Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Report

13 February 2012

QSAR Studies on AhR Ligands

Early stage researcher:
Alexander Safanyaev

Project supervisor:
Dr. Igor V. Tetko

Research Institution:
Helmholtz Zentrum Muenchen

Introduction

Three classes of aryl hydrocarbon receptor ligands were under investigation. Namely, it was polysubstituted dibenzo-p-dioxins (PSDDs), polysubstituted dibenzofuranes (PSDFs) and polybrominated dipenyl ethers (PBDEs), which are well-known as hazardous environmental contaminants. Different QSAR approaches — molecular field topology analysis (MFTA)[1], and seven machine learning methods implemented into OCHEM platform were used[2]. Separate models for each class of compounds and hybrid models have been created. Statistical parameters and graphical visualisation of derived results indicated their fine quality, absence (or minimal occurence) of chance correlations and high predictive power. MFTA topological maps allowed to get some inside into ligand-AhR interactions. Using OCHEM tools, the applicability domain (AD) of constructed models was stated. It gave the possibility to detect outliers, improve the quality of models and to distinguish reliable and unreliable predictions. It also allowed to make suggestions about inaccurate experimental measurements of AhR binding affinities for a few compounds and to propose structural explanations for some compounds with unreliable predictions. Comparison of created models with another QSAR studies on AhR ligands showed that proposed models were in a good qualitative accordance with the previous results and were comparable or even better in a statistical sense.

Experimental Data

AhR binding affinities for PSDDs, PSDFs and PBDEs were reported previously[3][4][5][6][7][8]. In all calculations we used EC_{50} values converted into logarithm of inverse concentration ($pEC_{50} = \log(1/EC_{50})$). This data is collected in **Table 1**.

Results and discussion

We have constructed seven types of models using OCHEM tools and MFTA approach: three separate models for each class of AhR ligands, then three hybrid models (including compounds from two classes in each model), and finally, one model which contained compounds from all classes. This information is summarized in **Tables 2-6**. It should be noted that the 4th, 6th and 7th types of models were built only by OCHEM tools. The reason is that in current version of MFTAWin it is not possible to unify both dibenzo-p-dioxin and dibenzofuran (and even dibenzofuran and biphenyl ethers) scaffolds in one MSG. We will analyze all results one after another, and then compare different groups of models and used approaches.

Table 2. Separate and hybrid models

type of model	*n*	classes included in model	*N*
separate	**I**	PSDDs	49
	II	PSDFs	53
	III	PBDEs	18
hybrid	**IV**	PSDDs + PSDFs	102
	V	PSDDs + PBDEs	67
	VI	PSDFs + PBDEs	71
	VII	PSDDs + PSDFs + PBDEs	120

N is the number of compounds in training set; *n* is the number of dataset

Table 3. Statistics for MFTA models

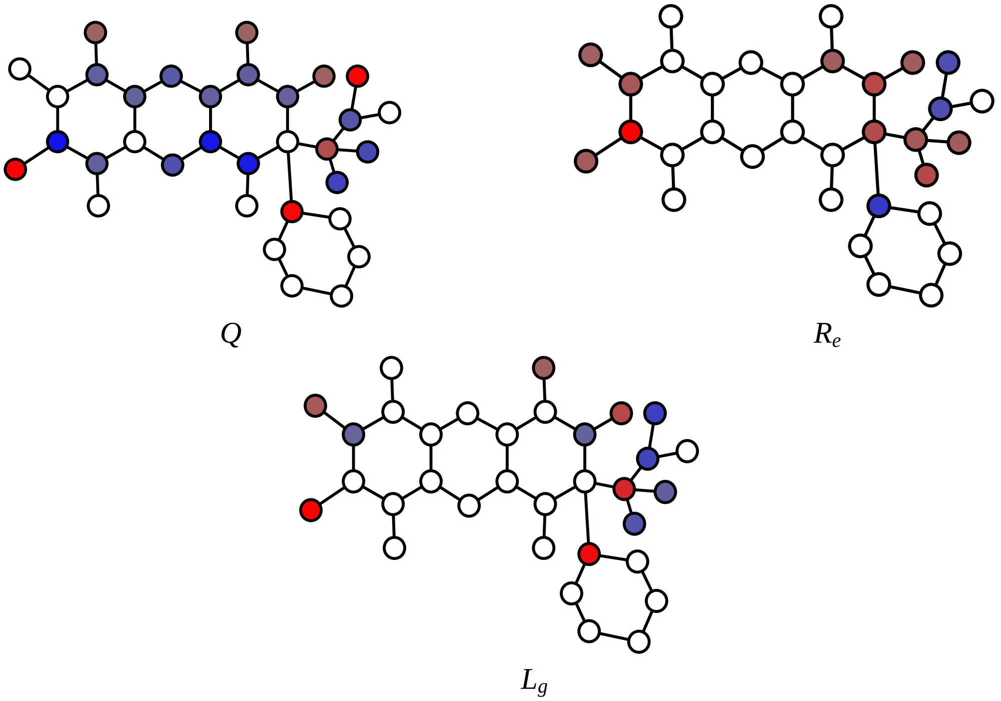
n	descriptors	Statistical parameters of model			
		N_F	R_Y	Q^2	AvgErr
I	Q	6	0.87	0.19	0.474
	R_e	2	0.74	0.40	0.649
	L_q	1	0.69	0.29	
	Q,R_e	6	0.92	0.53	0.391
	Q,R_e,L_q	6	0.95	0.70	0.293
	Q,R_e,L_q,H_a,H_d	6	0.96	0.72	0.273
II	Q	4	0.93	0.75	0.363
	R_e	4	0.91	0.72	0.379
	L_q	4	0.92	0.74	
	Q,R_e	4	0.92	0.74	0.359
	Q,R_e,L_q	4	0.93	0.77	0.335

	Q,*R*$_e$,*L*$_g$,*H*$_a$,*H*$_d$		4	0.94	0.79	0.311
III	*Q*		3	0.87	0.32	0.320
	P$_a$		2	0.87	0.36	0.321
	R$_e$		2	0.86	0.31	0.321
	L$_g$		1	0.73	0.20	
	Q,*P*$_a$		3	0.88	0.30	0.318
	P$_a$,*R*$_e$		2	0.86	0.33	0.321
III*	*Q*		3	0.88	0.50	0.307
	P$_a$		2	0.86	0.60	0.342
	R$_e$		3	0.89	0.47	0.308
	L$_g$		1	0.74	0.38	
	Q,*P*$_a$		2	0.85	0.39	0.329
	P$_a$,*R*$_e$		3	0.89	0.51	0.304
V	*Q*	general	5	0.88	0.38	0.469
		PSDDs	6	0.93	0.32	0.380
		PBDEs	1	0.74	0.37	0.395
	Q,*P*$_a$	general	5	0.87	0.37	0.492
		PSDDs	6	0.93	0.32	0.372
		PBDEs	1	0.76	0.38	0.390
	Q,*R*$_e$	general	6	0.92	0.63	0.381
		PSDDs	6	0.94	0.71	0.324
		PBDEs	1	0.67	0.30	0.443
	Q,*R*$_e$,*P*$_a$	general	6	0.92	0.63	0.382
		PSDDs	6	0.95	0.72	0.301
		PBDEs	1	0.66	0.28	0.446
V*	*Q*	general	6	0.90	0.63	0.423
		PSDDs	6	0.93	0.65	0.362
		PBDEs	1	0.82	0.55	0.386
	Q,*P*$_a$	general	6	0.91	0.58	0.397
		PSDDs	6	0.94	0.62	0.333
		PBDEs	1	0.80	0.54	0.383
	Q,*R*$_e$	general	4	0.91	0.68	0.386
		PSDDs	5	0.94	0.73	0.335
		PBDEs	1	0.81	0.49	0.334
	Q,*R*$_e$,*P*$_a$	general	6	0.93	0.69	0.373
		PSDDs	6	0.96	0.78	0.287
		PBDEs	1	0.80	0.48	0.351

N_F is the optimal number of factors in PLS-model, R_Y is the correlation coefficient for activity matrix (without cross-validation), Q^2 is the cross-validation parameter, AvgErr is the average error; local descriptors: *Q* is the partial atomic charge calculated by electronegativity equalization approach proposed by Oliferenko et. al., *R*$_e$ is the van der Waals radius of first environment (atom + neighbours), *L*$_g$ is the group lipophilicity (atom + hydrogens), *H*$_d$ (*H*$_a$) is the ability of an atom in a given environment to be a donor (acceptor) of a hydrogen bond characterized by the binding constants (Abraham approach), P_a and P_b are the site occupancy factors for atoms (P_a) and bonds (P_b) (which have the value 1 if a given feature is present in the structure and 0 otherwise); when the number of dataset is marked with asterisk, it means that genetic algorithm for selection of descriptors was used

For PSDDs (dataset **I**), the statistical parameters are presented in **Tables 4** (for OCHEM models) and **3** (for MFTA models). Analysing values of Q^2 and RMSE in OCHEM models one can see that the best results were derived using neural networks ANN, ASNN and FSMLR in conjunction with CDK and ADRIANA descriptors. High values of Q^2 indicate that we successfully eliminated the chance correlations. The plot with comparison of experimental and predicted pEC_{50} values for the best model (FSMLR, ADRIANA descriptors; $R^2 = 0.76$, $Q^2 = 0.75$, RMSE = 0.614, MAE = 0.500) is shown in **Figure 1** (*a*). It also confirms the fine correlation between selected descriptors and AhR binding affinities. To consider AD of this model one can use **Figure 2** (*a*). In accordance with the plot, the least reliable predictions were obtained for the compounds with the values of DM more than 0.9. Among these dioxins there were 2,3-dichloro-7-substituted- or 2-substituted -3,7,8-trichlorodioxins, where 7th and 2d substituents were another than halogens, e.g. ester, nitro, amide groups (compounds **35**, **36**, **39**, **43**, **48**, **49**). We proposed that relatively low reliability of predictions for mentioned dioxins might be connected with non-homogenicity of the training set: all of its compounds had atoms of halogens as substituents, and only the minority had another substituents in one position. We should also stress that, in spite of good predictions for 9 compounds (corresponding dots on the plot are inside the rectangle with DM > 0.9, absolute error (Y axis) < 0.7), it can be due to a chance. This observation indicates that we did not exclude all random correlations from the model. That is why one should carefully estimate results derived for similar compounds with the help of this model.

Using MFTA approach, we found a very important divergence in quality of models with different sets of descriptors. When only one electrostatic descriptor was included, we got unapplicable model with $Q^2 = 0.19$ ($N_F = 6$). Addition of steric and hydrophobic descriptors dramatically improved this situation (especially including of R_e) and lead to the highly predictive model with Q^2 equal to 0.70. The models with only steric (R_e) or hydrophobic (L_g) descriptors have much better quality and less number of PLS-components (for R_e: $Q^2 = 0.40$, $N_F = 2$; for L_g: $Q^2 = 0.29$, $N_F = 1$), than that one with electrostatic descriptor. It indicates that the steric and hydrophobic interactions are of prime importance for dioxins binding. This conclusion is in a good accordance with previous studies[3][4][9] and the fact, that there were topological and geometrical descriptors filtered and presented in the best OCHEM models considered above. We can analyse the contributions of local descriptors using topological maps (**Figure 3**).



Q

R_e

L_g

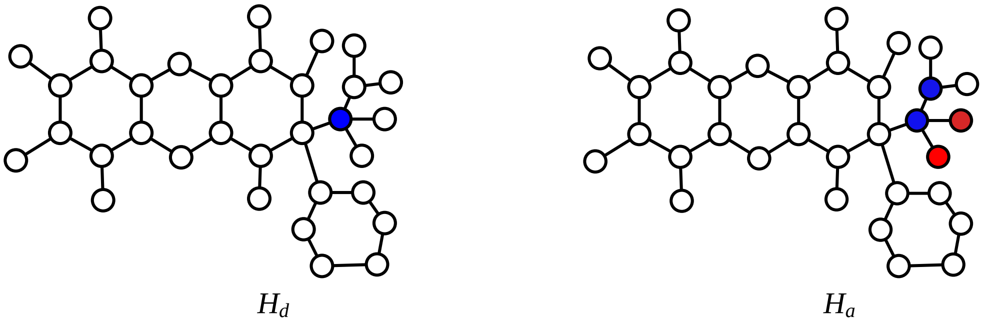


Figure 3. Contributions of electrostatic (Q), steric (R_e), hydrophobic (L_g) descriptors, and the abilities of an atom in a given environment to be a donor (H_d) or acceptor (H_a) of a hydrogen bond in PSDDs affinity to AhR in MFTA model (dataset **I**). Increasing the value of any descriptor in vertices with tints of red (and decreasing in those with tints of blue) leads to higher activity

Steric contributions clearly demonstrate that addition of bulky substituents (considering only monoatomic substituents) in all lateral (2-, 3-, 7- and 8-) positions is favourable for AhR affinity. Activity of PSDDs increases in the following series: 2,3-dichloro-dibenzo-p-dioxin < 2,3,7-trichloro-dibenzo-p-dioxin < TCDD < 2-iodo-3,7,8-trichloro-dibenzo-p-dioxin < TBDD more than 2.5 order of magnitude in logarithmic scale. The hydrophobic factor acts at the same way: increasing of lipophilicity in lateral positions also enlarges the binding affinity. Electrostatic contributions show that the increasing values of atomic charges on substituents in lateral positions and enrichment of aromatic dibenzo-p-dioxin system with electron density are preferable for dioxins activity. One can note that in the case of halogens all these contributions act at the same direction: moving from fluorine to iodine, the atomic radius, lipophilicity and electropositivity enlarge. Also, polarizability of molecule and degree of enrichment of the dibenzo-p-dioxin ring system with electron density increase for heavy halogens. It can be explained using the obvious fact, that positive mesomeric effects for halogens prevail much more than negative inductive effects in this case. We may suggest that the distribution of electron density in the dioxins system play an important role in stacking interactions with the receptor. On the other hand, insertion of bulky substituents in non-lateral positions decreases the activity. We also found the interesting fact: in one lateral position preferable functional groups (polyatomic) should have descriptors' distribution similar to trifluoromethyl group. It allows us to propose that the subtle balance between electrostatic, steric and hydrophobic factors takes place for substituents in lateral position: 2-trifluoromethyl-3,7,8-dibenzo-p-dioxin has a very high affinity (pEC_{50} = 8.495), but on the other hand, activities of 2-methyl-3,7,8-dibenzo-p-dioxin (pEC_{50} = 6.886) and 2-hydroxy-3,7,8-dibenzo-p-dioxin (pEC_{50} = 5.495) are much lower. It may be connected both with strong electrostatic interactions of trifluoromethyl group with positively charged residue of amino acid of AhR binding domain (alternatively, this group can serve as a hydrogen bond acceptor), and with proper van der Waals radius and lipophilicity of this group. This suggestion is in accordance with the fact, that amino and hydroxyl groups (i.e., donors of hydrogen bond; these substituents have qualitatively opposite charge distribution than trifluoromethyl group) in lateral position dramatically decrease the affinity (compare pEC_{50} values for compounds **40**, **47** and **49**). The distribution of H_a and H_d descriptors also confirms it.

Among previous studies on PSDDs, we should mention CoMFA model proposed by Waller and McKinney[3]. Their 4-component model had a fine statistical parameters (Q^2 = 0.72, R^2 = 0.92, SE = 0.450). They included in the training set only 25 polyhalogenated dioxins (with chlorine and bromine atoms as a substituents); on the one hand, we can conclude that we extended the scope of applicability of this model; on the other hand, 5-fold cross validation, bagging procedure and Stable-CV procedure in our studies seem to give more reliable statistical results in comparison with the leave-one-out procedure used in this CoMFA modelling.

For PSDFs (datasets **II, II***), statistics is presented in Tables **4** (OCHEM modelling) and **3** (MFTA modelling). The best results were derived using ANN, ASNN and LibSVM methods in conjunction with Spectrophores. Firstly, compounds **79** and **92** were recognized as outliers. It looks good at **Figures 1** (*b*) and **2** (*b*). Then we excluded these compounds and rebuilt 3 best models. It significantly improved statistical parameters; e.g., for ASNN model Q^2 value enlarged from 0.69 to 0.81, RMSE value decreased from 0.623 to 0.477. Because there were not a great structural diversity between these two compounds and another compounds from the training set, and also they had a low values of DM (i.e., they were inside the AD), we proposed that unreliable predictions might be connected with uncorrect experimental values of EC_{50}. Only two compounds (**83**, **102**) had the values of DM more than 0.4, and binding affinities for another two compounds inside the AD (**52**, **85**) were predicted with the absolute error more than 1 logarithmic unit. These facts, and the exellent statistical parameters (e.g., for ANN: $R^2 = 0.82$, $Q^2 = 0.81$, RMSE = 0.475, MAE = 0.370) indicate that we built highly predictive models.

In MFTA modelling, we also constructed fine models with high values of Q^2 and four PLS-components. Models with one electrostatic, steric and hydrophobic descriptors had the Q^2 values equal to 0.75, 0.72 and 0.74, respectively. Futher creating of model with 5 descriptors slightly improved these values ($Q^2 = 0.79$). It indicates that electrostatic, steric and hydrophobic factors are prevailing and give approximately the same contribution in the AhR-PSDFs interactions. Let us consider some details (**Figure 4**).

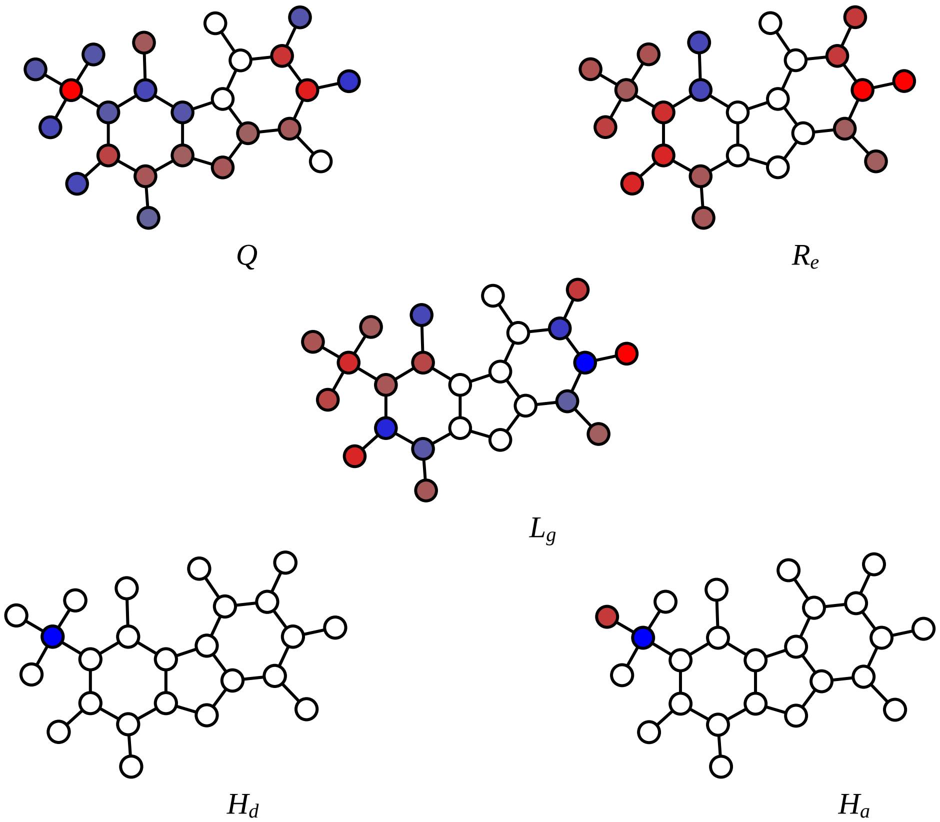


Figure 4. Contributions of electrostatic (Q), steric (R_e), hydrophobic (L_g) descriptors, and the abilities of an atom in a given environment to be a donor (H_d) or acceptor (H_a) of a hydrogen bond in PSDFs affinity to AhR in MFTA model (dataset **II**). Increasing the value of any descriptor in vertices with tints of red (and decreasing in those with tints of blue) leads to higher activity

Contributions of steric descriptors demonstrate that bulky substituents in all lateral (2-, 3-, 7-, 8-) and two non-lateral (4-, 6-) positions are favourable for AhR affinity. It should be stressed that the increasing of bulk in the 3d and 7th positions gives the most important contribution into activity. Activity enhances in the following series: dibenzofuran < 2-chlorodibenzofuran < 3-chlorodibenzofuran < 2,3,7,8-tetrachlorodibenzofuran (TCDF) more than 4.3 order of magnitude in logarithmic scale. In contrast, enlarging steric bulk in the other two non-lateral positions (1, 9) leads to a lower binding affinity. These results are in a a good qualitative accordance with those reported previously by Safe et al.[4] Authors used a simple plus/minus notations to represent the contributions of substituents in each position. Considering the hydrophobic factors, one can see that they act almost in the same way (like for dioxins): increasing of lipophilicity in all lateral and two non-lateral (4, 6) positions is preferable for activity, whereas in the other two non-lateral positions it leads to a lower affinity. Analyzing both abilities of atoms to be a hydrogen bond donor (acceptor) and electrostatic contributions, one can find the intriguing feature: at the 8th position, the trifluoromethyl (and bromomethyl) group leds to a relatively high affinity. 8-trifluoromethyl- and 8-(bromomethyl)-2,3,4-2,3,4-trichlorodibenzofurans are one of the most active congeners among the furan series (pEC_{50} = 7.060 and 6.577, respectively). So, like in analysis of dioxins' affinity, we can propose that the very presize balance between steric and electrostatic factors may take place for this substituent. High AhR binding affinity can be caused not only by enhancement of van der Waals radius of this group, but also by Coulombic interactions of fluorines (bromine) with the positively charged amino asid residue (alternatively, by forming of a hydrogen bond, where halogen would be an acceptor; it is in accordance with contributions of H_a and H_d descriptors).

We can compare created models with CoMFA study on PSDFs performed by Waller and McKinney[3]. They built 5-component model with good correlation (Q^2 = 0.74, R^2 = 0.86, SE = 0.539, 39 compounds). Again, we extended the training set (by including PSDFs not only with clorine atoms as a substituents), and, consequently, the AD of our model. We also used validation procedures mentioned above; it gives more reliable results that leave-one-out procedure applied in this CoMFA study.

The statistical results for PBDEs (datasets **III**, **III***) activity studies are summarized in **Tables 4** (OCHEM modelling) and **3** (MFTA modelling). It is obvious that the quality of these models is dramatically lower in comparison with the models constructed for PSDDs and PSDFs. The best one was obtained using LibSVM method in conjunction with GSFrag descriptors (R^2 = 0.41, Q^2 = 0.39, RMSE = 0.599, MAE = 0.473). Corresponding plot (**Figure 1**, *c*) also confirms the significant dispersion of predicted pEC_{50} values. Our failure may be explained by two factors. On the one hand, this training set (18 compounds) was the least representative one, and such a little amount of data would not be enough to reveal correlations between structure and activity of PBDEs. On the other hand, one can propose the low accuracy in the experimental measurements. These suggestions were corroborated after construction of hybrid models (see below). We should also stress another QSAR studies on PBDEs[7], where authors derived different models with Q^2_{LOO} ranged from 0.29 to 0.91. The important remark is that leave-one-out (LOO) teqnique can lead to a great overestimation of Q^2 value. Papa et. al. used MLRA in conjunction with DRAGON descriptors. They obtained a model with a good statistics (R^2 = 0.90, Q^2_{LOO} = 0.79; Q^2_{EXT} = 0.76, R^2_{EXT} = 0.73) but the dataset was splited into a training (10 compounds) and test set (8 compounds). From our viewpoint, apart from applying LOO teqnique, it is not reasonable to split this small dataset of 18 compounds into training and validation sets. The data contained in 10 compounds is not enough to create highly predictive model. We decided to find another solutions of this problem: first, genetic selection of descriptors was applied in MFTA approach; also, hybrid models were constructed (see below).

One can compare MFTA models for PBDEs with (**Table 3**, **III***) and without (**Table 3**, **III**) application of genetic algorithm for selection of descriptors. The number of PLS-components is almost the same in both cases, but the values of Q^2 are much better when genetic algorithm applied.

It is interesting that the best result ($Q^2 = 0.60$) was obtained using only one descriptor (occupancy factor, P_a). It was even better than all models created by OCHEM tools.

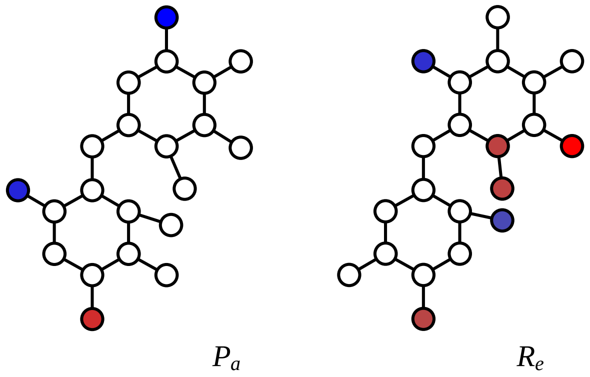


Figure 5. Contributions of steric descriptors (R_e) and occupancy factors (P_a) in PSDFs affinity to AhR in MFTA model (dataset **III***). Increasing the value of any descriptor in vertices with tints of red (and decreasing in those with tints of blue) leads to higher activity

Because only bromine atoms were substituents in diphenyl ether system, it is useful to analyze the contributions of occupancy factors P_a and steric descriptors R_e (**Figure 5**). Insertion of more than two bromine atoms in the ortho-position (or especially occurrence of two ortho-bromine atoms in one ring) decreases binding affinity. It may be connected with the different conformational behaviours of ortho-substituted diphenyl ethers. Additional bromine atoms lead to the distortion of planar structure. Activity increases in the following series: 2,2′,4,4′,5,6′-hexabrominated DE < 2,3′,4,4′,6-pentabrominated DE < 3,3′,4,4′-tetrabrominated DE. On the other hand, all active compounds have two substituents in para-position of both rings.

We performed hybrid modelling for PSDDs and PSDFs (**datasets IV**, **IV***) to extend the sphere of applicability of previous models. The statistical results are presented in **Table 5**. Apart from general statistics, we also calculated it separately for each class to compare results with those for splited models. The best correlations were derived using neural networks

ANN and ASNN in combination with CDK and ADRIANA descriptors. Compounds **40** and **79** were recognized as outliers and excluded from the training set. After rebilding results became slightly better. AD (**Figure 2**, *c*) and visual plot (**Figure 1**, *d*) of the finest model (ASNN, ADRIANA descriptors; general: R^2 = 0.75, Q^2 = 0.75, RMSE = 0.601, MAE = 0.470; PSDDs: Q^2 = 0.74, RMSE =0.627; PSDFs: Q^2 = 0.72, RMSE =0.576) also illustrate its high quality and give bases for exclusion of mentioned compounds. It is interesting that 1,3,8-trichlorodibenzofuran **79** was also removed from the training set of separate model as outlier. 2-trifluoromethyl-3,7,8-trichlorodibenzo-p-dioxin **40** lies very far from AD of constructed model (DM = 0.75); it may be due to the trifluoromethyl group, which is quite "exotic" substituent among another compounds of the training set. We should mention compounds with the values of DM more than 0.55 (**5**, **8**, **27**, **32**, **62**, **102**); again, three of them have trifluoromethyl or phenyl group as a substituent, and high values of DM for these compounds can be explained by non-homogenicity of the training set (the same suggestion was considered above in discussion of the separate model for PSDDs). Compounds 4, 11, 84 and 92 have the most significant absolute errors (more than 1.3 logarithmic unit); these predictions should be considered as unreliable. In general, this hybrid model has a high predictive power and allows to predict binding affinities for both classes. The scope of applicability was extended without loss of statistical quality (in comparison with separate models). AD of our model was also significantly extended in comparison with the CoMFA model proposed by Waller and McKinney[9] (7 PLS-components, Q^2 = 0.71, R^2 = 0.879, SEP = 0.971, 64 compounds). Comparison of validation procedures in CoMFA and in our approaches is mentioned above.

The next type of hybrid models was created for PSDDs and PBDEs (datasets **V, V***). The goal was not only to make dataset more representative, but also to obtain better correlations for PBDEs. The statistics is summarized in **Tables 5** (OCHEM modelling) and **3** (MFTA modelling). We built two models in OCHEM using LibSVM and MLRA methods in conjunction with Chemaxon and ADRIANA descriptors, respectively. General statistics was fine with Q^2 equal to 0.69 for both models. Statistical parameters for PSDDs also were good ($Q^2 = 0.64$), but those for PBDEs improved insignificantly ($Q^2 = 0.37$ and 0.42). Analysing AD of LibSVM model, we found that one compound — 1,2,3,4,6,7,8,9-octachlorodibenzo-p-dioxin **13** lies very far from the AD (DM = 2.27). Among another compounds with the values of DM > 1 (**15**, **38**, **39**, **48**, **104**, **118**), there were hydroxy-, amino- and acetamino- dioxin derivatives; it can be explained again by non-homogenicity of the training set. Also, trifluoromethyl-, nitro- and aminosubstituted dioxins were found among compounds with the highest absolute errors of prediction (**14**, **40**, **43**, **49**, **113**; absolute error > 1.2).

Using MFTA approach for hybrid modelling (without genetic selection of descriptors) gave good results only with two sets of descriptors Q, R_e and Q, R_e, P_a with general Q^2 equal to 0.63. It is interesting that separate statistics for PSDDs was improved ($Q^2 = 0.71$; compare with the same set of descriptors — Q, R_e in the model for only PSDDs with Q^2 equal to 0.53) while statistical parameters for diphenyl ethers were slightly lower that in separate model for PBDEs. Situation changed when we applied genetic algorithm for selection of descriptors. Statistics for both PSDDs and PBDEs was fine with all used sets of descriptors. Number of PLS-components in each class were almost the same as in the separate modelling. It indicates that constructed models were stable and highly predictive. Besides, the dataset became more representative, and these models could be applied for PSDDs and PBDEs.

Hybrid modelling with PSDFs and PBDEs (datasets **VI**, **VI***) is seemed to be the most successful one. The statistical parameters are presented in **Table 6**. The finest correlations were derived using ANN, ASNN and regression methods in conjunction with Chemaxon descriptors. The plot with comparison of experimental vs predicted pEC_{50} values (**Figure 1**, *f*) and AD (**Figure 2**, *e*) of one model (ANN, Chemaxon descriptors; general: $R^2 = 0.83$, $Q^2 = 0.82$, RMSE = 0.439, MAE = 0.351; for PSDFs: $Q^2 = 0.82$, RMSE = 0.458; for PBDEs: $Q^2 = 0.66$, RMSE =0.375) illustrate perfect results after removal of four outliers: two dibenzofurans (**79**, **92**), and two diphenyl ethers (**109**, **113**). These compounds had significantly higher absolute error than another representatives of the training set. After their exclusion statistical parameters became practically two-times better for PBDEs and much better for PSDFs. As far as these outliers were inside the AD, and because previous models also gave unreliable predictions for them, our suggestion about inaccurate experimental measurements is confirmed again. We should mention another compounds with unreliable predictions: **58**, **71**, **74**, **85** (DM > 0.4) and **77** (absolute error > 1); all of them were PSDFs (including two hydroxy derivatives). This model can be considered as the best one both for PSDFs and, especially, for PBDEs.

Finally, we created one general model with PSDDs, PSDFs and PBDEs (dataset **VII**) using LibSVM approach in combination with Chemaxon descriptors. Statistics is summarized in **Table 6**. **Figure 1**, *g* illustrates the fine correlation between experimental and predicted pEC_{50} values. Statistical parameters for PSDFs ($Q^2 = 0.53$, RMSE = 0.766) and PBDEs ($Q^2 = 0.43$, RMSE = 0.578) were slightly worse that those in previous models, but still good. AD (**Figure 2**, *f*) allows to reveal unreliable predictions for dibenzofurans **58**, **59**, **85**, **93**, **99** (DM > 1) and compounds **11**, **40**, **79**, **92**, **95** (absolute error > 1.2). In spite of an unsignificant loss of quality for PSDFs and PBDEs, this general model can be used for predictions of a new compounds from all three classes; it is the most representative one.

References

[1] Palyulin, Radchenko, and Zefirov, “Molecular field Topology analysis method in QSAR studies of organic compounds,” *J Chem Inf Comput Sci*, vol. 40, no. 3, pp. 659-667, May 2000.

[2] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.-Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, and I. V. Tetko, “Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information,” *J. Comput. Aided Mol. Des.*, vol. 25, no. 6, pp. 533-554, Jun. 2011.

[3] C. L. Waller and J. D. McKinney, “Comparative molecular field analysis of polyhalogenated dibenzo-p-dioxins, dibenzofurans, and biphenyls,” *J. Med. Chem.*, vol. 35, no. 20, pp. 3660-3666, Oct. 1992.

[4] S. H. Safe, “Comparative toxicology and mechanism of action of polychlorinated dibenzo-p-dioxins and dibenzofurans,” *Annu. Rev. Pharmacol. Toxicol.*, vol. 26, pp. 371-399, 1986.

[5] M. A. Denomme, K. Homonoko, T. Fujita, T. Sawyer, and S. Safe, “Effects of substituents on the cytosolic receptor-binding avidities and aryl hydrocarbon hydroxylase induction potencies of 7-substituted 2,3-dichlorodibenzo-p-dioxins. A quantitative structure-activity relationship analysis,” *Mol. Pharmacol.*, vol. 27, no. 6, pp. 656-661, Jun. 1985.

[6] Vedani, McMasters, and Dobler, “[Genetic algorithms in 3D-QSAR: Predicting the toxicity of dibenzodioxins, dibenzofurans and biphenyls],” *ALTEX*, vol. 16, no. 1, pp. 9-14, 1999.

[7] E. Papa, S. Kovarich, and P. Gramatica, “QSAR modeling and prediction of the endocrine-disrupting potencies of brominated flame retardants,” *Chem. Res. Toxicol.*, vol. 23, no. 5, pp. 946-954, May 2010.

[8] M. A. Denomme, K. Homonko, T. Fujita, T. Sawyer, and S. Safe, “Substituted polychlorinated dibenzofuran receptor binding affinities and aryl hydrocarbon hydroxylase induction potencies--a QSAR analysis,” *Chem. Biol. Interact.*, vol. 57, no. 2, pp. 175-187, Feb. 1986.

[9] C. L. Waller and J. D. McKinney, “Three-dimensional quantitative structure-activity relationships of dioxins and dioxin-like compounds: model validation and Ah receptor characterization,” *Chem. Res. Toxicol.*, vol. 8, no. 6, pp. 847-858, Sep. 1995.

Supporting Information

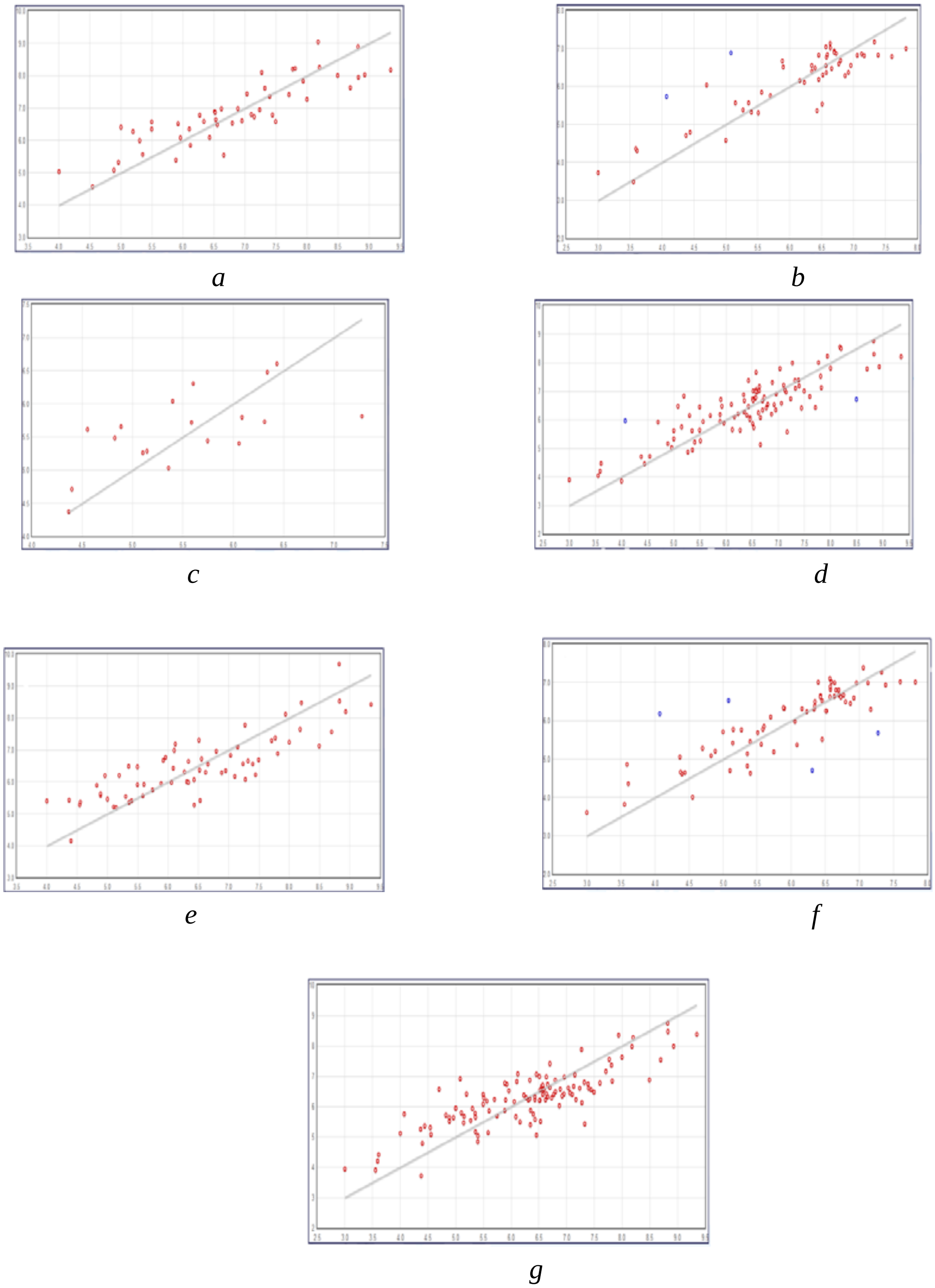


Figure 1. Experimental vs predicted by OCHEM tools pEC_{50} values: *a* – dataset **I** (FSMLR, *Adriana*), *b* – dataset **II*** (ANN, *Spectrofores*), *c* – dataset **III** (LibSVM, *GSFrag*), *d* – dataset **IV*** (ASNN, *Adriana*), *e* – dataset **V** (LibSVM, *ChemaxonDescriptors 7.4*), *f* – dataset **VI*** (ANN, *ChemaxonDescriptors 7.4*), *g* – dataset **VII** (LibSVM, *ChemaxonDescriptors 7.4*). Excluded compounds are marked as blue points

Table 4. Statistics for separate OCHEM models (datasets **I**, **II**, **III**)

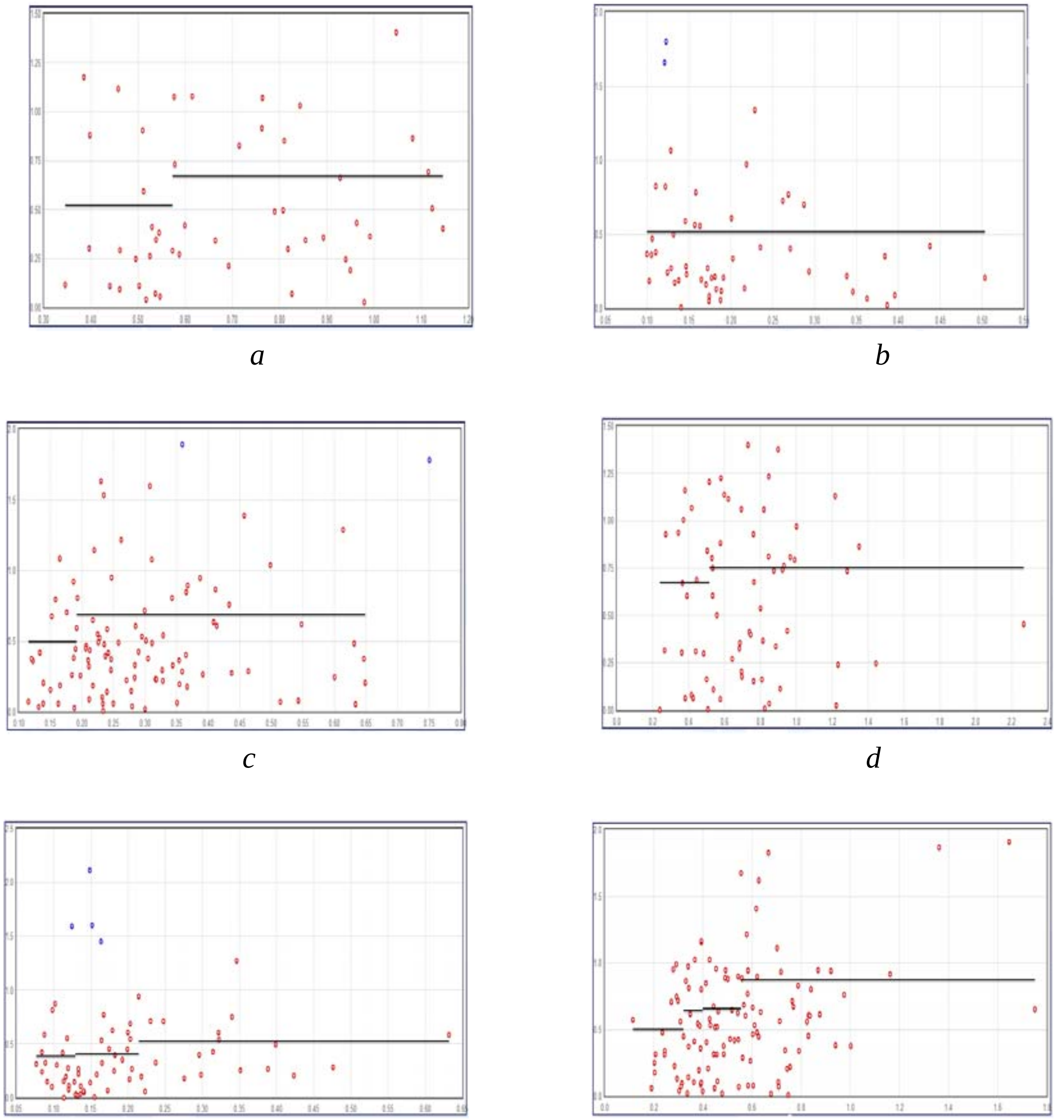
n of dataset	Descriptors	Machine learning method													
		ANN		ASNN		KNN		LibSVM		FSMLR		MLRA		PLS	
		Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*
I	*CDK*	0.70	0.680	0.73	0.646	0.63	0.754	0.55	0.832					0.64	0.746
	Dragon6 (blocks: 1-29)	0.55	0.833	0.58	0.802			0.53	0.845						
	Fragmentor (Length 2-4)	0.58	0.807	0.59	0.792	0.54	0.842	0.58	0.804						
	GSFrag	0.67	0.715	0.70	0.682			0.52	0.859						
	Chemaxon Descriptors (7.4)	0.59	0.797	0.62	0.759	0.51	0.870	0.51	0.865						
	Adriana	0.74	0.637	0.75	0.622	0.61	0.772			0.75	0.614			0.65	0.736
	Spectrophores	0.53	0.850	0.53	0.846										
II	*Fragmentor (Length 2-4)*	0.51	0.782					0.58	0.723						
	GSFrag	0.55	0.744	0.54	0.756										
	Chemaxon Descriptors (7.4)	0.60	0.702	0.62	0.686					0.59	0.714				
	Inductive Descriptors	0.59	0.716	0.63	0.673										
	Adriana	0.66	0.645	0.68	0.630	0.60	0.705	0.57	0.727						
	Spectrophores	0.71	0.604	0.69	0.623	0.62	0.685	0.70	0.612	0.60	0.704	0.62	0.690	0.61	0.698
II*	*Spectrophores*	0.81	0.475	0.81	0.477			0.78	0.513						
III	*Chemaxon Descriptors (7.4)*	0.27	0.655												
	Adriana	0.31	0.641			0.28	0.653								
	GSFrag							0.39	0.599						
	Inductive Descriptors													0.26	0.660

Table 5. Statistics for hybrid OCHEM models (datasets **IV**, **V**)

n of dataset	Descriptors	Statistics	Machine learning method									
			ANN		ASNN		KNN		LibSVM		MLRA	
			Q^2	*RMSE*	*Q*^{2}	*RMSE*	*Q*^{2}	*RMSE*	*Q*^{2}	*RMSE*	*Q*^{2}	*RMSE*
IV	*CDK*	general	0.72	0.652	0.73	0.640	0.65	0.736	0.57	0.808		
		PSDDs	0.70	0.684	0.70	0.675	0.59	0.794	0.65	0.732		
		PSDFs	0.69	0.621	0.71	0.604	0.63	0.677	0.39	0.873		
		PSDDs	0.64	0.748	0.65	0.729	0.54	0.844	0.41	0.954		
		PSDFs	0.60	0.701	0.63	0.677	0.63	0.675	0.67	0.639		
	GSFrag	general	0.67	0.713	0.67	0.713			0.59	0.795		
		PSDDs	0.72	0.656	0.72	0.656			0.48	0.894		
		PSDFs	0.53	0.762	0.53	0.762			0.62	0.689		
	Adriana	general	0.71	0.662	0.73	0.638	0.63	0.748				
		PSDDs	0.68	0.705	0.70	0.680	0.65	0.729				
		PSDFs	0.69	0.619	0.71	0.596	0.53	0.765				
IV*	*CDK*	general	0.72	0.645	0.74	0.617						
		PSDDs	0.69	0.687	0.71	0.662						
		PSDFs	0.69	0.604	0.73	0.571						
	Adriana	general	0.74	0.616	0.75	0.601						
		PSDDs	0.73	0.641	0.74	0.627						
		PSDFs	0.70	0.593	0.72	0.576						
V	*Chemaxon Descriptors (7.4)*	general							0.69	0.710		
		PSDDs							0.64	0.741		
		PBDEs							0.37	0.610		
	Adriana	general									0.69	0.700
		PSDDs									0.64	0.743
		PBDEs									0.42	0.585

Table 6. Statistics for hybrid OCHEM models (datasets **VI**, **VII**)

n of dataset	Descriptors	Statistics	Machine learning method											
			ANN		ASNN		LibSVM		FSMLR		MLRA		PLS	
			Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*	Q^2	*RMSE*
VI	*CDK*	general	0.61	0.659	0.63	0.648								
		PSDFs	0.66	0.647	0.67	0.638								
		PBDEs	0.19	0.693	0.22	0.678								
	Chemaxon Descriptors (7.4)	general	0.67	0.606	0.69	0.588			0.64	0.632	0.65	0.624	0.61	0.664
		PSDFs	0.71	0.599	0.73	0.580			0.67	0.639	0.68	0.631	0.63	0.677
		PBDEs	0.34	0.626	0.37	0.612			0.37	0.609	0.38	0.603	0.34	0.625
		PSDFs	0.56	0.741	0.57	0.726								
		PBDEs	0.34	0.627	0.33	0.627								
VI*	*CDK*	general	0.70	0.573	0.72	0.551								
		PSDFs	0.72	0.573	0.75	0.545								
		PBDEs	0.20	0.572	0.20	0.571								
	Chemaxon Descriptors (7.4)	general	0.82	0.439	0.83	0.432			0.77	0.507	0.74	0.538	0.72	0.552
		PSDFs	0.82	0.458	0.83	0.448			0.76	0.539	0.72	0.573	0.72	0.580
		PBDEs	0.66	0.375	0.66	0.374			0.63	0.390	0.59	0.408	0.49	0.456
	Inductive Descriptors	general	0.59	0.674	0.60	0.665								
		PSDFs	0.57	0.715	0.59	0.703								
		PBDEs	0.32	0.527	0.32	0.530								
VII	*Chemaxon Descriptors (7.4)*	general					0.67	0.705						
		PSDDs					0.70	0.678						
		PSDFs					0.53	0.766						
		PBDEs					0.43	0.578						



a

b

c

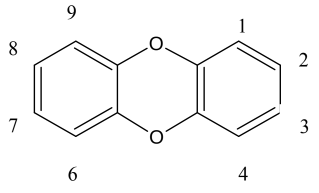
d

e

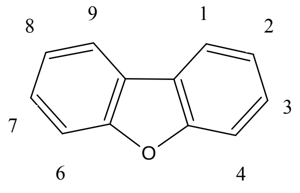
f

Figure 2. The plot shows the predictions for the training set against the “distance to model” (DM, selected type is BAGGING-STDEV); *a* – dataset **I** (FSMLR, *Adriana*), *b* – dataset **II*** (ANN, *Spectrofores*), *c* - dataset **IV*** (ASNN, *Adriana*), *d* – dataset **V** (LibSVM, *ChemaxonDescriptors 7.4*), *e* – dataset **VI*** (ANN, *ChemaxonDescriptors 7.4*), *f* – dataset **VII** (LibSVM, *ChemaxonDescriptors 7.4*). Excluded compounds are marked as blue points

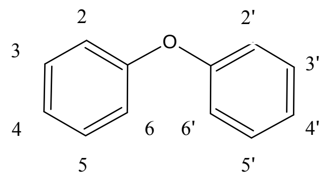
Table 1. Structures of PSDDs, PSDFs and PBDEs and their AhR binding affinities



A



B



C

no.	structure	R	pEC$_{50}$	no.	structure	R	pEC$_{50}$
1	A	2,3,7,8-Cl_4	8.000	**61**	B	8-Br, 2,3,4-Cl_3	6.577
2	A	1,2,3,7,8-Cl_5	7.102	**62**	B	8-CF_3, 2,3,4-Cl_3	7.060
3	A	2,3,6,7-Cl_4	6.796	**63**	B	8-I, 2,3,4-Cl_3	6.575
4	A	2,3,6-Cl_4	6.658	**64**	B	8-F, 2,3,4-Cl_3	6.230
5	A	1,2,3,4,7,8-Cl_6	6.553	**65**	B	8-Me, 2,3,4-Cl_3	6.870
6	A	1,3,7,8-Cl_4	6.102	**66**	B	8-i-Pr, 2,3,4-Cl_3	6.730
7	A	1,2,4,7,8-Cl_5	5.959	**67**	B	8-Et, 2,3,4-Cl_3	6.799
8	A	1,2,3,4-Cl_4	5.886	**68**	B	8-t-Bu, 2,3,4-Cl_3	6.592
9	A	2,3,7-Cl_3	7.149	**69**	B	2,3,4-Cl_3	5.561
10	A	2,8-Cl_2	5.495	**70**	B	8-OMe, 2,3,4-Cl_3	5.900
11	A	1,2,3,4,7-Cl_5	5.194	**71**	B	8-OH, 2,3,4-Cl_3	5.270
12	A	1,2,4-Cl_3	4.886	**72**	B	8-CH_2Br, 2,3,4-Cl_3	6.635
13	A	1,2,3,4,6,7,8,9-Cl_8	5.000	**73**	B	2-Cl	3.553
14	A	1-Cl	4.000	**74**	B	3-Cl	4.377
15	A	2,3,7,8-Br_4	8.824	**75**	B	4-Cl	3.000
16	A	2,3-Br_2, 7,8-Cl_2	8.830	**76**	B	2,6-Cl_2	3.609
17	A	2,8-Br_2, 3,7-Cl_2	9.350	**77**	B	2,8-Cl_2	3.590
18	A	2-Br, 3,7,8-Cl_3	7.939	**78**	B	1,3,6-Cl_3	5.357
19	A	1,3,7,9-Br_4	7.032	**79**	B	1,3,8-Cl_3	4.071
20	A	1,3,7,8-Br_4	8.699	**80**	B	2,6,7-Cl_3	6.347
21	A	1,2,4,7,8-Br_5	7.770	**81**	B	2,3,4,6-Cl_4	6.456
22	A	1,2,3,7,8-Br_5	8.180	**82**	B	2,3,7,8-Cl_4	7.387
23	A	2,3,7-Br_3	8.932	**83**	B	1,2,4,8-Cl_4	5.000
24	A	2,7-Br_2	7.810	**84**	B	1,2,4,6,7-Cl_5	7.169
25	A	2-Br	6.530	**85**	B	1,2,4,7,9-Cl_5	4.699
26	A	2,3-Cl_2, 7-F	6.951	**86**	B	1,2,3,4,8-Cl_5	6.921
27	A	2,3-Cl_2, 7-CF_3	7.710	**87**	B	1,2,3,7,8-Cl_5	7.128
28	A	2,3-Cl_2, 7-OMe	6.510	**88**	B	1,2,4,7,8-Cl_5	5.886
29	A	2,3-Cl_2, 7-Br	7.320	**89**	B	2,3,4,7,8-Cl_5	7.824
30	A	2,3-Cl_2, 7-I	7.270	**90**	B	1,2,3,4,7,8-Cl_6	6.638
31	A	2,3-Cl_2, 7-CN	5.921	**91**	B	1,2,3,6,7,8-Cl_6	6.569
32	A	2,3-Cl_2, 7-Ph	6.620	**92**	B	1,2,4,6,7,8-Cl_6	5.081
33	A	2,3-Cl_2, 7-t-Bu	6.520	**93**	B	2,3,4,6,7,8-Cl_6	7.328
34	A	2,3-Cl_2, 7-Me	6.429	**94**	B	2,3,6,8-Cl_4	6.658
35	A	2,3-Cl_2, 7-NO_2	6.337	**95**	B	1,2,3,6-Cl_4	6.456
36	A	2,3-Cl_2, 7-COOMe	6.270	**96**	B	1,2,3,7-Cl_4	6.959

37	A	2,3-Cl_2, 7-H	6.120	**97**	B	1,3,4,7,8-Cl_5	6.699
38	A	2,3-Cl_2, 7-OH	5.350	**98**	B	2,3,4,7,9-Cl_5	6.699
39	A	2,3-Cl_2, 7-NH_2	4.541	**99**	B	1,2,3,7,9-Cl_5	6.398
40	A	2-CF3, 3,7,8-Cl_3	8.495	**100**	B		3.000
41	A	2-I, 3,7,8-Cl_3	8.201	**101**	B	2,3,4,7-Cl_4	7.602
42	A	2-OMe, 3,7,8-Cl_3	7.495	**102**	B	1,2,4,6,8-Cl_5	5.509
43	A	2-NO_2, 3,7,8-Cl_3	7.444	**103**	C	4-Br	5.102
44	A	2-F, 3,7,8-Cl_3	7.398	**104**	C	4,4□-Br_2	5.585
45	A	2-CN, 3,7,8-Cl_3	7.237	**105**	C	2,2□,4-Br_3	5.357
46	A	2-Me, 3,7,8-Cl_3	6.886	**106**	C	2,4,4□-Br_3	6.086
47	A	2-OH, 3,7,8-Cl_3	5.495	**107**	C	2,2□,4,4□-Br_4	5.745
48	A	2-CH_3CONH, 3,7,8-Cl_3	5.301	**108**	C	2,2□,4,5□-Br_4	4.824
49	A	2-NH_2, 3,7,8-Cl_3	4.959	**109**	C	2,3□,4,4□-Br_4	6.310
50	B	8-t-Bu, 2,3-Cl_2	6.570	**110**	C	2,3□,4□,6-Br_4	4.553
51	B	8-F, 2,3-Cl_2	5.150	**111**	C	2,4,4□,6-Br_4	5.602
52	B	8-I, 2,3-Cl_2	6.429	**112**	C	3,3□,4,4□-Br_4	6.337
53	B	8-i-Pr, 2,3-Cl_2	6.520	**113**	C	2,2□,3,4,4□-Br_5	7.276
54	B	8-Br, 2,3-Cl_2	6.350	**114**	C	2,2□,4,4□,5-Br_5	5.143
55	B	2,3,8-Cl_3	6.160	**115**	C	2,2□,4,4□,6-Br_5	4.886
56	B	8-Me, 2,3-Cl_2	5.699	**116**	C	2,3□,4,4□,6-Br_4	6.056
57	B	8-OMe, 2,3-Cl_2	6.510	**117**	C	3,3□,4,4□,5-Br_5	6.432
58	B	8-OH, 2,3-Cl_2	4.440	**118**	C	2,2□,4,4□,5,5□-Br_6	4.398
59	B	2,3-Cl_2	5.401	**119**	C	2,2□,4,4□,5,6□-Br_6	4.367
60	B	2,3,4,8-Cl_4	6.770	**120**	C	2,2□,3,4,4□,5□,6-Br_7	5.398