inorganics

MDPI

Article

QSPR and Nano-QSPR: Which One Is Common? The Case of Fullerenes Solubility

Alla P. Toropova [1,*], Andrey A. Toropov [1] and Natalja Fjodorova [2]

[1] Laboratory of Environmental Chemistry and Toxicology, Istituto Di Ricerche Farmacologiche Mario Negri, IRCCS, Via Mario Negri, 2, 20156 Milano, Italy; andrey.toropov@marionegri.it

[2] Laboratory for Chemoinformatics, Theory Department, National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia; natalja.fjodorova@ki.si

* Correspondence: alla.toropova@marionegri.it; Tel.: +39-02-3901-4595

Abstract: Background: The system of self-consistent models is an attempt to develop a tool to assess the predictive potential of various approaches by considering a group of random distributions of available data into training and validation sets. Considering many different splits is more informative than considering a single model. **Methods**: Models studied here build up for solubility of fullerenes C60 and C70 in different organic solvents using so-called quasi-SMILES, which contain traditional simplified molecular input-line entry systems (SMILES) incorporated with codes that reflect the presence of C60 and C70. In addition, the fragments of local symmetry (FLS) in quasi-SMILES are applied to improve the solubility's predictive potential (expressed via mole fraction at 298′K) models. **Results**: Several versions of the Monte Carlo procedure are studied. The use of the fragments of local symmetry along with a special vector of the ideality of correlation improves the predictive potential of the models. The average value of the determination coefficient on the validation sets is equal to 0.9255 ± 0.0163. **Conclusions**: The comparison of different manners of the Monte Carlo optimization of the correlation weights has shown that the best predictive potential was observed for models where both fragments of local symmetry and the vector of the ideality of correlation were applied.

Keywords: nano-QSPR; fullerene; solubility; validation; system of self-consistent models; Monte Carlo method; CORAL software

check for updates

Citation: Toropova, A.P.; Toropov, A.A.; Fjodorova, N. QSPR and Nano-QSPR: Which One Is Common? The Case of Fullerenes Solubility. *Inorganics* **2023**, *11*, 344. https://doi.org/10.3390/inorganics11080344

Academic Editors: Zdeněk Slanina, Kyriakos Porfyrakis, Nikos Tagmatarchis and Alexander Novikov

Received: 30 June 2023
Revised: 4 August 2023
Accepted: 18 August 2023
Published: 21 August 2023

1. Introduction

Like the world of real material movements, in which all events that are visible and tangible to us in everyday life, such as wind, rain, and the movement of clouds, take place, there is a world of probabilistic actions, accidents, and tendencies that influence each other. However, these are not visible and not tangible to us. Perhaps quantitative structure-property/activity relationships (QSPR/QSAR) allow one to look into this world of accidents and trends that affect each other.

There is no mysticism here, but the phenomena occurring in such a space are not always described ideally and reliably. In other words, encountering situations that defy logic is possible. For example, the quality of calculations (models) can be affected by the collection of substances, which are available in the database, as well as priorities and criteria selected in the software used for QSPR/QSAR simulation.

However, in any case, it remains an indisputable axiom that models of random events are knowledge only when they are understandable and allow the possibility of verification by establishing and confirming their reproducibility.

Traditional QSPR were initially based on molecular structure [1–3] and later became involved in an extended set of descriptors that included information not only on molecular architecture but also on the magnitudes of various physicochemical properties [4]. This would be a good fit for nano-QSPR, if not for the lack of a clear relationship between the molecular structure and the pleasant/useful/dangerous nano-physicochemical properties of the respective nanomaterials [5–11]. It cannot be said that the molecular architecture does not affect the physicochemical properties of nano-substances in any way, but this influence is very sophisticated for nano-substances. That is, if, for small organic molecules, the modifications of the geometry/topology arrangement of a pair of atoms necessarily change the physicochemical parameters, then for fullerenes, and even more so, for multilayer nanotubes, changes in the arrangement of a pair of substituent atoms are very difficult to establish and/or measure experimentally. Naturally, simple homologous series, which formed the basis of the first QSPR experiments of organic compounds [1–3] for nano-substances, are extremely rare due to the high cost and weak motivation for experimental work designed to provide the corresponding numerical data on the physicochemical parameters of homologous series of fullerenes.

How do we obtain information on all promised abilities to apply nanomaterials? How do we select and use the unique potentials of nanomaterials? Hints, hypotheses, and intuition must be transformed into knowledge.

Can a model be knowledge?

Knowledge is a tool. It is preferable if knowledge is convenient for use in solving practical problems. Consequently, a model can be a way to reach knowledge when all excess is removed from the model and only the necessary remains. Nothing is surprising in that a brief instruction may be more useful than an excessively detailed one. That is why most researchers profess the principle that "to understand is to simplify".

Taking into account the absence of large databases on various nanomaterials and the availability of sufficiently large arrays of experimental data on the interaction of individual nanomaterials with different organic substances (for example, with solvents [12]), one should look for the possibility of constructing models of the behaviour of nanomaterials in interaction with "traditional" organic substances.

In the case of the QSPR study of solubility C60- and C70-fullerenes [12], the traditional paradigm of QSPR/QSAR simulation is represented as

$$S = F\ (M) \tag{1}$$

This is maybe extended as

$$S = F\ (M,\ Fullerene) \tag{2}$$

where S = solubility, F = mathematical function, and M = molecular structure.

The transition from the model expressed by Equation (1) to the model expressed by Equation (2) is essentially a transition from traditional QSPR to nano-QSPR.

It should be noted that the model expressed by Equation (2) must (as well as the model expressed by Equation (1)) comply with the requirements for the QSPR formulated as well-known OECD principles [13]:

1. A defined endpoint (including experimental protocol);
2. An unambiguous algorithm;
3. A defined domain of applicability;
4. Appropriate measures of goodness-of-fit, robustness, and predictive power;
5. A mechanistic interpretation, when it is possible.

One can use these principles for nano-QSPR expressed by Equation (2). Can the OECD principles be improved? Latent attempts to do this can be seen in many studies [14–21].

The approach considered here is that each object (solvent = SMILES, fullerene = [C60] or [C70]) is represented by a character string. The program divides the symbols into special groups, for which the so-called correlation weights (some coefficients) are found. The descriptor for each object is the sum of the correlation weights. The Monte Carlo method is used to find such correlation weights that provide the maximum value of the objective function. This optimization is carried out on the basis of partitioning the available data into special subsets: an active training set (its task is to develop a model), a passive training set (its task is to check the objectivity of the current model), a calibration set (its task is to detect the start of the overtraining), and the validation set to assess the predictive potential of the final model.

2. Results

The three schemes for constructing models of the solubility of fullerenes C60 and C70 in organic solvents were evaluated.

First Scheme

The models were constructed using new components of the model, which are named correlation weights of fragments of local symmetry (FLS). However, the Monte Carlo optimization of the extended set of quasi-SMILES codes was planned without using the correlation idealization vector, which has two components: the index of ideality of correlation (*IIC*) and the correlation intensity index (*CII*).

Second Scheme

The models were constructed via the Monte Carlo optimization of the set of quasi-SMILES codes, without correlation weights of FLS, using the above-mentioned vector of the ideality of correlation.

Third Scheme

The models were built using the Monte Carlo optimization of an extended list of the correlation weights, including FLS, along with using the vector of the ideality of correlation.

Figure 1 contains the graphical representation of the simulation processes observed for the three schemes in the case of split 1. One can see that the third scheme seems to have the most perspective.

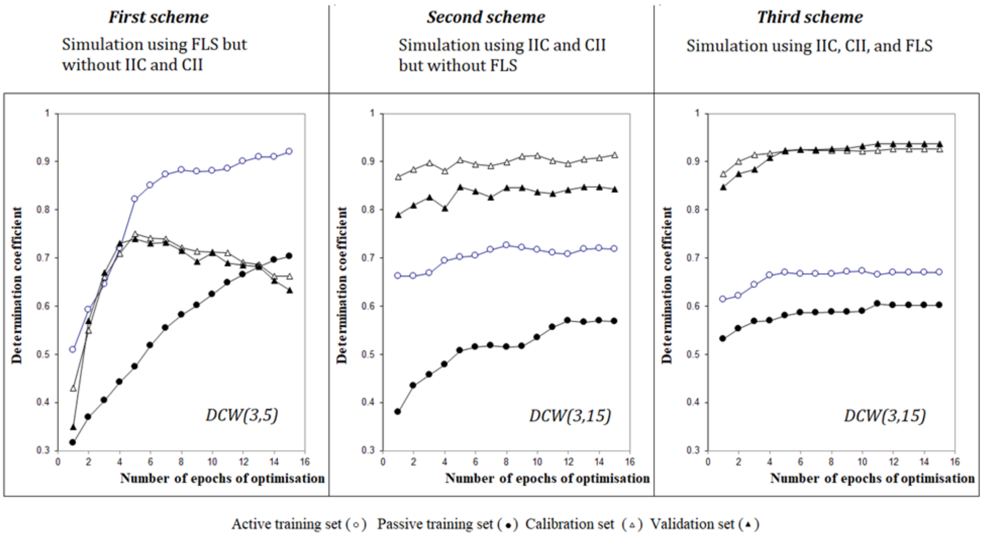


Figure 1. The histories of the Monte Carlo optimization using different target functions.

In addition, one can see that the practically reasoned optimal descriptor for the first scheme is *DCW*(*3*,*5*). In contrast, the preferable optimal descriptor for the second and third schemes is *DCW*(*3*,*15*).

According to the principle "QSAR/QSPR is a random event", it is necessary to study the statistical quality of models observed under different distributions in the training set (here, the set is structured into three components: active training, passive training, and calibration sets). Table 1 contains the results of applying the first scheme on splits 1–10.

One can see the determination coefficients for the active training, passive training, and calibration sets as a rule equivalent or even a little larger than the determination coefficient of the validation set. However, in the case of continued optimization, the determination coefficients for the active and passive training samples will increase. In contrast, for the external control sample, the determination coefficient will decrease (Figure 1).

Table 1. The statistical characteristics of the models were built without using *IIC* and *CII* but using correlation weights of FLS (first scheme). The average determination coefficient of the validation sets for the observed ten models follows 0.7742 ± 0.0713.

Split	Set *	n	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*MAE*	F
1	A	52	0.8276	0.9056	0.7797	0.9089	0.8119	0.716	0.586	240
	P	53	0.7569	0.8687	0.8561	0.8812	0.7362	0.630	0.492	159
	C	51	0.7428	0.8020	0.4311	0.8367	0.7231	0.859	0.667	142
	V	50	0.7406	-	-	-	-	0.794	-	-
2	A	51	0.7236	0.8396	0.8179	0.8495	0.6967	0.879	0.718	128
	P	51	0.7216	0.8153	0.6959	0.8431	0.6794	0.790	0.621	127
	C	51	0.8579	0.9084	0.9109	0.9175	0.8395	0.575	0.437	296
	V	53	0.8348	-	-	-	-	0.618	-	-
3	A	52	0.8036	0.8911	0.8301	0.8830	0.7845	0.654	0.498	205
	P	50	0.6151	0.7085	0.7008	0.7917	0.5817	0.983	0.800	77
	C	52	0.8910	0.9250	0.4079	0.9232	0.8839	0.460	0.334	409
	V	52	0.8762	-	-	-	-	0.458	-	-
4	A	50	0.7992	0.8884	0.7615	0.8647	0.7820	0.715	0.504	191
	P	53	0.7990	0.8811	0.7767	0.8669	0.7838	0.724	0.557	203
	C	50	0.8139	0.8192	0.4604	0.8856	0.7945	0.734	0.547	210
	V	53	0.8328	-	-	-	-	0.645	-	-
5	A	50	0.7798	0.8762	0.8151	0.8745	0.7592	0.715	0.575	170
	P	52	0.7798	0.8478	0.5541	0.8771	0.7549	0.889	0.686	177
	C	52	0.8522	0.8985	0.9073	0.9087	0.8405	0.524	0.381	288
	V	52	0.8411	-	-	-	-	0.585	-	-
6	A	50	0.8871	0.9402	0.8694	0.9299	0.8761	0.517	0.403	377
	P	50	0.8862	0.9293	0.6925	0.9328	0.8760	0.614	0.503	374
	C	53	0.6340	0.6938	0.7221	0.8138	0.5901	1.04	0.770	88
	V	53	0.6686	-	-	-	-	1.182	-	-
7	A	50	0.8230	0.9029	0.8374	0.8807	0.8022	0.657	0.480	223
	P	53	0.8076	0.8959	0.8748	0.8847	0.7896	0.678	0.530	214
	C	50	0.8283	0.8610	0.7257	0.8999	0.8110	0.684	0.544	232
	V	53	0.6583	-	-	-	-	0.760	-	-
8	A	51	0.7471	0.8553	0.6557	0.8584	0.7253	0.799	0.594	145
	P	52	0.7199	0.8250	0.7774	0.8430	0.6965	0.836	0.613	129
	C	51	0.8100	0.8941	0.8478	0.8945	0.7947	0.497	0.416	209
	V	52	0.7854	-	-	-	-	0.583	-	-
9	A	50	0.8187	0.9003	0.7109	0.8917	0.8003	0.717	0.541	217
	P	53	0.8068	0.8897	0.8728	0.8839	0.7897	0.730	0.589	213
	C	53	0.6486	0.7302	0.7753	0.7840	0.6130	0.917	0.678	94
	V	50	0.7649	-	-	-	-	0.832	-	-
10	A	53	0.7235	0.8396	0.7039	0.8572	0.6981	0.875	0.708	133
	P	51	0.6071	0.7545	0.5784	0.7947	0.5760	0.841	0.629	76
	C	51	0.7786	0.8462	0.6447	0.8545	0.7643	0.628	0.486	172
	V	51	0.7690	-	-	-	-	0.658	-	-

* A, P, C, and V denote active training, passive training, calibration, and validation sets, respectively.

The second scheme (Table 2) is characterized by a significant decrease in the statistical quality for the active and passive training sets, accompanied by a noticeable increase in the coefficient of determination for the validation set. This confirms the observed influence of *IIC* and *CII* described in the literature [18]; *IIC* and *CII* improve the statistical quality of the

QSPR/QSAR models for the validation set, but to the detriment of the statistical quality of the model for the training set.

Table 2. The statistical characteristics of the models were obtained using *IIC* and *CII* without correlation weights of fragments of local symmetry (second scheme). The average determination coefficient of the validation sets for the observed ten models follows 0.8832 ± 0.0273.

Split	**Set ***	*n*	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*MAE*	*F*
1	A	52	0.6643	0.7983	0.6986	0.8428	0.6323	0.984	0.822	99
	P	53	0.5997	0.7187	0.6328	0.7861	0.5688	0.958	0.805	76
	C	51	0.9023	0.9498	0.9496	0.9437	0.8942	0.365	0.293	453
	V	50	0.9075	-	-	-	-	0.342	-	-
2	A	51	0.6622	0.7968	0.7825	0.8162	0.6314	0.972	0.824	96
	P	51	0.5412	0.6758	0.3856	0.8273	0.4867	1.09	0.932	58
	C	51	0.9128	0.9445	0.9548	0.9498	0.8974	0.435	0.327	513
	V	53	0.9185	-	-	-	-	0.416	-	-
3	A	52	0.6182	0.7641	0.7863	0.8035	0.5877	0.912	0.748	81
	P	50	0.4222	0.5886	0.5358	0.7507	0.3527	1.17	0.997	35
	C	52	0.9144	0.9398	0.9544	0.9451	0.9080	0.357	0.285	534
	V	52	0.8822	-	-	-	-	0.399	-	-
4	A	50	0.5512	0.7107	0.6853	0.7685	0.4975	1.07	0.830	59
	P	53	0.6112	0.6850	0.6404	0.7733	0.5791	1.04	0.879	80
	C	50	0.7550	0.8602	0.8688	0.8762	0.7195	0.521	0.384	148
	V	53	0.8491	-	-	-	-	0.411	-	-
5	A	50	0.5951	0.7462	0.7714	0.8024	0.5615	0.970	0.809	71
	P	52	0.5972	0.7572	0.6266	0.7966	0.5627	1.02	0.817	74
	C	52	0.8523	0.9209	0.9202	0.9277	0.8311	0.395	0.320	288
	V	52	0.8816	-	-	-	-	0.404	-	-
6	A	50	0.3899	0.5611	0.4522	0.7307	0.3218	1.20	0.966	31
	P	50	0.6585	0.6472	0.6957	0.8256	0.6253	1.14	0.987	93
	C	53	0.7680	0.8347	0.8512	0.8634	0.6777	0.491	0.392	169
	V	53	0.8654	-	-	-	-	0.387	-	-
7	A	50	0.6543	0.7910	0.6890	0.8271	0.6215	0.918	0.759	91
	P	53	0.5105	0.7100	0.6265	0.7749	0.4656	1.11	0.916	53
	C	50	0.8562	0.9250	0.9244	0.9157	0.8442	0.418	0.326	286
	V	53	0.8619	-	-	-	-	0.389	-	-
8	A	51	0.6199	0.7654	0.7571	0.8057	0.5836	0.980	0.788	80
	P	52	0.5475	0.6955	0.6491	0.7697	0.5116	1.05	0.843	60
	C	51	0.8821	0.9381	0.9370	0.9386	0.8693	0.353	0.288	367
	V	52	0.8455	-	-	-	-	0.441	-	-
9	A	50	0.5886	0.7410	0.5556	0.7920	0.5543	1.08	0.927	69
	P	53	0.5540	0.7425	0.7153	0.7762	0.5145	1.05	0.871	63
	C	53	0.8186	0.9038	0.9041	0.9014	0.8035	0.418	0.333	230
	V	50	0.8894	-	-	-	-	0.370	-	-
10	A	53	0.6650	0.7988	0.6252	0.8247	0.6322	0.963	0.768	101
	P	51	0.4433	0.6448	0.6243	0.7746	0.4035	0.995	0.807	39
	C	51	0.8538	0.9129	0.9237	0.9243	0.8407	0.463	0.363	286
	V	51	0.9306	-	-	-	-	0.368	-	-

* A, P, C, and V denote active training, passive training, calibration, and validation sets, respectively.

The statistical quality, as well as the general logic of the models obtained using the third scheme (Table 3) are very similar, but not identical, concerning the results obtained using the second scheme.

Table 3. The statistical characteristics of the models were obtained using *IIC*, *CII*, and correlation weights of FLS (third scheme). The average determination coefficient of the validation sets for the observed ten models follows 0.9170 ± 0.0117.

Split	Set *	*n*	R^2	*CCC*	*IIC*	*CII*	Q^2	*RMSE*	*MAE*	*F*
1	A	52	0.5733	0.7288	0.7571	0.8361	0.5278	1.13	0.934	67
	P	53	0.4732	0.6379	0.4299	0.7750	0.4355	1.01	0.803	46
	C	51	0.9179	0.9456	0.9579	0.9566	0.9096	0.355	0.283	548
	V	50	0.9365	-	-	-	-	0.306	-	-
2	A	51	0.7079	0.8289	0.7479	0.8333	0.6823	0.904	0.743	119
	P	51	0.5492	0.6832	0.3766	0.8277	0.4973	1.07	0.895	60
	C	51	0.8844	0.9319	0.9402	0.9490	0.8711	0.472	0.387	375
	V	53	0.9115	-	-	-	-	0.447	-	-
3	A	52	0.8214	0.9020	0.7769	0.8947	0.8040	0.624	0.492	230
	P	50	0.5127	0.6948	0.6822	0.7602	0.4599	1.08	0.887	51
	C	52	0.9139	0.9460	0.9555	0.9456	0.9069	0.393	0.320	531
	V	52	0.9108	-	-	-	-	0.408	-	-
4	A	50	0.7237	0.8397	0.8507	0.8341	0.6961	0.838	0.632	126
	P	53	0.7083	0.8106	0.6961	0.8228	0.6859	0.887	0.684	124
	C	50	0.8863	0.8999	0.9380	0.9355	0.8765	0.498	0.401	374
	V	53	0.9056	-	-	-	-	0.360	-	-
5	A	50	0.6889	0.8158	0.8300	0.8392	0.6589	0.850	0.719	106
	P	52	0.7014	0.8092	0.6337	0.8391	0.6740	0.981	0.823	117
	C	52	0.9383	0.9623	0.9681	0.9623	0.9316	0.298	0.228	761
	V	52	0.9322	-	-	-	-	0.360	-	-
6	A	50	0.5614	0.7191	0.7492	0.8023	0.5244	1.02	0.850	61
	P	50	0.7684	0.7680	0.8519	0.8535	0.7503	0.955	0.828	159
	C	53	0.8415	0.9132	0.9165	0.8969	0.8292	0.416	0.337	271
	V	53	0.9160	-	-	-	-	0.318	-	-
7	A	50	0.7424	0.8522	0.7954	0.8554	0.7181	0.792	0.644	138
	P	53	0.6792	0.8182	0.7000	0.8529	0.6447	0.882	0.725	108
	C	50	0.8920	0.9362	0.9425	0.9259	0.8840	0.410	0.327	397
	V	53	0.9072	-	-	-	-	0.357	-	-
8	A	51	0.6460	0.7849	0.7144	0.8109	0.6122	0.945	0.755	89
	P	52	0.6078	0.7413	0.7690	0.7930	0.5763	0.984	0.811	77
	C	51	0.8807	0.9369	0.9368	0.9400	0.8644	0.371	0.291	362
	V	52	0.9041	-	-	-	-	0.324	-	-
9	A	50	0.7199	0.8372	0.6144	0.8507	0.6955	0.891	0.709	123
	P	53	0.6734	0.8159	0.5683	0.8307	0.6442	0.902	0.748	105
	C	53	0.8874	0.9304	0.9406	0.9293	0.8794	0.385	0.320	402
	V	50	0.9337	-	-	-	-	0.396	-	-
10	A	53	0.7013	0.8244	0.7477	0.8629	0.6707	0.909	0.749	120
	P	51	0.4248	0.6333	0.4521	0.7334	0.3807	1.07	0.836	36
	C	51	0.8766	0.9237	0.9342	0.9362	0.8653	0.429	0.345	348
	V	51	0.9122	-	-	-	-	0.403	-	-

* A, P, C, and V denote active training, passive training, calibration, and validation sets, respectively.

Figure 2 contains the graphical representations of the models observed in the cases of applying second and third schemes for split 1.

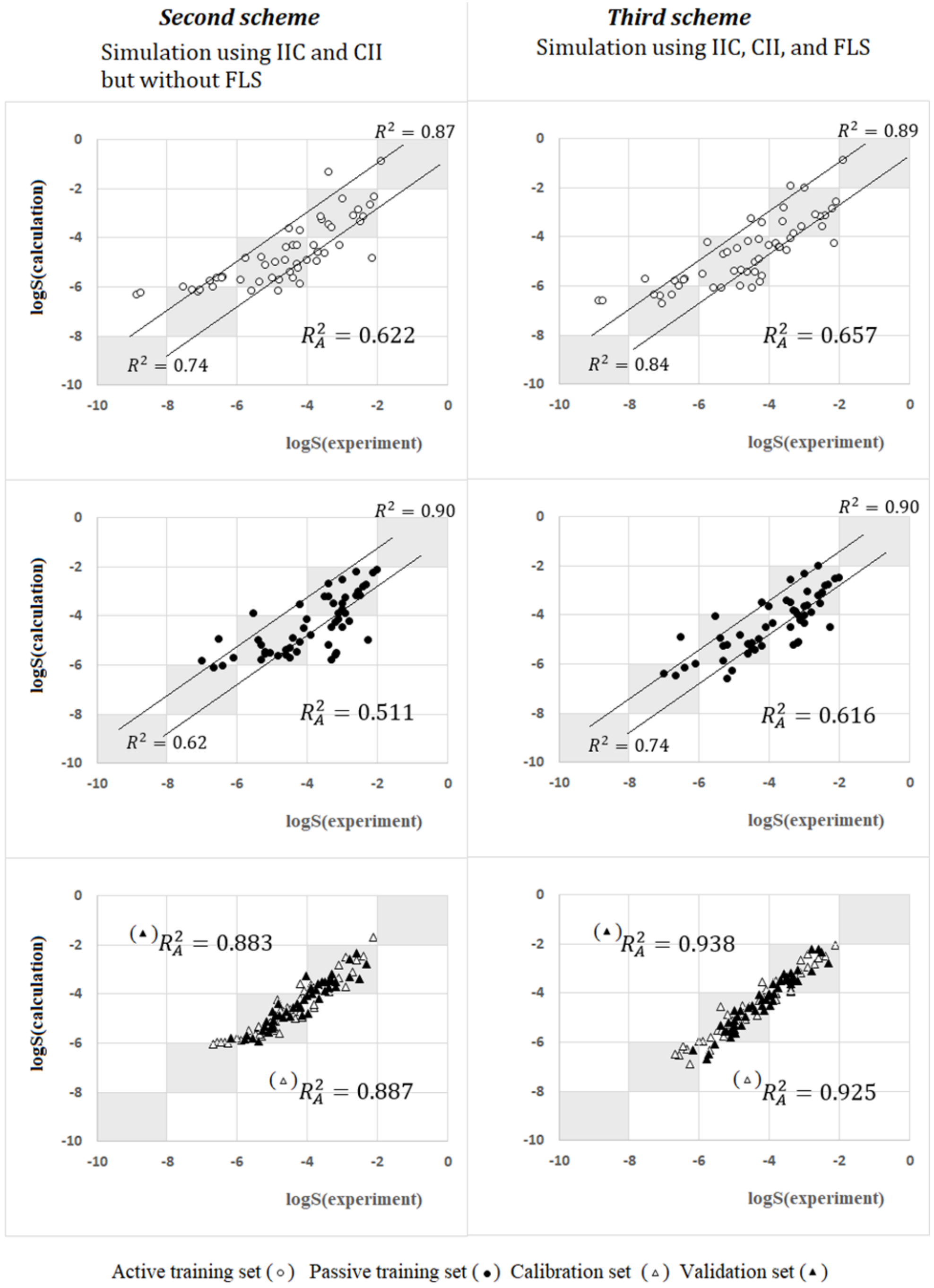


Figure 2. $R^2{}_A$ indicates all the points in the coordinates "experiment—calculation"; R^2 "without A" denote the points of the upper and lower clusters, and finally for the calibration and validation sets, the corresponding values are marked with black or uncoloured triangles (bottom of the figure).

Figure 2 shows an example of the models obtained using the second and third schemes for split 1. It should be noted that despite the statistical quality of the model for the active and passive training sets being low, these sets contain two latent correlations (Figure 2). Apparently, this is the effect of exposure to the vector of the ideality of correlation. Analogical pairs of correlations were observed in computer experiments described in the literature [20,21]. Figure 2 indicates that latent correlations on active and passive training sets are statistically more significant than total correlations on these sets.

It is under these circumstances that the problem arises regarding how to distinguish between the two approaches (second and third schemes). Which approach is more efficient, more precise, and more reliable?

Figure 1 shows some improvement in the statistical quality of the model for the case of the third scheme compared with the results observed in the case of the second scheme. However, it is related to split 1. Will this conclusion/hypothesis be true for splits 2, 3, . . ., 10?

2.1. System of Self-Consistent Models Observed for the Second Scheme

Table 4 contains the test results of the predictive potential of models with external validation sets that did not involve quasi-SMILES in constructing the tested models.

Table 4. System of self-consistent models built without the correlation weights of FLS.

		s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**	**s8**	**s9**	**s10**	$\overline{x_*}$	Δx_*
m1 *	N_v^*		12	17	21	19	19	19	23	21	22	19.2	3.1
	R_*^2		0.8818	0.8608	0.9199	0.9561	0.9642	0.8657	0.8844	0.9447	0.9145	0.9102	0.0369
m2	N_v^*	12		16	22	23	20	23	20	21	18	19.4	3.4
	R_*^2	0.9473		0.9024	0.9474	0.9656	0.8744	0.8915	0.9470	0.9494	0.9414	0.9296	0.0298
m3	N_v^*	17	16		19	17	16	20	19	20	19	18.1	1.5
	R_*^2	0.8358	0.8952		0.8289	0.8508	0.9573	0.8919	0.9064	0.9349	0.9239	0.8917	0.0424
m4	N_v^*	21	22	19		27	21	21	22	22	25	22.2	2.3
	R_*^2	0.7574	0.8989	0.7975		0.8782	0.8856	0.7983	0.8812	0.8742	0.8730	0.8493	0.0478
m5	N_v^*	19	23	17	27		18	20	19	17	16	19.6	3.3
	R_*^2	0.8055	0.9305	0.9025	0.9158		0.9459	0.9295	0.9191	0.9483	0.9617	0.9176	0.0432
m6	N_v^*	19	20	16	21	18		23	23	28	22	21.1	3.3
	R_*^2	0.8914	0.7225	0.8831	0.7803	0.8665		0.8635	0.9043	0.8815	0.9276	0.8577	0.0264
m7	N_v^*	19	23	20	21	20	23		20	26	24	21.8	2.2
	R_*^2	0.8910	0.8761	0.8585	0.8896	0.9050	0.9094		0.9390	0.9108	0.9194	0.8999	0.0226
m8	N_v^*	23	20	19	22	19	23	20		24	18	20.9	2.0
	R_*^2	0.8662	0.9003	0.9158	0.9358	0.8875	0.9022	0.8981		0.9171	0.9099	0.9037	0.0186
m9	N_v^*	21	21	20	22	17	28	26	24		24	22.6	3.1
	R_*^2	0.8469	0.9132	0.8995	0.8551	0.9176	0.9000	0.8932	0.9033		0.9015	0.8923	0.0232
m10	N_v^*	22	18	19	25	16	22	24	18	24		20.9	3.0
	R_*^2	0.9130	0.9226	0.9457	0.9273	0.9454	0.9498	0.9159	0.9614	0.9481		0.9366	0.0162

* m1–m10 denote the models from 1 to 10; s1–s10 denote the splits from 1 to 10; $\overline{x_*}$ is the average value of n* or R^2v*; Δx_* is the dispersion value of n* or R^2v*.

It can be seen that the results of applying the models to different test sets after removing the quasi-SMILES participating in the construction of the corresponding models are far from being the same. However, in all cases, there is a good predictive potential. The average value of determination coefficients for external validation sets is $\overline{R_v^2}$ = 0.8989 ± 0.0267.

2.2. System of Self-Consistent Models Observed for the Third Scheme

Table 5 contains the test results of the predictive potential of models with external validation sets that did not involve quasi-SMILES in the construction of the tested models.

Table 5. System of self-consistent models built using the correlation weights of FLS.

		s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**	**s8**	**s9**	**s10**	$\overline{x_*}$	Δx_*
m1 *	N_v^*		12	17	21	19	19	19	23	21	22	19.2	3.1
	R_*^2		0.9481	0.9025	0.9381	0.9539	0.9619	0.9384	0.9491	0.9483	0.9291	0.9410	0.0163
m2	N_v^*	12		16	22	23	20	23	20	21	18	19.4	3.4
	R_*^2	0.9176		0.9380	0.9414	0.9520	0.8814	0.8853	0.9482	0.9635	0.9168	0.9271	0.0273
m3	N_v^*	17	16		19	17	16	20	19	20	19	18.1	1.5
	R_*^2	0.8514	0.9264		0.9147	0.9196	0.9461	0.8722	0.9098	0.9123	0.9587	0.9123	0.0314
m4	N_v^*	21	22	19		27	21	21	22	22	25	22.2	2.3
	R_*^2	0.8856	0.9089	0.9185		0.9190	0.8723	0.8395	0.9201	0.9369	0.9325	0.9037	0.0300
m5	N_v^*	19	23	17	27		18	20	19	17	16	19.6	3.3
	R_*^2	0.9386	0.9539	0.9325	0.9397		0.9144	0.9501	0.9501	0.9492	0.9654	0.9438	0.0138
m6	N_v^*	19	20	16	21	18		23	23	28	22	21.1	3.3
	R_*^2	0.9619	0.8626	0.9248	0.9009	0.9391		0.9051	0.9227	0.9227	0.9349	0.9194	0.0264
m7	N_v^*	19	23	20	21	20	23		20	26	24	21.8	2.2
	R_*^2	0.8412	0.9526	0.8420	0.8539	0.9322	0.9185		0.9390	0.8842	0.9074	0.8968	0.0406
m8	N_v^*	23	20	19	22	19	23	20		24	18	20.9	2.0
	R_*^2	0.9117	0.9259	0.9402	0.9468	0.9780	0.9530	0.9364		0.9731	0.9457	0.9456	0.0198
m9	N_v^*	21	21	20	22	17	28	26	24		24	22.6	3.1
	R_*^2	0.9250	0.9553	0.9367	0.9450	0.9522	0.9290	0.9136	0.9467		0.9463	0.9389	0.0131
m10	N_v^*	22	18	19	25	16	22	24	18	24		20.9	3.0
	R_*^2	0.9136	0.9302	0.9430	0.9418	0.9288	0.8891	0.9137	0.9544	0.9201		0.9261	0.0185

* m1–m10 denote the models from 1 to 10; s1–s10 denote the splits from 1 to 10; $\overline{x_*}$ is the average value of n^* or R^2v^*; Δx_* is the dispersion value of n^* or R^2v^*.

It can be seen again that the results of applying the models to different test sets after removing the quasi-SMILES from the construction of the corresponding models are far from being the same. However, in all cases, there is a good predictive potential. The average value of determination coefficients for external validation sets is $\overline{R_v^2}$ = 0.9255 ± 0.0163.

2.3. The Comparison of Second and Third Schemes

The predictive potential of models built using the third scheme is better than that of models built using the second scheme. The dispersion in the determination coefficient values for the third scheme is less than one compared to models obtained using the second scheme.

Figure 3 shows the difference in the predictive potential of models obtained using the second and third schemes. One can see the preferable predictive potential for the second scheme for splits #2, #7, and #10. However, all other splits demonstrate the advantage of using the third scheme.

2.4. What Do QSAR/QSPR and Nano-QSAR/QSPR Have in Common?

First, QSAR/QSPR and nano-QSAR/QSPR are random events.

Second, the predictive potential in both cases can change markedly depending on the distribution of available data into training and validation sets.

Third, both QSAR/QSPR and nano-QSAR/QSPR cannot replace a natural experiment in measuring the values of various "usual" and nano-endpoints.

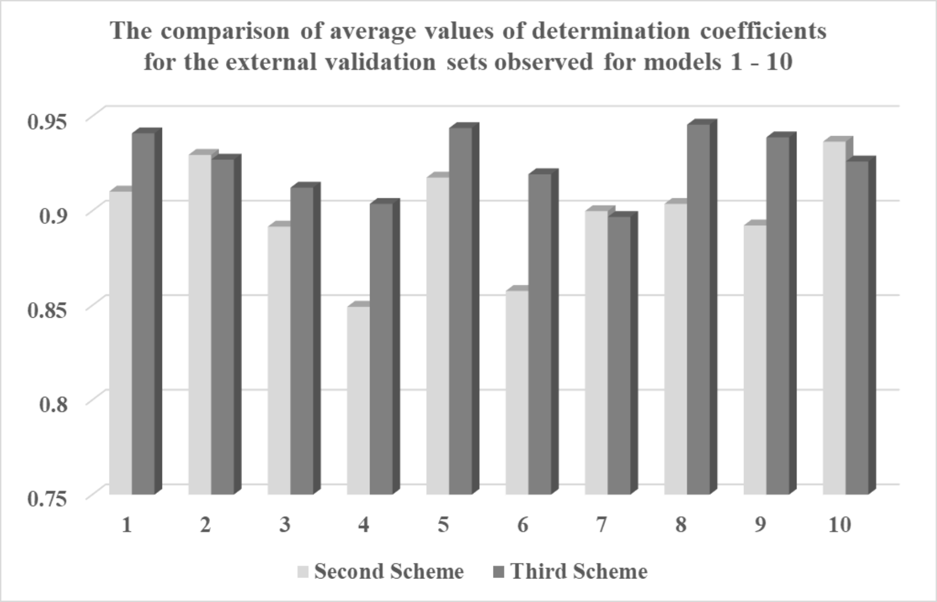


Figure 3. The comparison of the predictive potential of models 1–10 in the cases of applying the second and the third schemes.

3. Discussion

The principle "QSAR/QSPR is a random event" is confirmed in the results obtained in this study: completely homogeneous distributions in the training and control subsystems, for the same approach of the simulation of solubility fullerenes in organic solvents, provide different values for the statistical characteristics of the models (Tables 1–3).

For the proposed approach, which provides a certain "mathematical expectation" for the models obtained using the second and third schemes, it becomes possible to compare the average values, based on which it is possible to put forward a fairly reasonable hypothesis that the third scheme provides the best models compared to the models obtained using the second scheme. It is appropriate to note that the reliable criteria for the quality of models are not only the average values of the coefficients of determination but also their variances, which was observed in previous studies where systems of self-consistent models were used [20,21].

The FLS described here may not be a universal tool for developing arbitrary models, but it is only a technique that has proven successful for this task (i.e., for developing a model for the solubility of fullerenes in organic solvents). However, the vector of the ideality of correlations (or maybe the *IIC* and the *CII*, separately) perhaps can be recognized as useful and versatile tools for testing, and maybe even for improving, the predictive potential of traditional QSAR/QSPR and nano-QSAR/QSPR.

In this study, a quite simple version of quasi-SMILES has been applied to develop the models. However, one can easily extend the list of codes for quasi-SMILES to express more detailed and complex experimental conditions. In other words, one can hope that the quasi-SMILES serve as a language of communication between "classic" experimentalists who study nanomaterials and developers of nano-QSAR/QSPR models. A certain trend towards recognizing this language and even some experience in the practical use of this language have already been outlined [22].

4. Materials and Methods

4.1. Data

The experimental solubility values of C60 and C70 fullerenes in diverse solvents were reported in mole fraction determined at 298 K [12]. Table 6 contains the list of pairs of duplicates observed in [12]. Of each pair of duplicates, only one was left for further analysis.

After this removal, 206 quasi-SMILES representing various pairs of fullerenes (C60 or C70) and solvents were used for further computational experiments.

Table 6. The list of duplicated quasi-SMILES observed in [12].

CAS of Solvent	Quasi-SMILES	Mole Fraction	Comment
493-01-6	C1CCC2CCCCC2C1[C60]	−3.300	Deleted
493-02-7	C1CCC2CCCCC2C1[C60]	−3.500	Involved
6876-23-9	CC1CCCCC1C[C60]	−4.600	Deleted
2207-01-4	CC1CCCCC1C[C60]	−4.600	Involved
74-97-5	C(Cl)Br[C60]	−4.200	Deleted
74-97-5	C(Cl)Br[C60]	−4.200	Involved
540-49-8	C(=CBr)Br[C60]	−3.700	Deleted
540-49-8	C(=CBr)Br[C60]	−3.670	Involved
2586-62-1	CC1=C(C2=CC=CC=C2C=C1)Br[C60]	−2.100	Deleted
2586-62-1	CC1=C(C2=CC=CC=C2C=C1)Br[C60]	−2.130	Involved
112-71-0	CCCCCCCCCCCCCCBr[C60]	−2.590	Deleted
112-89-0	CCCCCCCCCCCCCCBr[C60]	−2.530	Involved

To this end, these quasi-SMILES were randomly distributed into the following subsets: (i) active training set (25%); (ii) passive training set (25%); (iii) calibration set (25%); and (iv) validation set (25%). Ten splits obtained corresponding to the above proportions are presented here. Table 7 contains the measures of identity for ten such splits examined in this study.

Table 7. The percentage of identity for random splits examined in this study.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	100	33.0	38.5	41.2	37.3	33.3	45.1	50.5	41.2	47.6
2	31.1	100	29.1	45.5	41.6	35.6	27.7	39.2	39.6	46.2
3	39.2	30.5	100	37.3	29.4	45.1	31.4	31.1	37.3	41.9
4	40.8	41.5	36.2	100	42.0	36.0	36.0	43.6	36.0	46.6
5	35.3	43.8	32.7	51.4	100	32.0	38.0	53.5	38.0	42.7
6	40.8	37.7	30.5	39.6	34.3	100	40.0	35.6	40.0	42.7
7	44.7	43.4	38.1	39.6	38.1	43.4	100	37.6	40.0	44.7
8	51.0	38.1	36.5	41.9	36.5	43.8	38.1	100	31.7	46.2
9	52.0	40.8	39.2	42.7	33.3	54.4	50.5	47.1	100	35.0
10	43.6	34.6	36.9	48.1	31.1	42.3	46.2	35.0	47.5	100

If $i > j$, then the matrix element [i, j] refers to the percentage of identity for the active training sets; if $i < j$, then the matrix element [i, j] refers to the percentage of identity for the validation sets (external sets). The i and j indicate the numbering of the 10 splits examined.

Each of the above sets has a defined task. The active training set is used to build the model. Molecular features extracted from quasi-SMILES of the active training set are involved in the process of Monte Carlo optimization aimed to provide correlation weights for the above features, which provide the maximal target function value, which is calculated using descriptors (it is calculated as the sum of the correlation weights of all the components of quasi-SMILES) and endpoint values on the active training set. The task of the passive training set is to certify if the model obtained for the active training set is satisfactory for quasi-SMILES, which were not involved in the active training set. The calibration set should detect the start of the overtraining (overfitting). The optimization must stop if overtraining starts. After stopping the optimization procedure, the validation set is used to assess the predictive potential of the obtained model.

4.2. Optimal Descriptor

The model of fullerene solubility in organic solvents studied here is as follows:

$$logS = C_0 + C_1 \times DCW(T, N) \tag{3}$$

where *DCW*(*T*,*N*) is the optimal descriptor.

The optimal descriptor is the basis for calculating the model value of the solubility of fullerenes in organic solvents from the correlation weights of quasi-SMILES codes representing the "fullerene-solvent" systems. The quasi-SMILES reflect the presence of nano-features by two codes, indicated as [C60] and [C70], which indicate the fullerene C60 and C70, respectively. From the traditional SMILES representing the solvent, data on the atomic composition of the solvent (denoted as S) and interatomic bonds (denoted as SS) are extracted. It should be noted that atoms indicate SMILES-atoms, which is one symbol (e.g., 'C', 'N', '=') or a group of symbols that cannot be considered separately (e.g., 'Cl', %11). In this study, the so-called fragments of local symmetry (FLS) are additionally used. Three types of FLS are considered as follows: (i) XYX; (ii) XYYX; and (iii) XYZYX, where X and Y are arbitrary symbols, but X is not equal to Y. FLS are characteristics of the SMILES/quasi-SMILES strings. Generally, they are not reflections of molecular features that are somehow correlated with traditional symmetry. Nevertheless, as SMILES or quasi-SMILES features, they can be useful participants in the described optimization procedure since they improve the predictive potential of the models obtained using the approach considered here.

The above-listed features extracted from quasi-SMILES have so-called correlation weights (*CW*) obtained via the Monte Carlo optimization. Thus, the optimal descriptor is calculated as follows:

$$\boldsymbol{DCW(T,N) = \sum CW(S) + \sum CW(SS) + CW(XYX) + CW(XYYX) + CW(XYXYX)} \tag{4}$$

where *T* is the threshold, i.e., an integer to separate codes into two categories. If a code has a frequency in the active training set less than *T*, it is considered rare and removed from the simulating process. If the code has a frequency in the active training set larger than *T*, it is considered active and involved in the simulating process. *N* is the number of epochs of the Monte Carlo optimization.

4.3. The Monte Carlo Optimization

The correlation weights necessary to calculate the optimal descriptors *DCW*(*T*,*N*) are calculated using the Monte Carlo optimization based on special target functions.

Equation (4) needs the numerical data on the above correlation weights. The Monte Carlo optimization is a tool to calculate these correlation weights. Here, two target functions for the Monte Carlo optimization are examined:

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \tag{5}$$

$$TF_1 = TF_0 + (IIC \ + CII) \times 0.3 \tag{6}$$

The $\boldsymbol{r_{AT}}$ and $\boldsymbol{r_{PT}}$ are correlation coefficients between the observed and predicted endpoints for the active and passive training sets, respectively; *IIC* is the index of ideality of correlation [14,15]; and *CII* is the correlation intensity index [14,15].

Figure 1 shows the history of the optimization process for various options for the optimal descriptor and the objective function. A comparison of the results presented in Figure 1 indicates that the most promising option for obtaining the best predictive potential is the option where *IIC*, *CII*, and FLS are used (third Scheme). Table 8 contains the correlation weights for quasi-SMILES codes for the model (split 1).

Table 8. Correlation weights of the codes of quasi-SMILES used to build the model of solubility of fullerene C60 and C70 in organic solvents (split 1, third Scheme).

Code	*CW* (Code)	Frequency of Code in Active Training Set	Frequency of Code in Passive Training Set	Frequency of Code in Calibration Set	Statistical Defect of Code	Code Is Involved in Simulation
#...........	0.0	3	1	0	1.0000	FALSE
(...(.......	−0.4129	6	4	6	0.0053	TRUE
(...........	−0.1268	30	34	28	0.0020	TRUE
1...(.......	0.1602	10	17	10	0.0069	TRUE
1...........	−0.3386	17	29	18	0.0069	TRUE
2...(.......	0.0	1	0	0	1.0000	FALSE
2...........	−0.1768	5	3	2	0.0114	TRUE
2...1.......	0.0	1	0	0	1.0000	FALSE
3...........	0.0	1	0	0	1.0000	FALSE
3...2.......	0.0	1	0	0	1.0000	FALSE
=...(.......	−0.3789	13	14	7	0.0075	TRUE
=...........	0.4685	20	30	17	0.0069	TRUE
=...1.......	0.0963	12	22	13	0.0078	TRUE
=...2.......	0.0538	5	2	1	0.0191	TRUE
=...3.......	0.0	1	0	0	1.0000	FALSE
C...#.......	0.0	3	1	0	1.0000	FALSE
C...(.......	−0.4347	28	33	26	0.0026	TRUE
C...........	0.1797	50	52	50	0.0003	TRUE
C...1.......	−0.4957	17	29	18	0.0069	TRUE
C...2.......	0.3629	5	3	2	0.0114	TRUE
C...3.......	0.0	1	0	0	1.0000	FALSE
C...=.......	−0.3702	15	24	16	0.0060	TRUE
C...C.......	0.1677	40	43	43	0.0012	TRUE
F...(.......	0.0	1	0	2	1.0000	FALSE
F...........	0.0	1	0	2	1.0000	FALSE
Br..(.......	0.0	2	6	7	1.0000	FALSE
Br..........	−0.4858	8	7	9	0.0037	TRUE
Br..C.......	−0.4007	6	1	4	0.0175	TRUE
I...(.......	0.0	2	3	0	1.0000	FALSE
I...........	0.0	3	6	0	1.0000	FALSE
I...C.......	0.0	1	4	0	1.0000	FALSE
Cl..(.......	0.4380	8	10	10	0.0030	TRUE
Cl..........	−0.0828	9	11	10	0.0023	TRUE
Cl..1.......	0.0	1	0	0	1.0000	FALSE
Cl..2.......	0.0	0	1	0	1.0000	FALSE
Cl..C.......	0.0	3	1	2	1.0000	FALSE
N...#.......	0.0	2	1	0	1.0000	FALSE
N...(.......	0.0	0	2	0	1.0000	FALSE
N...........	0.0	3	7	2	1.0000	FALSE
N...1.......	0.0	0	1	0	1.0000	FALSE
N...2.......	0.0	0	0	1	1.0000	FALSE

Table 8. *Cont.*

Code	*CW* (Code)	Frequency of Code in Active Training Set	Frequency of Code in Passive Training Set	Frequency of Code in Calibration Set	Statistical Defect of Code	Code Is Involved in Simulation
N...=.......	0.0	0	2	1	1.0000	FALSE
N...C.......	0.0	1	5	1	1.0000	FALSE
O...(.......	−0.1669	9	4	5	0.0108	TRUE
O..........	0.2254	16	13	14	0.0029	TRUE
O...1.......	0.0	0	2	0	1.0000	FALSE
O...=.......	0.3200	7	6	3	0.0095	TRUE
O...C.......	0.4474	8	5	9	0.0075	TRUE
S...(.......	0.0	0	3	0	1.0000	FALSE
S..........	0.0	0	5	0	1.0000	FALSE
S...1.......	0.0	0	1	0	1.0000	FALSE
S...=.......	0.0	0	2	0	1.0000	FALSE
S...C.......	0.0	0	2	0	1.0000	FALSE
[C60].......	−0.3578	45	49	39	0.0024	TRUE
[C70].......	−0.1348	7	4	12	0.0139	TRUE
[CH2].......	0.0	0	2	1	1.0000	FALSE
[CH]........	0.0	1	0	1	1.0000	FALSE
[Ge]........	0.0	0	1	0	1.0000	FALSE
[N+]........	0.0	3	0	1	1.0000	FALSE
[O-]........	0.0	3	0	1	1.0000	FALSE
[Si]........	0.0	0	0	1	1.0000	FALSE
[Sn]........	0.0	2	0	0	1.0000	FALSE
[xyx0]......	0.3834	15	13	13	0.0021	TRUE
[xyx1]......	0.2474	18	14	19	0.0043	TRUE
[xyx2]......	−0.4400	6	8	6	0.0036	TRUE
[xyx3]......	−0.1233	5	8	10	0.0087	TRUE
[xyx4]......	0.0	2	6	1	1.0000	FALSE
[xyx5]......	0.0	2	2	1	1.0000	FALSE
[xyx6]......	0.0	3	1	1	1.0000	FALSE
[xyx7]......	0.0	1	0	0	1.0000	FALSE
[xyx9]......	0.0	0	1	0	1.0000	FALSE
[xyyx0].....	0.2019	44	34	40	0.0035	TRUE
[xyyx1].....	0.0	3	11	5	1.0000	FALSE
[xyyx2].....	0.0	2	8	6	1.0000	FALSE
[xyyx3].....	0.0	2	0	0	1.0000	FALSE
[xyyx4].....	0.0	1	0	0	1.0000	FALSE
[xyzyx0]....	−0.4334	45	44	47	0.0013	TRUE
[xyzyx1]....	−0.1116	5	8	4	0.0085	TRUE
[xyzyx2]....	0.0	2	0	0	1.0000	FALSE
[xyzyx3]....	0.0	0	1	0	1.0000	FALSE

Table 9 contains quasi-SMILES, split into active (A) and passive (P) training sets, calibration (C), and validation (V) sets, and the experimental and calculated values of fullerene C60 and C70 solubility in an organic solvent. Table 10 shows an example of the *DCW*(3,15) calculation.

Table 9. Quasi-SMILES encode a set of solutions of fullerene C60 and C70 in organic solvents along with the values of the optimal descriptor, experimental (Expr), and calculated (Calc) values of molar fraction and applicability domain (AD). The case of split 1 (third scheme). The regression formula is as follows: logS = −7.608 + 0.4426 × *DCW*(3,15).

Set	CAS *	Quasi-SMILES	*DCW*(3,15)	Expr	Calc	The Statistical Defect of Quasi-SMILES	AD
P	109-66-0	CCCCC[C60]	3.6226	−6.1000	−6.0050	0.0174	YES
V	110-54-3	CCCCCC[C60]	4.0950	−5.1000	−5.7958	0.0189	YES
C	111-65-9	CCCCCCCC[C60]	5.0400	−5.2000	−5.3776	0.0217	YES
P	26635-64-3	CC(C)CC(C)(C)C[C60]	5.3953	−5.2000	−5.2203	1.0549	YES
V	124-18-5	CCCCCCCCCC[C60]	5.9849	−4.7000	−4.9594	0.0246	YES
A	112-40-3	CCCCCCCCCCCC[C60]	6.9298	−3.5000	−4.5411	0.0275	YES
V	493-02-7	C1CCC2CCCCC2C1[C60]	9.3895	−3.5000	−3.4524	0.1283	YES
A	137-43-9	C1CCC(C1)CBr[C60]	9.4279	−4.2000	−3.4354	0.0891	YES
V	542-18-7	C1CCC(CC1)Cl[C60]	7.5453	−4.1000	−4.2687	0.0717	YES
C	108-85-0	C1CCC(CC1)Br[C60]	8.2381	−3.4000	−3.9620	1.0701	YES
P	626-62-0	C1CCC(CC1)I[C60]	8.3333	−2.8000	−3.9199	2.0664	YES
P	5401-62-7	C1CCC(C(C1)Br)Br[C60]	9.9420	−2.6000	−3.2079	4.0855	No
C	110-83-8	C1CCC=CC1[C60]	7.5129	−3.8000	−4.2830	0.0692	YES
C	108-87-2	CC1CCCCC1[C60]	7.0015	−4.5000	−4.5094	0.0536	YES
P	75-09-2	C(Cl)Cl[C60]	4.5883	−4.6000	−5.5775	0.0320	YES
A	56-23-5	C(Cl)(Cl)(Cl)Cl[C60]	5.8391	−4.4000	−5.0239	0.0672	YES
V	74-95-3	C(Br)Br[C60]	6.9622	−4.5000	−4.5268	3.0236	YES
P	75-25-2	C(Br)(Br)Br[C60]	8.1953	−3.2000	−3.9810	5.0366	No
A	74-88-4	CI[C60]	4.5584	−4.2000	−5.5908	2.0096	YES
C	74-96-4	CCBr[C60]	6.2277	−5.2000	−4.8519	0.0323	YES
P	75-03-6	CCI[C60]	5.0308	−4.5000	−5.3816	2.0110	YES
P	79-34-5	C(C(Cl)Cl)(Cl)Cl[C60]	7.6012	−3.1000	−4.2439	0.0718	YES
A	107-06-2	C(CCl)Cl[C60]	5.0251	−5.0000	−5.3842	1.0318	YES
C	71-55-6	CC(Cl)(Cl)Cl[C60]	5.6861	−4.7000	−5.0916	0.0510	YES
A	540-54-5	C[CH]CCl[C60]	3.4486	−5.6000	−6.0820	2.0155	YES
P	107-08-4	CCCI[C60]	5.5033	−4.6000	−5.1725	2.0124	YES
A	75-29-6	CC(C)Cl[C60]	4.7185	−5.9000	−5.5199	0.0298	YES
C	75-26-3	CC(C)Br[C60]	4.9687	−5.4000	−5.4091	1.0289	YES
A	75-30-9	CC(C)I[C60]	5.0639	−4.8000	−5.3670	2.0252	YES
C	78-87-5	CC(CCl)Cl[C60]	5.4975	−4.9000	−5.1751	1.0333	YES
C	142-28-9	C([CH]CCl)Cl[C60]	6.3111	−4.8000	−4.8149	2.0297	YES
V	78-75-1	CC(CBr)Br[C60]	8.0234	−4.3000	−4.0570	2.0476	YES
P	627-31-6	C(CI)CI[C60]	7.0158	−3.4000	−4.5030	5.0240	No
C	96-11-7	C(C(CBr)Br)Br[C60]	10.5461	−2.9000	−2.9405	4.0637	No
A	96-18-4	C(C(CCl)Cl)Cl[C60]	7.4126	−4.0000	−4.3274	1.0541	YES
V	513-36-0	CC(C)CCl[C60]	5.1553	−5.4000	−5.3266	1.0297	YES
P	513-38-2	CC(C)CI[C60]	5.8766	−4.3000	−5.0073	2.0274	YES

Table 9. *Cont.*

Set	CAS *	Quasi-SMILES	*DCW*(3,15)	Expr	Calc	The Statistical Defect of Quasi-SMILES	AD
V	507-19-7	CC(C)(C)Br[C60]	4.8092	−5.0000	−5.4797	1.0437	YES
C	127-18-4	C(=C(Cl)Cl)(Cl)Cl[C60]	8.1332	−3.8000	−4.0085	0.0852	YES
P	513−37−1	CC(=CCl)C[C60]	5.5615	−4.5000	−5.1467	1.0486	YES
V	71-43-2	C1=CC=CC=C1[C60]	8.1013	−4.0000	−4.0226	1.0973	YES
P	95-47-6	CC1=CC=CC=C1[CH2][C60]	9.0053	−2.9000	−3.6224	2.0987	YES
V	108-38-3	CC1=CC(=CC=C1)C[C60]	9.2607	−3.3000	−3.5094	1.1166	YES
C	526-73-8	CC1=C(C(=CC=C1)C)C[C60]	9.9717	−3.1000	−3.1947	2.1265	YES
A	95-63-6	CC1=CC(=C(C=C1)C)C[C60]	9.1010	−2.5000	−3.5801	1.1372	YES
P	108-67-8	CC1=CC(=CC(=C1)C)C[C60]	9.4917	−3.5000	−3.4071	0.1388	YES
A	527-53-7	CC1=CC(=C(C(=C1)C)C)C[C60]	10.1268	−2.4000	−3.1260	2.1422	YES
P	119-64-2	C1CCC2=CC=CC=C2C1[C60]	10.1800	−2.5000	−3.1025	2.1722	YES
C	103-65-1	CCCC1=CC=CC=C1[C60]	9.5187	−3.5000	−3.3952	1.1016	YES
A	98-82-8	CCC1=CC=CC=C1C(C)C[C60]	10.8426	−3.6000	−2.8092	2.1187	YES
V	104-51-8	CCCCC1=CC=CC=C1[C60]	9.9911	−3.4000	−3.1861	1.1030	YES
V	98-06-6	CC(C)(C)C1=CC=CC=C1[C60]	9.2698	−3.7000	−3.5054	2.1248	YES
C	462-06-6	C1=CC=C(C=C1)F[C60]	8.4784	−4.1000	−3.8557	3.1246	YES
P	108-90-7	C1=CC=C(C=C1)Cl[C60]	8.8771	−3.0000	−3.6792	1.1299	YES
V	108-86-1	C1=CC=C(C=C1)Br[C60]	9.5699	−3.3000	−3.3726	2.1283	YES
P	95-50-1	C1=CC=C(C(=C1)Cl)Cl[C60]	10.8547	−2.4000	−2.8039	1.1392	YES
P	108-36-1	C1=CC(=CC(=C1)Br)Br[C60]	12.6461	−2.6000	−2.0109	3.1299	YES
C	694-80-4	C1=CC=C(C(=C1)Cl)Br[C60]	11.5475	−2.4000	−2.4972	2.1375	YES
P	108-37-2	C1=CC(=CC(=C1)Br)Cl[C60]	11.9533	−3.0000	−2.3176	2.1315	YES
V	120-82-1	C1=CC(=C(C=C1Cl)Cl)Cl[C60]	12.1777	−2.8000	−2.2183	1.1482	YES
V	100-42-5	C=CC1=CC=CC=C1[C60]	9.2708	−3.2000	−3.5049	1.1230	YES
V	98-95-3	C1=CC=C(C=C1)[N+](=O)[O-][C60]	7.4812	−3.9000	−4.2970	3.1715	No
P	100-47-0	C1=CC=C(C=C1)CCN[C60]	9.3119	−4.2000	−3.4867	5.1274	No
P	100-66-3	COC1=CC=CC=C1[C60]	7.7096	−3.1000	−4.1959	1.1205	YES
C	100-52-7	C1=CC=C(C=C1)C=O[C60]	7.9172	−4.2000	−4.1041	1.1527	YES
P	103-71-9	C1=CC=C(C=C1)N=C=O[C60]	9.3125	−3.4000	−3.4865	5.1544	No
A	99-08-1	CC1=CC(=CC=C1)[N+](=O)[O-][C60]	8.0163	−3.4000	−4.0602	3.1614	YES
P	108-98-5	C1=CC=C(C=C1)S[C60]	8.0975	−3.0000	−4.0243	3.1246	YES
C	100-39-0	C1=CC=C(C=C1)CBr[C60]	11.2321	−3.1000	−2.6368	1.1487	YES
A	30583-33-6	CC1=CC(=C(C=C1Cl)Cl)Cl[C60]	12.6502	−3.0000	−2.0091	1.1496	YES
A	90-12-0	CC1=CC=CC2=CC=CC=C12[C60]	10.7555	−2.2000	−2.8478	2.1924	YES
A	28804-88-8	CC1=CC2=C(C=C1)C=C(C=C2)C[C60]	11.3352	−2.1000	−2.5912	3.2245	No
A	605-02−7	C1=CC=C(C=C1)C2=CC=CC3=CC=CC=C32[C60]	15.1824	−1.9000	−0.8883	8.2616	No

Table 9. *Cont.*

Set	CAS *	Quasi-SMILES	*DCW*(3,15)	Expr	Calc	The Statistical Defect of Quasi-SMILES	AD
A	64-17-5	CCO[C60]	2.7640	−7.1000	−6.3850	0.0214	YES
C	71-36-3	CCCCO[C60]	3.7089	−5.9000	−5.9667	0.0242	YES
C	71-41-0	CCCCCO[C60]	4.1814	−5.3000	−5.7576	0.0257	YES
P	67-64-1	CC(=O)C[C60]	2.7551	−7.0000	−6.3889	0.0603	YES
P	68-12-2	CN(C)C=O[C60]	3.8967	−5.3000	−5.8836	3.0493	YES
P	110-01-0	C1CCSC1[C60]	5.9939	−5.4000	−4.9554	3.0474	YES
V	110-02-1	C1=CSC=C1[C60]	6.5278	−4.4000	−4.7191	3.0862	YES
P	554-14-3	CC1=CC=CS1[C60]	7.3816	−3.0000	−4.3412	4.0719	No
P	872-50-4	CN1CCCC1=O[C60]	7.3505	−3.9000	−4.3549	3.0688	YES
P	110-86-1	C1=CC=NC=C1[C60]	8.9043	−4.0000	−3.6671	4.0906	No
C	91-22-5	C1=CC=C2C(=C1)C=CC=N2[C60]	11.6924	−2.9000	−2.4331	4.1982	No
V	62-53-3	C1=CC=C(C=C1)N[C60]	9.0161	−3.9000	−3.6177	3.1246	YES
C	100-61-8	CNC1=CC=CC=C1[C60]	9.2792	−3.8000	−3.5012	4.1027	No
V	121-69-7	CN(C)C1=CC=CC=C1[C60]	10.3144	−3.2000	−3.0430	4.1147	No
C	4904-61-4	C1CC=CCCC=CCCC=C1[C60]	10.8090	−2.7000	−2.8241	1.1096	YES
A	629-59-4	CCCCCCCCCCCCCC[C60]	7.8747	−4.3000	−4.1229	0.0303	YES
A	110-82-7	C1CCCCC1[C60]	6.5290	−5.3000	−4.7185	0.0521	YES
C	591-49-1	CC1=CCCCC1[C60]	8.1394	−3.8000	−4.0057	0.0653	YES
A	2207-01-4	CC1CCCCC1C[C60]	7.7404	−4.6000	−4.1823	0.0651	YES
C	1678-91-7	CCC1CCCCC1[C60]	7.4740	−4.3000	−4.3003	0.0550	YES
V	67-66-3	C(Cl)(Cl)Cl[C60]	5.2137	−4.8000	−5.3007	0.0496	YES
V	106-93-4	C(CBr)Br[C60]	7.5510	−4.2000	−4.2662	2.0461	YES
A	106-94-5	CCCBr[C60]	6.7002	−5.2000	−4.6427	0.0337	YES
C	109-64-8	C(CBr)CBr[C60]	9.2132	−4.2000	−3.5304	1.0665	YES
A	78-77-3	CC(C)CBr[C60]	7.0735	−4.9000	−4.4775	0.0486	YES
C	507-20-0	CC(C)([CH2])Cl[C60]	2.9118	−5.7000	−6.3196	1.0451	YES
A	558-17-8	CC(C)(C)I[C60]	4.9044	−4.4000	−5.4376	2.0400	YES
A	79-01-6	C(=C(Cl)Cl)Cl[C60]	7.5078	−3.8000	−4.2853	0.0676	YES
C	108-88-3	CC1=CC=CC=C1[C60]	8.5737	−3.4000	−3.8135	1.0987	YES
A	106-42-3	CC1=CC=C(C=C1)C[C60]	8.4511	−3.3000	−3.8678	2.1202	YES
P	488-23-3	CC1=C(C(=C(C=C1)C)C)C[C60]	10.2939	−2.9000	−3.0521	2.1422	YES
V	100-41-4	CCC1=CC=CC=C1[C60]	9.0462	−3.4000	−3.6043	1.1002	YES
V	135-98-8	CCC(C)C1=CC=CC=C1[C60]	9.9017	−3.6000	−3.2257	2.1115	YES
V	591-50-4	C1=CC=C(C=C1)I[C60]	9.6651	−3.5000	−3.3304	3.1246	YES
P	541-73-1	C1=CC(=CC(=C1)Cl)Cl[C60]	11.3457	−3.4000	−2.5865	0.1361	YES
V	583-53-9	C1=CC=C(C(=C1)Br)Br[C60]	12.1551	−2.6000	−2.2283	4.1329	No
C	88-72-2	CC1=CC=CC=C1[N+](=O)[O-][C60]	8.3836	−3.4000	−3.8976	3.1473	YES
V	100-44-7	C1=CC=C(C=C1)CCl[C60]	9.3139	−3.4000	−3.4859	2.1297	YES

Table 9. *Cont.*

Set	CAS *	Quasi-SMILES	*DCW*(3,15)	Expr	Calc	The Statistical Defect of Quasi-SMILES	AD
P	90-13-1	C1=CC=C2C(=C1)C=CC=C2Cl[C60]	11.5646	−2.0000	−2.4897	3.2094	No
P	71-23-8	CCCO[C60]	3.2365	-6.4000	−6.1758	0.0228	YES
V	111-27-3	CCCCCCO[C60]	4.6538	−5.1000	−5.5485	0.0271	YES
V	111-87-5	CCCCCCCCO[C60]	5.5988	−5.0000	−5.1303	0.0300	YES
A	107-13-1	C=CCCN[C60]	4.2477	−6.4000	−5.7282	4.0323	No
P	111-96-6	COCCOCCOC[C60]	2.2569	−5.2000	−6.6094	2.0611	YES
C	111-84-2	CCCCCCCCC[C60]	5.5124	−4.9200	−5.1685	0.0232	YES
C	79-00-5	C(C(Cl)Cl)Cl[C60]	6.9758	−4.7800	−4.5208	0.0542	YES
A	109-65-9	CCCCBr[C60]	7.1726	−3.7400	−4.4336	0.0351	YES
P	629-04-9	CCCCCCCBr[C60]	8.5900	−3.3000	−3.8063	0.0394	YES
A	111-83-1	CCCCCCCCBr[C60]	9.0625	−3.0900	−3.5971	0.0408	YES
V	112-89-0	CCCCCCCCCCCCCCBr[C60]	11.8972	−2.5300	−2.3424	0.0494	YES
A	67-56-1	CO[C60]	2.2916	−8.8700	−6.5941	0.0199	YES
A	143-08-8	CCCCCCCCCO[C60]	6.0712	−4.2900	−4.9211	0.0314	YES
C	112-30-1	CCCCCCCCCCO[C60]	6.5437	−4.1500	−4.7120	0.0328	YES
V	112-42-5	CCCCCCCCCCCO[C60]	7.0161	−3.9900	−4.5029	0.0343	YES
P	67-63-0	CC(C)O[C60]	2.5223	−6.6500	−6.4920	0.0390	YES
C	78-92-2	CCC(C)O[C60]	2.9947	−6.3400	−6.2828	0.0404	YES
V	6032-29-7	CCCC(C)O[C60]	3.4672	−5.5700	−6.0737	0.0418	YES
A	584-02-1	CCC(CC)O[C60]	3.4672	−5.3600	−6.0737	0.0418	YES
A	504-63-2	C(CO)CO[C60]	1.9748	−7.0500	−6.7343	0.0556	YES
C	110-63-4	C(CCO)CO[C60]	2.4473	−6.5700	−6.5252	0.0571	YES
V	111-29-5	C(CCO)CCO[C60]	2.9197	−6.1900	−6.3161	0.0585	YES
A	102-04-5	C1=CC=C(C=C1)CC(=O)CC2=CC=CC=C2[C60]	12.8645	−3.4000	−1.9143	2.2878	YES
P	104-92-7	COC1=CC=C(C=C1)Br[C60]	9.1904	−2.5400	−3.5405	3.1377	YES
A	2398-37-0	COC1=CC(=CC=C1)Br[C60]	10.0001	−2.5500	−3.1821	2.1341	YES
P	573-98-8	CC1=C(C2=CC=CC=C2C=C1)C[C60]	11.4939	−2.1200	−2.5209	2.2303	YES
A	75-05-8	CCCN[C60]	4.3075	−7.5400	−5.7018	4.0110	No
V	109-99-9	C1CCOC1[C60]	5.4793	−5.1700	−5.1831	0.0652	YES
V	108-75-8	CC1=CC(=NC(=C1)C)C[C60]	10.1468	−2.8000	−3.1172	3.1314	YES
C	64-19-7	CC(=O)O[C60]	1.5955	−6.2700	−6.9022	0.0712	YES
V	79-09-4	CCC(=O)O[C60]	2.0679	−5.7900	−6.6931	0.0726	YES
V	107-92-6	CCCC(=O)O[C60]	2.5404	−5.7400	−6.4840	0.0740	YES
P	109-52-4	CCCCC(=O)O[C60]	3.0128	−5.0500	−6.2748	0.0754	YES
A	142-62-1	CCCCCC(=O)O[C60]	3.4853	−4.5000	−6.0657	0.0769	YES
A	111-14-8	CCCCCCC(=O)O[C60]	3.9577	−4.2600	−5.8566	0.0783	YES
V	124-07-2	CCCCCCCC(=O)O[C60]	4.4302	−4.9800	−5.6475	0.0797	YES

Table 9. *Cont.*

Set	CAS *	Quasi-SMILES	*DCW*(3,15)	Expr	Calc	The Statistical Defect of Quasi-SMILES	AD
P	112-05-0	CCCCCCCCC(=O)O[C60]	4.9026	−4.4100	−5.4384	0.0812	YES
A	76-13-1	C(C(F)(Cl)Cl)(F)(F)Cl[C60]	7.6642	−5.7700	−4.2161	9.0770	No
V	540-49-8	C(=CBr)Br[C60]	9.2824	−3.6700	−3.4998	2.0618	YES
C	1649-08-7	C(C(F)(F)Cl)Cl[C60]	6.9149	−5.3800	−4.5477	6.0501	No
P	123-91-1	C1COCCO1[C60]	5.3115	−5.3100	−5.2574	2.0652	YES
P	95-48-7	CC1=CC=CC=C1O[C60]	8.0059	−5.5400	−4.0648	2.1016	YES
P	287-92-3	C1CCCC1[C60]	6.0566	−6.5200	−4.9276	0.0507	YES
P	75-11-6	C(I)I[C60]	6.2755	−4.8200	−4.8307	5.0183	No
A	79-24-3	CC[N+](=O)[O-][C60]	4.1130	−6.7000	−5.7879	2.0553	YES
P	74-97-5	C(Cl)Br[C60]	5.2811	−4.2000	−5.2709	1.0304	YES
P	109-73-9	CCCCN[C60]	5.3981	−3.3000	−5.2191	2.0139	YES
V	583-57-3	C[C@@H]1CCCC[C@H]1C[C60]	6.7955	−4.6000	−4.6006	0.0623	YES
A	106-96-7	CCCCBr[C60]	4.9448	−4.6400	−5.4197	3.0347	YES
V	10026-04-7	[Si](Cl)(Cl)(Cl)Cl[C60]	6.5498	−4.8200	−4.7093	1.0622	YES
P	10038-98-9	Cl[Ge](Cl)(Cl)Cl[C60]	6.9654	−4.1000	−4.5254	1.0499	YES
A	7646-78-8	Cl[Sn](Cl)(Cl)Cl[C60]	7.2152	−3.7000	−4.4148	1.0499	YES
V	13465-77-5	[Si]([Si](Cl)(Cl)Cl)(Cl)(Cl)Cl[C60]	7.9273	−4.0500	−4.0996	2.0975	YES
C	7789-66-4	[Si](Br)(Br)(Br)Br[C60]	9.0656	−3.8900	−3.5958	8.0467	No
A	7789-67-5	Br[Sn](Br)(Br)Br[C60]	9.8161	−4.5200	−3.2636	7.0374	No
C	107-83-5	CCCC(C)C[C60]	4.7527	−5.4900	−5.5047	0.0354	YES
C	96-14−0	CCC(C)CC[C60]	4.7527	−5.3500	−5.5047	0.0354	YES
V	142-82-5	CCCCCCC[C60]	4.5675	−5.0100	−5.5867	0.0203	YES
A	75-52-5	C[N+](=O)[O-][C60]	3.6406	−4.8200	−5.9970	2.0538	YES
P	75-15-0	C(=S)=S[C60]	5.5718	−3.1800	−5.1422	5.0450	No
A	110-89-4	C1CCNCC1[C60]	7.5213	−2.1400	−4.2793	3.0488	YES
P	123-75-1	C1CCNC1[C60]	7.0489	−2.2700	−4.4884	3.0474	YES
C	2586-62-1	CC1=C(C2=CC=CC=C2C=C1)Br[C60]	12.5551	−2.1300	−2.0512	3.2311	No
A	109-66-0	CCCCC[C70]	3.6495	−6.5720	−5.9930	0.0289	YES
C	110-54-3	CCCCCC[C70]	4.1220	−5.6830	−5.7839	0.0304	YES
C	142-82-5	CCCCCCC[C70]	4.5944	−5.0830	−5.5748	0.0318	YES
C	111-65-9	CCCCCCCC[C70]	5.0669	−5.0950	−5.3657	0.0332	YES
V	124-18-5	CCCCCCCCCC[C70]	6.0118	−4.9130	−4.9474	0.0361	YES
C	112-40-3	CCCCCCCCCCCC[C70]	6.9567	−4.5780	−4.5292	0.0390	YES
V	110-82-7	C1CCCCC1[C70]	6.5560	−4.9870	−4.7066	0.0636	YES
A	67-64-1	CC(=O)C[C70]	2.7820	−6.7700	−6.3770	0.0717	YES
C	67-63-0	CC(C)O[C70]	2.5492	−6.6990	−6.4800	0.0505	YES
C	56-23-5	C(Cl)(Cl)(Cl)Cl[C70]	5.8660	−4.8570	−5.0120	0.0787	YES
P	106-42-3	CC1=CC=C(C=C1)C[C70]	8.4780	−3.2360	−3.8558	2.1317	YES

Table 9. *Cont.*

Set	CAS *	Quasi-SMILES	*DCW*(3,15)	Expr	Calc	The Statistical Defect of Quasi-SMILES	AD
A	108-67-8	CC1=CC(=CC(=C1)C)C[C70]	9.5187	−3.6130	−3.3952	0.1503	YES
V	108-88-3	CC1=CC=CC=C1[C70]	8.6007	−3.7500	−3.8015	1.1102	YES
V	71-43-2	C1=CC=CC=C1[C70]	8.1282	−3.8590	−4.0107	1.1088	YES
P	75-15-0	C(=S)=S[C70]	5.5987	−3.1510	−5.1303	5.0565	No
V	75-09-2	C(Cl)Cl[C70]	4.6152	−5.2150	−5.5656	0.0435	YES
P	95-50-1	C1=CC=C(C(=C1)Cl)Cl[C70]	10.8816	−2.3160	−2.7919	1.1507	YES
P	95-47-6	CC1=CC=CC=C1[CH2][C70]	9.0323	−2.6500	−3.6105	2.1102	YES
C	541-73-1	C1=CC(=CC(=C1)Cl)Cl[C70]	11.3726	−2.5950	−2.5746	0.1476	YES
A	119-64-2	C1CCC2=CC=CC=C2C1[C70]	10.2069	−2.6970	−3.0906	2.1837	YES
A	67-56-1	CO[C70]	2.3185	−8.7420	−6.5822	0.0314	YES
A	64-17-5	CCO[C70]	2.7910	−7.2720	−6.3730	0.0329	YES
C	71-23-8	CCCO[C70]	3.2634	−6.4570	−6.1639	0.0343	YES
C	71-36-3	CCCCO[C70]	3.7359	−6.0230	−5.9548	0.0357	YES
A	71-41-0	CCCCCO[C70]	4.2083	−6.4350	−5.7457	0.0372	YES
V	111-27-3	CCCCCCO[C70]	4.6808	−5.2740	−5.5366	0.0386	YES
C	111-87-5	CCCCCCCCO[C70]	5.6257	−5.0500	−5.1183	0.0415	YES
V	111-70-6	CCCCCCCO[C70]	5.1532	−5.0040	−5.3275	0.0400	YES
C	143-08-8	CCCCCCCCCO[C70]	6.0981	−4.3590	−4.9092	0.0429	YES
V	112-30-1	CCCCCCCCCCO[C70]	6.5706	−4.1520	−4.7001	0.0443	YES
C	112-42-5	CCCCCCCCCCCO[C70]	7.0431	−4.2160	−4.4910	0.0458	YES

* CAS is related to the corresponding solvent; A, P, C, and V denote active training, passive training, calibration, and validation sets, respectively.

Table 10. Quasi-SMILES CCCCC[C60] is the code of the solution for fullerene C60 in pentane.

Code of Quasi-SMILES	*CW* (Code)	Frequency of Code in Active Training Set	Frequency of Code in Passive Training Set	Frequency of Code in Calibration Set
[C60].......	0.4203	45	49	39
C...........	0.3047	50	52	50
C...........	0.3047	50	52	50
C...........	0.3047	50	52	50
C...........	0.3047	50	52	50
C...........	0.3047	50	52	50
C...C.......	0.1677	40	43	43
C...C.......	0.1677	40	43	43
C...C.......	0.1677	40	43	43
C...C.......	0.1677	40	43	43
[xyx1]......	0.2474	18	14	19
[xyyx0].....	0.2019	44	34	40
[xyzyx0]....	0.5584	45	44	47
∑	3.6226			

4.4. The Applicability Domain

The applicability domain is considered in many studies devoted to QSPR/QSAR analysis [16]. The main question is, "Can the resulting model be applied to a given/interest substance?". However, the counter-question is also logical. Is it not better to determine for which substances the model being developed is intended before developing it [17]? Can the model's applicability domain change if one changes the distribution of available data into training and validation sets?

It should be noted that for the approach studied here, the applicability domain for different splits slightly changes.

The applicability domain for the described CORAL models are defined via the so-called statistical defects of codes used in quasi-SMILES. These defects are calculated as follows:

$$d_k = \frac{|P(S_k) - P'(S_k)|}{N(S_k) + N'(S_k)} + \frac{|P(S_k) - P''(S_k)|}{N(S_k) + N''(S_k)} + \frac{|P'(S_k) - P''(S_k)|}{N'(S_k) + N''(S_k)} \quad (7)$$

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

$$\boldsymbol{D_j = \sum\nolimits_{k=1}^{NA} d_k} \quad (8)$$

where *NA* is the number of non-blocked codes in quasi-SMILES.

A quasi-SMILES falls in the applicability domain, if

$$\boldsymbol{Dj} < 2*\boldsymbol{\overline{D}} \quad (9)$$

where $\boldsymbol{\overline{D}}$ is the average statistical defect for the active training set.

4.5. Mechanistic Interpretation

With the numerical data on the correlation weights of codes applied in quasi-SMILES, which was observed in several runs of the Monte Carlo optimization, one can extract three categories of these codes:

i. Codes that have a positive value of the correlation weight in all runs. These are promoters of endpoint increase;

ii. Codes with a negative correlation weight value in all runs. These are promoters of endpoint decrease;

iii. Codes with negative and positive correlation weight values in different optimization runs. These codes have unclear roles (one cannot classify these features as promoters of increase or decrease for endpoint).

4.6. System of Self-Consistent Models

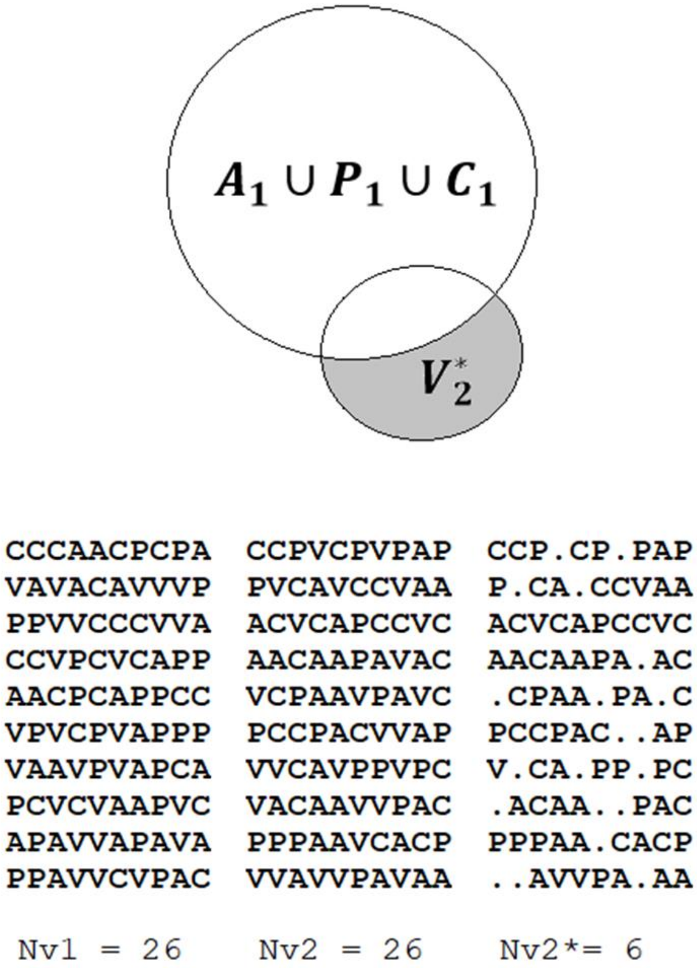
The reliability of an approach can be assessed by the so-called system of self-consistent models [18,19]. The main idea of such a system is to test the performance of an approach on many random splits of the available data into training and validation subsets. This task can be represented by a matrix of determination coefficients related to applying the model built using split 1 to the validation set observed for split 2. Suppose some quasi-SMILES, which are allocated to the validation set of split 2, are present in the training or the calibration sets of split 1 at the same time. In that case, they may improve the statistical quality of model 1 for the split 2 validation set.

$$\begin{matrix} R^2_{1,1} & R^2_{1,2} & \cdots & R^2_{10,1} \\ R^2_{2,1} & R^2_{2,2} & \cdots & R^2_{10,2} \\ \vdots & \vdots & & \vdots \\ R^2_{10,1} & R^2_{10,2} & \cdots & R^2_{10,10} \end{matrix} \quad (10)$$

In order for the assessment of the statistical quality of model 1 for the validation set of split 2 to be adequate, it is necessary to remove the abovementioned quasi-SMILES from consideration. It can be expressed as the following:

$$\begin{matrix} R^2_{1,1} & R^{*2}_{1,2} & \cdots & R^{*2}_{10,1} \\ R^{*2}_{2,1} & R^2_{2,2} & \cdots & R^{*2}_{10,2} \\ \vdots & \vdots & & \vdots \\ R^{*2}_{10,1} & R^{*2}_{10,2} & \cdots & R^2_{10,10} \end{matrix} \quad (11)$$

Figure 4 indicates the essence of asterisks in the matrix (11). It is clear that the principles for selecting quasi-SMILES in the validation set of split 2 to assess the predictive potential of model 1 can be clearly translated for the arbitrary pairs of the *i*-th model vs. the *j*-th split ($i \neq j$).



```
CCCAACPCPA  CCPVCPVPAP  CCP.CP.PAP
VAVACAVVVP  PVCAVCCVAA  P.CA.CCVAA
PPVVCCCVVA  ACVCAPCCVC  ACVCAPCCVC
CCVPCVCAPP  AACAAPAVAC  AACAAPA.AC
AACPCAPPCC  VCPAAVPAVC  .CPAA.PA.C
VPVCPVAPPP  PCCPACVVAP  PCCPAC..AP
VAAVPVAPCA  VVCAVPPVPC  V.CA.PP.PC
PCVCVAAPVC  VACAAVVPAC  .ACAA..PAC
APAVVAPAVA  PPPAAVCACP  PPPAA.CACP
PPAVVCVPAC  VVAVVPAVAA  ..AVVPA.AA

 Nv1 = 26    Nv2 = 26    Nv2*= 6
```

Figure 4. For the demonstration scheme of the assessment of a model: let 100 quasi-SMILES be used and distributed into active training (A1), passive training (P1), and calibration (C1) sets, which are used to build model 1. The subset of the validation set of split 2, denoted as V2*, is used to assess the predictive potential of model 1. One can see that, instead of 26 quasi-SMILES (Nv2), only 6 are involved in assessing.

4.7. Comparison with Other Models

The strange influence of *IIC* and *CII* on the simulation process via improving the statistical quality of the model for the calibration sets leads to the temptation to compare different models in terms of their quality for the external validation set. Table 11 contains the comparisons of models for the solubility of fullerenes in various solvents.

Table 11. The comparison of different approaches for simulation of solubility of fullerenes in different solvents.

Approach	Set	n	R^2	References
MLR *	Training set	92	0.861	[23]
	Validation set	30	0.903	
PLS	Training set	80	0.674	[24]
	Validation set	28	0.692	
SVM	Training set	92	0.871	[25]
	Validation set	30	0.940	
DTB	Training set	145	0.970	[12]
	Validation set	36	0.964	
Monte Carlo	Training set	55	0.947	[26]
	Validation set	35	0.915	
DFT	Training set	44	0.73	[27]
	Validation set	15	0.74	
Quantum-mechanical descriptors	Training set	44	0.76	[28]
	Validation set	15	0.70	
CODESSA software	Training set	21	0.745	[29]
	Validation set	6	0.801	
Self-consistent models	Training set	≈100	≈0.73	In this study
	Validation set	19-22	0.84–0.94	

* MLR = multiple linear regression; PLS = partial least square regression; SVM = support vector machine; DTB = decision tree boost; DFT = density functional theory.

5. Conclusions

A model observed for a single distribution of the available data into a training and a validation set can be either too good or too bad. It is preferable to consider a set of models built on sufficiently diverse distributions of the available data in the training and validation sets to obtain reliable information about the suitability of the chosen approach. The two-component vector of the ideality of correlation based on the use of the *IIC* and *CII* for the Monte Carlo optimization improves the predictive potential of the model. However, the paradoxical effect of the mentioned vector is to reduce the determination coefficient values for the active and passive training sets. However, if the main aim of the simulation is to obtain a satisfactory prediction for the external validation set, then the effect of the vector of the ideality of correlation which leads to improving the statistical quality of a model for the external validation set, even in the detriment the training set, the result rather useful rather than adverse. Using the proposed fragments of local symmetry (FLS) significantly improves the predictive potential of the solubility model of fullerenes C60 and C70 in organic solvents. It would be wrong to claim that FLS is related to traditional classical symmetry. However, it is clear that FLS contains some information that can improve the statistical quality and possibly the interpretability of the models. QSPR and nano-QSPR are random events since the appearance of new experimental data may challenge the already created models. Therefore, each model should be considered useful only temporarily and should be prepared for the need for radical alteration. The quasi-SMILES technique offers the possibility of the fast modification of models taking into account new conditions/circumstances.

Author Contributions: Conceptualization, A.A.T., A.P.T. and N.F.; Data curation, A.A.T., A.P.T. and N.F.; Writing—original draft, A.A.T., A.P.T. and N.F.; Review and editing, A.A.T., A.P.T. and N.F. All authors have read and agreed to the published version of the manuscript.

Funding: A.P.T. and A.A.T. acknowledge EFSA for the financial contribution to the project sOFT-ERA, OC/EFSA/IDATA/2022/02.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Technical details on five models are available on request.

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

References

1. Wiener, H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636–2638. [CrossRef]
2. Wiener, H. Structural Determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20. [CrossRef] [PubMed]
3. Wiener, H. Influence of interatomic forces on paraffin properties. *J. Chem. Phys.* **1947**, *15*, 766. [CrossRef]
4. Venkatapathy, R.; Wang, C.Y.; Bruce, R.M.; Moudgal, C. Development of quantitative structure-activity relationship (QSAR) models to predict the carcinogenic potency of chemicals. I. Alternative toxicity measures as an estimator of carcinogenic potency. *Toxicol. Appl. Pharmacol.* **2009**, *234*, 209–221. [CrossRef] [PubMed]
5. Harris, P.J.F. Fullerene-related structure of commercial glassy carbons. *Philos. Mag.* **2004**, *84*, 3159–3167. [CrossRef]
6. Olmstead, M.M.; Costa, D.A.; Maitra, K.; Noll, B.C.; Phillips, S.L.; Van Calcar, P.M.; Balch, A.L. Interaction of curved and flat molecular surfaces. The structures of crystalline compounds composed of fullerene (C_{60}, $C_{60}O$, C_{70}, and $C_{120}O$) and metal octaethylporphyrin units. *J. Am. Chem. Soc.* **1999**, *121*, 7090–7097. [CrossRef]
7. Johnson, R.D.; Bethune, D.S.; Yannoni, C.S. Fullerene structure and dynamics: A magnetic resonance potpourri. *Acc. Chem. Res.* **1992**, *25*, 169–175. [CrossRef]
8. Liu, X.; Schmalz, T.G.; Klein, D.J. Favorable structures for higher fullerenes. *Chem. Phys. Lett.* **1992**, *188*, 550–554. [CrossRef]
9. Rao, C.N.R.; Seshadri, R.; Govindaraj, A.; Sen, R. Fullerenes, nanotubes, onions and related carbon structures. *Mater. Sci. Eng. R Rep.* **1995**, *15*, 209–262. [CrossRef]
10. Dinadayalane, T.C.; Leszczynski, J. Remarkable diversity of carbon-carbon bonds: Structures and properties of fullerenes, carbon nanotubes, and graphene. *Struct. Chem.* **2010**, *21*, 1155–1169. [CrossRef]
11. Mikheev, I.V.; Verkhovskii, V.A.; Byvsheva, S.M.; Volkov, D.S.; Proskurnin, M.A.; Ivanov, V.K. Simultaneous quantification of fullerenes C_{60} and C_{70} in organic solvents by excitation–emission matrix fluorescence spectroscopy. *Inorganics* **2023**, *11*, 136. [CrossRef]
12. Gupta, S.; Basant, N. Predictive modeling: Solubility of C_{60} and C_{70} fullerenes in diverse solvents. *Chemosphere* **2018**, *201*, 361–369. [CrossRef]
13. Prana, V.; Fayet, G.; Rotureau, P.; Adamo, C. Development of validated QSPR models for impact sensitivity of nitroaliphatic compounds. *J. Hazard. Mater.* **2012**, *235–236*, 169–177. [CrossRef]
14. Toropova, A.P.; Toropov, A.A.; Fjodorova, N. Quasi-SMILES for predicting toxicity of Nano-mixtures to *Daphnia Magna*. *NanoImpact* **2022**, *28*, 100427. [CrossRef] [PubMed]
15. Toropov, A.A.; Toropova, A.P. Correlation intensity index: Building up models for mutagenicity of silver nanoparticles. *Sci. Total Environ.* **2020**, *737*, 139720. [CrossRef] [PubMed]
16. Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504. [CrossRef] [PubMed]
17. Toropov, A.A.; Toropova, A.P.; Benfenati, E. Additive SMILES-based carcinogenicity models: Probabilistic principles in the search for robust predictions. *Int. J. Mol. Sci.* **2009**, *10*, 3106–3127. [CrossRef]
18. Toropov, A.A.; Toropova, A.P. The system of self-consistent models for the uptake of nanoparticles in PaCa2 cancer cells. *Nanotoxicology* **2021**, *15*, 995–1004. [CrossRef]
19. Toropova, A.P.; Toropov, A.A. The system of self-consistent 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.* **2021**, *15*, 709–722. [CrossRef]
20. Toropov, A.A.; Toropova, A.P.; Roncaglioni, A.; Benfenati, E. The system of self-consistent models for pesticide toxicity to *Daphnia magna*. *Toxicol. Mech. Methods*, 2023; *in press*. [CrossRef]
21. Toropov, A.A.; Toropova, A.P.; Achary, P.G.R.; Rašková, M.; Raška, I. The searching for agents for Alzheimer's disease treatment via the system of self-consistent models. *Toxicol. Mech. Methods* **2022**, *32*, 549–557. [CrossRef]

22. Toropova, A.P.; Toropov, A.A. (Eds.) *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES. Challenges and Advances in Computational Chemistry and Physics*; Springer: Cham, Switzerland, 2023; Volume 33, pp. 1–467. [CrossRef]
23. Petrova, T.; Rasulev, B.F.; Toropov, A.A.; Leszczynska, D.; Leszczynski, J. Improved model for fullerene C_{60} solubility in organic solvents based on quantumchemical and topological descriptors. *J. Nanopart. Res.* **2011**, *13*, 3235–3247. [CrossRef]
24. Ghasemi, J.B.; Salahinejad, M.; Rofouei, M.K. Alignment independent 3DQSAR modeling of fullerene (C_{60}) solubility in different organic solvents. *Fuller. Nanotub. Carbon Nanostruct.* **2013**, *21*, 367–380. [CrossRef]
25. Cheng, W.-D.; Cai, C.-Z. Accurate model to predict the solubility of fullerene C_{60} in organic solvents by using support vector regression. *Fuller. Nanotub. Carbon Nanostruct.* **2017**, *25*, 58–64. [CrossRef]
26. Toropova, A.P.; Toropov, A.A. QSPR and nano-QSPR: What is the difference? *J. Mol. Struct.* **2019**, *1182*, 141–149. [CrossRef]
27. Roy, J.K.; Kar, S.; Leszczynski, J. Optoelectronic properties of C_{60} and C_{70} fullerene derivatives: Designing and evaluating novel candidates for efficient P3HT polymer solar cells. *Materials* **2019**, *12*, 2282. [CrossRef] [PubMed]
28. Kar, S.; Sizochenko, N.; Ahmed, L.; Batista, V.S.; Leszczynski, J. Quantitative structure-property relationship model leading to virtual screening of fullerene derivatives: Exploring structural attributes critical for photoconversion efficiency of polymer solar cell acceptors. *Nano Energy* **2016**, *26*, 677–691. [CrossRef]
29. Pourbasheer, E.; Aalizadeh, R.; Ardabili, J.S.; Ganjali, M.R. QSPR study on solubility of some fullerenes derivatives using the genetic algorithms-Multiple linear regression. *J. Mol. Liq.* **2015**, *204*, 162–169. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.