**Modeling of adipose/blood partition coefficient for environmental chemicals**

**K.C. Papadaki1, S.P. Karakitsios1,2,3, D.A. Sarigiannis1,2,3**

1Aristotle University of Thessaloniki, Department of Chemical Engineering, Environmental Engineering Laboratory, University Campus, Thessaloniki 54124, Greece

2 HERACLES Research Center on the Health and Exposome, Center for Interdisciplinary Research and Innovation, Balkan Center, Bldg. B, 10th km Thessaloniki-Thermi Road, 57001, Greece

3 School for Advanced Study (IUSS), Piazza della Vittoria 15, Pavia 27100, Italy

**\*CORRESPONDING AUTHOR**

Prof. Dimosthenis Sarigiannis

Director,

Environmental Engineering Laboratory

Department of Chemical Engineering

Aristotle University of Thessaloniki

University Campus, Bldg. D, Rm 201

54124 Thessaloniki, Greece

tel. +30-2310-994562

e-mail: denis@eng.auth.gr

**Abstract**

Α Quantitative Structure Activity Relationship (QSAR) model was developed in order to predict the adipose/blood partition coefficient of environmental chemical compounds. The first step of QSAR modeling was the collection of inputs. Input data included the experimental values of adipose/blood partition coefficient and two sets of molecular descriptors for 67 organic chemical compounds; a) the descriptors from Linear Free Energy Relationship (LFER) and b) the PaDEL descriptors. The datasets were randomly split to training and prediction set and were analysed using two statistical methods; Genetic Algorithm based Multiple Linear Regression (GA-MLR) and Artificial Neural Networks (ANN). The models with LFER and PaDEL descriptors, coupled with ANN, produced satisfying performance results. The fitting performance (R2) of the models, using LFER and PaDEL descriptors, was 0.94 and 0.96, respectively. The Applicability Domain (AD) of the models was assessed and then the models were applied to a large number of chemical compounds with unknown values of adipose/blood partition coefficient. In conclusion, the proposed models were checked for fitting, validity and applicability. It was demonstrated that they are stable, reliable and capable to predict the values of adipose/blood partition coefficient of “data poor” chemical compounds that fall within the applicability domain.

**Highlights**

* QSARs accurately predict adipose/blood partition coefficient for environmental chemicals
* QSARs analysed by ANN outperform the ones analysed by MLR
* QSARsfill the data gaps of “data poor” chemicals inside applicability domain

**Keywords:** QSARs, adipose/blood partition coefficient, Multiple Linear Regression, Artificial Neural Networks, Linear Free Energy Relationship, PaDEL descriptors

**Nomenclature**

|  |  |
| --- | --- |
| **QSARs** | Quantitative Structure Activity Relationships |
| **PBTK** | Physiologically Based Toxicokinetic models |
| **ADMET** | Absorption, Distribution, Metabolism, Excretion and Toxicity |
| **ECHA** | European Chemicals Agency |
| **REACH** | Registration, Evaluation, Authorisation and Restriction of Chemicals |
| **OECD** | Organisation for Economic Co-operation and Development |
| **LFER** | Linear Free Energy Relationship |
| **SP** | Biological property |
| **E** | Excess molar refractivity (cm3/mol/10) |
| **S** | Dipolarity/polarizability |
| **A** | Effective or summation hydrogen bond acidity |
| **B** | Effective or summation hydrogen bond basicity |
| **V** | McGowan characteristic volume (cm3/mol/100) |
| **e** | Measure of the tendency of a chemical to interact with solute π- and n- electrons |
| **s** | Measure of the polarizability of a chemical |
| **a** | Measure of the hydrogen bond acidity of a chemical |
| **b** | Measure of the hydrogen bond basicity of a chemical |
| **v** | Measure of the lipophilicity of a chemical |
| **Padipose/blood** | Adipose/blood partition coefficient |
| **PCA** | Principal Component Analysis |
| **PCs** | Principal Components |
| **GA** | Genetic Algorithm |
| **GA-MLR** | Genetic Algorithm based Multiple Linear Regression |
| **OLS** | Ordinary Least Squares |
| **MCDM** | Multi-Criteria Decision Making |
| **ANN** | Artificial Neural Networks |
| **LM** | Levenberg-Marquardt algorithm |
| **SCG** | Scaled Conjugate Gradient algorithm |
| **BFGS** | Broyden–Fletcher–Goldfarb–Shanno Quasi-Newton algorithm |
| **AD** | Applicability Domain |
| **PDF** | Probability Density Function |
| **LOO** | Leave One Out validation |
| **LMO** | Leave Many Out validation |
| **MSE** | Mean Squared Error |
| **RMSE** | Root Mean Squared Error |
| **kNN** | k Nearest Neighbors |
| **PFCs** | Perfluorinated Compounds |
| **PCBs** | Polychlorinated Biphenyls |
| **BTEX** | Benzene, Toluene, Ethylbenzene, Xylene compounds |
| **VOCs** | Volatile Organic Compounds |
| **PAHs** | Polycyclic Aromatic Hydrocarbons |
| **PBDEs** | Polybrominated Diphenyl Ethers |

# Introduction

In recent years, there is an increasing interest for the development and application of Physiologically Based Toxicokinetic (PBTK) models in toxicity testing and health risk assessment. PBTK models providequantitative descriptions of the Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) processes of environmental or pharmaceutical chemicals ([WHO, 2010](#_ENREF_55)). This is achieved by simulating the fate of these substances in the body using mathematical equations regarding mass transport, fluid dynamics, and biochemistry.

The species-specific inputs required for solving the PBTK models are physiological, anatomical and biochemical data that can be obtained from literature ([Arms et al., 1988](#_ENREF_6); [Brown et al., 1997](#_ENREF_10); [Davies and Morris, 1993](#_ENREF_11)). The chemical-specific inputsinclude the kinetic parameters for metabolic processes and the physicochemical ones, such as tissue/air, blood/air and tissue/blood partition coefficients. These parameters can be obtained by in vitro and in vivo measurements or by computational algorithms, known as “in silico” methods([Lipscomb et al., 2012](#_ENREF_29); [Peyret and Krishnan, 2011](#_ENREF_40)).

Quantitative Structure – Activity Relationships (QSARs) are regression or classification models, included in “in silico” approaches, that form a relationship between the biological effects and the chemical structure of a chemical compound ([Puzyn et al., 2010](#_ENREF_45)). QSARs have been widely used as methodology for assisting governments and industry to manage risks posed by chemicals. Theyhave been suggested by European Chemicals Agency (ECHA) and promoted in the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation regarding health and safety of chemical compounds([European Commission, 2006](#_ENREF_13)).

Adipose/blood partition coefficient is considered as one of the most significant input parameters of PBTK models. The partitioning of chemical compounds into adipose tissue and blood provide information regarding distribution and toxicological effects of these substances ([Pierce et al., 1996](#_ENREF_42)).Several QSAR models have been developed for the prediction of tissue/blood partition coefficientswith respect to drugs and environmental chemical compounds. Most of the studies refer to animal species, especially rats ([Béliveau et al., 2003](#_ENREF_9); [Paixão et al., 2014](#_ENREF_37); [Peyret et al., 2010](#_ENREF_41); [Rodgers and Rowland, 2006](#_ENREF_46)), while some of them involve human ([Baláž and Lukáčová, 1999](#_ENREF_8); [Liu et al., 2005](#_ENREF_30); [Zhan and Zhang, 2006](#_ENREF_58); [Zhang, 2004](#_ENREF_59)).

The aim of this study was to develop reliable QSAR models, which could serve as a tool for the prediction of tissue/blood partition coefficient of environmental chemical compounds for the human adipose tissue.The models were trained and validated internally and externally, while their domain of applicability was assessed, developing in this way models according to the Organisation for Economic Co-operation and Development (OECD) Principles for the regulatory acceptance of QSARs ([OECD, 2014](#_ENREF_35)).

# Material and methods

## Overall Approach

The methodological approach presented in this studyis based on the development of QSAR models to predict adipose/blood partition coefficient for environmental chemical compounds. The necessary input data for the models to be trainedconsisted of a) the values of adipose/blood partition coefficient and b) the values of descriptors for67 environmental chemicals of the initial set.The input descriptors were divided into two individual sets; the Linear Free Energy Relationship (LFER) and the PaDEL descriptors. The sets were analysed using two different statistical techniques; the Genetic Algorithm based Multiple Linear Regression (GA-MLR) and the Artificial Neural Networks (ANN). The applicability domain (AD) of the models with the best statistical performance was exploredto define the structural and physicochemical space, where the models’ predictions can be reliable.In conclusion, the developed modelswereapplied to a large number of chemical compounds that fall within the applicability domain and for which values of adipose/blood partition coefficientwere missing.

## LFER Descriptors

## The first set of descriptors characterized theLFER equationproposed by Abraham ([1993](#_ENREF_1)). This relationship describes the equilibrium concentrationof a chemical from the liquid phase to solvents or other condensed phases andis defined as follows:

 (1)

Where *SP* is a biological property for the chemical in a given system. The independent descriptors represent the properties of the examined chemicals. Specifically,*E* is the excess molar refractivity of the chemical, *S* is the chemical’s dipolarity/polarizability, *A* and *B* are the chemical’s effective or summation hydrogen bond acidity and basicity, respectively, and *V* is the McGowan characteristic volume of the chemical. The coefficients *c*, *e*, *s*, *a*, *b* and *v*describe the contribution of these properties, so *e* corresponds to the tendency of the chemical to interact with solute π- and n- electrons, *s* corresponds to the chemical’s dipolarity/polarizability, *a* and *b* corresponds to the chemical’s hydrogen bond basicity and acidity, respectively, and *v* is a measure of chemical’s lipophilicity([Abraham, 1993](#_ENREF_1); [Payne and Kenny, 2002](#_ENREF_38)).

## PaDEL Descriptors

The second dataset consisted of 1444 1D and 2D descriptors, known as PaDEL descriptors. These descriptors are related to the molecular structure of the chemicals and characterized as constitutional, topological, geometrical or electronic([Yap, 2011](#_ENREF_57)). Constitutional descriptors reflect the chemical composition of a compound, while topological and geometrical descriptors are derived from the graph representation and the three-dimensional structure of the molecule, respectively. Electronic descriptors are obtained by expressing the electronic and geometrical interactions into the molecules using quantum chemical methods ([Karelson et al., 1996](#_ENREF_27); [Todeschini and Consonni, 2008](#_ENREF_54)).

## Data Collection

The experimental and computational values of input parametersof the examined environmental chemicals were required for the analysis of the models. The experimental values of adipose/blood partition coefficient were collected from literature ([Baláž and Lukáčová, 1999](#_ENREF_8); [DeJongh et al., 1997](#_ENREF_12); [Pelekis and Krishnan, 2004](#_ENREF_39)) and are listed in. The computational values of LFER descriptors were obtained from Abraham and co-workers ([Abraham et al., 1994a](#_ENREF_2); [Abraham et al., 1994b](#_ENREF_3); [Abraham et al., 1999](#_ENREF_4); [Sprunger et al., 2008](#_ENREF_53)) and from the online database “Open Notebook Science” ([Open Notebook Science, 2016](#_ENREF_36)). PaDEL descriptorswere derived from the molecular structure of the chemicals. The 3D chemical structures were drawn using theACD/ChemSketch ([Advanced Chemistry Development Inc., 2012](#_ENREF_5)) and were transferred into the PaDEL-Descriptor ([Yap, 2011](#_ENREF_57))software for the calculation of PaDEL descriptors.The values of LFER and PaDEL descriptors are provided in Supplementary file S1.

## Data Preparation

Data preparation includeda)prereduction and dimensionality reduction of the initial descriptors, b) distribution and categorization of the chemicals using Principal Component Analysis (PCA), c) data splitting.

The prereduction process was followed for the LFER and PaDEL descriptors in order to avoid the semi-constant (>80%) and intercorrelated (>95%) ones. One LFER and 1016 PaDEL descriptors were excluded from the analysis and development of the models. The dimensionality reduction was necessary because of the large number of PaDEL inputs, which could cause overfitting problems. Overfitting occurs when there are too many parameters for a sample. It is important about 10 to 15 observations for each term be considered in a regression model([Frost, 2015](#_ENREF_15)).

PCA was used for further reduction of PaDEL inputs, as well as for the distribution of chemical compounds. Data mapping to a lower-dimensional space was performed in such a way that the variance of the data in the low-dimensional representation was maximized. The scores of data representation, called Principal Components,whose eigenvalues were found to be equal to 1 or greater, were retained as the most significant for data analysis([Kaiser, 1960](#_ENREF_26)).Variance maximizing (varimax) rotation was then followed to calculate the loadings of the important principal components for each of the original descriptors. The purpose of varimax rotation was to investigate the underlying characteristics in scores’ structure by detecting the contribution of the initial descriptors to principal components ([Girden, 2011](#_ENREF_19)).

## Statistical Analysis

The dataset of 67 environmental chemical compounds, with the experimental values of adipose/blood partition coefficient and the computed descriptors, was randomly divided into the training set, containing 70% of the total data, the validation and the test set, each one representing 15% of the total set. The datasets were analysed using the statistical methodsof GA-MLR and ANN. Genetic Algorithm (GA) was implemented for feature selection in PaDEL descriptors set. The GA-MLR technique was implemented in QSARINS software([Gramatica et al., 2014](#_ENREF_20); [Gramatica et al., 2013](#_ENREF_21)), while the ANN technique was implemented in MATLAB®(version R2016a, Mathworks Inc)using the Neural Network Toolbox.

### Genetic Algorithm basedMultiple Linear Regression

The statistical method of MLR, combined with Ordinary Least Squares (OLS), was used for the model development in QSARINS software. Several models were developed and ordered according to their fitting performance. Internal (Leave One Out (LOO), Leave Many Out (LMO)) and external validation was applied for the evaluation of models’ validity. GA was used for the selection of the optimal set of descriptors for each of the models, creating several populations of different possible solutions until the acceptable result was reached([Gharagheizi, 2008](#_ENREF_18); [Liu and Long, 2009](#_ENREF_31)).The fitness function evaluated the goodness of solutions and resulted in the best combination of descriptors using the Tournament Selection method ([Haupt and Haupt, 2004](#_ENREF_23)). The selection of the best model was based on the Multi-Criteria Decision Making (MCDM) value, which summarized the fitting, cross validation and external validation criteria.

### Artificial Neural Networks

The relationship between the two datasets of descriptorsand adipose/blood partition coefficient was mapped by a neural network. A two-layer feed-forward network with sigmoid hidden neurons and linear output neurons was developed for the data fitting. The network was trained with several algorithms including Levenberg-Marquardt (LM), Scaled Conjugate Gradient (SCG) and BFGS Quasi-Newton (BFG), while Mean Squared Error (MSE) was used as the performance function. The training set was used to teach the network. Training continued as long as the network continued improving on the validation set. The validation set was used to check that the network is generalizing and to stop training before overfitting (internal validation). The test set provided a completely independent test of network accuracy (external validation).

#### Hidden Neurons Optimization

The number of neurons in the hidden layer is of crucial importancein designingan ANN structure and depends on the number of inputs and outputs of the analysis. There are several techniques for selecting the number of neurons while designing an ANN topology ([Azoff, 1994](#_ENREF_7); [Freisleben, 1992](#_ENREF_14); [Gençay, 1999](#_ENREF_17); [Heaton, 2008](#_ENREF_24); [Man-Chung et al., 2000](#_ENREF_32)). In this study, different ANNs with different number of hidden neurons were constructedto determine the optimal number of hidden neurons.MSE on the test set was used as the evaluation criterion of each ANN. The minimum number of neurons that produced the network with the lowest value of MSE was selected for the network topology([Sheela and Deepa, 2013](#_ENREF_50)).

## Applicability Domain

The applicability of QSARs towards reliable predictions is restricted in a chemical space that includes chemical compounds structurally similar to the ones used to train the model ([Nikolova-Jeliazkova and Jaworska, 2005](#_ENREF_34); [Sheridan et al., 2004](#_ENREF_51); [Worth et al., 2004](#_ENREF_56)). The AD of the developed models, which defined thelimitations regardingtheir structural and targetdomain, was determined using several approached,such as bounding box, bounding box on PCs, convex hull, leverage, distance to centroid, k Nearest Neighbors (kNN) approach with fixed k, k Nearest Neighbors (kNN) approach with variable k, Probability Density Function (PDF)based methods ([Sahigara et al., 2014](#_ENREF_48); [Sahigara et al., 2012](#_ENREF_49)). These strategies were implemented using the Applicability Domain Toolbox in MATLAB®(version R2016a, Mathworks Inc).

# Results

## Dimensionality Reduction

The distribution of chemical compounds and the dimensionality reduction of the initial PaDEL descriptors was performed using the principal components derived from PCA. Score plot () shows thecategorizationof the chemicals by involving the projection of input data onto the first two principal components. The numbers in correspond to those of the chemical compounds in.



Figure 1. Score plot derived from PCA.

As presented in, chemical compounds were divided in four distinctive groups depending on the similarity of their structural properties. The first one included the Perfluorinated Compounds (PFCs) of PFOA and PFOS (ID 58, 59), while the second one included the polyfluorinated ethers of fluroxene (ID 46) and “flurane” family (ID 42, 49, 52, 62, 64), as well as the halogenated (fluorine) compounds of 1-Chloro-2, 2-difluoroethene and 1-Chloro-2, 2, 2-trifluoroethane (ID 2, 3). The third group contained the Polychlorinated Biphenyls (PCBs) (ID 8-22, 24-32, 35) and p p'-DDE (ID 56). The fourth bigger group included the remained chemicals, which belong to benzene, toluene, ethylbenzene, xylene (BTEX), acyclic saturated and unsaturated hydrocarbons, cyclic hydrocarbons and alcohols.

PCA also revealed the most significant principal components for the analysis, based on the eigenvalue-one criterion proposed by Kaiser ([1960](#_ENREF_26)). The percentage of variance explained by principal components is given in .

Table 1. Variance explained for the significant principal components from PCA.

|  |  |  |  |
| --- | --- | --- | --- |
| Principal Component | Initial eigenvalues and loadings | | |
| Total | % of Variance | Cumulative % |
| 1 | 119.174 | 27.8 | 27.8 |
| 2 | 71.044 | 16.6 | 44.4 |
| 3 | 43.660 | 10.2 | 54.6 |
| 4 | 27.629 | 6.5 | 61.1 |
| 5 | 21.745 | 5.1 | 66.2 |
| 6 | 16.534 | 3.9 | 70.0 |
| 7 | 14.089 | 3.3 | 73.3 |
| 8 | 13.428 | 3.1 | 76.5 |
| 9 | 12.832 | 3.0 | 79.5 |
| 10 | 10.469 | 2.4 | 81.9 |
| 11 | 9.879 | 2.3 | 84.2 |
| 12 | 8.778 | 2.1 | 86.3 |
| 13 | 7.810 | 1.8 | 88.1 |
| 14 | 6.409 | 1.5 | 89.6 |
| 15 | 5.447 | 1.3 | 90.9 |
| 16 | 4.953 | 1.2 | 92.0 |
| 17 | 4.703 | 1.1 | 93.1 |
| 18 | 3.954 | 0.9 | 94.1 |
| 19 | 3.661 | 0.9 | 94.9 |
| 20 | 2.960 | 0.7 | 95.6 |
| 21 | 2.605 | 0.6 | 96.2 |
| 22 | 2.428 | 0.6 | 96.8 |
| 23 | 1.925 | 0.4 | 97.2 |
| 24 | 1.725 | 0.4 | 97.6 |
| 25 | 1.699 | 0.4 | 98.0 |
| 26 | 1.277 | 0.3 | 98.3 |
| 27 | 1.246 | 0.3 | 98.6 |
| 28 | 1.126 | 0.3 | 98.9 |

28 out of 428 principal components were obtained with eigenvalues greater than 1, summing almost 99% of the total variance in the PaDEL descriptors dataset ().

The loadings of the most correlated descriptors on the first two principal components (PC1, PC2), as were extracted by varimax rotation, are illustrated in. The greatest factor loadings are highlighted in bold. The explanation of chemical features, as encoded in the above descriptors, is presented in Supplementary file S2.

Table 2. Varimax rotated factor matrix of the most correlated PaDEL descriptors in PC1 and PC2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Descriptors | PC1 | PC2 | Descriptors | PC1 | PC2 |
| AATS0p | **0.87** | 0.01 | MAXDP2 | 0.15 | **0.94** |
| AATS1s | -0.29 | **0.91** | minaasC | **0.91** | -0.19 |
| AATS3e | -0.08 | **0.97** | minsCl | **0.87** | -0.14 |
| AATS3p | **0.93** | -0.06 | MLFER\_S | **0.92** | -0.05 |
| AATS3s | -0.17 | **0.84** | piPC5 | **0.97** | 0.04 |
| AATS3v | **0.89** | 0.32 | SaasC | **0.88** | -0.06 |
| AATS4v | **0.90** | 0.24 | SCH-6 | **0.82** | -0.33 |
| AATS7i | **0.82** | 0.08 | SpMAD\_Dzp | -0.02 | **0.94** |
| AATS8p | **0.86** | 0.02 | SpMAD\_Dzs | **0.89** | 0.19 |
| AATSC0c | -0.25 | **0.95** | SpMAD\_Dzv | 0.39 | **0.84** |
| AATSC0s | -0.45 | **0.87** | SpMax1\_Bhs | -0.27 | **0.93** |
| ATS0m | **0.82** | 0.40 | SpMax2\_Bhi | **0.90** | 0.18 |
| ATS7m | **0.80** | 0.30 | SpMax2\_Bhm | **0.88** | 0.35 |
| ATSC0e | 0.18 | **0.93** | SpMax3\_Bhm | **0.91** | 0.28 |
| ATSC8i | **0.88** | 0.09 | SpMax3\_Bhs | -0.12 | **0.90** |
| AVP-6 | **0.83** | -0.20 | SpMax4\_Bhm | **0.95** | 0.22 |
| BCUTc-1h | -0.13 | **0.96** | SpMax4\_Bhp | **0.93** | -0.08 |
| BCUTp-1h | **0.96** | 0.00 | SpMax4\_Bhs | 0.08 | **0.89** |
| BCUTp-1l | **0.80** | -0.53 | SpMax5\_Bhi | **0.95** | 0.02 |
| BIC5 | **0.89** | 0.02 | SpMax5\_Bhm | **0.94** | 0.25 |
| ETA\_dAlpha\_A | **0.88** | -0.03 | SpMax5\_Bhs | 0.28 | **0.80** |
| ETA\_dEpsilon\_B | **0.90** | -0.09 | SpMax6\_Bhi | **0.94** | 0.10 |
| ETA\_dPsi\_A | -0.32 | **0.91** | SpMax6\_Bhp | **0.97** | -0.04 |
| ETA\_Epsilon\_2 | -0.14 | **0.97** | SpMax7\_Bhp | **0.93** | 0.14 |
| ETA\_Eta\_F\_L | **0.93** | 0.28 | SpMin1\_Bhs | **0.82** | -0.52 |
| ETA\_EtaP\_B | 0.13 | **0.96** | SpMin1\_Bhv | **0.92** | -0.13 |
| ETA\_Shape\_Y | **0.88** | 0.04 | SpMin2\_Bhi | **0.93** | -0.08 |
| FMF | **0.88** | -0.29 | SpMin2\_Bhp | **0.87** | 0.04 |
| GATS6v | **0.89** | 0.02 | SpMin2\_Bhs | **0.86** | -0.41 |
| GATS7e | **0.85** | 0.05 | SpMin4\_Bhp | **0.82** | 0.06 |
| GATS8i | **0.87** | 0.22 | SpMin4\_Bhv | **0.83** | 0.03 |
| GATS8p | **0.85** | 0.03 | SpMin5\_Bhv | **0.87** | 0.15 |
| gmax | -0.50 | **0.83** | SsCl | **0.86** | -0.03 |
| hmax | -0.32 | **0.89** | TIC1 | **0.91** | 0.28 |
| IC0 | 0.08 | **0.83** | TIC3 | **0.88** | 0.21 |
| JGI1 | -0.11 | **0.92** | topoRadius | **0.91** | 0.24 |
| JGI5 | **0.83** | 0.10 | TSRW | **0.82** | 0.46 |
| JGI6 | **0.91** | 0.13 | VC-3 | **0.87** | 0.38 |
| JGI7 | **0.90** | 0.18 | WTPT-2 | **0.91** | -0.16 |
| JGT | 0.36 | **0.88** | XLogP | **0.89** | 0.12 |
| Kier2 | **0.80** | 0.26 | ZMIC3 | **0.82** | 0.47 |
| MATS8s | **0.85** | -0.05 | ZMIC4 | **0.87** | 0.33 |

As shown in, the first and the second principal component, accounting for 27.8% and 16.6% of the total variance, respectively, were primarily correlated with molecular and topological descriptors. All the other principal components did not display any strong relationship to the descriptors.84 out of 428 PaDEL descriptors, with the highest percentage of contribution to the first two principal components’ matrix, were selected for further analysis ().

## Predictions using Multiple Linear Regression Analysis

Two linear models were derived using MLR technique; a) the model using LFER descriptors and b) the model using PaDEL descriptors as inputs. The regression coefficients, standard errors and p-values of the model with LFER descriptors are shown in .

Table 3. Results of the developed linear model using LFER descriptors and MLR analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Coefficient | Standard coefficient | Standard error | Confidence Interval (95%) | p-value |
| Intercept | 1.49 |  | 0.13 | 0.26 | <0.001 |
| E | -0.44 | -0.52 | 0.09 | 0.20 | <0.001 |
| A | -2.95 | -0.48 | 0.39 | 0.78 | <0.001 |
| B | -3.64 | -0.67 | 0.35 | 0.71 | <0.001 |
| V | 0.99 | 0.69 | 0.15 | 0.30 | <0.001 |

After the implementation of GA method for feature selection, the linear model with the best performance was achieved using five PaDEL descriptors that are presented in .

Table 4. Definition of the selected descriptors using GA method.

|  |  |  |
| --- | --- | --- |
| Descriptor | Type of Descriptor | Description |
| XLogP | Constitutional | Octanol/water partition coefficient |
| SpMin5\_Bhv | Topological | Smallest absolute eigenvalue of Burden modified matrix - n 5 / weighted by relative van der Waals volumes |
| BCUTc-1h | Molecular | nlow highest partial charge weighted BCUTS |
| hmax | Topological | Maximum H E-State |
| AATSC0s | Topological | Average centered Broto-Moreau autocorrelation - lag 0  / weighted by I-state |

The regression coefficients, standard errors and p-values of the linear model with PaDEL descriptors, obtained by GA-MLR, are shown in .

Table 5. Results of the developed linear model using PaDEL descriptors and MLR analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Coefficient | Standard coefficient | Standard error | Confidence Interval (95%) | p-value |
| Intercept | 1.62 |  | 0.14 | 0.28 | <0.001 |
| XLogP | 0.31 | 0.63 | 0.03 | 0.05 | <0.001 |
| SpMin5\_Bhv | -0.25 | -0.17 | 0.09 | 0.17 | 0.0046 |
| BCUTc-1h | 2.23 | 0.31 | 0.80 | 1.61 | 0.0074 |
| hmax | -3.08 | -1.60 | 0.22 | 0.45 | <0.001 |
| AATSC0s | 0.21 | 0.86 | 0.03 | 0.07 | <0.001 |

The multiple regression models, obtained from LFER and PaDEL descriptors’ analysis,are described by the following relationships according to the coefficients in and :

LFER:  (2)

PaDEL:  (3)

The linear regression models produced an estimate of the association between adipose/blood partition coefficient and LFER/PaDEL descriptors. According to equation 2, one unit increase in McGowan volume was related to 0.99 unit increase in the logarithmic values of adipose/blood partition coefficient, holding the remaining LFER parameters constant. An increase in refractivity, acidity and basicity, though, resulted in a decrease in the values of adipose/blood partition coefficient. In equation 3, it is indicated that the response was directly proportional to octanol-water partition coefficient and Broto-Moreau autocorrelation, while it was inversely proportional to the additional descriptors.

The standardised regression coefficients is a measure of variable importance. Regarding the absolute values of standardised coefficients, it is noted that the most important variables were McGowan volume and Hmax for LFER and PaDEL model, respectively. The estimated low P-values indicate that all descriptors were statistically significant and should have been included in models’ analysis.

A comparison of the observedlogarithmic values of adipose/blood partition coefficient and the predicted ones from the derived linear models is presented in .

Figure 2. Predicted versus observed values of adipose/blood partition coefficient using MLR.

The high values of coefficient of determination denoted a very good correlation between the target and the descriptors for both models. Specifically, MLR analysis resulted in R2 equal to 0.82 and 0.86 for LFER and PaDEL descriptors, respectively.

The residuals against the predicted values were plotted in Figure A.1 and Figure A.2 in Supplementary file S3 for the developed models in order to explore homoscedasticity and normality of residuals.The residual plots show that the residuals were normally distributed and randomly scattered around zero, without making a clear pattern and shape or indicating outliers.

## Predictions using Artificial Neural Networks Analysis

ANN analysis was implemented for the datasets of LFER and PaDEL descriptors using Levenberg-Marquardt, Scaled Conjugate Gradient and Quasi-Newton backpropagation algorithms. Hidden neurons optimization was necessary for estimating the most appropriate ANN structure (Figure B.1-Figure B.6 in Supplementary file S3).

- present the predicted values of adipose/blood partition coefficient using the LM, SCG and BFG based ANN models for LFER and PaDEL descriptors versus the observed ones.

Figure 3. Predicted versus observed values of adipose/blood partition coefficient using LM based ANN.

Figure 4. Predicted versus observed values of adipose/blood partition coefficient using SCG based ANN.

Figure 5. Predicted versus observed values of adipose/blood partition coefficient using BFG based ANN.

The correlation between the experimental and predicted values of adipose/blood partition coefficient was found to be exceptional for all the statistical methods and for both LFER and PaDEL descriptors, as the coefficients of determination were higher than 0.90. Despite the high values of coefficients of determination, and show that the predicted values were found constant for most of the chemicals. This means that SCG method coupled with PaDEL descriptors, as well as BFG method for both sets of descriptors failed to detect the differences arose from input data, leading to the same predicted values of the response.

The residual analysis is shown in Figure A.3-Figure A.8 in Supplementary file S3 in order to identify the appropriateness of the models using LFER and PaDEL descriptors coupled with LM, SCG and BFG methods. It is illustrated that all the residuals, except for those of the models using PaDEL descriptors and SCG method (Figure A.6), as well as those using BFG method (Figure A.7-Figure A.8) were randomly dispersed around the horizontal axis and the error variances were equal.

## 3.4 Comparison of the Statistical Methods of MLR and ANN

presents the statistical values for the fitting and validation performance of the developed models. These include the coefficients of determination, as well as the mean squared error for the training, validation and test set of the models.

Table 6. Statistical values of the developed models.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Descriptors | Method | R2 | Rtr2 | Qcv2 | Rext2 | MSEtr | MSEval | MSEtest |
| 1 | LFER | MLR | 0.82 | 0.84 | 0.72 | 0.85 | 0.12 | 0.21 | 0.16 |
| 2 | **LM-ANN** | **0.94** | **0.93** | **0.96** | **0.98** | **0.05** | **0.04** | **0.03** |
| 3 | SCG-ANN | 0.92 | 0.87 | 0.97 | 0.95 | 0.07 | 0.06 | 0.05 |
| 4 | BFG-ANN | 0.91 | 0.92 | 0.84 | 0.92 | 0.09 | 0.06 | 0.05 |
| 5 | PaDEL | MLR | 0.86 | 0.88 | 0.84 | 0.85 | 0.09 | 0.12 | 0.19 |
| 6 | **LM-ANN** | **0.96** | **0.96** | **0.95** | **0.94** | **0.04** | **0.05** | **0.03** |
| 7 | SCG-ANN | 0.92 | 0.91 | 0.93 | 0.93 | 0.07 | 0.05 | 0.06 |
| 8 | BFG-ANN | 0.94 | 0.95 | 0.89 | 0.90 | 0.04 | 0.06 | 0.06 |

As it is shown in , the models with the best statistical values are those which were obtained using LM based ANN method. PaDEL descriptors gave slightly better performance results compared to LFER descriptors in training set, while LFER descriptors led to more accurate predictionsin validation and test set.

The predicted values of adipose/blood partition coefficient using the derived best models, compared to the observed literature ones are illustrated in .

Table 7. Observed and predicted values of adipose/blood partition coefficient using the developed best models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Chemical compound | log (adipose/blood partition coefficient) | | |
| **Observed** | **Predicted by LFER model** | **Predicted by PaDEL model** |
| 1 | 1-Butanol | -0.14 | 0.00 | 0.06 |
| 2 | 1-Chloro-2,2-difluoroethene | 1.36 | 1.44 | 1.37 |
| 3 | 1-Chloro-2,2,2-trifluoroethane | 1.36 | 1.66 | 1.61 |
| 4 | 1-Propanol | -0.51 | -0.48 | -0.64 |
| 5 | 1,1,1-Trichloroethane | 1.88 | 1.73 | 1.99 |
| 6 | 2-Methylpentane | 2.33 | 2.27 | 2.27 |
| 7 | 2-Propanol | -0.60 | -0.71 | -0.37 |
| 8 | 2,2',3,3',4',5,5',6-Octachlorobiphenyl | 2.56 | 2.37 | 2.56 |
| 9 | 2,2',3,3',4,4',5-Heptachlorobiphenyl | 2.42 | 2.30 | 2.34 |
| 10 | 2,2',3,4',5,5'-Hexachlorobiphenyl | 2.43 | 2.25 | 2.35 |
| 11 | 2,2',3,4',5,5',6-Heptachlorobiphenyl | 2.20 | 2.32 | 2.42 |
| 12 | 2,2',3,4,4',5'-Hexachlorobiphenyl | 2.04 | 2.24 | 2.28 |
| 13 | 2,2',3,4,4',5',6-Heptachlorobiphenyl | 2.57 | 2.32 | 2.42 |
| 14 | 2,2',3,4,4',5,5'-Heptachlorobiphenyl | 2.48 | 2.31 | 2.41 |
| 15 | 2,2',3,5'-Tetrachlorobiphenyl | 1.78 | 2.09 | 2.11 |
| 16 | 2,2',3,5',6-Pentachlorobiphenyl | 1.90 | 2.18 | 2.17 |
| 17 | 2,2',4,4'-Tetrachlorobiphenyl | 2.43 | 2.09 | 2.09 |
| 18 | 2,2',4,4',5-Pentachlorobiphenyl | 2.38 | 2.17 | 2.15 |
| 19 | 2,2',4,4',5,5'-Hexachlorobiphenyl | 2.49 | 2.25 | 2.32 |
| 20 | 2,2',4,5-Tetrachlorobiphenyl | 1.70 | 2.09 | 2.02 |
| 21 | 2,2',4,5,5'-Pentachlorobiphenyl | 1.95 | 2.17 | 2.16 |
| 22 | 2,2',5,5'-Tetrachlorobiphenyl | 1.90 | 2.09 | 2.13 |
| 23 | 2,2-Dimethylbutane | 2.40 | 2.29 | 2.27 |
| 24 | 2,3',4',5-Tetrachlorobiphenyl | 1.85 | 2.09 | 2.13 |
| 25 | 2,3',4,4'-Tetrachlorobiphenyl | 2.20 | 2.09 | 2.12 |
| 26 | 2,3',4,4',5-Pentachlorobiphenyl | 2.25 | 2.17 | 2.17 |
| 27 | 2,3',4,4',5,5'-Hexachlorobiphenyl | 2.08 | 2.24 | 2.34 |
| 28 | 2,3,3',4,4'-Pentachlorobiphenyl | 2.11 | 2.18 | 2.17 |
| 29 | 2,3,3',4,4',5-Hexachlorobiphenyl | 2.49 | 2.24 | 2.23 |
| 30 | 2,3,3',4,5,5'-Hexachlorobiphenyl | 2.56 | 2.24 | 2.30 |
| 31 | 2,4,4'-Trichlorobiphenyl | 2.30 | 2.00 | 2.00 |
| 32 | 2,4,4',5-Tetrachlorobiphenyl | 2.30 | 2.09 | 2.05 |
| 33 | 3-Methylhexane | 2.33 | 2.46 | 2.48 |
| 34 | 3-Methylpentane | 2.38 | 2.27 | 2.26 |
| 35 | 4,4'-Dichlorobiphenyl | 1.95 | 1.94 | 1.97 |
| 36 | Benzene | 1.82 | 1.55 | 1.80 |
| 37 | Cyclohexane | 2.30 | 1.94 | 2.18 |
| 38 | Cyclopropane | 1.45 | 1.31 | 1.65 |
| 39 | Dichloromethane | 1.15 | 1.40 | 1.20 |
| 40 | Diethyl ether | 0.77 | 1.37 | 0.69 |
| 41 | Divinyl ether | 1.20 | 1.04 | 1.31 |
| 42 | Enflurane | 1.87 | 1.42 | 1.60 |
| 43 | Ethanol | -0.8 | -0.87 | -0.79 |
| 44 | Ethyl Benzene | 1.79 | 1.92 | 1.73 |
| 45 | Ethylene | 0.94 | 1.19 | 0.84 |
| 46 | Fluroxene | 1.39 | 1.52 | 1.59 |
| 47 | Heptane | 2.31 | 2.44 | 2.52 |
| 48 | Hexane | 2.11 | 2.26 | 2.30 |
| 49 | Isoflurane | 1.69 | 1.63 | 1.60 |
| 50 | m-xylene | 1.78 | 1.91 | 1.83 |
| 51 | Methanol | -0.85 | -1.18 | -0.93 |
| 52 | Methoxyflurane | 1.80 | 1.88 | 1.59 |
| 53 | Methylcyclopentane | 2.31 | 1.96 | 2.17 |
| 54 | o-xylene | 1.84 | 1.91 | 1.82 |
| 55 | p-xylene | 1.71 | 1.91 | 1.88 |
| 56 | p,p'-DDE | 2.23 | 2.42 | 2.24 |
| 57 | Pentane | 2.02 | 2.02 | 2.10 |
| 58 | PFOA | -0.33 | -0.39 | -0.07 |
| 59 | PFOS | -0.48 | -0.48 | -0.49 |
| 60 | Propanone | -0.36 | -0.07 | -0.42 |
| 61 | Propylene | 1.07 | 1.37 | 1.50 |
| 62 | Sevoflurane | 1.86 | 1.72 | 1.57 |
| 63 | Styrene | 1.70 | 1.79 | 1.82 |
| 64 | Teflurane | 1.52 | 1.64 | 1.61 |
| 65 | Toluene | 1.97 | 1.71 | 1.76 |
| 66 | Trichloroethene | 1.85 | 1.55 | 1.80 |
| 67 | Trichloromethane | 1.54 | 1.58 | 1.74 |

The relative importance of input descriptors for the development of the models ( and ) was estimated using Garson method, based on the partitioning of neural network connection weights([Garson, 1991](#_ENREF_16)).

|  |  |
| --- | --- |
| Figure 6. Relative importance of LFER descriptors for the developed model. | Figure 7. Relative importance of PaDEL descriptors for the developed model. |

It is found that the most influential descriptor was McGowan volume (V) and octanol-water partition coefficient (XlogP) for the model derived from LFER and PaDEL descriptors, respectively. Both parameters give a measure ofthe lipophilicity of chemical compounds. The lipophilic/hydrophilic nature of chemicals influences their distribution and transport through cell membranes, as well as their mechanism of metabolism([Rutkowska et al., 2013](#_ENREF_47)).

## 3.5 Applicability Domain of the Developed Models

In order to estimate the AD of the developed models, the input data of chemical compounds was divided into the set for training the models and the independent set for assessing the performance of the trained models. and show the chemicals that were included in the test set of QSAR models.

Table 8. Chemical compounds in the test set of the developed model using LFER descriptors.

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Chemical Compound | ID | Chemical Compound |
| 1 | 2-Methylpentane | **6** | 2.3'.4.4'.5-Pentachlorobiphenyl |
| 2 | **2-Propanol** | **7** | 2.4.4'-Trichlorobiphenyl |
| 3 | **2.2'.3.3'.4.4'.5-Heptachlorobiphenyl** | **8** | Benzene |
| 4 | 2,2',3,4',5,5',6-Heptachlorobiphenyl | **9** | **Fluroxene** |
| 5 | 2.2'.3.4.4'.5.5'-Heptachlorobiphenyl | **10** | Methoxyflurane |

Table 9. Chemical compounds in the test set of the developed model using PaDEL descriptors.

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Chemical Compound | ID | Chemical Compound |
| 1 | **1-Butanol** | **6** | 2,3',4,4',5,5'-Hexachlorobiphenyl |
| 2 | **2,2',3,3',4,4',5-Heptachlorobiphenyl** | **7** | **4,4'-Dichlorobiphenyl** |
| 3 | 2,2',4,4',5-Pentachlorobiphenyl | **8** | **Hexane** |
| 4 | 2,2',4,5,5'-Pentachlorobiphenyl | **9** | **Isoflurane** |
| 5 | 2,3',4,4',5-Pentachlorobiphenyl | **10** | **Pentane** |

The number of chemical compounds placed out of AD using range-based, distance-based and PDF-based approaches are shown in and , while the frequency with which they wereidentified as outliers is presented in and .

Table 10. Results of the AD of the LFER-LM model using various statistical approaches.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | Options | Test inside AD | Test outside AD | ID of test outside AD |
| Bounding box | --- | 10 | 0 | [] |
| Bounding box PCA | 4 PCs | 10 | 0 | [] |
| Convex hull | --- | 7 | 3 | [2 3 9] |
| Leverage | --- | 10 | 0 | [] |
| Distance from centroid | Euclidean | 10 | 0 | [] |
| Distance kNN - fixed k | Euclidean | 10 | 0 | [] |
| Distance kNN - variable k | Euclidean | 10 | 0 | [] |
| Potential functions | Gaussian | 10 | 0 | [] |

Table 11. Results of the AD of the PaDEL-LM model using various statistical approaches.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | Options | Test inside AD | Test outside AD | ID of test outside AD |
| Bounding box | --- | 10 | 0 | [] |
| Bounding box PCA | 5 PCs | 10 | 0 | [] |
| Convex hull | --- | 4 | 6 | [1 2 7 8 9 10] |
| Leverage | --- | 10 | 0 | [] |
| Distance from centroid | Euclidean | 10 | 0 | [] |
| Distance kNN - fixed k | Euclidean | 10 | 0 | [] |
| Distance kNN - variable k | Euclidean | 10 | 0 | [] |
| Potential functions | Gaussian | 9 | 1 | [1] |

|  |  |
| --- | --- |
| Figure 8. Frequency of test samples placed out of the AD of the LFER-LM model by the calculated methods. | Figure 9. Frequency of test samples placed out of the AD of the PaDEL-LM model by the calculated methods. |

shows that 2-Propanol, 2,2’,3,3’,4,4’,5-Heptachlorobiphenyl and fluroxene were marked as outliers by convex hull method for the developed model with LFER descriptors. The remaining methods did not showcase any chemical out of the model’s AD. Regarding the developed model with PaDEL descriptors, 1-Butanol, 2,2’,3,3’,4,4’,5-Heptachlorobiphenyl, 4,4’-Dichlorobiphenyl, hexane, isoflurane and pentane (ID 1, 2, 7, 8, 9, 10 as seen in) were excluded from AD by convex hull geometric method. These chemical compounds were not indicated as outliers by the other implemented methods, except from 1-Butanol, which was excluded from the probabilistic method of potential functions.

## Application of the developed QSAR modelson chemicals with unknown values

For the reliable application of the developed models to new chemical compounds, it is important to ensure that these compounds are inside AD of the models.

Table 12presents the results of ADestimation for a set of 143 chemical compounds with unknown values of adipose/blood partition coefficient.

Table 12. Results of the AD of the LFER/PaDEL – LM based ANN model applied to the new set of chemicals using various statistical approaches.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | LFER model | | PaDEL model | |
| **Test inside AD** | **Test outside AD** | **Test inside AD** | **Test outside AD** |
| Bounding box | 77 | 66 | 99 | 44 |
| Bounding box PCA | 82 | 61 | 132 | 11 |
| Convex hull | 8 | 135 | 33 | 110 |
| Leverage | 59 | 84 | 122 | 21 |
| Distance from centroid | 57 | 86 | 137 | 6 |
| Distance kNN - fixed k | 49 | 94 | 135 | 8 |
| Distance kNN - variable k | 48 | 95 | 132 | 11 |
| Potential functions | 41 | 102 | 31 | 112 |

For the LFER model, all methods, except from those of bounding box, indicated that most of the compounds do not belong to model’s AD. In contrast, all methods, except convex hull and potential functions, indicated that most of the compounds belong to AD of the PaDEL model. showsthe compounds that are most frequently estimated outside of AD.

Table 13. Frequency of the chemical compounds placed out of AD of the developed models by the calculated methods.

|  |  |  |
| --- | --- | --- |
| Freq. | Model | ID of test compounds () |
| 8 | LFER | [12534-384648495051535872747778 80 82 83 86 87 8990-92101-103109135136 138 139] |
| PaDEL | [86 118] |
| 7 | LFER | [1932335763-676984859399104-108110111114-116128129133141142] |
| PaDEL | [8487] |
| 6 | LFER | [1420-2430424471768188113119124126 127130132137] |
| PaDEL | [6591104] |
| 5 | LFER | [52100] |
| PaDEL | [46668998129] |
| 4 | LFER | [13555662687596143] |
| PaDEL | [495163648096112119130137] |

The chemical compounds that were found out of AD of the LFER model were atrazine and resorcinol derivatives and those included in chemical families, such asphthalates, phenols, parabens, Polycyclic Aromatic Hydrocarbons (PAHs) andhigher Polybrominated Diphenyl Ethers (PBDEs) with more than 5 bromine atoms per molecule([Siddiqi et al., 2003](#_ENREF_52)).The outliers of PaDEL model were found to be high-molecular-weight phthalates with more than 5 carbons in their backbone ([National Research Council et al., 2009](#_ENREF_33)), sulfur based and nitrogen based compounds, as well as PAHs containing bay or bay-like regions ([Harvey, 1985](#_ENREF_22)).

demonstrates the values of adipose/blood partition coefficient, which were predicted for both developed QSAR models.

Table 14. Predicted values of adipose/blood partition coefficient for the new set of chemical compounds.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Chemical compound | Log (Padipose/blood) | |  | Chemical compound | Log (Padipose/blood) | |  | Chemical compound | Log (Padipose/blood) | |
| **LFER** | **PaDEL** | **LFER** | **PaDEL** | **LFER** | **PaDEL** |
| 1 | 1-hydroxypyrene | 3.33 | 1.24 | **49** | Deisopropylhydroxy atrazine | -1.89 | -0.88 | **97** | γ-lindane | 2.78 | 2.55 |
| 2 | 1,2,3,4,6,7,8-Heptachlorooxanthrene | 2.32 | 1.74 | **50** | Atrazine | -1.59 | 1.51 | **98** | Hydrogen sulfide | 1.24 | -0.68 |
| 3 | 1,2,3,4,6,7,8-Heptachlorodibenzo[b,d]furan | 2.03 | 2.07 | **51** | Bromophenol | 4.33 | 0.08 | **99** | Heptyl paraben | 1.38 | 1.5 |
| 4 | 1,2,3,4,7,8-Hexachlorooxanthrene | 2.25 | 1.71 | **52** | 2,2',4,4',5,5'-Hexabromobiphenyl | 2.03 | 2.55 | **100** | Hexachlorobenzene | 2 | 2.45 |
| 5 | 1,2,3,4,7,8-Hexachlorodibenzo[b,d]furan | 1.94 | 1.85 | **53** | Benzyl butyl phthalate | 4.53 | 0.17 | **101** | Hydroxy atrazine | -1.87 | -0.11 |
| 6 | 1,2,3,4,7,8,9-Heptachlorodibenzofuran | 2.03 | 1.9 | **54** | 2,2′,4,4′,6-Pentabromodiphenyl ether | 2.19 | 1.82 | **102** | Hydroxydesethyl atrazine | -1.88 | 0.78 |
| 7 | 1,2,3,6,7,8-Hexachlorodibenzo-p-dioxin | 2.25 | 1.72 | **55** | 2,2',4,4',5,5'-Hexabromodiphenyl ether | 2.23 | 2.09 | **103** | Hydroxysimazine | -1.86 | 1.3 |
| 8 | 1,2,3,6,7,8-Hexachlorodibenzofuran | 1.95 | 1.86 | **56** | 2,2′,4,4′,5,6′-Hexabromodiphenyl ether | 2.23 | 2.12 | **104** | Indeno(1,2,3-c,d)pyrene | 1.31 | 3 |
| 9 | 1,2,3,7,8-Pentachlorooxanthrene | 2.17 | 1.69 | **57** | Octabromodiphenyl ether | 2.27 | 2.3 | **105** | Isobutyl Paraben | -0.64 | 1.61 |
| 10 | 1,2,3,7,8-Pentachlorodibenzo[b,d]furan | 1.87 | 1.82 | **58** | Decabromodiphenyl ether | 2.39 | 2.99 | **106** | Isopropyl Paraben | -1.02 | 1.6 |
| 11 | 1,2,3,7,8,9-Hexachlorooxanthrene | 2.25 | 1.72 | **59** | 2,4,4'-Tribromodiphenyl ether | 2.15 | 1.51 | **107** | Monobenzyl phthalate | 1.22 | 1.14 |
| 12 | 1,2,3,7,8,9-Hexachlorodibenzofuran | 1.94 | 1.86 | **60** | 2,2',4,4'-Tetrabromodiphenyl ether | 2.16 | 1.62 | **108** | Monocyclohexyl phthalate | 0.47 | 1.52 |
| 13 | 2-phenoxyphenol | 3.09 | 1.57 | **61** | 2,2′,4,4′,5-Pentabromodiphenyl ether | 2.16 | 1.62 | **109** | Mono(3-carboxypropyl) phthalate | -1.81 | 1.54 |
| 14 | 2,3-dimethylphenol | 2.01 | 1.58 | **62** | Benz(a)anthracene | 1.57 | 2.83 | **110** | Mono(2-ethylhexyl) phthalate | 1.51 | 1.09 |
| 15 | 2,3,3',4,4',5,5'-Heptachlorobiphenyl | 2.3 | 2.41 | **63** | Benzo(a)pyrene | 1.41 | 2.94 | **111** | Methyl Paraben | -1.41 | 1.39 |
| 16 | 2,3,4,6,7,8-Hexachlorodibenzo[b,d]furan | 1.95 | 1.97 | **64** | Benzo(b)fluoranthene | 1.41 | 2.94 | **112** | Mercury | 0.84 | -0.51 |
| 17 | 2,3,4,7,8-Pentachlordibenzo[b,d]furan | 1.87 | 1.82 | **65** | Benzo(ghi)perylene | 1.31 | 3 | **113** | Metolachlor | 4.44 | 1.59 |
| 18 | 2,3,7,8-Tetrachlorodibenzo[b,d]furan | 1.8 | 1.8 | **66** | Benzo(k)fluoranthene | 1.41 | 2.94 | **114** | Monoisobutyl phthalate | -0.57 | 1.58 |
| 19 | 2,4-Dichlorophenoxyacetic acid | 1.7 | 1.55 | **67** | Benzophenone-1 | 2.66 | 1.54 | **115** | Mono-n-butyl phthalate | -0.4 | 1.56 |
| 20 | 2,4-Dichlorophenol | 2.99 | 1.56 | **68** | Benzophenone-8 | 3.48 | 1.57 | **116** | Butyl Paraben | -0.44 | 1.61 |
| 21 | 2,4-Dimethylphenol | 2.01 | 1.57 | **69** | Benzyl Paraben | 1.31 | 1.37 | **117** | Naphthalane | 2.49 | 2.85 |
| 22 | 2,4,5-Trichlorophenol | 3.12 | 1.6 | **70** | β-Hexachlorocyclohexane | 2.78 | 2.55 | **118** | Ammonia | 1.09 | -0.94 |
| 23 | 2,4,6-Trichlorophenol | 2.18 | 1.6 | **71** | Benzophenone-3 | 3.73 | 1.47 | **119** | Nitric oxide | -1.74 | -0.81 |
| 24 | 2,5-Dichlorophenol | 2.93 | 1.57 | **72** | Bisphenol A | -0.52 | 1.26 | **120** | o,p-DDE | 2.42 | 2.19 |
| 25 | 2,5-Dimethylresorcinol | -1.65 | 1.01 | **73** | Chlordane | 2.75 | 2.76 | **121** | o,p-DDT | 2.55 | 2.11 |
| 26 | 2,6-Diethylaniline | 3.15 | 1.71 | **74** | Chlorpyrifos | -0.86 | 0.9 | **122** | Octachlorooxanthrene | 2.38 | 2.02 |
| 27 | 3,3',4,4'-Tetrachlorobiphenyl | 2.11 | 2.1 | **75** | Chrysene | 1.57 | 2.85 | **123** | Octachlorodibenzofuran | 2.13 | 2.34 |
| 28 | 3,3',4,4',5-Pentachlorobiphenyl | 2.18 | 2.15 | **76** | n-Butyl phthalate | 3.9 | 1.32 | **124** | p-Cresol | 1.39 | 1.55 |
| 29 | 3,3',4,4',5,5'-Hexachlorobiphenyl | 2.25 | 2.32 | **77** | Dibenzyl phthalate | 4.66 | 0.24 | **125** | 1,4-Dichlorobenzene | 1.63 | 1.82 |
| 30 | 3,4-Dimethylphenol | 2.01 | 1.57 | **78** | Dicyclohexyl phthalate | 4.54 | 0.99 | **126** | Phencyclidine | 4.18 | 2.03 |
| 31 | 3,4,4',5-Tetrachlorobiphenyl | 2.1 | 2.08 | **79** | DDT | 2.55 | 2.05 | **127** | Pentachlorophenol | 2.21 | 1.64 |
| 32 | 4-Chloro-2-methylphenoxyacetic acid | 1.96 | 1.59 | **80** | Bis(2-ethylhexyl) phthalate | 4.68 | 1.8 | **128** | Pentyl Paraben | 0.1 | 1.6 |
| 33 | 4-Phenoxyphenol | 2.23 | 1.52 | **81** | Diethyl phthalate | 2.25 | 1.59 | **129** | Perfluorononanoic acid | -0.31 | -0.85 |
| 34 | 5-Ethylresorcinol | -1.69 | 1.41 | **82** | Desethylatrazine | -1.59 | 0.61 | **130** | Perfluoroctylsulfonamide | 2.96 | -0.27 |
| 35 | 5-Methylresorcinol | -1.75 | 0.85 | **83** | Desisopropyl atrazine | -1.63 | -0.08 | **131** | Phenanthrene | 1.58 | 2.27 |
| 36 | Mono(5-carboxy-2-ethylpentyl) phthalate | -1.26 | 1.56 | **84** | Dibenz(a,h)anthracene | 1.57 | 3.06 | **132** | Phenol | 0.7 | 1.57 |
| 37 | Mono(2-ethyl-5-hydroxyhexyl) phthalate | -0.94 | 1.54 | **85** | Diisobutyl phthalate | 3.49 | 1.46 | **133** | Propyl Paraben | -0.87 | 1.59 |
| 38 | Mono(2-ethyl-5-oxohexyl) phthalate | -0.55 | 1.56 | **86** | Diisodecyl phthalate | 4.76 | 2.95 | **134** | Pyrene | 1.41 | 2.47 |
| 39 | Acenaphthene | 1.61 | 1.79 | **87** | Diisononyl phthalate | 4.72 | 2.65 | **135** | Resorcinol | -1.8 | 0.74 |
| 40 | Acenaphthylene | 1.51 | 1.94 | **88** | Dimethyl phthalate | 1.05 | 1.59 | **136** | Simazine | -1.66 | 1.19 |
| 41 | Acetaldehyde | -0.74 | -0.4 | **89** | Dioctyl phthalate | 4.68 | 2.15 | **137** | Sulfur dioxide | -1.71 | -0.58 |
| 42 | Acetochlor | 4.37 | 1.47 | **90** | Endosulfan | -0.83 | 0.71 | **138** | SPMA | -1.7 | 1.58 |
| 43 | Acrylonitrile | 0.34 | -0.53 | **91** | Et-PFOSA-AcOH | 0.23 | -0.5 | **139** | t.t-muconic acid | -1.87 | -0.57 |
| 44 | Alachlor | 4.37 | 1.45 | **92** | Ethyl 4-(2-(t-butylcarbonyloxy)butoxy) benzoate | 1.34 | 1.49 | **140** | 2,3,7,8-Tetrachlorodibenzo-p-dioxin | 2.09 | 1.68 |
| 45 | α-Hexachlorocyclohexane | 2.78 | 2.55 | **93** | EthylParaben | -1.19 | 1.55 | **141** | 3,5,6-Trichloro-2-pyridinol | 0.23 | 1.53 |
| 46 | Ammeline | -1.89 | -0.93 | **94** | Fluoranthene | 1.41 | 2.47 | **142** | Triclocarban | 3.12 | 1.58 |
| 47 | Anthracene | 1.58 | 2.23 | **95** | Fluorene | 1.66 | 1.49 | **143** | Triclosan | 2.59 | 1.58 |
| 48 | Atrazine Mercapturate | -1.84 | 0.18 | **96** | Formaldehyde | -1.28 | -0.65 |  |  |  |  |

# Discussion

In the current study, PCA was used for the categorization of chemical compounds and the reduction of the initial PaDEL descriptor set.The score plot () indicated that the molecules were clustered by structures. The chemical families which were involved in the analysis are PFCs, fluorinated compounds, PCBs, acyclic and cyclic hydrocarbons, alcohols and BTEX.

As far as the input reduction process is concerned,28out of the 428 principal components that were transformed using input descriptors were considered as significant from PCA (). The first two principal components covered almost 45% of the total variance and were found to be strongly correlated with 84molecular and topological descriptors.These descriptors were considered as the critical ones for the characterization of the chemicalsand the further analysis for the development of the models ().

As it was mentioned before, input data consisted of two sets of descriptors; a) the LFER and b) the PaDEL parameters. GA was used as the optimization method for the selection of the most appropriate PaDEL descriptors. From the set of 84 descriptors obtained by PCA, 5 descriptors were chosen for the development of the models (). It is observed that amongst the topological descriptors that were obtained, octanol/water partitioning was also selected for the analysis. Topological indices give important information about the orientation of atoms and bonds in a molecule, while octanol/water partition coefficient is a measure of its lipophilicity. Many studies have been focused on the influence of lipophilicity on adipose/blood exchange kinetics, which results in diffusion limitations regarding highly lipid soluble molecules([Levitt, 2010](#_ENREF_28); [Poulin and Krishnan, 1995](#_ENREF_43)).

The statistical methods of MLR and ANN based on LM, SCG and BFG algorithms were implemented for the analysis of input data. Regarding MLR analysis, two linear relationships (eq. 2, 3) were developed and were related to each of the descriptors’ set.P-values were found to be smaller than the defined limit of 0.05, which means that all the descriptors were statistically significant for MLR analysis. Furthermore, the ratio between confidence interval and regression coefficient was not greater than 1for all descriptors, which concludes that the models should be considered reliable (-). As far as ANN analysis is concerned, neurons optimization procedure indicated the appropriate number of hidden nodes for avoiding overfitting problem in the networks. The selection criterion for the network topology was the best performance on test set appeared on the minimum number of hidden neurons. The optimal number of neurons was estimated between the number of inputs and outputs of the models, which complies with the recommendations from literature ([Freisleben, 1992](#_ENREF_14); [Heaton, 2008](#_ENREF_24); [Man-Chung et al., 2000](#_ENREF_32)).

The developed models were evaluated regarding goodness of fit, robustness and predictivity. The coefficient of determination as well as the MSE of training, validation and test set were used as measures of performance. demonstrates that the overall performance of all models was excellent. However, satisfyingperformance results do not always imply robust and reliable models.A thorough analysis, as shown in and ,indicated that SCG based ANN model using PaDEL descriptors and BFG based ANN models failed to accurately predict the values of adipose/blood partition coefficient.This led to the appearance of systematic error and non-normal distribution of residuals (Figure A.6-Figure A.8 in Supplementary file S3). Through MLR and ANN comparison, results suggest that ANN models performed best, with Levenberg-Marquardt algorithm giving the most accuratepredictions and the lowest values of MSE not only for training but also for validation and test sets.This is a logical conclusion as ANN method has the advantage of “self-learning” and developing nonlinear, complex interactions between the independent and dependent variables, compared to MLR that forms only simple, linear relationships([Puri et al., 2015](#_ENREF_44)).

According to the performance criteria (), both LFER and PaDEL descriptors were found to be suitable for characterizing the distribution phenomena that influence adipose/blood partitioning.From relative importance analysis ( and ), it isindicatedthat McGowan volume and XlogP, as measures of lipophilicity, contributed the most to the models’ development. This finding is reasonable as lipophilicity influences the chemical, toxicokinetic and toxicodynamic profile of chemicalsand it is of crucial importance in environmental modeling, design and toxicology of substances.

Regarding applicability domain of the models, the range-based, distance-based and PDF-based methods did not highlight any chemical compound as outlier, except from 1-Butanol which was marked outside of AD of PaDEL model by potential functions approach. On the contrary, the geometric method of convex hull marked outliers for both models. As it is shown combining the results from , and , the predicted values for all outliers, estimated by the statistical methods for AD, were close enough to the experimental ones. The prediction regarding 1-Butanol, which was indicated as an outlier by two methods, was found to varyfrom the corresponding observedvalue. The predictive ability of the models for the outliers by convex hull questions the efficiency of this method for AD identification. As reported by several studies, convex hull is considered inefficient at AD estimationfor number of descriptors bigger than 3. Furthermore, set boundaries are estimated without considering data distribution ([Jaworska et al., 2005](#_ENREF_25)). Hence, the results for AD by this method cannot be representative.

The developed models, which were analysed by LM based ANN using LFER and PaDEL descriptors as inputs, were applied to chemical compoundswith unknown values of adipose/blood partition coefficient. It was investigated which of those chemical compounds fall inside AD of the models by determining their properties. shows that much more chemicals were estimated as outliers for LFER model than those for PaDEL model. This means that the structural space covered by the first model is strictly limited to chemicals whose parameters do not differ much from the parameters of the training set. The excluded chemicals from the AD of PaDEL model were mainly those with remarkably different characteristics from the ones of the training set, such as inorganic compounds. Therefore, PaDEL model offers an enlarged applicability domain.As it is demonstrated in , the values of adipose/blood partition coefficient, predicted by both models,for the compounds inside AD ofthe models,were found similar. Generally, the developed QSAR models can be considered applicable to make reliable predictions for the compounds that belong to the chemical space determined by the characteristics of their training set.

Finally, the developed QSARs provide a successful prediction of adipose/blood partition coefficient, which is of major importance for the parameterization of PBTK models. Specifically, the partitioning of environmental chemical compounds into blood and adipose tissue influences their distribution as well as their toxicological profile. Adipose tissue is considered as a storage compartment for lipophilic compounds, mediating a part of their metabolic effects. On the other hand, it may protect other organs and tissues from pollutants’ overload. Thus, it is obvious that adipose/blood partition coefficient may seriously affect the calculation of internal body concentrations of chemicals or their metabolites that is provided by PBTK models.With regard to simple PK models, distribution volume is one of the very few, but key parameters for providing first cut estimates on toxicokinetics. This in turn is greatly affected by the presence of adipose tissue, especially for highly lipophilic compounds, rendering the accurate prediction of adipose:blood partition coefficient highly important.

# Conclusions

Thestudypresented herein followsan advanced methodological framework for QSAR modeling. Two QSAR models were developed to predict adipose/blood partitioning for environmental chemical compounds. The models were successfully evaluated for their goodness of fit, robustness, predictive power and reliability using a series of statistical measures. The models, obtained using LFER/ PaDEL descriptors and LM based ANN approach, showed strong fitting and predictive ability for many chemical families. This is the result of the selection of the optimal descriptors, combined with ANN,which provide the suitable computational scheme for interpreting thephysicochemical interactions influencing the distribution of chemicals into adipose tissue. Consequently,the developed models could be successfully applied in order to fill the data gaps of “data poor” chemicals which fall within their AD. In this way, the application of PBTK modeling, as well as the “safe by design” concept for environmental chemical compounds could be supported, promoting green chemistry and cost saving of product development.

# Acknowledgments

The authors gratefully acknowledge the financial support of CEFIC under the CEFIC-LRI project B11: Integrated External and Internal Exposure Modelling Platform (INTEGRA) and of the European Commission under grants: CROME-LIFE, which was funded from the LIFE+ program (LIFE12 ENV/GR/001040) and HEALS, which was funded from the 7th RTD Framework Programme of the European Union (Grant agreement no: 603946).

# References

Abraham, M.H., 1993. Application of solvation equations to chemical and biochemical processes. Pure and Applied Chemistry 65, 2503-2512.

Abraham, M.H., Chadha, H.S., Mitchell, R.C., 1994a. Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. Journal of Pharmaceutical Sciences 83, 1257-1268.

Abraham, M.H., Chadha, H.S., Whiting, G.S., Mitchell, R.C., 1994b. Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the Δlog P parameter of seiler. Journal of Pharmaceutical Sciences 83, 1085-1100.

Abraham, M.H., Martins, F., Mitchell, R.C., Salter, C.J., 1999. Hydrogen bonding. 47. Characterization of the ethylene glycol-heptane partition system: Hydrogen bond acidity and basicity of peptides. Journal of Pharmaceutical Sciences 88, 241-247.

Advanced Chemistry Development Inc., 2012, ACD/ChemSketch (Freeware), [www.acdlabs.com](http://www.acdlabs.com),

Arms, A.D., Travis, C.C., Health, U.S.E.P.A.O.o., Assessment, E., 1988. Reference Physiological Parameters in Pharmacokinetic Modeling. U.S. Environmental Protection Agency, Office of Health and Environmental Assessment.

Azoff, E.M., 1994. Neural Network Time Series Forecasting of Financial Markets. John Wiley & Sons, Inc.

Baláž, S., Lukáčová, V., 1999. A model-based dependence of the human tissue/blood partition coefficients of chemicals on lipophilicity and tissue composition. Quantitative Structure-Activity Relationships 18, 361-368.

Béliveau, M., Tardif, R., Krishnan, K., 2003. Quantitative structure-property relationships for physiologically based pharmacokinetic modeling of volatile organic chemicals in rats. Toxicology and Applied Pharmacology 189, 221-232.

Brown, R.P., Delp, M.D., Lindstedt, S.L., Rhomberg, L.R., Beliles, R.P., 1997. Physiological parameter values for physiologically based pharmacokinetic models. Toxicology and industrial health 13, 407-484.

Davies, B., Morris, T., 1993. Physiological parameters in laboratory animals and humans. Pharm Res 10, 1093-1095.

DeJongh, J., Verhaar, H.J.M., Hermens, J.L.M., 1997. A quantitative property-property relationship (QPPR) approach to estimate in vitro tissue blood partition coefficients of organic chemicals in rats and humans. Archives of Toxicology 72, 17-25.

European Commission, 2006, Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), https://ec.europa.eu/growth/sectors/chemicals/reach\_en,

Freisleben, B., 1992. Stock market prediction with backpropagation networks. Proceeding of the 5th International Conference Industrial and Engineering Applications of Artificial Intelligence and Expert System.

Frost, J., 2015, The Danger of Overfitting Regression Models, <http://blog.minitab.com/blog/adventures-in-statistics/the-danger-of-overfitting-regression-models>,

Garson, G.D., 1991. Interpreting neural-network connection weights. AI Expert 6, 47-51.

Gençay, R., 1999. Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules. Journal of International Economics 47, 91-107.

Gharagheizi, F., 2008. QSPR studies for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized regression neural network. QSAR and Combinatorial Science 27, 165-170.

Girden, E.R., 2011. Evaluating research articles : from start to finish, in: R. Kabacoff (Ed.), 3rd ed. ed. SAGE, Los Angeles.

Gramatica, P., Cassani, S., Chirico, N., 2014. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. J Comput Chem 35, 1036-1044.

Gramatica, P., Chirico, N., Papa, E., Cassani, S., Kovarich, S., 2013. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. Journal of Computational Chemistry 34, 2121-2132.

Harvey, R.G., 1985. Polycyclic Hydrocarbons and Carcinogenesis. American Chemical Society.

Haupt, R.L., Haupt, S.E., 2004. Practical Genetic Algorithms, 2nd ed. Wiley-Interscience.

Heaton, J., 2008. Introduction to Neural Networks with Java. Heaton Research.

Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T., 2005. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. ATLA Alternatives to Laboratory Animals 33, 445-459.

Kaiser, H.F., 1960. The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement 20, 141-151.

Karelson, M., Lobanov, V.S., Katritzky, A.R., 1996. Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chem Rev 96, 1027-1044.

Levitt, D.G., 2010. Quantitative relationship between the octanol/water partition coefficient and the diffusion limitation of the exchange between adipose and blood. BMC Clinical Pharmacology 10, 1-1.

Lipscomb, J.C., Haddad, S., Poet, T., Krishnan, K., 2012. Physiologically-based pharmacokinetic (PBPK) models in toxicity testing and risk assessment. Advances in experimental medicine and biology 745, 76-95.

Liu, H.X., Yao, X.J., Zhang, R.S., Liu, M.C., Hu, Z.D., Fan, B.T., 2005. Prediction of the tissue/blood partition coefficients of organic compounds based on the molecular structure using least-squares support vector machines. Journal of Computer-Aided Molecular Design 19, 499-508.

Liu, P., Long, W., 2009. Current Mathematical Methods Used in QSAR/QSPR Studies. International journal of molecular sciences 10, 1978-1998.

Man-Chung, C., Chi-Cheong, W., Chi-Chung, L., 2000. Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization. Computing in Economics and Finance, Society for Computational Economics 61.

National Research Council, Division on Earth and Life Studies, Board on Environmental Studies and Toxicology, Phthalates., C.o.t.H.R.o., 2009. Phthalates and Cumulative Risk Assessment: The Tasks Ahead. National Academies Press.

Nikolova-Jeliazkova, N., Jaworska, J., 2005. An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. Alternatives to laboratory animals : ATLA 33, 461-470.

OECD, 2014. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Publishing.

Open Notebook Science, 2016, Abraham Descriptor Prediction from SMILES, <http://showme.physics.drexel.edu/onsc/models/AbrahamDescriptorsModel001.php>,

Paixão, P., Aniceto, N., Gouveia, L.F., Morais, J.A.G., 2014. Prediction of Drug Distribution in Rat and Humans Using an Artificial Neural Networks Ensemble and a PBPK Model. Pharmaceutical Research 31, 3313-3322.

Payne, M.P., Kenny, L.C., 2002. Comparison of models for the estimation of biological partition coefficients. Journal of toxicology and environmental health. Part A 65, 897-931.

Pelekis, M., Krishnan, K., 2004. Magnitude and mechanistic determinants of the interspecies toxicokinetic uncertainty factor for organic chemicals. Regulatory Toxicology and Pharmacology 40, 264-271.

Peyret, T., Krishnan, K., 2011. QSARs for PBPK modelling of environmental contaminants. SAR QSAR Environ Res 22, 129-169.

Peyret, T., Poulin, P., Krishnan, K., 2010. A unified algorithm for predicting partition coefficients for PBPK modeling of drugs and environmental chemicals. Toxicology and Applied Pharmacology 249, 197-207.

Pierce, C.H., Dills, R.L., Silvey, G.W., Kalman, D.A., 1996. Partition coefficients between human blood or adipose tissue and air for aromatic solvents. Scandinavian Journal of Work, Environment & Health 22, 112-118.

Poulin, P., Krishnan, K., 1995. An algorithm for predicting tissue : Blood partition coefficients of organic chemicals from n‐octanol: Water partition coefficient data. Journal of Toxicology and Environmental Health 46, 117-129.

Puri, M., Pathak, Y., Sutariya, V., Tipparaju, S., Moreno, W., 2015. Artificial Neural Network for Drug Design, Delivery and Disposition, 1st Edition ed.

Puzyn, T., Leszczynski, J., Cronin, M.T.D., 2010. Recent Advances in QSAR Studies. Springer Science+Business Media, New York.

Rodgers, T., Rowland, M., 2006. Physiologically based pharmacokinetic modelling 2: Predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. Journal of Pharmaceutical Sciences 95, 1238-1257.

Rutkowska, E., Pajak, K., Jozwiak, K., 2013. Lipophilicity--methods of determination and its role in medicinal chemistry. Acta poloniae pharmaceutica 70, 3-18.

Sahigara, F., Ballabio, D., Todeschini, R., Consonni, V., 2014. Assessing the validity of QSARs for ready biodegradability of chemicals: an applicability domain perspective. Current computer-aided drug design 10, 137-147.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17, 4791-4810.

Sheela, K.G., Deepa, S.N., 2013. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. Mathematical Problems in Engineering 2013, 11.

Sheridan, R.P., Feuston, B.P., Maiorov, V.N., Kearsley, S.K., 2004. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J Chem Inf Comput Sci 44, 1912-1928.

Siddiqi, M.A., Laessig, R.H., Reed, K.D., 2003. Polybrominated Diphenyl Ethers (PBDEs): New Pollutants-Old Diseases. Clinical Medicine and Research 1, 281-290.

Sprunger, L.M., Gibbs, J., Acree Jr, W.E., Abraham, M.H., 2008. Correlation of human and animal air-to-blood partition coefficients with a single linear free energy relationship model. QSAR and Combinatorial Science 27, 1130-1139.

Todeschini, R., Consonni, V., 2008. Handbook of Molecular Descriptors. Wiley-VCH Verlag GmbH.

WHO, 2010. Characterization and application of Physiologically Based PharmacoKinetic models in risk assessment., IPCS Harmonization Project Document No. 9, Geneva, Switzerland.

Worth, A.P., Van Leeuwen, C.J., Hartung, T., 2004. The prospects for using (Q)SARs in a changing political environment - High expectations and a key role for the european commission's joint research centre. SAR and QSAR in Environmental Research 15, 331-343.

Yap, C.W., 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 32, 1466-1474.

Zhan, H., Zhang, Y., 2006. Convenient nonlinear model for predicting the tissue/blood partition coefficients of seven human tissues of neutral, acidic, and basic structurally diverse compounds. Journal of Medicinal Chemistry 49, 5815-5829.

Zhang, H., 2004. A new nonlinear equation for the tissue/blood partition coefficients of neutral compounds. Journal of Pharmaceutical Sciences 93, 1595-1604.