Journal of Carbon Research

MDPI

Article

Monte Carlo Simulation of Aromatic Molecule Adsorption on Multi-Walled Carbon Nanotube Surfaces Using Coefficient of Conformism of a Correlative Prediction (CCCP)

Alla P. Toropova *, Andrey A. Toropov, Alessandra Roncaglioni and Emilio Benfenati

Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milan, Italy; andrey.toropov@marionegri.it (A.A.T.); alessandra.roncaglioni@marionegri.it (A.R.); emilio.benfenati@marionegri.it (E.B.)
* Correspondence: alla.toropova@marionegri.it; Tel.: +39-02-3901-4595

Abstract: Using the Monte Carlo technique via CORAL-2024 software, models of aromatic substance adsorption on multi-walled nanotubes were constructed. Possible mechanistic interpretations of such models and the corresponding applicability domains were investigated. In constructing the models, criteria of the predictive potential such as the iIndex of Ideality of Correlation (IIC), the Correlation Intensity Index (CII), and the Coefficient of Conformism of a Correlative Prediction (CCCP) were used. It was assumed that the CCCP could serve as a tool for increasing the predictive potential of adsorption models of organic substances on the surface of nanotubes. The developed models provided good predictive potential. The perspectives on the improvement of the nano-QSPR/QSAR were discussed.

Keywords: QSPR; Monte Carlo method; multi-walled carbon nanotubes (MWCNTs); adsorption; coefficient of conformism of a correlative prediction (CCCP); CORAL software

check for updates

Academic Editor: Miguel A. Montes Morán

Received: 18 December 2024
Revised: 3 January 2025
Accepted: 8 January 2025
Published: 14 January 2025

Citation: Toropova, A.P.; Toropov, A.A.; Roncaglioni, A.; Benfenati, E. Monte Carlo Simulation of Aromatic Molecule Adsorption on Multi-Walled Carbon Nanotube Surfaces Using Coefficient of Conformism of a Correlative Prediction (CCCP). *C* **2025**, *11*, 7. https://doi.org/10.3390/c11010007

1. Introduction

Nanotechnology has opened up new foundations for scientific disciplines and technology due to its ability to demonstrate extraordinary properties of materials. Nanotechnology is an interdisciplinary subject that combines engineering and manufacturing concepts at a molecular level, while also providing new possibilities in medicine [1]. Nanotechnology is mainly concerned with the study of physical processes and the application of these nanostructures in real-world applications [2–7]. In recent decades, nanomaterials have been synthesized, characterized, and widely applied in various fields. This technology has had an impact on several fields, including electronics, chemistry, biology, and biomedicine. Although different sectors create different types of nanomaterials, nanotechnology is becoming increasingly relevant in environmental engineering to address environmental pollution and contamination issues.

The development of new materials and technologies for their production and processing is currently recognized as the so-called "key" or "critical" aspect of the foundation of economic power [8]. In fact, two of the priority areas of development in modern materials science are nanomaterials and nanotechnology. The development of fundamental and applied concepts of nanomaterials and nanotechnology in the coming years can lead to fundamental changes in many areas of human activity: materials science, energy, electronics, computer science, mechanical engineering, medicine, agriculture, and ecology. Along with computer information technology and biotechnology, nanotechnology is the basis of scientific and technological revolution. It is difficult to imagine modern science without the

widespread use of mathematical modeling, which consists of replacing the original object with its "image"—a mathematical model—and further studying this model using computer systems. This method combines many advantages of both theory and experiment. Working not with the object itself, but with its model, allows us to painlessly and relatively quickly study its properties and behavior in any conceivable situation, without significant costs. At the same time, computational experiments with object models, relying on the power of modern computing tools in informatics, allow us to study objects in sufficient detail and depth, which is inaccessible to purely theoretical approaches. The use of experimental methods leads to large amounts of time, financial, and labor costs. Therefore, the use of methods for modeling the properties of nanostructures will significantly reduce these costs.

The use of computer modeling for nanosystems has fundamental difficulties. Firstly, there is no long-range order, characteristic of crystals, allowing us to reduce the number of independent degrees of freedom of the system; secondly, the short-range order, characteristic of liquids, does not allow us to determine all the functional properties of nanomaterials. Thirdly, there are technical difficulties associated with modeling macro-objects at the atomic level. Direct modeling of such systems in the approximation of molecular dynamics and, especially, quantum mechanics is difficult even with the use of modern supercomputer technology. A solution may be the use of a hierarchical multiscale approach to modeling, when, at each lower level, the parameters and variables necessary for constructing upper-level models are calculated. At present, experimental methods are mainly used in the development of nanomaterials with specified properties, which does not always allow us to find the optimal solution, and increases the cost of development, so it is advisable to more actively involve mathematical modeling methods that allow predicting the composition, characteristics, and properties of future nanomaterials. To implement the mathematical modeling of physicochemical processes, it is necessary to have mathematical models based on certain theoretical approaches. Computational nanotechnology is of crucial importance for prototyping nanomaterials, devices, systems, and various applications. At the same time, it can be used not only to understand and characterize systems obtained as a result of experiments but also to predict the properties of new materials, since there is a close relationship between structural, mechanical, chemical, and electrical properties in the nanoscale region [8,9].

A special place among nanomaterials is occupied by multi-walled nanotubes (MWCNTs), which have found numerous applications. Among the applications of nanotubes, a special place is occupied by their applications for cleaning and blocking various man-made emissions that pose an environmental problem [10].

The measure of the ability of organic compounds to be blocked before causing harm to ecology or human beings is their ability to adsorb on the surface of nanotubes. Thus, knowledge of this parameter for various organic compounds (potential industrial pollutants) is useful information from the point of view of risk assessment. The list of organic compounds capable of participating in technological cycles and, therefore, being potential pollutants of the environment is large and constantly growing.

Since the experimental determination of the adsorption capacity of various organic compounds on nanotubes is a rather complex and expensive problem, the development of models of this physicochemical parameter in the form of a mathematical function of the molecular structure is a very relevant theoretical and practical problem. Quantitative structure–property/activity relationships (QSPRs/QSARs) are widely used tools to solve the problem, at least for substances, which are not nanomaterials [11]. It should be noted, however, that the search for nano-versions of QSPRs/QSARs is also taking place in the stream of work aimed at developing models of various endpoints [12–15].

Here, the possibility and efficiency of using the Monte Carlo method to solve the given problem is considered. Previously, the so-called index of ideality correlation (IIC) [13] and the correlation intensity index (CII) [14] were suggested as means of improving the operation of the optimization procedure carried out using the Monte Carlo method.

The *IIC* is a value sensitive to both the value of the correlation coefficient and the value of the average absolute error [13]. The *CII* is a value defined by the presence of structures in the set that support the correlation; the removal of such a structure, while leaving all others in the set under consideration leads to a decrease in the correlation coefficient [14]. The so-called coefficient of conformism of a correlative prediction (CCCP) [15] is the ratio of the sum of 'supporters' and the sum of 'oppositionists' of the correlation in a set [15]. An oppositionist is a structure removed, which leads to an increase in the correlation coefficient.

These special criteria of predictive potential can serve as a tool to improve it. It was assumed that the *CCCP* could serve as a tool for increasing the predictive potential of models of adsorption of organic substances on the surface of nanotubes. Since, when calculating the *CCCP*, the influence of both 'supporters' and 'opponents' of the correlation is actually taken into account, there is a possibility that the *CCCP* will be more effective (at least more informative) than the *CII* value, which does not take into account the influence of 'opponents' of the correlation.

Three levels of computer simulation can be applied that relate to the endpoints of nanomaterials or the physicochemical phenomena associated with nanomaterials. The first level relates to models reached by traditional tools used in the traditional (without nanomaterials) QSPRs/QSARs. The second level, or hybrid, consists of models reached partially with traditional tools of QSPRs/QSARs analysis but, in addition, using new descriptors/algorithms related solely to nano-reality. The third level consists of models reached with exclusive tools related solely to nano-reality. From this point of view, the given work considered a QSPR analysis of the first level, despite MWCNTs being part of the wider phenomena of nano-reality.

Each of the above-mentioned levels has advantages and disadvantages. The first level is poor in the context of the nano-reality. However, it can be a convenient tool in practice. The second level may be quite intriguing in practical terms, but the interaction between nano-reality and the traditional QSPRs/QSARs is unpredictable. The third level, being the most obvious, is at the same time the most innovative, and therefore the most unpredictable.

2. Materials and Methods

2.1. Data

Experimental data on the adsorption of organic substances on MWCNTs, as well as simplified molecular input-line entry systems (SMILES) [16] representing the mentioned organic substances, were borrowed from the literature [11,12]. The surface area normalized adsorption coefficients of organic compounds on MWCNTs are considered here as the endpoint. There were two duplicates of SMILES in the work [11]. Thus, the total number of compounds considered is n = 68.

The model for the endpoint calculated via the descriptor of the correlation weights (DCW) of SMILES attributes is as follows:

$$\log K = C_0 + C_1 \times DCW(T, N) \quad (1)$$

The *DCW*(*T*,*N*) is the optimal SMILES-based descriptor calculated with the correlation weights of all the statistically significant molecular features extracted from SMILES [13].

T and N are parameters of the Monte Carlo method that provide the correlation weights used to calculate the $DCW(T,N)$ values. T is the threshold to define statistically significant (non-rare) molecular features. A feature is not significant if its frequency on the training set is smaller than T. N is the number of epochs of the Monte Carlo optimization. One epoch is a random sequence of modifications for all the statistically significant molecular features extracted from SMILES. LogK is the logarithm of the surface area normalized adsorption coefficients of organic compounds on MWCNTs.

The choice of values of T and N is made empirically. However, it is obvious: (i) that choosing too large a value of T will result in too few parameters being optimized, which will make the model more primitive; and (ii) too large a value of N will make the calculation longer, and a significant part of the epochs may be in vain (the same achieved value of the objective function will be repeated).

2.2. Optimal SMILES-Based Descriptor

The optimal SMILES-based descriptor used in Equation (1) is calculated using the following equation:

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

S_k is a SMILES atom, i.e., an undivided fragment of SMILES (e.g., 'C', 'O', 'Br', 'Cl', etc.). SS_k is a pair of SMILES atoms, which are neighbors in the SMILES line. From the two possible configurations of SMILES atoms, we selected those according to the ASCII codes [17] of the corresponding symbols.

2.3. Monte Carlo Method

The numerical data on the correlation weights necessary to calculate the optimal descriptor are obtained using the four different target functions listed below.

$$\mathrm{TF_0} = \mathrm{R_A^2} + \mathrm{R_P^2} + \left|\mathrm{R_A^2} - \mathrm{R_P^2}\right| \times 0.1 \quad (3)$$

$$\mathrm{TF_{IIC}} = \mathrm{R_A^2} + \mathrm{R_P^2} + \left|\mathrm{R_A^2} - \mathrm{R_P^2}\right| \times 0.1 + \mathrm{IIC} \times 0.25 \quad (4)$$

$$\mathrm{TF_{CII}} = \mathrm{R_A^2} + \mathrm{R_P^2} + \left|\mathrm{R_A^2} - \mathrm{R_P^2}\right| \times 0.1 + \mathrm{CII} \times 0.25 \quad (5)$$

$$\mathrm{TF_{CCCP}} = \mathrm{R_A^2} + \mathrm{R_P^2} + \left|\mathrm{R_A^2} - \mathrm{R_P^2}\right| \times 0.1 + \mathrm{CCCP} \times 0.25 \quad (6)$$

The Monte Carlo optimization includes the following stages. First, distribution into the active training set, the passive training set, the calibration set, and the validation set. The distribution of the entire data set into the specified sets is carried out randomly, but in equal shares (i.e., approximately 25%). Second, stochastic modifications of the correlation weights aimed to reach a maximum of the selected target function.

In Equations (3)–(6), $R^2{}_A$ and $R^2{}_P$ are determination coefficients related to the active and passive training sets.

The index of ideality of correlation (IIC) [15] is calculated using data on the calibration set as follows:

$$IIC = r\frac{\min\left({}^{-}MAE_C, {}^{+}MAE_C\right)}{\max\left({}^{-}MAE_C, {}^{+}MAE_C\right)} \quad (7)$$

$$\min(x,y) = \begin{cases} x, \ if \ x < y \\ y, otherwise \end{cases} \quad (8)$$

$$\max(x,y) = \begin{cases} x, \ if \ x > y \\ y, otherwise \end{cases} \quad (9)$$

$$^{-}MAE = \frac{1}{^{-}N}\sum|\Delta_k|,\ ^{-}N \text{ is the number of } \Delta_k < 0 \quad (10)$$

$$^{+}MAE = \frac{1}{^{+}N}\sum|\Delta_k|,\ ^{+}N \text{ is the number of } \Delta_k \geq 0 \quad (11)$$

$$\Delta_k = observed_k - calculated_k \quad (12)$$

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

The correlation intensity index (*CII*) is calculated as follows [16]:

$$CII_C = 1 - \sum Protest_k \quad (13)$$

$$Protest_k = \begin{cases} R_k^2 - R^2, & if\ R_k^2 - R^2 > 0 \\ 0, & otherwise \end{cases} \quad (14)$$

R^2 is the correlation coefficient for a set that contains n substances. $R^2{}_k$ is the correlation coefficient for $n - 1$ substances of a set after removing the k-th substance. Hence, if $\Delta R = (R^2{}_k - R^2)$ is larger than zero, the k-th substance is an "opponent" for the correlation between experimental and predicted values of the set. A small sum of "protests" means a higher correlation [14].

The cCoefficient of Conformism of a Correlative Prediction (*CCCP*) is a new criterion of the predictive potential [15]. The CCCP is calculated the follows:

$$CCCP = \frac{\Delta R_{oppositionist}}{\Delta R_{supporter}} \quad (15)$$

$\Delta R_{supporter}$ is the antagonist to $\Delta R_{oppositionist}$, i.e., $\Delta R = (R^2{}_k - R^2)$, which is smaller than zero.

The main idea of the *IIC* is an attempt to combine the statistical quality of the model, transmitted through the values of the coefficients of determination, and the statistical quality of the model, transmitted through the values of the average absolute error. Instead of the latter, one can use the root mean square error, but the use of the absolute error turned out to be more effective in terms of the results of the stochastic process of the optimization using the Monte Carlo method. The basic idea of *CII* is to evaluate how individual compounds influence the overall correlation in the set. In the case of *CII*, the overall total contribution of all opponents of the correlation is examined, that is, those compounds whose removal from consideration leads to an improvement in the correlation. The *CCCP* should be considered an attempt to improve the information content of the *CII* by taking into account the influence of the "supporters" of the correlation (compounds whose removal from consideration leads to a decrease in the coefficient of determination).

There is a variety of SMILES. Some are called canonical because they follow a standardized procedure for the representation of the molecular structure; in practice, there may be several canonical SMILES for the same substance. Therefore, it is better to consider SMILES obtained with a single program, without pretending that canonical SMILES obtained with different programs will be identical.

3. Results

To assess the statistical quality of the models, the following criteria have been used: the number of compounds in a set; the determination coefficient R^2; the concordance correlation coefficient CCC [18]; the cross-validated correlation coefficient Q^2 [19], the root means square error (RMSE), and the Fischer F-ratio [19]. In addition, the IIC, CII, and CCCP were used for comparing the quality of the models. Five different splits into active training (A), passive training (P), calibration (C), and validation sets (V) have been analyzed here.

3.1. Monte Carlo Optimization Without Considering Calibration Set Status

Table 1 contains the statistical characteristics of the models for logK observed in the case of the target function TF_0 (as seen in Equation (3)). One can see that the validation sets have several "unstable" statistical characteristics (for instance, the determination coefficient ranges from 0.41 to 0.74). The average and dispersion of the determination coefficient of the validation sets are 0.59 ± 0.12.

Table 1. The statistical characteristics of models build up using TF_0 on five random splits. The best model is indicated in bold.

	n *	R^2	*CCC*	*IIC*	*CII*	Q^2	*CCCP*	*RMSE*	*F*
A	18	0.9696	0.9846	0.6266	0.9790	0.9632	0.9591	0.245	510
P	17	0.9303	0.8920	0.2280	0.9384	0.9151	0.8044	0.625	200
C	18	0.0805	0.1647	0.1036	0.7526	0	−0.5554	1.29	1
V	16	0.5055	-	-	-	-	-	1.00	-
A	**18**	**0.8026**	**0.8905**	**0.4480**	**0.8732**	**0.7530**	**0.5573**	**0.553**	**65**
P	**16**	**0.8029**	**0.8670**	**0.6314**	**0.8477**	**0.7599**	**0.4188**	**0.803**	**57**
C	**18**	**0.7991**	**0.8743**	**0.6235**	**0.8464**	**0.7460**	**0.0355**	**0.318**	**64**
V	**17**	**0.4082**	**-**	**-**	**-**	**-**	**-**	**0.62**	**-**
A	17	0.7921	0.8840	0.7911	0.8387	0.7538	−0.0397	0.702	57
P	18	0.9667	0.2240	0.3461	0.9731	0.9595	0.9091	1.23	465
C	16	0.1597	0.2261	0.1199	0.5083	0	−0.7493	0.566	3
V	18	0.7376	-	-	-	-	-	0.53	-
A	18	0.8744	0.9330	0.9351	0.9163	0.8489	0.7761	0.429	111
P	16	0.9539	0.8740	0.6130	0.9587	0.9430	0.7264	0.511	290
C	18	0.2744	0.4070	0.3395	0.6712	0	−0.0965	0.859	6
V	17	0.6498	-	-	-	-	-	0.73	-
A	18	0.9363	0.9671	0.7741	0.9517	0.9208	0.8812	0.331	235
P	17	0.7197	0.5935	0.3515	0.7897	0.6671	0.2012	2.22	39
C	18	0.4443	0.5533	0.4044	0.6672	0.1452	−0.2561	0.956	13
V	16	0.6514	-	-	-	-	-	0.71	-

* *n* = 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 R^2; *CCCP* = coefficient of conformism of a correlative prediction; *RMSE* = root mean square error; *F* = Fischer F-ratio.

3.2. Monte Carlo Optimization Using TF_{IIC}

Table 2 contains the statistical characteristics of the models for logK observed in the case of the target function TF_{IIC} (as seen in Equation (4)). One can see that the validation sets have quite "unstable" statistical characteristics (the determination coefficient ranges from 0.04 to 0.76). The average and dispersion of the determination coefficient of the validation sets are 0.57 ± 0.28.

Table 2. The statistical characteristics of models build up using TF_{IIC} on five random splits. The best model is indicated in bold.

	n *	R^2	*CCC*	*IIC*	*CII*	Q^2	*CCCP*	*RMSE*	*F*
A	18	0.7928	0.8844	0.8904	0.8380	0.7509	0.5566	0.298	61
P	17	0.9234	0.8023	0.4737	0.9359	0.9092	0.6593	0.650	181
C	18	0.5022	0.6026	0.7085	0.6930	0.3673	−0.4570	0.891	16

Table 2. *Cont.*

	n *	R^2	*CCC*	*IIC*	*CII*	Q^2	*CCCP*	*RMSE*	*F*
V	16	0.0373	-	-	-	-	-	1.07	-
A	**18**	**0.5856**	**0.7387**	**0.7653**	**0.7861**	**0.4718**	**0.0819**	**0.822**	**23**
P	**16**	**0.6605**	**0.6277**	**0.2436**	**0.7873**	**0.5807**	**0.3494**	**1.08**	**27**
C	**18**	**0.7235**	**0.8090**	**0.8499**	**0.8490**	**0.5394**	**0.7073**	**0.433**	**42**
V	**17**	**0.7629**	**-**	**-**	**-**	**-**	**-**	**0.67**	**-**
A	17	0.6099	0.7577	0.6942	0.7737	0.5289	0.3759	0.833	23
P	18	0.8177	0.7779	0.7673	0.8835	0.7801	0.4544	0.681	72
C	17	0.7959	0.8583	0.8901	0.8291	0.7384	−0.2418	0.374	58
V	17	0.7524	-	-	-	-	-	0.89	-
A	18	0.7297	0.8437	0.5436	0.8128	0.6720	0.1829	0.681	43
P	18	0.6606	0.7949	0.5372	0.7728	0.5816	0.2076	0.918	31
C	16	0.7444	0.8432	0.8599	0.8272	0.6332	0.3673	0.380	41
V	17	0.7503	-	-	-	-	-	0.34	-
A	18	0.6722	0.8040	0.6559	0.7892	0.6091	0.2275	0.699	33
P	17	0.7685	0.7756	0.6137	0.8239	0.7136	0.3726	0.748	50
C	17	0.7881	0.8733	0.8871	0.8735	0.7416	0.7421	0.367	56
V	17	0.5267	-	-	-	-	-	0.44	-

* *n* = 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 R^2; *CCCP* = coefficient of conformism of a correlative prediction; *RMSE* = root mean square error; *F* = Fischer F-ratio.

3.3. Monte Carlo Optimization Using TF_{CII}

Table 3 contains the statistical characteristics of the models for logK observed in the case of the target function TF_{CII} (as seen in Equation (5)). One can see that the validation sets have different statistical characteristics (the determination coefficient ranges from 0.49 to 0.84). The average and dispersion of the determination coefficient of the validation sets are 0.68 ± 0.14. Thus, there is an improvement, compared to the situation observed in Tables 1 and 2.

Table 3. The statistical characteristics of models build up using TF_{CII} on five random splits. The best model is indicated in bold.

	n *	R^2	*CCC*	*IIC*	*CII*	Q^2	*CCCP*	*RMSE*	*F*
A	17	0.8800	0.9362	0.6567	0.9063	0.8473	0.7954	0.339	110
P	17	0.8480	0.5642	0.5567	0.8940	0.7680	0.7925	0.561	84
C	18	0.3884	0.4909	0.3401	0.7357	0.2762	0.0219	1.59	10
V	17	0.5362	-	-	-	-	-	1.32	-
A	**18**	**0.7863**	**0.8803**	**0.8867**	**0.8299**	**0.7503**	**0.1435**	**0.611**	**59**
P	**16**	**0.7802**	**0.7517**	**0.5756**	**0.8234**	**0.7380**	**−0.5862**	**1.09**	**53**
C	**18**	**0.5537**	**0.7299**	**0.5065**	**0.8456**	**0.4091**	**0.6445**	**0.633**	**17**
V	**17**	**0.8415**	**-**	**-**	**-**	**-**	**-**	**0.43**	**-**
A	17	0.8530	0.9207	0.8210	0.8682	0.8257	−0.9189	0.393	87
P	18	0.8523	0.7918	0.3440	0.8741	0.8298	0.5431	1.11	92
C	17	0.8714	0.7608	0.4424	0.8984	0.8469	0.6512	0.547	102
V	17	0.7516	-	-	-	-	-	0.81	-

Table 3. *Cont.*

	n *	R^2	*CCC*	*IIC*	*CII*	Q^2	*CCCP*	*RMSE*	*F*
A	18	0.7263	0.8415	0.8522	0.8051	0.6724	−0.2712	0.627	42
P	16	0.8237	0.8691	0.5583	0.8598	0.7841	0.6580	0.556	65
C	18	0.3674	0.4141	0.2882	0.8012	0.2145	0.1111	0.837	9
V	17	0.4929	-	-	-	-	-	0.94	-
A	18	0.9287	0.9630	0.9637	0.9439	0.9147	0.8443	0.350	208
P	18	0.7132	0.5873	0.4697	0.7954	0.6608	0.2863	2.14	40
C	16	0.8820	0.5160	0.8183	0.9037	0.8579	0.6819	1.49	105
V	17	0.8069	-	-	-	-	-	2.08	-

* *n* = 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 R^2; *CCCP* = coefficient of conformism of a correlative prediction; *RMSE* = root mean square error; *F* = Fischer F-ratio.

3.4. Monte Carlo Optimization Using TF_{CCCP}

Table 4 contains the statistical characteristics of the models for logK observed in the case of the target function TF_{CCCP} (as seen in Equation (6)). One can see that the validation sets have different but quite similar statistical characteristics (the smallest determination coefficient is 0.84 and the largest one is 0.87). The average and dispersion of the determination coefficient of the validation sets are 0.85 ± 0.01. This target function is more advanced compared to the previous ones since it incorporates several optimization steps.

Figure 1 contains the comparison of the best models observed in the cases of the considered target functions for split 2. The target function preferable according to the determination coefficient value (Table 4) of the validation set is TF_{CCCP}. The consideration of the average values confirms this.

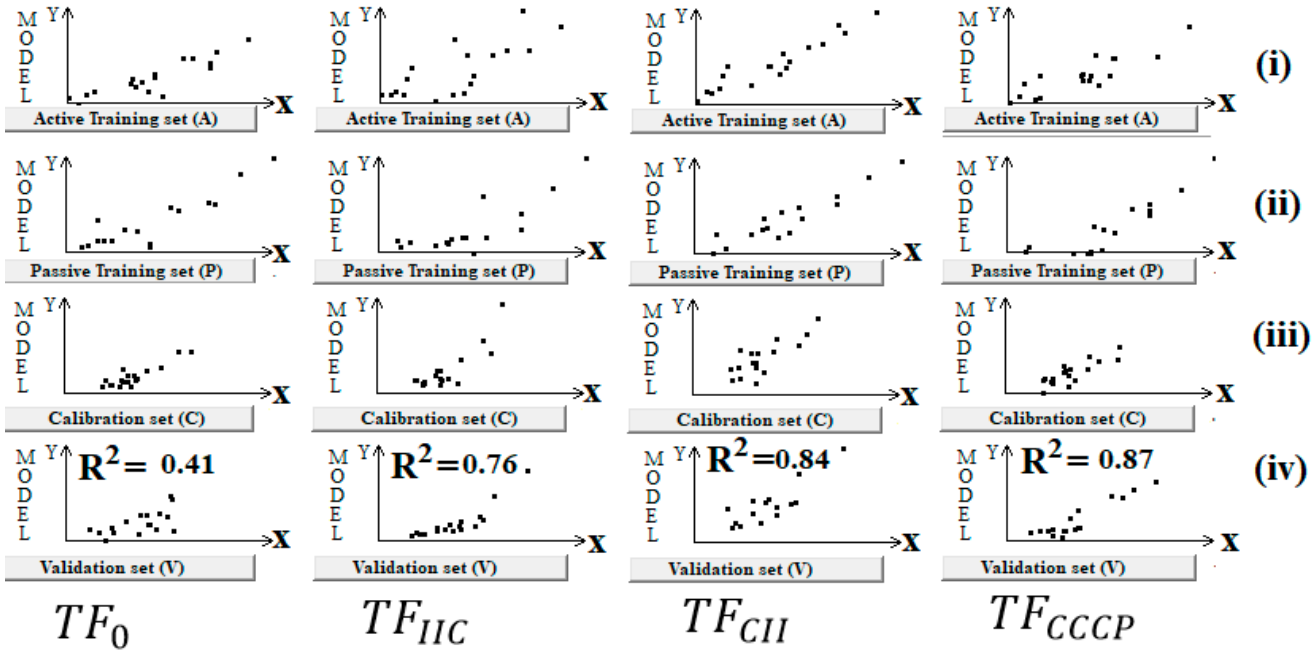

Figure 1. Comparison of the models for split 2 obtained with different target functions: TF_0, TF_{IIC}, TF_{CII}, and TF_{CCCP}. The statistical status of the models in the experiment (abscissa)—calculated model (ordinate) coordinates is presented separately for (**i**) the active training set; (**ii**) the passive training set; (**iii**) the calibration set; and (**iv**) the validation set.

Figure 1 contains the comparison of the best models observed in the cases of the considered target functions for split 2. The target function preferable according to the determination coefficient value (Table 4) of the validation set is TF_{CCCP}. It is confirmed by the average values.

Table 4. The statistical characteristics of models build up using TF_{CCCP} on five random splits. The best model is indicated in bold.

	n *	R^2	*CCC*	*IIC*	*CII*	Q^2	*CCCP*	*RMSE*	*F*
A	18	0.7775	0.8748	0.7054	0.8024	0.7430	−0.4113	0.551	56
P	18	0.6895	0.7424	0.7674	0.7852	0.6157	−0.7758	0.790	36
C	16	0.2700	0.5142	0.3941	0.5092	−1.0093	1.1379	0.801	5
V	17	0.8432	-	-	-	-	-	0.47	-
A	**18**	**0.7080**	**0.8291**	**0.8414**	**0.7762**	**0.6376**	**−0.3017**	**0.566**	**39**
P	**16**	**0.8308**	**0.8231**	**0.2762**	**0.8667**	**0.8015**	**0.2512**	**0.917**	**79**
C	**18**	**0.7013**	**0.7369**	**0.1887**	**0.8478**	**0.5838**	**0.6864**	**0.400**	**33**
V	**17**	**0.8696**	**-**	**-**	**-**	**-**	**-**	**0.46**	**-**
A	18	0.6995	0.8232	0.6691	0.7666	0.6472	0.1098	0.756	37
P	18	0.9350	0.6279	0.1702	0.9526	0.9176	0.8963	0.785	230
C	17	0.7652	0.7661	0.2206	0.8839	0.6057	0.9142	0.378	49
V	16	0.8498	-	-	-	-	-	0.77	-
A	18	0.7010	0.8242	0.5328	0.8040	0.6470	0.3868	0.766	38
P	16	0.5630	0.6925	0.5178	0.7620	0.4662	0.0715	1.06	18
C	17	0.7572	0.8614	0.7692	0.8895	−0.0759	6.7240	0.266	47
V	18	0.8668	-	-	-	-	-	0.34	-
A	17	0.7624	0.8652	0.7761	0.8021	0.7254	−0.1380	0.652	48
P	18	0.7340	0.6887	0.7656	0.8275	0.6721	0.4612	1.36	44
C	17	0.7127	0.7032	0.3162	0.8847	0.0267	2.0307	0.527	37
V	17	0.8388	-	-	-	-	-	1.40	-

* *n* = 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 R^2; *CCCP* = coefficient of conformism of the correlative prediction; *RMSE* = root mean square error; *F* = Fischer F-ratio.

The best model has been observed by considering the computational experiment obtained with the target function TF_{CCCP} (split 2), demonstrated in the Supplementary Materials section (Table S1). The model is as follows:

$$\log K = -3.193 + 0.3878 \times DCW(1,15) \quad (16)$$

Table 5 contains the correlation weights to calculate the optimal descriptor using Equation (6).

Figure 2 contains the comparison of the average and dispersion of determination coefficients observed for the models obtained with different target functions. One can see that the largest value observed for the model is obtained with the target function TF_{CCCP}. Furthermore, the approach (building up models using the target function TF_{CCCP}) is characterized by the smallest value of dispersion of the determination coefficient of validation sets.

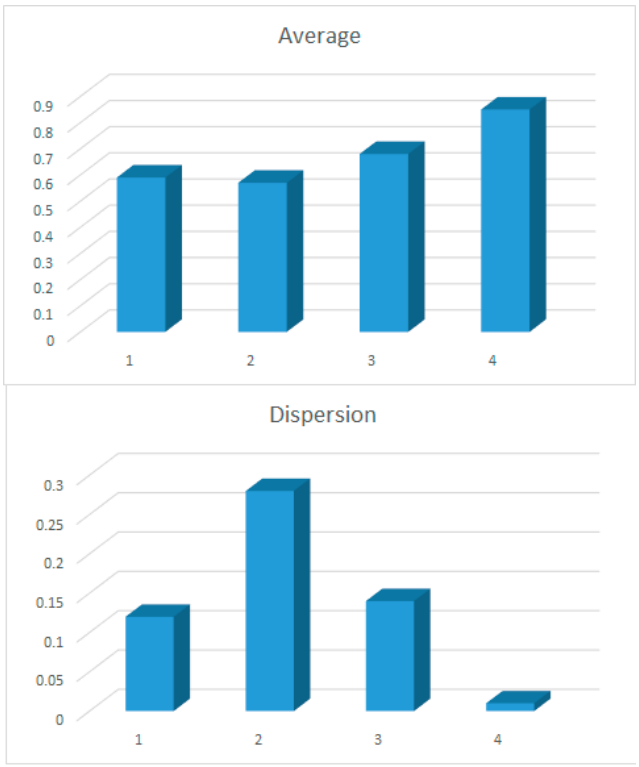

Figure 2. Statistical parameters for the models obtained with different target functions: TF_0 (1), TF_{IIC} (2), TF_{CII} (3), and TF_{CCCP} (4).

Table 5. The correlation weights of SMILES attributes (SA) were observed in the case of the Monte Carlo optimization with the target function TF_{CCCP}.

SA_k	$CW(SA_k)$ *	NA	NP	NC	DEFECT of SA_k
(...(.....	−0.4067	3	3	2	0.0104
(.....	0.1922	17	18	14	0.0051
1...(.....	0.1304	11	15	10	0.0123
1.....	−0.1853	18	18	16	0.0000
2.....	−0.2441	3	6	2	0.0379
=...(.....	−0.4736	12	12	10	0.0025
=.....	0.0237	18	18	16	0.0000
=...1.....	−0.1245	18	15	16	0.0068
=...2.....	0.0228	2	4	1	0.0456
C...(.....	0.3560	17	18	14	0.0051
C.....	−0.4398	18	18	16	0.0000
C...1.....	−0.0480	18	18	16	0.0000
C...2.....	−0.1285	3	6	2	0.0379
C... = ...	−0.4080	17	18	16	0.0022
C...C.....	0.2619	18	16	16	0.0044
Cl..(.....	0.2425	7	4	4	0.0222
Cl.....	0.2188	7	4	4	0.0222
N...(.....	−0.4308	3	4	1	0.0399
N.....	−0.1021	4	4	1	0.0355
N...1.....	−0.2286	2	0	1	0.0741

Table 5. *Cont.*

SA_k	$CW(SA_k)$ *	NA	NP	NC	DEFECT of SA_k
O...(.....	−0.1898	10	10	9	0.0005
O.....	−0.2509	10	14	11	0.0127
O... = ...	0.4619	6	6	3	0.0194
O...C.....	−0.1823	3	4	2	0.0216
[N+].....	−0.1048	2	2	2	0.0046
[O−].....	0.4524	2	2	2	0.0046

* CWs = correlation weights; NA, NP, and NC are the numbers of SMILES attributes in the active training set, passive training set, and calibration set, respectively.

3.5. Mechanistic Interpretation

The influence of SMILES atoms and their connected pairs can be estimated by running a computer experiment running a Monte Carlo optimization. As a result of stochastic optimization of the correlation weights of the specified SMILES attributes, two groups of SMILES attributes can be distinguished. First, those that have positive values of the correlation weight in all runs. Second, those that have negative values of the correlation weight in all runs. These two groups do not exhaust all possible situations, since SMILES attributes that have both positive and negative correlation weights can be observed in the specified study, depending on the split. However, for determining the mechanistic interpretation of models, such SMILES attributes are less reliable than the two groups with stable positive and negative correlation weights. Table 6 contains a collection of SMILES attributes with stable positive and stable negative weights (TF_{CCCP} is the target function for the study). It should be noted that the prevalence of the corresponding attributes in the active and passive training sets and the calibration set also needs to be taken into account.

Table 6. SMILES attributes (S_k and SS_k) promoters of increase (all correlation weights > 0) or decrease (all correlation weights < 0) for logK.

SMILES Attributes	CWs * Run 1	CWs Run 2	CWs Run 3	CWs Run 4	CWs Run 5	NA	NP	NC
1.....	0.6723	1.0064	1.7081	0.8739	1.8681	18	18	16
=...1.....	1.0086	0.5197	3.3561	2.1575	0.4241	18	15	16
C... = ...	0.6019	0.0956	0.3528	0.2506	0.2050	17	18	16
Cl..(.....	0.2058	0.1434	1.2270	0.7832	0.3002	7	4	4
2.....	1.0902	1.0887	1.3159	0.5377	0.4817	3	6	2
O...C.....	0.3976	0.6187	0.7488	0.4841	0.4969	3	4	2
=...2.....	1.0285	1.3549	2.4826	1.2312	0.9400	2	4	1
[N+].....	1.5490	1.7140	3.0811	0.7048	1.4949	2	2	2
[O−].....	0.9550	1.9835	2.2483	1.4910	1.8871	2	2	2
C...1.....	−1.5184	−1.4368	−2.8522	−1.9809	−1.1727	18	18	16
O.....	−0.1404	−0.2616	−0.2447	−0.3083	−0.4433	10	14	11
N.....	−0.3332	−0.7337	−1.2481	−0.3569	−0.5047	4	4	1

* CWs = correlation weights; NA, NP, and NC are the numbers of SMILES attributes in the active training set, passive training set, and calibration set, respectively.

3.6. Applicability Domain

The outliers for the models considered are defined via the so-called statistical defect for corresponding SMILES [20]. The total number of outliers for the best model is seven. Only one outlier occurs in the validation set. Figure 3 contains the structures of the outliers.

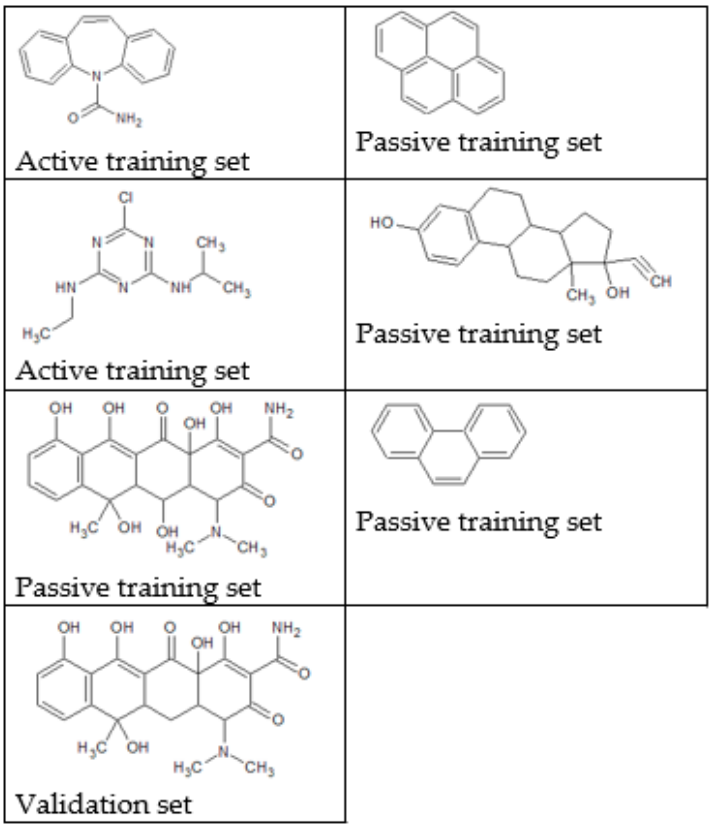

Figure 3. Structures of outliers according to the statistical defects [20].

The backside of the OECD principle, which proclaimed the need for an applicability domain, is the requirement of representativeness. Table 7 contains the numbers of active SMILES attributes together with the total quantities of SMILES attributes observed for each split. One can see that Table 7 confirms the representativeness of the considered models.

Table 7. The representativeness of SMILES attributes involved in the building up models on five splits.

Split	Number of Active SMILES Attributes	Total Number of SMILES Attributes
1	30	53
2	28	50
3	31	54
4	34	51
5	28	53

3.7. Comparison with Models logK from the Literature

There are two works where the endpoint is considered. Table 8 contains the comparison of the statistical characteristics of our models for validation sets and those in the literature. Our model has values in the highest range, and it refers to the highest number of substances used for validation.

Table 8. The comparison of the predictive potential of models suggested for logK values.

n	R^2	Method	Reference
8	0.61	Genetic algorithm	[11]
8	0.92	Radial basis function neural network	[11]
30	0.83	Support vector machines	[21]
30	0.78	Artificial Neural Networks	[21]
30	0.51	Multiple regression	[21]
17	0.87	Monte Carlo optimization	This work

4. Discussion

The proposed modeling scheme allows us to consider stochastic natural processes by means of modeling, based on available eclectic data. In particular, the adsorption of organic substances on multi-walled carbon nanotubes is a rather complex polymorphic process and it is unlikely that highly accurate and universal quantitative models will be obtained. However, qualitative models oriented to separated classes of chemicals can be used too. The approach used in this study can provide certain clues for the practice of predicting the adsorption of organic compounds on nanotubes, using exclusive data on the structure of organic molecules (represented via SMILES).

A useful feature of the approach under consideration is the ability to attract and evaluate external additional data as a basis for modifying the optimal descriptor by adding new components to the list of correlation weights. For example, it is possible to include hybrid descriptors as well as quasi-SMILES (i.e., SMILES supplemented with code-transmitting experimental conditions).

The above indicated that the hypothesis used here (the possibility of using SMILES to predict the adsorption ability of organic compounds) is confirmed by the described computational experiments.

The philosophy of technology continues in the twenty-first century thanks to the emergence of nanotechnology, which has sufficient capabilities for this, and a number of specific features that distinguish it from past technologies and require deep understanding. The evolution associated with nanotechnology is faster than what occurred with past technologies. One of the main problems of our time is the convergence of technologies. The opposite of divergence.

The terms convergence and divergence are used in various natural and humanitarian sciences. Nanotechnology acts as a "root technology" and in this, it opens new scenarios and new issues, not only technical ones, and this has occurred in the past with other new sectors, in all sciences. In particular, there is the problem of modernization of the proven and tested technique of QSPRs/QSARs analysis, providing the possibility of extending classical methods of the QSPRs/QSARs and improving their use in relation to the nanomaterials. The conceptualization of this modernization process requires further studies, intending to work on the nano-QSPRs/QSARs.

Three levels of the conceptualization of modeling of nano-reality phenomena were mentioned above. From this point of view, this study is in preparation for the second level, i.e., it proceeds towards the convergence of traditional methods of QSPR/QSAR analysis into methods of nano-QSPRs/QSARs analysis.

The evaluation of the capabilities of the three criteria of predictive potential considered here is possible only based on a large number of corresponding computational experiments. It is clear that the impact of these criteria on stochastic processes of optimization using the Monte Carlo method is different. It is shown that for some data arrays, the IIC becomes better than the CII [22], and for others, the CII becomes better than the IIC [14]. Furthermore,

this criterion is able to affect cooperation [22]. The same is expected for the *CCCP*; it can probably also be both preferable and less effective in comparison with the IIC and the CII. In any case, these criteria can be used only for sufficiently large data arrays (at least 100). It should be noted that broad prospects for the application of the discussed criteria are opened for the construction of models using quasi-SMILES [14], i.e., by taking into account codes of molecular structure together with codes of experimental conditions including the described stochastic process of the Monte Carlo optimization.

5. Conclusions

The considered concept of stochastic modeling makes models comparable to those published in the corresponding literature. In the case we addressed here, we showed a quite effective model for its ability to predict the adsorption of organic substances on multi-wall carbon tubes. The model has been obtained with the Monte Carlo algorithm, optimizing parameters using the new criterion of the predictive potential of the *CCCP*. The mechanistic interpretation and applicability domain for the developed model have been described.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/c11010007/s1, Table S1: Technical details on split 2 with target function TF_{CCCP}.

Author Contributions: Conceptualization, A.P.T., A.A.T., A.R. and E.B.; methodology, A.P.T., A.A.T., A.R. and E.B.; software, A.A.T.; validation, A.P.T., A.A.T., A.R. and E.B.; formal analysis, A.P.T.; data curation, A.P.T. and A.A.T.; writing—original draft preparation, A.P.T., A.A.T., A.R. and E.B.; writing—review and editing, A.P.T., A.A.T., A.R. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are available in the article or its Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sadiq, S.; Khan, S.; Khan, I.; Khan, A.; Humayun, M.; Wu, P.; Usman, M.; Khan, A.; Alanazi, A.F.; Bououdina, M. A critical review on metal-organic frameworks (MOFs) based nanomaterials for biomedical applications: Designing, recent trends, challenges, and prospects. *Heliyon* **2024**, *10*, e25521. [CrossRef] [PubMed]
2. Jafari, K.; Fatemi, M.H. Application of nano-quantitative structure–property relationship paradigm to develop predictive models for thermal conductivity of metal oxide-based ethylene glycol nanofluids. *J. Therm. Anal. Calorim.* **2020**, *142*, 1335–1344. [CrossRef]
3. Kumar, P.; Kumar, A.; Sindhu, J.; Lal, S. Quasi-SMILES as a basis for the development of QSPR models to predict the CO_2 capture capacity of deep eutectic solvents using correlation intensity index and consensus modelling. *Fuel* **2023**, *345*, 128237. [CrossRef]
4. Ahmadi, S.; Ketabi, S.; Qomi, M. CO_2 uptake prediction of metal-organic frameworks using quasi-SMILES and Monte Carlo optimization. *New J. Chem.* **2022**, *46*, 8827–8837. [CrossRef]
5. Jafari, K.; Fatemi, M.H. A new approach to model isobaric heat capacity and density of some nitride-based nanofluids using Monte Carlo method. *Adv. Powder Technol.* **2020**, *31*, 3018–3027. [CrossRef]
6. Azimi, A.; Ahmadi, S.; Javan, M.J.; Rouhani, M.; Mirjafary, Z. QSAR models for the ozonation of diverse volatile organic compounds at different temperatures. *RSC Adv.* **2024**, *14*, 8041–8052. [CrossRef]
7. Laudone, G.M.; Jones, K.L. A Grand Canonical Monte Carlo Simulation for the Evaluation of Pore Size Distribution of Nuclear-Grade Graphite from Kr Adsorption Isotherms. *C-J. Carbon Res.* **2023**, *9*, 86. [CrossRef]
8. Dong, C. Effective Elastic Modulus of Wavy Single-Wall Carbon Nanotubes. *C-J. Carbon Res.* **2023**, *9*, 54. [CrossRef]
9. Soave, R.; Cargnoni, F.; Trioni, M.I. Thermodynamic Stability and Electronic Properties of Graphene Nanoflakes. *C-J. Carbon Res.* **2024**, *10*, 5. [CrossRef]
10. Gautam, S.; Cole, D. Ethane-CO_2 Mixture Adsorption in Silicalite: Influence of Tortuosity and Connectivity of Pores on Selectivity. *C-J. Carbon Res.* **2023**, *9*, 116. [CrossRef]

11. Hassanzadeh, Z.; Kompany-Zareh, M.; Ghavami, R.; Gholami, S.; Malek-Khatabi, A. Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface. *J. Mol. Struct.* **2015**, *1098*, 191–198. [CrossRef]
12. Xia, X.-R.; Monteiro-Riviere, N.A.; Riviere, J.E. An index for characterization of nanomaterials in biological systems. *Nat. Nanotechnol.* **2010**, *5*, 671–675. [CrossRef] [PubMed]
13. 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. [CrossRef] [PubMed]
14. Toropov, A.A.; Toropova, A.P. Correlation intensity index: Building up models for mutagenicity of silver nanoparticles. *Sci. Total Environ.* **2020**, *737*, 139720. [CrossRef]
15. Toropova, A.P.; Toropov, A.A. The coefficient of conformism of a correlative prediction (CCCP): Building up reliable nano-QSPRs/QSARs for endpoints of nanoparticles in different experimental conditions encoded via quasi-SMILES. *Sci. Total Environ.* **2024**, *927*, 172119. [CrossRef]
16. 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. [CrossRef]
17. Mukhopadhyay, S.K.; Ahmad, M.O.; Swamy, M.N.S. ASCII-character-encoding based PPG compression for tele-monitoring system. *Biomed. Signal Process. Control* **2017**, *31*, 470–482. [CrossRef]
18. Lin, L.I.-K. Assay validation using the concordance correlation coefficient. *Biometrics* **1992**, *48*, 599–604. [CrossRef]
19. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678. [CrossRef]
20. Toropova, A.P.; Toropov, A.A. Hybrid optimal descriptors as a tool to predict skin sensitization in accordance to OECD principles. *Toxicol. Lett.* **2017**, *275*, 57–66. [CrossRef]
21. Wang, Q.; Apul, O.G.; Xuan, P.; Luoc, F.; Karanfil, T. Development of a 3D QSPR model for adsorption of aromatic compounds by carbon nanotubes: Comparison of multiple linear regression, artificial neural network and support vector machine. *RSC Adv.* **2013**, *3*, 23924–23934. [CrossRef]
22. Toropov, A.A.; Toropova, A.P. The unreliability of the reliability criteria in the estimation of QSAR for skin sensitivity: A pun or a reliable law? *Toxicol. Lett.* **2021**, *340*, 133–140. [CrossRef]