J Math Chem (2009) 46:1060–1073
DOI 10.1007/s10910-008-9491-3

ORIGINAL PAPER

QSPR modeling of octanol water partition coefficient of platinum complexes by InChI-based optimal descriptors

A. A. Toropov · A. P. Toropova · E. Benfenati

Received: 7 October 2008 / Accepted: 17 October 2008 / Published online: 18 November 2008

Abstract Comparison of the quantitative structure—property relationships (QSPR) based on optimal descriptors calculated with the International Chemical Identifier (InChI) and QSPR based on optimal descriptors calculated with simplified molecular input line entry system has shown that the InChI-based optimal descriptors give more accurate prediction for the logarithm of octanol/water partition coefficient of platinum complexes.

Keywords QSPR · InChI · SMILES · Platinum complexes · Octanol/water partition coefficient

1 Introduction

Platinum complexes are effective anti-cancer drugs [1] (Fig. 1). The logarithm of octanol/water partition coefficient ($\log P$) of a drug is related to its ability to cross cell membranes by means of diffusion. Hence, the data on the $\log P$ values for platinum complexes is important information from point of view of biochemistry and drug design.

Quantitative structure-property/activity relationship (QSPR/QSAR) is a tool for the estimation of unavailable numerical data on endpoints of interest by means of correlations 'descriptor-endpoint.' The descriptor is a numerical index of the molecular structure that can be calculated with tools such as molecular graphs [2–6] or simplified molecular input line entry system (SMILES) [7–12]. SMILES is a sequence of symbols which are images of molecular fragments [13–15]. Under such circumstances, one can develop an approach similar to the Free-Wilson scheme [16] with SMILES-based optimal descriptors [11].

A. A. Toropov (✉) · A. P. Toropova · E. Benfenati
Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milan, Italy
e-mail: aatoropov@yahoo.com

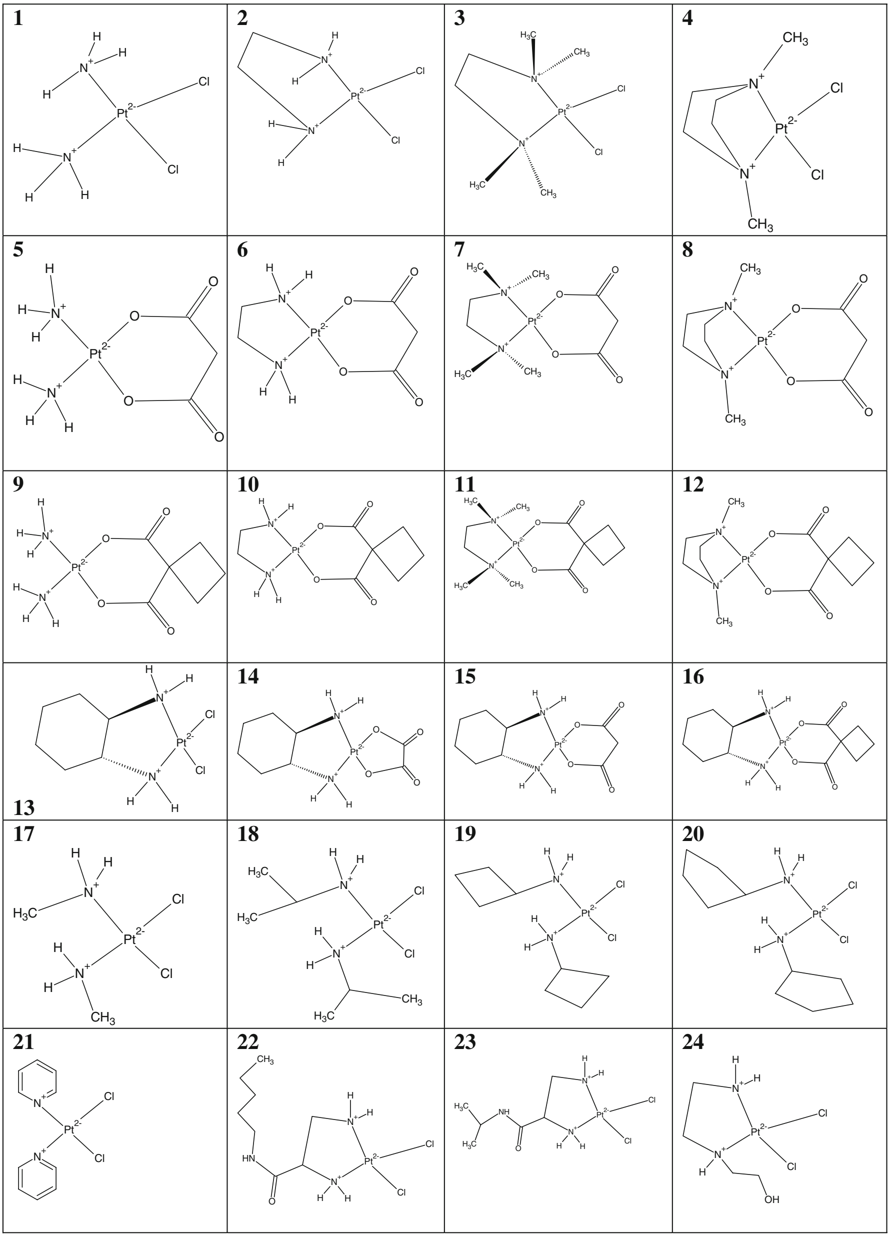
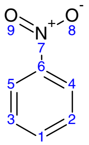


Fig. 1 Structures and ID of the platinum complexes

The International Chemical Identifier (InChI) [17,18] is an alternative of SMILES. The InChI is an automatically created unique string of symbols for a chemical compound. This string is generated using chemical properties such as atom types, nature of bonds, chirality, and atomic charge of the molecule. This string is stored using text that is organized in several "layers" corresponding to different varieties of structural

Fig. 2 Example of the InChI layers clarification from [18]



InChI=1/C6H5NO2/c8-7(9)6-4-2-1-3-5-6/h1-5H

Description	**Layers**
Formula	C6H5NO2
Connectivity	8-7(9)6-4-2-1-3-5-6
Hydrogen atoms	h1-5H

Table 1 Statistical characteristics of the SMILES-based and InChI-based models of the octanol/water partition coefficients of platinum complexes

		Training set, $n = 17$			Test set, $n = 7$		
Split	Probe of the Monte Carlo optimization	r^2	s	F	r^2	s	F
SMILES							
Split1	1	0.9629	0.180	389	0.9176	0.615	56
	2	0.9627	0.181	387	0.9202	0.621	58
	3	0.9626	0.181	386	0.9178	0.633	56
Split2	1	0.9728	0.156	536	0.3375	0.706	3
	2	0.9717	0.158	515	0.3694	0.678	3
	3	0.9722	0.157	525	0.3540	0.685	3
Split3	1	0.9767	0.153	629	0.1050	0.833	1
	2	0.9770	0.152	638	0.1078	0.818	1
	3	0.9761	0.155	613	0.0995	0.809	1
InChI							
Split1	1	0.9999	0.011	109,619	0.9641	0.177	134
	2	0.9999	0.011	111,315	0.9647	0.176	137
	3	0.9999	0.011	104,605	0.9675	0.172	149
Split2	1	0.9994	0.023	24,138	0.9658	0.170	141
	2	0.9994	0.023	24,139	0.9665	0.179	144
	3	0.9994	0.023	24,148	0.9661	0.179	142
Split3	1	0.9989	0.033	14,096	0.8783	0.182	36
	2	0.9989	0.033	14,098	0.9040	0.183	47
	3	0.9989	0.033	14,093	0.9052	0.170	48

information. This layered arrangement of InChI not only allows the software to gradually build the chemical identifiers in a series of well-defined steps, but also allows a user to selectively utilize these layers for data annotation and navigation purposes. The basic layers are formula, connectivity, and hydrogen atoms (Fig. 2). The basic layers may be extended by layers of stereo, charges, etc. [17–19].

Table 2 Correlation weights of the three probes of the Monte Carlo optimization: split1, SMILES-based model

No.	A_k	$CW(A_k)$	$CW(A_k)$	$CW(A_k)$	NA_{TRN}	NA_{TST}
1	(	0.1857265	0.1083756	0.1102056	84	50
2	+	3.3534278	2.5043207	2.8134087	30	12
3	1	−0.2289646	−0.2911812	−0.3042447	30	12
4	2	−0.9271422	−0.9462265	−0.8750283	39	18
5	3	0.8097876	0.7988719	0.7226090	12	8
6	4	−2.2170057	−2.1839837	−1.9882386	2	2
7	@@	0.3716691	0.6376347	0.4458076	4	2
8	@	−1.4299967	−1.4416197	−1.2352659	6	2
9	Cl	0.8452517	0.4412330	0.8892076	22	4
10	C	1.9107654	1.9079033	1.7101038	99	58
11	H	0.6416134	0.5597121	0.6588478	30	12
12	O=	0.3668589	−0.0393390	0.4719766	13	11
13	N	−0.8588510	−0.1594860	−0.6923705	29	13
14	Pt−2	4.1280849	3.0732377	4.3632689	17	7
15	O	−3.3914429	−3.2887640	−2.8874995	13	10
16	[N+]	0.3767157	0.1485647	0.1771194	4	2
17	[	−0.2868340	−0.3384192	−0.3335381	106	42
18	c	1.2620401	1.0494473	0.8541268	10	0
19	n	0.1337099	1.7773360	1.5757864	2	0

NA_{TRN}: the number of attributes (A_k) in the training set; NA_{TST}: the number of attributes (A_k) in the test set

The InChI can replace the SMILES in construct of the optimal descriptors. The aim of this study is comparison of the SMILES-based and InChI-based model of the octanol/water partition coefficient of platinum complexes.

2 Method

Twenty-four platinum complexes which have been examined are shown in Fig. 1. Numerical data on the octanol water partition coefficient ($\log P$) for the complexes was taken from [1]. Three random splits into training and test sets have been examined. The test set of the split1 contains platinum complexes 6, 9, 11, 12, 15, 18, 22; the test set of the split2 contains the complexes 3, 5, 8, 10, 13, 15, 21; the test set of the split3 contains complexes 7, 10, 12, 13, 16, 18, 23.

Optimal descriptors used for the QSPR-modeling of the $\log P$ are calculated as the following

$$DCW = \sum CW(A_k) \tag{1}$$

where A_k is an attribute of the SMILES or the InChI.

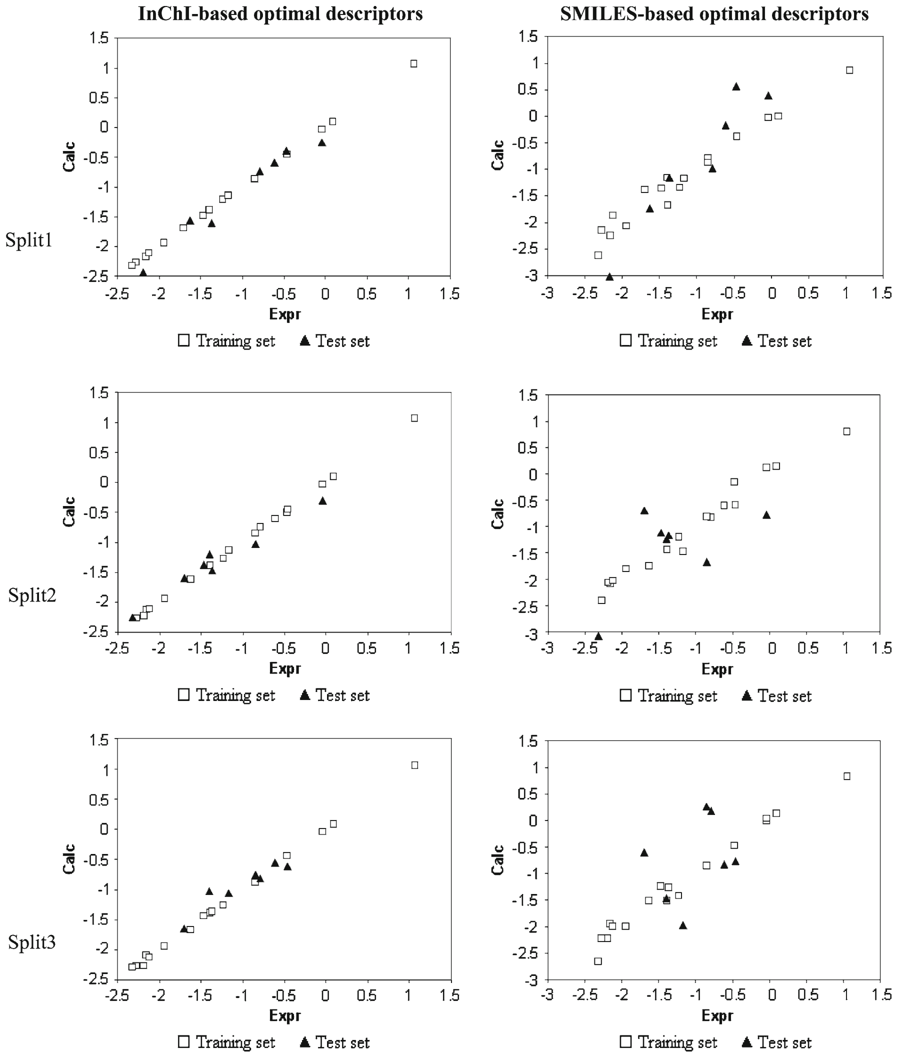


Fig. 3 Comparison of the SMILES-based and InChI-based models for the octanol/water partition coefficient (log P) of platinum complexes

SMILES and InChI for the platinum complexes have been obtained with ACD/ChemSketch [19].

The SMILES attributes [8–12] used for calculation of the DCW with Eq. 1 are the following: ‘(’, ‘+’, ‘1’, ‘2’, ‘3’, ‘4’, ‘@@’, ‘@’, ‘Cl’, ‘C’, ‘H’, ‘O=’, ‘N’, ‘Pt-2’, ‘O’, ‘[N+]’, ‘[’, c’, and ‘n’.

The InChI attributes used for calculation of the DCW with Eq. 1 are the following: ‘(10’, ‘(12’, ‘(2’, ‘(3’, ‘(4’, ‘(5’, ‘(6’, ‘(8’, ‘(9’, ‘(’, ‘*’, ‘+2’, ‘,10’, ‘,11’, ‘,12’, ‘,1’, ‘,2’, ‘,3’, ‘,4’, ‘,5’, ‘,6’, ‘,7’, ‘,8’, ‘,9’, ‘,’, ‘−10’, ‘−11’, ‘−1’, ‘−2’, ‘−3’, ‘−4, ‘−5’, ‘−6’, ‘−7’, ‘−8’, ‘−9, ‘−’, ‘.’, ‘/p−2’, ‘/’, ‘0’, ‘1’, ‘2’, ‘3’, ‘4’, ‘5’, ‘6’, ‘7’, ‘8’, ‘9’, ‘;;;;’, ‘;;;’, ‘;;’, ‘;’, ‘C2’, ‘C3’, ‘C4’, ‘C5’, ‘C6’, ‘C8’, ‘Cl’, ‘C’, ‘H11’, ‘H12’,

Table 3 Correlation weights of the three probes of the Monte Carlo optimization: split1, InChI-based model

No.	A_k	$CW(A_k)$	$CW(A_k)$	$CW(A_k)$	NA_{TRN}	NA_{TST}
1	(10	1.3206161	1.4993008	1.0708496	1	1
2	(12	0.9964627	0.9980989	0.9973323	0	1
3	(2	3.0961008	3.3249078	3.1593685	5	3
4	(3	2.8724164	3.0298989	2.8154052	2	1
5	(4	0.2851281	0.2388510	0.3635758	1	0
6	(5	−0.1956206	−0.2775790	0.0129074	9	6
7	(6	0.3142350	0.0334462	0.1496712	3	2
8	(8	0.2121796	0.0275897	0.0100814	3	3
9	(9	0.1485928	−0.2258932	0.1453974	2	3
10	(	0.2658955	0.1767416	0.2548518	52	42
11	*	1.5975147	1.7004242	1.6342472	21	5
12	+2	3.4632553	4.0701863	2.6756009	17	7
13	,10	−0.3642350	−0.4602278	−0.2044626	3	3
14	,11	1.0002657	1.0042323	1.0049945	0	1
15	,12	0.9996177	1.0049551	0.9958180	0	1
16	,1	0.5585562	0.4895263	0.6909545	12	5
17	,2	0.9988274	0.9963947	0.9958674	0	1
18	,3	0.0790021	0.0744291	−0.0252232	2	0
19	,4	0.2960958	−0.0762848	−0.1754969	4	3
20	,5	0.5282956	0.9039751	0.4888024	5	2
21	,6	1.0861147	0.9032168	1.0700031	8	3
22	,7	−0.0773815	−0.0043819	−0.1261481	9	6
23	,8	−0.5737810	0.0159991	0.1639409	2	3
24	,9	0.0523319	−0.0331767	0.2142049	3	4
25	,	−0.1355397	−0.1217153	−0.2374156	6	6
26	−10	1.0004889	0.9985008	0.9971112	0	1
27	−11	1.0028956	0.9951694	1.0045491	0	1
28	−1	1.4003937	1.1891649	1.5540471	12	5
29	−2	−0.6665529	−0.5515321	−0.5579771	16	7
30	−3	0.8077284	0.9794195	1.0090460	17	12
31	−4	3.2526992	3.0529112	3.3709538	23	10
32	−5	3.7628422	3.7142905	3.6648123	11	3
33	−6	−0.0101552	0.1495479	−0.1850035	17	9
34	−7	−0.1244988	−0.3010750	−0.1356518	9	3
35	−8	1.0580352	1.4576941	1.8006212	8	4
36	−9	0.9971270	1.0031248	1.0013629	0	1
37	−	0.1630124	0.0265314	0.2634748	6	2
38	.	−0.5638372	−0.7585328	−0.6338418	41	16
39	/p−2	3.3159769	3.6113501	3.7657206	17	7
40	/	−0.2799090	−0.4147596	−0.2956186	59	24

Table 3 continued

No.	A_k	CW(A_k)	CW(A_k)	CW(A_k)	NA_{TRN}	NA_{TST}
41	0	0.1795508	0.1968443	−0.2024467	2	3
42	1	1.0546427	0.9794899	1.1633653	30	16
43	2	1.4333024	1.2076727	1.2327514	37	11
44	3	0.8837821	0.8974788	0.6966316	4	2
45	4	2.7329826	2.7040623	2.8498811	6	4
46	5	2.4374728	2.8277112	2.5383670	6	1
47	6	2.1422646	2.4419443	2.4644654	6	5
48	7	−0.1275738	0.1397108	−0.1268069	8	7
49	8	0.3842439	0.8723362	0.6749603	3	1
50	9	1.3408698	0.9833707	1.5291936	1	0
51	;;;;	2.3041245	1.9102967	2.0761148	5	1
52	;;;	0.9632186	0.9547658	0.9208272	19	5
53	;;	0.8987676	0.7241649	0.7541002	7	5
54	;	−0.0907247	0.1132245	−0.0275953	44	22
55	C2	0.7174790	1.1403123	1.5405578	3	1
56	C3	−0.3739093	0.1045629	0.2391302	3	3
57	C4	1.4834392	1.3966765	1.3093074	2	0
58	C5	3.5657617	3.8705702	3.8456625	2	0
59	C6	1.1134160	1.1842723	0.8117528	10	6
60	C8	1.0001066	0.9994745	1.0036292	0	1
61	Cl	3.1717073	3.2740691	3.7668924	11	2
62	C	−0.9867901	−0.8894183	−1.0836099	1	0
63	H11	3.8145257	3.8489565	3.6911728	1	0
64	H12	−0.7583521	−0.7095466	−0.6990216	1	0
65	H14	1.1576299	1.0388721	0.5515735	5	2
66	H15	1.1151642	1.5922360	1.2029192	1	0
67	H16	3.2548013	2.6384030	3.3491705	2	1
68	H19	1.0020406	1.0009543	1.0039107	0	1
69	H2	−0.5878628	−0.5528193	−0.6981382	20	11
70	H3	1.6452714	1.6826133	1.8295348	10	6
71	H4	0.2899596	−0.1739670	0.1358263	3	2
72	H5	0.5172829	0.5756826	0.6336017	2	0
73	H8	1.0897919	1.0630220	0.7000046	4	4
74	H9	3.6842099	3.4018512	3.6093229	1	1
75	H	1.8852123	2.1350464	1.7708095	43	18
76	N2	1.0225276	0.9881997	0.5421019	10	4
77	N3	1.7493303	1.1670076	1.1897725	1	1
78	N	2.5655161	3.0420343	2.8237259	6	2
79	O4	−0.2524798	−0.3365038	−0.3407151	6	5
80	O	−0.4485221	−0.5505830	−0.3750065	2	1

Table 3 continued

No.	A_k	$CW(A_k)$	$CW(A_k)$	$CW(A_k)$	NA_{TRN}	NA_{TST}
81	Pt	4.4762621	3.9019305	3.1353682	17	7
82	c	0.6892123	1.0408012	1.2252149	16	7
83	h1	2.3718609	2.6079459	2.0033952	3	2
84	h2	1.9008503	2.2715106	1.7594282	5	1
85	h3	2.0983134	1.9379604	1.9522401	2	1
86	h4	0.9896445	1.2454485	1.4019823	1	0
87	h5	0.7460415	0.9296181	0.5238512	5	2
88	h6	−0.7024369	−0.6132135	−0.6722979	1	0
89	h7	1.0044924	1.0025959	1.0005819	0	1
90	m1	0.4489353	0.3083152	0.3460628	3	1
91	q	4.1785871	3.5594385	3.6961600	17	7
92	s1	0.0913054	0.5535280	0.6740237	3	1
93	t5	0.4885586	0.6105282	0.3847979	3	1

NA_{TRN}: the number of attributes (A_k) in the training set; NA_{TST}: the number of attributes (A_k) in the test set

'H14', 'H15', 'H16', 'H19', 'H2', 'H3', 'H4', 'H5', 'H8', 'H9', 'H', 'N2', 'N3', 'N', 'O4', 'O', 'Pt', 'c', 'h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'h7', 'm1', 'q', 's1', and 't5'. Each InChI-attribute reflects some molecular properties, e.g., '−1,', '−2', ..., '−11' are attributes of the connectivity layer; 'C2', 'C3', ..., 'C8', 'H2', 'H3', ..., 'H19' are component of the formula layer (Fig. 2), 't5', 's1', 'm1' are component of the stereochemical layer; 'q', ';', ';;', ';;;', and ';;;;' are attributes of the electronic charge layer [17–19]. The beginning of InChI strings (i.e., 'InChI = 1/') is not used for the DCW calculation.

The correlation weights of SMILES-attributes and InChI-attributes were calculated by the Monte Carlo optimization with the correlation coefficient between the DCW and $\log P$ for the training set as the target function (i.e., the optimization is a search for maximum of the correlation coefficient for the training set).

Having the correlation weights which provide the maximum of the correlation coefficient for the training set, one can calculate the model:

$$\log P = C0 + C1 \cdot DCW. \tag{2}$$

The predictive potential of the model calculated with Eq. 2, one can estimate with an external test set [8–12].

3 Results

Table 1 shows the statistical characteristics of the SMILES-based and the InChI-based models for $\log P$. One can see from Table 1 that InChI-based models are better for all three splits into training and test sets and more robust and reproducible. Figure 3 shows the predicted versus experimental values for these models, graphically. Tables 2 and

Table 4 Example of DCW calculation for the InChI-based optimal descriptor: `InChI ="InChI=1/2ClH.2H3N.Pt/h2*1H;2*1H3;/q;;;;+2/p−2"; DCW = 40.1725272`

A_k	$CW(A_k)$ in probe 1
2	1.4333024
Cl	3.1717073
H	1.8852123
.	−0.5638372
2	1.4333024
H3	1.6452714
N	2.5655161
.	−0.5638372
Pt	4.4762621
/	−0.2799090
h2	1.9008503
*	1.5975147
1	1.0546427
H	1.8852123
;	−0.0907247
2	1.4333024
*	1.5975147
1	1.0546427
H3	1.6452714
;	−0.0907247
/	−0.2799090
q	4.1785871
;;;;	2.3041245
+2	3.4632553
/p−2	3.3159769

3 show the numerical data on the correlation weights for calculation of the SMILES- and the InChI-based models, obtained in three runs of the Monte Carlo optimization for the first splits into training and test sets.

The first run of the Monte Carlo optimization for the InChI-based model (split1) is the following (Fig. 3):

$$\begin{aligned} \log P &= -6.6515(\pm 0.0017) + 0.1091(\pm 0.00003) * \mathrm{DCW} \qquad (3)\\ n &= 17, r^2 = 0.9999, s = 0.011, F = 109619\,(\text{training set})\\ n &= 7, r^2 = 0.9641, s = 0.176, F = 134\,(\text{test set}) \end{aligned}$$

Table 1 shows that other InChI-based model for the logarithm of the octanol water partition coefficient of platinum complexes have similar statistical quality, but results

Table 5 Experimental and calculated values of the octanol/water partition coefficient for the platinum complexes (split1, the Monte Carlo optimization probe 1)

No.	SMILES	DCW	log *P* Expr	log *P* Calc
Training set				
1	InChI = 1/2ClH.2H3N.Pt/h2*1H;2*1H3;/q;;;;+2/p−2	40.1725272	−2.270	−2.269
2	InChI = 1/C2H8N2.2ClH.Pt/c3-1-2-4;;;/h1-4H2;2*1H;/q;;;+2/p−2	41.0985921	−2.160	−2.168
3	InChI = 1/C6H16N2.2ClH.Pt/c1-7(2)5-6-8(3)4;;;/h5-6H2,1-4H3;2*1H;/q;;;+2/p−2	53.0058569	−0.850	−0.869
4	InChI = 1/C6H14N2.2ClH.Pt/c1-7-3-5-8(2)6-4-7;;;/h3-6H2,1-2H3;2*1H;/q;;;+2/p−2	49.8841288	−1.230	−1.209
5	InChI = 1/C3H4O4.2H3N.Pt/c4-2(5)1-3(6)7;;;/h1H2,(H,4,5)(H,6,7);2*1H3;/q;;;+2/p−2	39.7064937	−2.320	−2.320
7	InChI = 1/C6H16N2.C3H4O4.Pt/c1-7(2)5-6-8(3)4;4-2(5)1-3(6)7;/h5-6H2,1-4H3;1H2,(H,4,5)(H,6,7);/q;;+2/p−2	50.4494181	−1.170	−1.147
8	InChI = 1/C6H14N2.C3H4O4.Pt/c1-7-3-5-8(2)6-4-7;4-2(5)1-3(6)7;/h3-6H2,1-2H3;1H2,(H,4,5)(H,6,7);/q;;+2/p−2	47.3276900	−1.470	−1.488
10	InChI = 1/C6H8O4.C2H8N2.Pt/c7-4(8)6(5(9)10)2-1-3-6;3-1-2-4;/h1-3H2,(H,7,8)(H,9,10);1-4H2;/q;;+2/p−2	45.4849078	−1.700	−1.689
13	InChI = 1/C6H14N2.2ClH.Pt/c7-5-3-1-2-4-6(5)8;;;/h5-6H,1-4,7-8H2;2*1H;/q;;;+2/p−2/t5-,6-;;;/m1.../s1	48.2273118	−1.400	−1.390
14	InChI = 1/C6H14N2.C2H2O4.Pt/c7-5-3-1-2-4-6(5)8;3-1(4)2(5)6;/h5-6H,1-4,7-8H2;(H,3,4)(H,5,6);/q;;+2/p−2/t5-,6-;;/m1../s1	8.2383774	−1.390	−1.389
16	InChI = 1/C6H14N2.C6H8O4.Pt/c7-5-3-1-2-4-6(5)8;7-4(8)6(5(9)10)2-1-3-6;/h5-6H,1-4,7-8H2;1-3H2,(H,7,8)(H,9,10);/q;;+2/p−2/t5-,6-;;/m1../s1	53.1130137	−0.850	−0.857
17	InChI = 1/2CH5N.2ClH.Pt/c2*1-2;;;/h2*2H2,1H3;2*1H;/q;;;;+2/p−2	43.1985305	−1.940	−1.939
19	InChI = 1/2C4H9N.2ClH.Pt/c2*5-4-2-1-3-4;;;/h2*4H,1-3,5H2;2*1H;/q;;;;+2/p−2	61.8076825	0.090	0.092

Table 5 continued

No.	SMILES	DCW	log *P* Expr	log *P* Calc
20	InChI = 1/2C5H11N.2ClH.Pt/c2*6-5-3-1-2-4-5;;;/h2*5H,1-4,6H2;2*1H;/q;;;;+2/p−2	70.7053779	1.060	1.062
21	InChI = 1/2C5H5N.2ClH.Pt/c2*1-2-4-6-5-3-1;;;/h2*1-5H;2*1H;/q;;;;+2/p−2	60.6180206	−0.040	−0.038
23	InChI = 1/C6H15N3O.2ClH.Pt/c1-4(2)9-6(10)5(8)3-7;;;/h4-5H,3,7-8H2,1-2H3,(H,9,10);2*1H;/q;;;+2/p−2	56.7688854	−0.460	−0.458
24	InChI = 1/C4H12N2O.2ClH.Pt/c5-1-2-6-3-4-7;;;/h6-7H,1-5H2;2*1H;/q;;;+2/p−2	41.5497662	−2.120	−2.118
Test set				
6	InChI = 1/C3H4O4.C2H8N2.Pt/c4-2(5)1-3(6)7;3-1-2-4;/h1H2,(H,4,5)(H,6,7);1-4H2;/q;;+2/p−2	38.5421533	−2.190	−2.447
9	InChI = 1/C6H8O4.2H3N.Pt/c7-4(8)6(5(9)10)2-1-3-6;;;/h1-3H2,(H,7,8)(H,9,10);2*1H3;/q;;;+2/p−2	46.6492482	−1.630	−1.562
11	InChI = 1/C6H16N2.C6H8O4.Pt/c1-7(2)5-6-8(3)4;7-4(8)6(5(9)10)2-1-3-6;/h5-6H2,1-4H3;1-3H2,(H,7,8)(H,9,10);/q;;+2/p−2	57.3921726	−0.470	−0.390
12	InChI = 1/C6H14N2.C6H8O4.Pt/c1-7-3-5-8(2)6-4-7;7-4(8)6(5(9)10)2-1-3-6;/h3-6H2,1-2H3;1-3H2,(H,7,8)(H,9,10);/q;;+2/p−2	54.2704445	−0.790	−0.731
15	InChI = 1/C6H14N2.C3H4O4.Pt/c7-5-3-1-2-4-6(5)8;4-2(5)1-3(6)7;/h5-6H,1-4,7-8H2;1H2,(H,4,5)(H,6,7);/q;;+2/p−2/t5-,6-;;/m1../s1	46.1702592	−1.370	−1.614
18	InChI = 1/2C3H9N.2ClH.Pt/c2*1-3(2)4;;;/h2*3H,4H2,1-2H3;2*1H;/q;;;;+2/p−2	55.5128334	−0.610	−0.595
22	InChI = 1/C8H19N3O.2ClH.Pt/c1-2-3-4-5-11-8(12)7(10)6-9;;;/h7H,2-6,9-10H2,1H3,(H,11,12);2*1H;/q;;;+2/p−2?	58.6262171	−0.040	−0.255

Table 6 Changes of statistical characteristics of the InChI-based models in cases of removing of the complexes from the training set (split1)

ID of removed complex	Training set, $n = 16$			Test set, $n = 7$		
	r^2	s	F	r^2	s	F
1	0.9998	0.011	90,752	0.9680	0.171	152
2	**0.9999**	**0.010**	**113,232**	**0.8423**	**0.396**	**27**
3	1.0000	0.005	490,994	0.9611	0.176	124
4	**1.0000**	**0.005**	**595,239**	**0.9537**	**0.203**	**106**
5	0.9999	0.011	98,851	0.9655	0.188	141
7	1.0000	0.005	539,523	0.9646	0.169	137
8	1.0000	0.005	642,221	0.9601	0.194	123
10	**0.9999**	**0.010**	**119,805**	**0.8474**	**0.335**	**28**
13	**0.9999**	**0.010**	**121,288**	**0.9449**	**0.205**	**86**
14	**0.9999**	**0.011**	**106,800**	**0.9485**	**0.201**	**95**
16	**0.9999**	**0.010**	**121,190**	**0.9456**	**0.232**	**87**
17	**0.9999**	**0.011**	**95,811**	**0.8988**	**0.297**	**46**
19	**0.9998**	**0.011**	**89,235**	**0.9215**	**0.267**	**61**
20	**0.9998**	**0.011**	**64,858**	**0.9435**	**0.210**	**84**
21	0.9998	0.011	92,317	0.9579	0.187	114
23	0.9999	0.011	99,974	0.9574	0.187	113
24	0.9999	0.011	98,783	0.9615	0.194	125

The average values of the r, s and F are represented. Dispersion of the r^2 and s (the test set) is about 0.003 and 0.005, respectively. Complexes which have significant informative contribution for the InChI-based model are indicated by bold

are variable. Table 4 contains an example of the DCW calculations. Table 5 contains the experimental and calculated with Eq. 3 $\log P$ values.

4 Discussion

SMILES notation is an informative representation of the molecular structure. This representation contains the chemical elements composition, data on double and triple bonds, stereochemical data, and other details of molecular architecture [13–15]. In fact, the InChI contains the same information. However, in addition, the InChI contains the connectivity layer (Fig. 2), that is absent in the SMILES notation. Moreover, the InChI contains data on the electronic charges at different atoms [17–19]. Thus, it is not surprisingly, that InChI-based model for the $\log P$ is better than SMILES-based model.

The number of the InChI-attributes is considerable, larger than the number of the SMILES-attributes. However, since the preferable InChI-based models take place for all splits, the InChI should be estimated as interesting alternative of SMILES in the QSPR/QSAR analyses.

The study of influence of removing of compounds of training set for the statistical characteristics of prediction can be useful information. If absence of a complex in the training set leads to decrease of the statistical quality of the model for an external test set, one can classify the complex as one that has significant informative contribution for the model. The influence of each complex of the training set (split1) is represented in Table 6.

One can see from the Table 6 that the absence of complexes 2, 4, 10, 13, 14, 16, 17, 19, and 20 had decreased the statistical quality of the $\log P$ prediction. One can also see from Table 6 that the influence of the absence of 1, 3, 5, 7, 8, 21, 23, and 24 for statistical characteristics of the $\log P$ model is weaker. It is to be noted that the absence of complex 1 leads to slight improvement of the $\log P$ model. Probably the informative contribution of this complex is not important for the InChI-based model.

The model described in [1] based on numerical data for the polarisability and dipolemoment (derived from density functional theory calculations) for the $\log P$ of 24 complexes (Table 4, Fig. 1) was characterized by $n = 24$, $r^2=0.952$, $s = 0.210$, and $F = 95$. Thus, the statistical characteristics of the InChI-based model that is calculated with Eq. 3 are better.

5 Conclusions

The SMILES- and InChI-based models for the octanol/water partition coefficient of platinum complexes are different (Fig. 3). The statistical characteristics of the InChI-based predictions are better for three splits into training and test sets, which have been examined. The informative contributions of complexes of the training set to quality of the $\log P$ prediction is different (Table 6). Statistical characteristics of the InChI-based models can be classified as stable ones.

Acknowledgements The authors thank the Marie Curie Fellowship for financial support (contract ID 39036, CHEMPREDICT).

References

1. J.A. Platts, S.P. Oldfield, M.M. Reif, A. Palmucci, E. Gabano, D. Osella, J. Inorg. Biochem. **100**, 1199–1207 (2006)
2. E. Estrada, N. Guevara, I. Gutman, J. Chem. Inf. Comput. Sci. **38** , 428–431 (1998)
3. P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. **16**, 7944–7955 (2008)
4. K. Roy, I. Sanyal, G. Ghosh, QSAR Comb. Sci. **26**, 629–646 (2007)
5. J.A. Castillo-Garit, Y. Marrero-Ponce, J. Escobar, F. Torrens, R. Rotondo, Chemosphere **73**, 415–427 (2008)
6. A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Molec. Diversity **10**, 405–414 (2006)
7. D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. **45**, 386–393 (2005)
8. A.A. Toropov, A.P. Toropova, D.V. Mukhamedzhanova, I. Gutman, Ind. J. Chem. A **44**, 1545–1552 (2005)
9. A.A. Toropov, D. Leszczynska, J. Leszczynski, Chem. Phys. Lett. **441**, 119–122 (2007)
10. A.A. Toropov, A.P. Toropova, I. Raska Jr., Eur. J. Med. Chem. **43**, 714–740 (2008)
11. A.A. Toropov, E. Benfenati, Bioorg. Med. Chem. **16**, 4801–4809 (2008)

12. A.A. Toropov, A.P. Toropova, E. Benfenati, Chem. Phys. Lett. **461**, 343–347 (2008)
13. D. Weininger, J. Chem. Inf. Comput. Sci. **28**, 31–36 (1988)
14. D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. **29**, 97–101 (1989)
15. D. Weininger, J. Chem. Inf. Comput. Sci. **30**, 237–243 (1990)
16. K. Hasegawa, N. Yokoo, K. Watanabe, M. Hirata, Y. Miyashita, S.-I. Sasaki, Chemometr. Intel. Lab. **33**, 63–69 (1996)
17. http://wwmm.ch.cam.ac.uk/inchifaq/
18. http://www.warr.com/
19. ACD/ChemSketch Freeware, version 11.00 (Advanced Chemistry Development, Inc., Toronto, ON, Canada, 2007), http://www.acdlabs.com