]. The log-scaled blood–brain barrier penetration (logBBB) is defined as follows:

$$logBBB = \log\left(\frac{C_{Brain}}{C_{Blood}}\right) \quad (1)$$

where, C_{brain} and C_{blood} are concentrations of the substances in the brain and blood, respectively. The total set of available data on logBBB was split into the active training (≈25%), passive training (≈25%), calibration (≈25%), and validation sets (≈25%). Five random splits were examined. Table S1 confirms that these splits are far from being identical (Table S1, *Supporting Information*).

Optimal descriptor

The optimal descriptor is the sum of the correlation weights of molecular attributes extracted from a simplified molecular input-line entry system (SMILES) [17]. Correlation weights are calculated through optimization, providing the maximum values of the objective function calculated from data on a structured training set consisting of active training, passive training, and calibration sets. Having the numerical data on the correlation weights, one can calculate the optimal descriptor for each SMILES according to the formula:

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

where, S_k is SMILES-atom, i.e. one symbol (e.g. 'C', 'N', 'O') or a group of symbols which cannot be examined separately (e.g. 'Cl', [N+], [O-]). The T and N are parameters of the optimization procedure. The T is the threshold, i.e. an integer used to separate SMILES-atoms into two classes (i) rare and (ii) non-rare. The rare SMILES-atoms are not involved in building up a model. The non-rare SMILES atoms are included in the list of molecular features involved in building a model. Thus, having the numerical data on correlation weights, the model for *logBBB* can be calculated by the traditional least squares method as the following:

$$logBBB = C_0 + C_1 * DCW(T,N) \quad (3)$$

The Monte Carlo optimization

Equation 2 requires an application of the numerical data on the above correlation weights. Monte Carlo optimization is a tool to calculate those correlation weights using the CORAL software (http://www.insilico.eu/coral). Here, two target functions for the Monte Carlo optimization are examined:

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (4)$$

$$TF_1 = TF_0 + (IIC_C + CII_C) \times 0.3 \quad (5)$$

The coefficients 0.1 and 0.3 were defined empirically.

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

The IIC_C is the index of ideality of correlation [20, 21]. The correlation intensity index (*CII*), similarly to the above *IIC*, was developed as a tool to improve the quality of the Monte Carlo optimization aimed at building up QSPR/QSAR models [22].

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

Applicability domain

The applicability domain for the described model defines via the so-called statistical defects of codes used in SMILES. These defects can be calculated as follows:

$$d_k = \frac{|P(S_k) - P'(S_k)|}{N(S_k) + N'(S_k)} + \frac{|P(S_k) - P''(S_k)|}{N(S_k) + N''(S_k)} + \frac{|P'(S_k) - P''(S_k)|}{N'(S_k) + N''(S_k)} \quad (6)$$

where $P(S_k)$, $P'(S_k)$ $P''(S_k)$ are the probability of S_k in the active training, passive training, and calibration sets, respectively; $N(S_k)$, $N'(S_k)$, and $N''(S_k)$ are frequencies of S_k in the active training, passive training, and calibration sets, respectively. The statistical defects of SMILES (D_j) are calculated as:

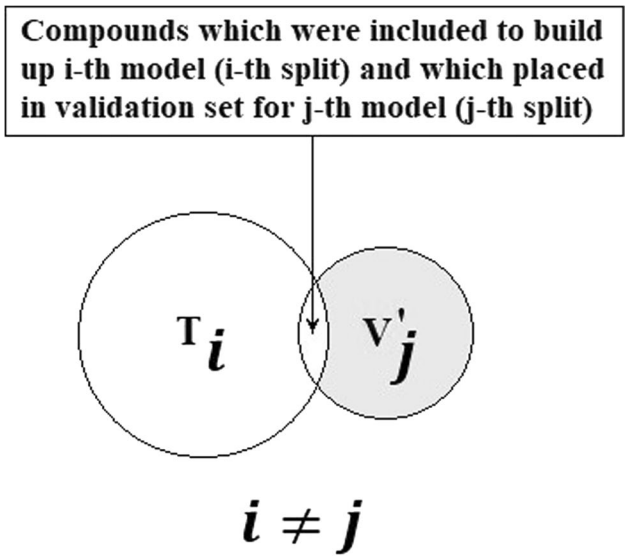

Fig. 1 The scheme for selecting compounds to check up the i-th model using an external set of j-th model. The T is the training set and the V is the validation set

$$D_j = \sum_{k=1}^{NA} d_k \quad (7)$$

where NA is the number of non-rare codes in SMILES.

A SMILES falls in the applicability domain if:

$$Dj < 2 * \overline{D} \quad (8)$$

The $\overline{D}$ is an average defect over the active training set.

The system of self-consistent models

Each i-th model has an i-th validation set. As it is demonstrated (Table S1), the validation sets are far from identical. To have a model which can be used in a general way, it is important that the arbitrary model can be used for an arbitrary validation set. If we achieve this condition, the different models should be considered self-consistent.

The measurement of self-consistency is carried out as the average and dispersion of the correlation coefficients on group of different validation sets. The corresponding computational experiments are represented by the matrix:

Table 1 The statistical quality of models observed for splits 1–5. The best model observes for split 4

Split	*Set*	*n*	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*F*
1	A	59	0.4012	0.5727	0.6123	0.7574	0.3566	0.529	38
	P	57	0.3177	0.4663	0.4539	0.7240	0.2768	0.684	26
	C	56	0.7350	0.8552	0.8557	0.8268	0.7157	0.233	150
	V	59	0.8037					0.216	
	AD	54	0.7833					0.213	
2	A	58	0.3780	0.5487	0.5739	0.7436	0.3389	0.626	34
	P	59	0.2657	0.4420	0.4556	0.7549	0.2174	0.621	21
	C	56	0.7987	0.8916	0.8936	0.8830	0.7835	0.226	214
	V	58	0.7808					0.236	
	AD	56	0.7592					0.238	
3	A	60	0.2570	0.4089	0.4742	0.7966	0.2122	0.647	20
	P	59	0.2459	0.4387	0.3835	0.7407	0.1981	0.619	19
	C	56	0.7471	0.8353	0.8641	0.8551	0.7247	0.213	159
	V	56	0.7726					0.272	
	AD	55	0.7789					0.273	
4	A	56	0.4342	0.6055	0.6135	0.7355	0.3949	0.542	41
	P	58	0.3176	0.4755	0.4540	0.7420	0.2813	0.664	26
	C	59	0.8191	0.8981	0.9048	0.8987	0.8057	0.190	258
	V	**58**	**0.8527****					**0.191**	
	AD	**50**	**0.8589**					**0.189**	
5	A	58	0.3881	0.5592	0.5814	0.7334	0.3534	0.570	36
	P	57	0.3726	0.5243	0.4214	0.7403	0.3310	0.639	33
	C	59	0.7956	0.8908	0.8915	0.8741	0.7822	0.204	222
	V	57	0.7039					0.263	
	AD	52	0.7325					0.246	

**n*, the number of quasi-SMILES in a set; *A*, active training set; *P*, passive training set; *C*, calibration set; *V*, validation set; *AD*, applicability domain; R^2, determination coefficient; *CCC*, concordance correlation coefficient; *IIC*, index of ideality of correlation; *CII*, correlation intensity index; Q^2, cross validated correlation coefficient; *RMSE*, root mean squared error; *F*, Fischer F-ratio. **The best result indicated by bold

$$\begin{bmatrix} \left(M_1 : V_1 \rightarrow Rv_{11}^2\right) & \cdots & \left(M_5 : V\prime_1 \rightarrow R'v_{51}^2\right) \\ \vdots & & \vdots \\ \left(M_1 : V\prime_5 \rightarrow R'v_{15}^2\right) & \cdots & \left(M_5 : V_5 \rightarrow Rv_{55}^2\right) \end{bmatrix} \quad (9)$$

the 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; the Rv_{ij}^2 is the correlation coefficient observed for the *j-th* validation set if applied *i-th* model.

Checking up on the predictive potential

Having models built using different i-th splits ($i = 1,5$) into active training, passive training, calibration, and validation sets, it is possible to estimate the average value of the coefficient of determination for j-th external sets ($j = 1,5$) without the compounds that are included in the j-th active training set, j-th passive training set, and j-th calibration set (i≠j). Figure 1 shows the scheme of selection compounds for the (i,j)-checking up. The mean should be as large as possible and the variance as small as possible.

Results and discussion

QSAR models

Five models obtained for different splits are the following:

$$\text{LogBBB} = 0.2406(\pm 0.0098) + 0.1935(\pm 0.0044) * \text{DCW}(1,15) \quad (10)$$

$$\text{LogBBB} = 0.4135(\pm 0.0164) + 0.1134(\pm 0.0023) * \text{DCW}(1,15) \quad (11)$$

$$\text{LogBBB} = 0.0747(\pm 0.0113) + 0.0845(\pm 0.0021) * \text{DCW}(1,15) \quad (12)$$

$$\text{LogBBB} = 0.3066(\pm 0.0110) + 0.1130(\pm 0.0022) * \text{DCW}(1,15) \quad (13)$$

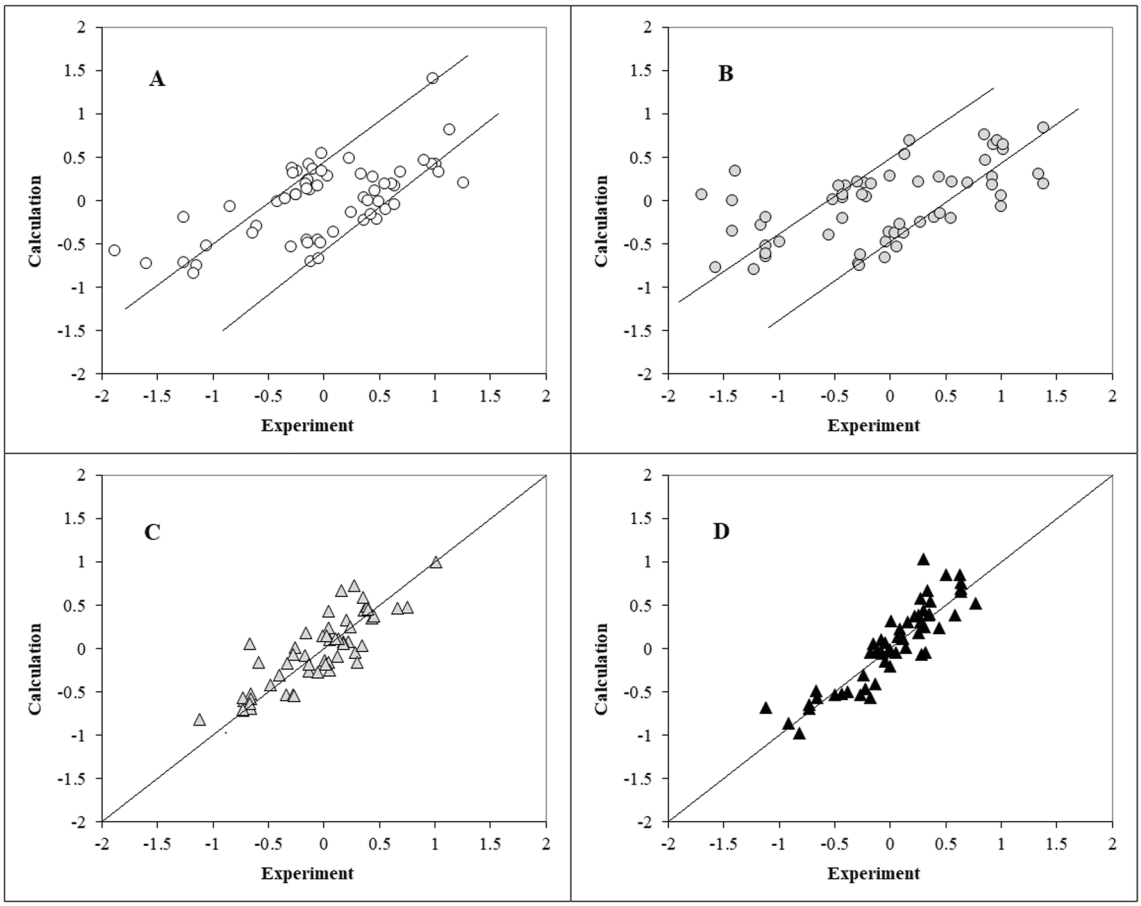

Fig. 2 The "latent" correlations observed for the active training set (**A**) and passive training set (**B**). The "usual" correlation for the calibration set (**C**) and validation set (**D**)

$$\text{LogBBB} = 0.2942(\pm 0.0110) + 0.1023(\pm 0.0018) * \text{DCW}(1, 15) \quad (14)$$

Table 1 shows data related to the statistical quality of these models.

The *IIC* and *CII* improve the statistical quality of models on the calibration and validation sets but they detriment active and passive training sets. Nonetheless, the active and passive training sets contain latent separated correlations. Figure 2 demonstrates these correlations (observed for the model based on split-4).

It is generally accepted that for good models, the coefficient of determination has high values for all sets. However, application of *IIC* and *CII* indices can lead to unusual models in which there is a stratification of correlations on active and passive training sets [23]. The physicochemical and/or biochemical behaviour of substances as participants in the construction of the model can be typical (average) or special (non-standard). Correlation weights calculated by the CORAL program at the first stage of model construction consider average molecules most of all. However, there inevitably comes a moment of information saturation when the average molecules cease to be informative for the construction of the model. From this point on, the model is built using the resources of molecules with non-standard behaviour. This leads to overfitting, which causer, an improvement in the model for the visible training set but with a deterioration in the model for the invisible external set [24]. However, the application of the *IIC* and *CII* apparently leads to unexpected geometry of the above-mentioned classes of substances. Figure 3 contains three examples of the latent correlations together with corresponding values of the *IIC* and *CII*. Since the *CII* has maximal values compared to the determination and concordance correlation coefficients, one can suggest that *CII* leads to these double-hidden correlations. In any case, the described version of the Monte Carlo optimization improves the models' predictive potential.

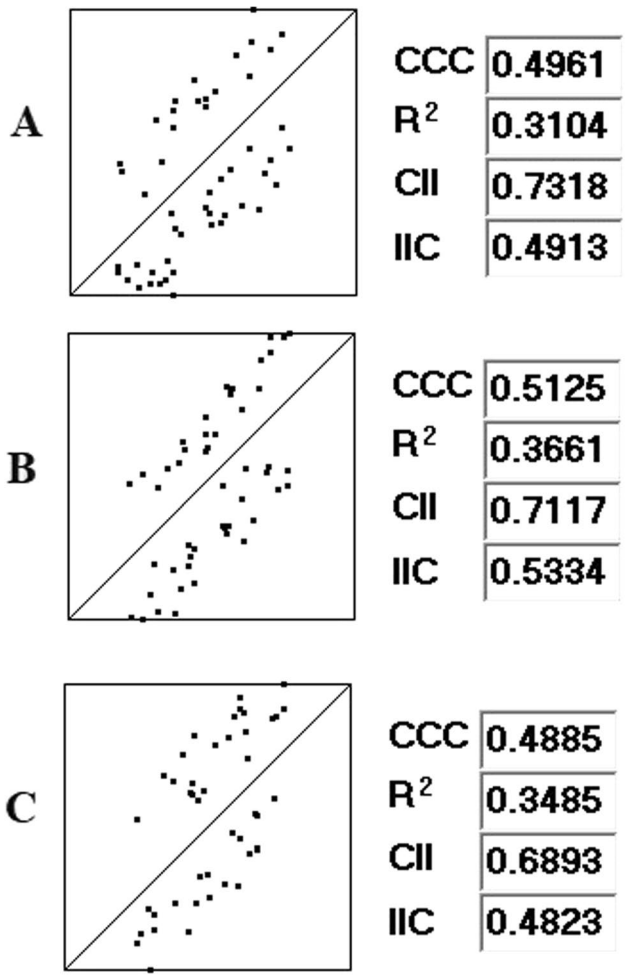

Fig. 3 Examples of latent correlations and their evaluation via concordance correlation coefficient, determination coefficient, index of ideality of correlation, and correlation intensity index

The estimation of the predictive potential of models

In order to assess the reliability of the constructed models, their verification was carried out on the compounds that were included in the external sets used for other models. Compounds that were used during the development of the i-th model were excluded from consideration. Figure 2 shows the verification scheme of the i-th model with the compounds used as an external set for the j-th model (i≠j). Table 2 contains the results of such verifications. From Table 2, it can be seen that there is an approximate similarity in the predictive potential of all five models built on completely different partitions of the available compounds into training (i.e. active training, passive training, and calibration sets) and validation set.

Applicability domain

According to inequality 8, there are eight suspected compounds in the validation set of split 4. Removing these compounds improves the statistical quality of the model; the determination coefficient becomes 0.8589, and RMSE becomes 0.189 (Table 1) after their elimination.

Mechanistic interpretation

The developed QSAR models allow for mechanical interpretation of the studied phenomena. Having the numerical data on the correlation weights of features which are included in several runs of the Monte Carlo optimization, one can extract three categories of these features:

Table 2 The system of self-consistency of models

Model	Split	1	Split	2	Split	3	Split	4	Split	5
	n'_V	Rt^2_V	n'_V	Rt^2_V	n'_V	Rt^2_V	n'_V	Rt^2_V	n'_V	Rt^2_V
1			28	0.855	17	0.919	29	0.852	24	0.843
2	28	0.858			22	0.830	29	0.879	18	0.721
3	17	0.834	22	0.841			27	0.861	23	0.782
4	29	0.874	29	0.881	27	0.880			23	0.903
5	24	0.819	18	0.841	23	0.854	23	0.845		

(i) Features which have a positive value of the correlation weight in all runs. These are promoters of endpoint increase;
(ii) Features which have a negative value of the correlation weight in all runs. These are promoters of endpoint decrease;
(iii) Features which have both negative and positive values of the correlation weight in different runs of the optimization. These are features with an unclear role (one cannot classify these features as a promoter of increase or decrease for endpoint).

Considering Table 3, it can be seen that the presence of carbon (sp^2 and sp^3 hybridization, conveyed in SMILES by uppercase 'c' and lowercase 'C', respectively) is a promoter of logBBB growth. The remaining SMILES attributes, characterized by positive correlation weights in several optimization runs, have little distribution in the training set. It can be seen that there are more promoters of the logBBB decrease; this is the branching of the molecular skeleton (expressed in SMILES via brackets), oxygen, nitrogen, as well as double bonds (=).

Comparison with models from the literature

The statistical quality of the models for logBBB suggested in the literature is the following: $n = 153$, $R^2 = 0.76$, $RMSE = 0.300$ (training set); and $n = 78$, $R^2 = 0.64$, $RMSE = 0.370$ (validation set) [12]. The best model suggested here for the validation set is characterized by $n = 50$, $R^2 = 0.8589$, and $RMSE = 0.189$. The applied approach results in the better statistics of the developed model than those reported in the literature. Furthermore, the approach we applied is simpler, since it only uses the SMILES format of the chemical substance.

Table 3 Promoters of increase or decrease for logBBB

No.	S_k	*CWs run 1*	*CWs run 2*	*CWs run 3*	N_A	N_P	N_C	d_k *Eq.14*
1	C.....	0.69080	0.74857	0.50292	56	58	58	0.0002
2	c.....	0.42069	0.49400	0.00676	23	22	25	0.0013
3	F.....	0.84471	1.42285	1.18168	4	6	7	0.0056
4	I.....	1.78015	5.27754	3.02432	2	1	5	0.0169
1	(.....	−0.35661	−0.81598	−0.41823	48	54	51	0.0010
2	O.....	−1.73406	−1.53694	−1.40457	44	42	39	0.0020
3	N.....	−0.43987	−1.25271	−1.06118	38	41	33	0.0026
4	=.....	−0.69390	−0.53892	−0.90548	36	45	42	0.0022
5	2.....	−0.23351	−0.31056	−0.09185	18	25	19	0.0035
6	S.....	−1.12558	−2.53046	−1.61434	7	5	11	0.0087
7	n.....	−1.31308	−2.76438	−1.71295	7	9	11	0.0046
8	Cl.....	−0.59814	−0.41532	−0.34454	6	10	5	0.0083
9	#.....	−2.80264	−1.75335	−1.93459	3	2	4	0.0074
10	4.....	−1.07672	−0.66505	−0.71514	2	2	5	0.0112
11	[nH].....	−1.25183	−1.68805	−1.69190	2	2	5	0.0112
12	o.....	−2.08765	−3.13159	−2.11409	2	3	2	0.0051
13	s.....	−1.09455	−1.00518	−1.35623	2	1	3	0.0112
14	-.....	−2.14315	−4.00805	−1.99609	1	5	1	0.0198

Last but not least

One can state, citing earlier papers that all models are wrong, but some are useful [25, 26]. The presented here models must obey two conditions:

- Firstly, the mentioned correlation ideality index and correlation intensity index should be used in optimizing the correlation weights on which the proposed logBBB models are calculated.
- Secondly, it is necessary to select distributions with the smallest possible sum of statistical SMILES defects (Eq. 7) involved in calculating the model. Compliance with these conditions helps not only for the "usefulness" of models for logBBB but also their reproducibility [27].

Conclusions

In this study, a rather convenient and dependable approach for logBBB modelling is proposed. The self-consistency of the models obtained with different distributions of available data into the training and validation sets confirms the reliability of this approach. Comparison with similar models from the literature indicates the acceptable accuracy of the constructed models. The approach we used is quite simple, since it only uses the SMILES format to represent the substance. Still, results are improved compared with those previously published.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00894-023-05632-2.

Acknowledgements AAT, APT, AR, and EB are acknowledge EFSA for the financial contribution within the project sOFT-ERA, OC/EFSA/IDATA/2022/0. DL and JL are thankful to the National Science Foundation for grant No. NSF/CREST HRD-1547754 support.

Data and software availability CORAL software (http://www.insilico.eu/coral)

Code availability Not applicable

Author contribution All authors contributed to the study's conception and design. Data collection and analysis were performed by A.P.T. and A.A.T.; the first draft of the manuscript was written by A.P.T., A.A.T., A.R., E.B., D.L., and J.L., and all authors commented on previous versions of the manuscript. E.B. and J.L.: supervising, reviewing, and editing. All authors read and approved the final manuscript.

Funding AAT, APT, AR, and EB are acknowledge EFSA for the financial contribution within the project sOFT-ERA, OC/EFSA/IDATA/2022/0. DL and JL are thankful to the National Science Foundation for grant No. NSF/CREST HRD-1547754 support.

Declarations

Ethics approval All authors approve the ethics.

Consent to participate All authors agree to participate in this investigation.

Consent for publication All authors give their consent for publication.

Competing interests The authors declare no competing interests.

References

1. Paun I, Pirvu F, Iancu VI, Chiriac FL (2022). Int J Environ Res Public Health 19(13):7777. https://doi.org/10.3390/ijerph19137777
2. Vega-Garcia P, Lok CSC, Marhoon A, Schwerd R, Johann S, Helmreich B (2022). Build Environ 216:108991. https://doi.org/10.1016/j.buildenv.2022.108991
3. Mata P, Coulon A, Filhol A, Pillet G (2022) InterCeram. Int Ceram Rev 71(1):18–27. https://doi.org/10.1007/s42411-021-0481-9
4. Koning JT, Bollmann UE, Bester K (2021). J Hazard Mater 406:124755. https://doi.org/10.1016/j.jhazmat.2020.124755
5. Schinkel AH, Smit JJM, van Tellingen O, Beijnen JH, Wagenaar E, van Deemter L, Mol CAAM, van der Valk MA, Robanus-Maandag EC, te Riele HPJ, Berns AJM, Borst P (1994). Cell 77(4):491–502. https://doi.org/10.1016/0092-8674(94)90212-7
6. Pardridge WM (2005). NeuroRx 2(1):3–14. https://doi.org/10.1602/neurorx.2.1.3
7. Medley CD, Gruenhagen J, Yehl P, Chetwyn NP (2014). Anal Methods 6(17):6635–6640. https://doi.org/10.1039/c4ay01418
8. Ghanem B, Haddadin RN (2018). Antimicrob Resist Infect Control 7(1):47. https://doi.org/10.1186/s13756-018-0339-8
9. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, Belk KE, Noyes NR, Morley PS (2020). Nucleic Acids Res 48(D1):D561–D569. https://doi.org/10.1093/nar/gkz1010
10. Mitusova K, Peltek OO, Karpov TE, Muslimov AR, Zyuzin MV, Timin AS (2022). Nanobiotechnology 20(1):412. https://doi.org/10.1186/s12951-022-01610-7
11. Wang J, Li Z, Pan M, Fiaz M, Hao Y, Yan Y, Sun L, Yan F (2022). Adv Drug Deliv Rev 190:114539. https://doi.org/10.1016/j.addr.2022.114539
12. Shin HK, Lee S, Oh H-N, Yoo D, Park S, Kim W-K, Kang M-G (2021). Chemosphere 277:130330. https://doi.org/10.1016/j.chemosphere.2021.130330
13. Wardlaw JM, Sandercock PAG, Dennis MS, Starr J (2003). Stroke 34(3):806–811. https://doi.org/10.1161/01.STR.0000058480.77236.B3
14. Toropov AA, Toropova AP, Beeg M, Gobbi M, Salmona M (2017). J Pharmacol Toxicol Methods 88:7–18. https://doi.org/10.1016/j.vascn.2017.04.014
15. Toropov AA, Toropova AP, Benfenati E, Dorne JL (2018). Chem-Biol Interact 290:1–5. https://doi.org/10.1016/j.cbi.2018.04.030
16. López-Ornelas A, Jiménez A, Pérez-Sánchez G, Rodríguez-Pérez CE, Corzo-Cruz A, Velasco I, Estudillo E (2022). Int J Mol Sci 23(17):10136. https://doi.org/10.3390/ijms231710136
17. Weininger D (1988). J Chem Inf Comput Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005
18. Toropova AP, Toropov AA, Begum S, Achary PGR (2018). Curr Neuropharmacol 16:769–785. https://doi.org/10.2174/1570159X15666171016163951

19. Toropov AA, Toropova AP, Benfenati E, Salmona M (2018). Toxicol Mech Methods 28(5):321–327. https://doi.org/10.1080/15376516.2017.1422579
20. Toropova AP, Toropov AA (2017). Sci Total Environ 586:466–472. https://doi.org/10.1016/j.scitotenv.2017.01.198
21. Toropov AA, Carbó-Dorca R, Toropova AP (2018). Struct Chem 29(1):33–38. https://doi.org/10.1007/s11224-017-0997-9
22. Toropov AA, Toropova AP (2020). Sci Total Environ 737:139720. https://doi.org/10.1016/j.scitotenv.2020.139720
23. Toropov AA, Toropova AP, Achary PGR, Raškova M, Raška I (2022). Toxicol Mech Methods 32(7):549–557. https://doi.org/10.1080/15376516.2022.2053918
24. Toropova AP, Toropov AA, Benfenati E, Gini G (2011). Chemom Intell Lab Syst 105(2):215–219. https://doi.org/10.1016/j.chemolab.2010.12.007
25. Box GEP (1976). J Am Stat Assoc 71(356):791–799. https://doi.org/10.1080/01621459.1976.10480949
26. Delatte N (2022). J Perform Constr Facil 36(3):04022025. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001730
27. Wilson MD, Boag RJ, Strickland L (2019). Comput Brain Behav 2(3-4):239–241. https://doi.org/10.1007/s42113-019-00054-x