Journal of Molecular Graphics and Modelling 29 (2010) 450–460

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

# Prediction of toxicity and data exploratory analysis of estrogen-active endocrine disruptors using counter-propagation artificial neural networks

Nataša Stojić[a], Slavica Erić[b], Igor Kuzmanovski[a,*]

[a] Institut za Hemija, PMF, Univerzitet "Sv. Kiril i Metodij", PO Box 162, 1001 Skopje, Macedonia
[b] Faculty of Pharmacy, University of Belgrade, Vojvode Stepe 450, 11000 Belgrade, Serbia

ARTICLE INFO

*Article history:*
Received 6 July 2010
Received in revised form 5 September 2010
Accepted 9 September 2010
Available online 17 September 2010

*Keywords:*
Endocrine disruptors toxicity
Estrogenic activity prediction
Data exploratory analysis
SAR
QSAR
Counter-propagation artificial neural networks

ABSTRACT

In this work, a novel algorithm for optimization of counter-propagation artificial neural networks has been used for development of quantitative structure–activity relationships model for prediction of the estrogenic activity of endocrine-disrupting chemicals. The search for the best model was performed using genetic algorithms. Genetic algorithms were used not only for selection of the most suitable descriptors for modeling, but also for automatic adjustment of their relative importance. Using our recently developed algorithm for automatic adjustment of the relative importance of the input variables, we have developed simple models with very good generalization performances using only few interpretable descriptors. One of the developed models is in details discussed in this article. The simplicity of the chosen descriptors and their relative importance for this model helped us in performing a detailed data exploratory analysis which gave us an insight in the structural features required for the activity of the estrogenic endocrine-disrupting chemicals.



## 1. Introduction

The endocrine disrupting chemicals (EDCs) are exogenous substances or mixture of substances capable of altering the function(s) of the endocrine system and consequently causing adverse health effects in an intact organism, its progeny or subpopulations. However, those environmental toxicants, which are proposed to cause adverse effects through disruption of the endocrine system, belong to diverse groups of chemicals with very dissimilar structures, biochemical properties and mechanisms of action. Thus, chemical class cannot be used to discriminate toxicants with the potential to act as endocrine disrupter. The identification of potential endocrine disrupters requires a step-wise, systematic approach which includes mechanism of action, structure–activity analysis, metabolism and pharmacokinetics [1].

Concerning the mechanism of action of estrogenic endocrine-disrupting chemicals (EEDCs), it is supposed that they influence the normal function of the endocrine system in several ways: (1) competing with the endogenous endocrines for the binding to estrogen receptors (ER); which represents the main mechanism for large number of EDCs; (2) influencing the synthesis of the endogenous endocrines via interaction with enzymes; (3) interfering with transport of the endogenous endocrines by plasma transport proteins; (4) altering the mechanism of the endogenous endocrines via interaction with the enzymes; and (5) involving in the regulation of various neural centers [2–4]. Even if binding affinities of EEDCs to ER is much smaller compared to 17$\beta$-estradiol, they can bind competitively for ER due to their high concentrations. The cumulative effect of EEDCs, as a consequence of their wide use for decades (as insecticides, pesticides, additives in packing, etc.), can significantly affect human and wildlife health [5,6]. People exposed to EEDCs can develop cancer, or they can suffer from altered immune and nervous systems function [7,8]. Considering the undesirable effect of the endocrine disruptors and, on the other hand, the fact that some of these substances are economically important chemicals, the governmental bodies are trying to regulate the use of potential endocrine disruptors [9]. Consequently, the recent research has been focused on *in vitro* assays for identification of chemicals that mimic the role of 17$\beta$-estradiol by binding to the human ER [10], structural examination of the estrogen disruptors [11], as well as the development of *in silico* methodologies for priority testing [12].

Computer-based prediction of toxicity of EEDCs is becoming widespread because of the increasing need for screening of a large number of chemicals for their adverse biological activity [13,14,15,16,17,18,19]. The growing importance of the computer-based predictions of toxicity is a result of the recently adopted REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) policy by the European Union, which states that quantitative structure–activity relationship (QSAR) studies are accepted for use in regulatory purposes. Because of this structure–property

* Corresponding author.
*E-mail address:* shigor@iunona.pmf.ukim.edu.mk (I. Kuzmanovski).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.09.001

relationships (SPR) and structure–activity relationships (SAR) methods today are often used for *in silico* predictions of toxicity. It is expected that this approach will contribute in reduction of *in vivo* tests [20,21]. The validation of the developed models should be performed in accordance with the established principles by OECD [22]. On the basis of SAR analysis, few criteria were found to be essential for estrogenic activity of EEDCs, such as the presence of specific structural features as well as their precise size and orientation [23]. However, correlation between structure and toxicity of potential EEDCs still remains a formidable task, since toxicants that mimic or block estrogen activity are structurally dissimilar, lacking the consistent structural characteristics.

Modern computer based tools have enabled the development of QSAR models for identifying steric and electrostatic features of a molecule in three-dimensional space related to estrogenic activity [24,25]. Quantum similarity methods have also been applied for development of models dividing whole set on several chemical classes [26]. Despite the good quality of the models presented, the majority of QSAR models developed to date are based on the activity of relatively small number of compounds with similar structural features. So, it is not certain whether such models could be applied for the prediction of estrogenic activity of all new and more structurally diverse estrogenic chemicals, which number increases rapidly. In addition, some of presented models are not completely interpretable, representing *black box models*, so they cannot be fully explored for understanding of structural features determining estrogenic activity of the chemicals. In order to better understand the structural requirements for ER binding, it is important to have a reliable and interpretable model, established with data obtained by consistent experimental assay and, at the same time, covering a broad range of chemical classes.

In this article, we tried to develop a reliable model based on diverse data set consisting of 188 EEDCs [27]. We used this data set because the log($IC_{50}$) values were obtained uniform experimental procedure. The modeling was performed using counter-propagation artificial neural networks (CPANN) [28,29,30,31,32,33,34]. In order to fulfill our goal to develop simple and, if possible, interpretable model which might give us an insight into different structural features of the EEDCs that are influencing the estrogen binding affinity, we used our recently developed algorithm for automatic adjustment of the relative importance of the input variables [35].

### 1.1. The data analysis and the algorithms used

#### 1.1.1. The data

Several papers have been published using the data obtained by a validated ER competitive binding assay [36,37,38]. In this study, we used the data set consisting of 188 structurally diverse ligands [27], originally divided into twelve groups: (1) steroidal estrogens; (2) synthetic estrogens; (3) antiestrogens; (4) other miscellaneous steroids; (5) alkylphenols; (6) diphenyl derivatives; (7) organochlorines; (8) pesticides; (9) alkylhydroxybenzoate preservatives (parabens); (10) phthalates; (11) benzophenone compounds and a number of (12) other miscellaneous compounds. As a dependent variable we used log($IC_{50}$) values presented in the original work [27].

Before we started the calculation of the descriptors, we removed the disjoined structures (in the original article [27] labeled with 34, 35, 161, 164 and 183) from the data set, since the Dragon 5.5 [39] program which was used for descriptor calculation does not recognize the disjoined structure as a single structure. We did not include the neutral forms of these structures instead, since descriptors obtained from neutral form might not be relevant for modeling of *in vitro* obtained log($IC_{50}$). Furthermore, the structures that appeared more than once in the data set were also removed. In this case, whenever possible, we kept the compounds which had the largest purity. Finally, after that, the data set was reduced down to 174 structures (Table 1).

As previously stated, for modeling purposes we calculated a number of descriptors using Dragon 5.5. For the development of the models which could be easier for interpretation, we tried to select descriptors which are as simple as possible. For the selection of a suitable subset of the descriptors which will be used for modeling, we first correlated all descriptors with log($IC_{50}$). The descriptors with the largest absolute value of the pair-wise correlation coefficient with log($IC_{50}$) were further used. Additionally, among each pair of the previously selected descriptors correlation coefficients were calculated. Among each pair of descriptors with absolute value of the correlation coefficients larger than 0.8 the one which was easier for interpretation was selected for further use (Table 2). The only exception of this rule is the descriptor connectivity index (X1), for which the correlation coefficient with nC was larger than 0.8. Since the meaning of these two correlated descriptors is different (nC belongs to constitutional and the X1 to topological descriptors), but also because of the fact the correlation coefficient between X1 and log($IC_{50}$) is the third largest in the data set this descriptor was included in the final data set. For the remaining descriptors presented in the Table 2, the absolute value of the correlation coefficient is smaller than 0.79.

For a proper external validation of the models, in accordance with OECD principles [22], the data set was divided into training and test set using Kennard–Stone algorithm [40]. The test set was composed of 1/4 of the molecules in the data set, while the remaining structures represented the training set.

#### 1.1.2. Counter-propagation artificial neural networks

This neural network algorithm has been previously used in our research during the last few years [34,35,41]. The details about the algorithm and its use in chemistry are well described in the literature [32,33]. Here we will only briefly describe how this neural network algorithm works.

CPANN are usually represented as consisting of two layers (Fig. 1). The first layer (called Kohonen layer) performs the mapping of the multidimensional input data into, most often, two-dimensional plane of neurons. The mapping is performed by use of *competitive learning* – often called winner-takes-it-all strategy.

The training process of the CPANN [32,33] is performed in similar way as the training process of self-organizing maps (SOM) [42,43]. This means that, in both cases, the vectors with the $N$ input variables ($x_s = x_{s1}, \ldots x_{si}, \ldots x_{sN}$) are compared only with the weights ($w_j = w_{j1}, \ldots w_{ji}, \ldots w_{jN}$) of the neurons in the Kohonen layer using, most often, the criterion for similarity between input variables and the weights:

$$\text{out}_c \leftarrow \min\left\{\sum_{i=1}^{N}(x_{si} - w_{ji})^2\right\} \quad (1)$$

Once the winning neuron $c$ is found using Eq. (1) the weights of both layers are simultaneously adjusted for the pairs of input and target vectors $(x, y)$ using following equations:

$$w_{ji}^{\text{new}} = w_{ji}^{\text{old}} + \eta(t)\cdot a(d_j - d_c)\cdot(x_i - w_{ji}^{\text{old}}) \quad (2)$$

$$u_{ji}^{\text{new}} = u_{ji}^{\text{old}} + \eta(t)\cdot a(d_j - d_c)\cdot(y_i - u_{ji}^{\text{old}}) \quad (3)$$

In these two equations $\eta(t)$ represents the learning rate. The learning rate is a non-increasing function and it defines the intensity of the changes of the weights during the training process. Beside the changes of the weights of the *central neuron*, the weights of the neighboring neurons are also corrected. The term $a(d_j - d_c)$ represents the neighborhood function which defines the intensity of the change of the weight of the neurons that surround central neuron,

**Table 1**
The data set of estrogenic endocrine-disrupting chemicals used in this study. The labels of the structures are the same as in the original article [27]. After the removal of the structures that appear more than once and the disjoined structures the data set is composed of total number of 174 structures. Types of binders: SB – strong binder; MB – moderate binder; WB – weak binder; NB – nonbinding substance.

| No.[a] | | Chemical name | Type of binder | log($IC_{50}$) | CAS numbers |
|---|---|---|---|---|---|
| Steroidal estrogens | | | | | |
| 1 | 1 | 17$\beta$-Estradiol | SB | −9.04 | 50-28-2 |
| 2 | 2 | Estra-1,3,5(10)-trien-3-ol | SB | −8.30 | 25517-69-5 |
| 4 | 4 | Estriol | SB | −8.03 | 50-27-1 |
| 5 | 5 | Estrone | SB | −7.91 | 53-16-7 |
| 6 | 6 | 17$\alpha$-Estradiol | SB | −7.53 | 57-91-0 |
| 7 | 7 | 1,3,5(10)-Estratrien-3, 6$\alpha$,17$\beta$-triol | MB | −6.89 | Not available |
| 8 | 8 | 3-Hydroxyestra-1,3,5(10)-trien-16-one | MB | −6.75 | 474-86-2 |
| 9 | 9 | 3-Deoxyestradiol | MB | −6.74 | 2529-64-8 |
| 10 | 10 | 16$\beta$-Hydroxy-16-methyl-3-methylether 17$\beta$ estradiol | MB | −5.56 | 5108-94-1 |
| 11 | 11 | 3-Methylestriol | MB | −5.39 | 3434-79-5 |
| 12 | 12 | 3-Deoxyestrone | WB | −4.84 | 53-45-2 |
| Synthetic estrogens | | | | | |
| 13 | 1 | Diethylstilbestrol (DES) | SB | −9.64 | 56-53-1 |
| 14 | 2 | Meso-hexestrol | SB | −9.52 | 84-16-2 |
| 15 | 3 | Ethinyl estradiol | SB | −9.32 | 57-63-6 |
| 16 | 4 | Dienestrol | SB | −8.61 | 84-17-3 |
| 17 | 5 | Diethylstilbestrol monomethyl ether | SB | −8.35 | 18839-90-2 |
| 18 | 6 | 3,3′-Dihydroxyl hexestrol | SB | −8.23 | 79199-51-2 |
| 19 | 7 | Dimethylstilbestrol | SB | −8.20 | 13366-36-4 |
| 20 | 8 | Moxestrol | SB | −8.18 | 34816-55-2 |
| 21 | 9 | 2,6-Dimethyl hexestrol | SB | −8.15 | Not available |
| 22 | 10 | Hexestrol, mono methyl ether | SB | −8.01 | 13026-26-1 |
| 23 | 11 | p-($\alpha$,$\beta$-Diethyl-*p*-methyl phenethyl)-meso phenol | SB | −7.64 | Not available |
| 24 | 12 | DL-Hexestrol | SB | −7.60 | 84-16-2 |
| 25 | 13 | Mestranol | SB | −7.40 | 72-33-3 |
| 26 | 14 | $\alpha$,$\alpha$-Dimethyl-$\beta$-ethyl allenolic acid | MB | −7.02 | 15372-37-9 |
| 27 | 15 | Diethylstilbestrol dimethyl ether | MB | −5.79 | 7773-34-4 |
| 28 | 16 | Doisynoestrol | WB | −4.30 | 15372-34-6 |
| Antiestrogens | | | | | |
| 29 | 1 | 4-Hydroxytamoxifen | SB | −9.28 | 68047-06-3 |
| 30 | 2 | ICI 182,780 (faslodex) | SB | −8.61 | 129453-61-8 |
| 31 | 3 | Droloxifene (3-hydroxytamoxifen) | SB | −8.22 | 82413-20-5 |
| 32 | 4 | ICI 164,384 | SB | −8.20 | 98007-99-9 |
| 36 | 8 | Nafoxidine | MB | −6.90 | 1845-11-0 |
| 37 | 9 | Triphenylethylene | WB | −4.26 | 58-72-0 |
| Other miscellaneous steroids | | | | | |
| 38 | 1 | Norethynodrel | WB | −6.39 | 68-23-5 |
| 40 | 3 | 5$\alpha$-Androstane-3$\beta$,17$\beta$-diol | WB | −6.12 | 571-20-0 |
| 41 | 4 | 5$\alpha$-Androstane-3$\alpha$,17$\beta$-diol | WB | −4.37 | 1852-53-5 |
| 42 | 5 | 5$\alpha$-Dihydrotestosterone | NB | −3 | 121520-97-6 |
| 43 | 6 | Aldosterone | NB | −4 | 52-39-1 |
| 44 | 7 | Cholesterol | NB | −3 | 57-88-5 |
| 45 | 8 | Corticosterone | NB | −4 | 50-22-6 |
| 46 | 9 | Dexamethasone | NB | −4 | 50-02-2 |
| 47 | 10 | Epitestosterone | NB | −4 | 481-30-1 |
| 48 | 11 | Etiocholan-17$\beta$-ol-3-one | NB | −4 | 571-22-2 |
| 49 | 12 | Progesterone | NB | −3 | 57-83-0 |
| 50 | 13 | Testosterone | NB | −3 | 58-22-0 |
| Alkylphenols | | | | | |
| 51 | 1 | 4-Nonylphenol | MB | −5.61 | 104-40-5 |
| 56 | 6 | 4-Dodecylphenol | MB | −5.31 | 27459-10-5 |
| 57 | 7 | 4-*tert*-Octylphenol | MB | −5.22 | 140-66-9 |
| 58 | 8 | 4-Octylphenol | WB | −4.70 | 1806-26-4 |
| 59 | 9 | 4-*n*-Nonylphenol | WB | −4.55 | 104-40-5 |
| 60 | 10 | 4-*tert*-Amylphenol | WB | −3.78 | 80-46-6 |
| 61 | 11 | 4-*sec*-Butylphenol | WB | −3.67 | 99-71-8 |
| 62 | 12 | 4-Chloro-3-methylphenol | WB | −3.66 | 59-50-7 |
| 63 | 13 | 2-*sec*-Butylphenol | WB | −3.50 | 89-72-5 |
| 64 | 14 | 4-*tert*-Butylphenol | WB | −3.43 | 98-54-4 |
| 65 | 15 | 2-Chloro-4-methylphenol | WB | −3.38 | 6640-27-3 |
| 66 | 16 | 4-Chloro-2-methylphenol | WB | −3.37 | 1570-64-5 |
| 67 | 17 | 3-Ethylphenol | WB | −3.18 | 620-17-7 |
| 68 | 18 | 4-Ethylphenol | WB | −2.87 | 123-07-9 |
| 69 | 19 | 2-Ethylphenol | NB | −3 | 90-00-6 |
| 70 | 20 | Eugenol | NB | −3 | 97-53-0 |
| 71 | 21 | Isoeugenol | NB | −4 | 97-54-1 |
| Diphenyl derivatives | | | | | |
| 72 | 1 | 2,2-Bis-(4-hydroxyphenyl)-butane (*bisphenol B*) | MB | −5.97 | 77-40-7 |
| 73 | 2 | Bisphenol A | WB | −4.93 | 80-05-7 |
| 74 | 3 | 2,2′-Methylenebis (4-chlorophenol) | WB | −4.59 | 106-48-9 |
| 75 | 4 | Bis (4-hydroxyphenyl)-methane | WB | −4.02 | 620-92-8 |
| 76 | 5 | 4,4′-Sulfonyldiphenol | WB | −3.97 | 80-09-1 |

Table 1 ( *Continued* )

| No.[a] | | Chemical name | Type of binder | log($IC_{50}$) | CAS numbers |
|---|---|---|---|---|---|
| 77 | 6 | Diphenolic acid | WB | −3.92 | 126-00-1 |
| 78 | 7 | 4,4′-Methylenebis (2,6-di-*tert*-butylphenol) | NB | −4 | 118-82-1 |
| 79 | 8 | Bis (2-hydroxyphenyl)-methane | NB | −5 | 2467-02-9 |
| 80 | 9 | 4,4′-Dihydroxystilbene | MB | −6.49 | 659-22-3 |
| 81 | 10 | 2,2′,4,4′-Tetrahydroxybenzil | MB | −6.36 | 5394-98-9 |
| 82 | 11 | 4, 4′-Ethylene diphenol | MB | −5.61 | 6052-84-2 |
| 83 | 12 | 4-Phenethylphenol | WB | −4.35 | 6335-83-7 |
| 84 | 13 | 4-Stilbenol | NB | −4 | 3839-46-1 |
| 85 | 14 | 4-Phenylphenol | WB | −4.01 | 92-69-3 |
| 86 | 15 | 3-Phenylphenol | WB | −3.61 | 580-51-8 |
| 87 | 16 | 2-Phenylphenol | NB | −4 | 90-43-7 |
| Organochlorines | | | | | |
| 88 | 1 | o,p′-DDT | WB | −4.19 | 789-02-6 |
| 89 | 2 | o,p′-DDD | NB | −3.52 | 53-19-0 |
| 90 | 3 | p,p′-DDD | NB | −4 | 72-54-8 |
| 91 | 4 | o,p′-DDE | NB | −3.30 | 3424-82-6 |
| 92 | 5 | p,p′-DDE | NB | −4 | 72-55-9 |
| 93 | 6 | p,p′-DDT | NB | −3 | 50-29-3 |
| 94 | 7 | Dihydroxymethoxychlor olefin | SB | −7.46 | 14868-03-2 |
| 95 | 8 | Dihydroxymethoxychlor (*HPTE*) | MB | −6.44 | 2971-36-0 |
| 96 | 9 | Monohydroxymethoxychlor olefin | MB | −6.40 | 75938-34-0 |
| 97 | 10 | Monohydroxymethoxychlor | MB | −6.16 | 28463-03-8 |
| 99 | 12 | Methoxychlor | NB | −4 | 72-43-5 |
| 100 | 13 | Methoxychlor olefin | NB | −4 | 2132-70-9 |
| 101 | 14 | 2′,3′,4′,5′-Tetrachloro-4-biphenylol | MB | −6.40 | 67651-34-7 |
| 102 | 15 | 2′,5′-Dichloro-4-biphenylol | MB | −5.60 | 53905-28-5 |
| 103 | 16 | 4-Chloro-4′-biphenylol | WB | −4.86 | 28034-99-3 |
| 104 | 17 | 2-Chloro-4-biphenylol | WB | −4.27 | 92-04-6 |
| 105 | 18 | 3,3′,5,5′-Tetrachloro-4,4′-biphenyldiol | WB | −3.79 | 13049-13-3 |
| 106 | 19 | 2,4′-Dichlorobiphenyl | WB | −3.43 | 34883-43-7 |
| 107 | 20 | 2,2′,4,4′-Tetrachlorobiphenyl | NB | −4 | 2437-79-8 |
| 108 | 21 | 3,3′,4,4′-Tetrachlorobiphenyl | NB | −3.52 | 32598-13-3 |
| 109 | 22 | 4,4′-Dichlorobiphenyl | NB | −3.52 | 2050-68-2 |
| Pesticides | | | | | |
| 110 | 1 | Kepone | MB | −5.15 | 143-50-0 |
| 111 | 2 | 2,4,5-T | NB | −3 | 93-76-5 |
| 112 | 3 | 2,4-Dichlorophenoxyacetic acid (*2,4-D*) | NB | −4 | 94-75-7 |
| 113 | 4 | α-Chlordane (mix of isomers) | NB | −3 | 5103-71-9 |
| 114 | 5 | Alachlor | NB | −4 | 15972-60-8 |
| 115 | 6 | Aldrin | NB | −3.22 | 309-00-2 |
| 116 | 7 | Atrazine | NB | −4 | 1912-24-9 |
| 117 | 8 | Carbaryl | NB | −4 | 63-25-2 |
| 118 | 9 | Carbofuran | NB | −4 | 1563-66-2 |
| 119 | 10 | Dieldrin | NB | −4 | 60-57-1 |
| 121 | 12 | Endosulfan | NB | −3 | 115-29-7 |
| 122 | 13 | Heptachlor | NB | −4 | 76-44-8 |
| 123 | 14 | Hexachlorobenzene | NB | −3 | 118-74-1 |
| 124 | 15 | Lindane | NB | −4 | 58-89-9 |
| 125 | 16 | Metolachlor | NB | −4 | 51218-45-2 |
| 126 | 17 | Mirex | NB | −4 | 2385-85-5 |
| 127 | 18 | Prometon | NB | −3 | 1610-18-0 |
| 128 | 19 | Simazine | NB | −4.47 | 122-34-9 |
| 129 | 20 | Vinclozolin | NB | −4 | 83792-61-4 |
| Alkylhydroxybenzoate preservatives (parabens) | | | | | |
| 130 | 1 | 2-Ethylhexyl 4-hydroxybenzoate | MB | −5.30 | 5153-25-3 |
| 131 | 2 | Heptyl 4-hydroxybenzoate | WB | −4.95 | 1085-12-7 |
| 132 | 3 | Benzyl 4-hydroxybenzoate | WB | −4.50 | 94-18-8 |
| 133 | 4 | Butyl 4-hydroxybenzoate | WB | −3.97 | 94-26-8 |
| 134 | 5 | Propyl 4-hydroxybenzoate | WB | −3.82 | 94-13-3 |
| 135 | 6 | Ethyl 4-hydroxybenzoate | WB | −3.82 | 120-47-8 |
| 136 | 7 | Methyl 4-hydroxybenzoate | WB | −3.61 | 99-76-3 |
| Phthalates | | | | | |
| 137 | 1 | Benzylbutyl phthalate | NB | −3 | 85-68-7 |
| 138 | 2 | Bis(2-ethylhexyl) phthalate | NB | −3 | 117-81-7 |
| 139 | 3 | Dibutyl phthalate | NB | −3 | 84-74-2 |
| 140 | 4 | Diethyl phthalate | NB | −3 | 84-66-2 |
| 141 | 5 | Di-isobutyl phthalate | NB | −3 | 84-69-5 |
| 142 | 6 | Di-isononyl phthalate | NB | −3 | 68515-48-0 |
| 143 | 7 | Dimethylphthalate | NB | −3 | 131-11-3 |
| 144 | 8 | *n*-Dioctyl phthalate | NB | −3 | 117-84-0 |
| Benzophenone compounds | | | | | |
| 145 | 1 | 4,4′-Dihydroxybenzophenone | WB | −4.59 | 611-99-4 |
| 146 | 2 | 2,4′-Dihydroxybenzophenone | WB | −4.44 | 131-56-6 |
| 147 | 3 | 2,2′-Dihydroxy-4-methoxybenzophenone | NB | −4 | 131-53-3 |
| 148 | 4 | 2,2′-Dihydroxybenzophenone | NB | −4 | 835-11-0 |
| 149 | 5 | 2-Hydroxy-4-methoxybenzophenone | NB | −4 | 131-57-7 |

Table 1 ( *Continued* )

| No.[a] | | Chemical name | Type of binder | log($IC_{50}$) | CAS numbers |
|---|---|---|---|---|---|
| Other miscellaneous compounds | | | | | |
| 150 | 1 | Castor oil | NB | −4 | 8001-79-4 |
| 151 | 2 | Cinnamic acid | NB | −3 | 621-82-9 |
| 152 | 3 | Folic acid | NB | −4 | 59-30-3 |
| 153 | 4 | Suberic acid | NB | −4 | 505-48-6 |
| 154 | 5 | 1,8-Octanediol | NB | −4 | 629-41-4 |
| 155 | 6 | Benzyl alcohol | NB | −2 | 100-51-6 |
| 156 | 7 | Hexyl alcohol | NB | −2 | 111-27-3 |
| 157 | 8 | Nerolidol | NB | −3 | 40716-66-3 |
| 158 | 9 | 2-Furaldehyde (*furfural* ) | NB | −3 | 98-01-1 |
| 159 | 10 | Heptanal | NB | −2 | 111-71-7 |
| 160 | 11 | Vanillin | NB | −4 | 121-33-5 |
| 162 | 13 | 4,4′-Methylenebis (N,N-dimethylaniline) | NB | −3 | 101-61-1 |
| 164 | 15 | 4-Aminophenylether (*4,4′-oxydianiline*) | NB | −3 | 101-80-4 |
| 165 | 16 | Butyl-4-aminobenzoate | NB | −4 | 94-25-7 |
| 166 | 17 | 4-Heptyloxyphenol | WB | −4.17 | 13037-86-0 |
| 167 | 18 | 4-Benzyloxyphenol | WB | −3.60 | 103-16-2 |
| 168 | 19 | Cineole | NB | −2 | 470-82-6 |
| 169 | 20 | Dibenzo-18-crown-6 | NB | −5 | 14187-32-7 |
| 170 | 21 | Ethyl cinnamate | NB | −3 | 103-36-6 |
| 171 | 22 | 1,6-Dimethylnaphthalen | NB | −4 | 575-43-9 |
| 172 | 23 | Benzo[a]fluorene | NB | −4.47 | 238-84-6 |
| 173 | 24 | *sec*-Butylbenzene | NB | −3 | 135-98-8 |
| 174 | 25 | Chrysene | NB | −5 | 218-01-9 |
| 175 | 26 | *n*-Butylbenzene | NB | −3.69 | 104-51-8 |
| 176 | 27 | *trans*,*trans*-1,4-Diphenyl-1,3-butadiene | NB | −4 | 538-81-8 |
| 177 | 28 | Aurin | MB | −5.55 | 603-45-2 |
| 178 | 29 | Nordihydroguairaretic acid | MB | −5.54 | 500-38-9 |
| 179 | 30 | Phenolphthalein | MB | −5.17 | 77-09-8 |
| 180 | 31 | Phenol red | WB | −3.80 | 143-74-8 |
| 181 | 32 | Phenolphthalin | WB | −3.37 | 81-90-3 |
| 182 | 33 | 2-Chlorophenol | NB | −3.70 | 95-57-8 |
| 184 | 35 | Caffeine | NB | −4 | 58-08-2 |
| 185 | 36 | Dopamine | NB | −4 | 51-61-6 |
| 186 | 37 | Melatonin | NB | −4 | 73-31-4 |
| 187 | 38 | Thalidomide | NB | −3 | 50-35-1 |
| 188 | 39 | Triphenyl phosphate | NB | −4 | 115-86-6 |

[a] The labels for the structures are the same as in the original article [27].

and the difference $d_j - d_c$ is the topological distance between the winning neuron $c$ and the neuron $j$ which weights are adjusted. $w_{ji}^{old}$ and $w_{ji}^{new}$ represent the weights of the Kohonen layer before and after its adjustments were performed, while $u_{ji}^{old}$ and $u_{ji}^{new}$ are the weights of the output layer before and after the performed adjustments.

At the end of this part we should mention that from the programming point of view the main difference between CPANN (an algorithm for supervised learning) and SOM (an algorithm for unsupervised leaning) is in the fact that, while training CPANN, the winning neuron is selected *only* by comparing the independent variables with the weights of the corresponding weight levels from the Kohonen layer [32–34]. After that, the weights of the winning neuron and neighboring neurons are adjusted in both layers according to the distances between the weights of the corresponding variables (independent and dependent ones) of a particular object.

In this work we chose CPANN over SOM algorithm because it permits us not only to perform the mapping on the Kohonen layer, and consequently a clustering the structures according to their similarity, but it also permits the prediction of the log($IC_{50}$) values. Additional advantage is the supervised character of the CPANN algorithm which gives us an opportunity to check the log($IC_{50}$) could serve as a *guide* to perform automated selection of the descriptors, network size and the epochs used for training of the CPANN using, for example, genetic algorithm.

**Table 2**
The descriptors used for the optimization of the CPANN using genetic algorithms.

| No. | Symbol | Description |
|---|---|---|
| 1 | nC | Number of C atoms in the molecule |
| 2 | Vindex | Balaban V index |
| 3 | X1 | Connectivity index chi-1 |
| 4 | nArOH | Number of aromatic hydroxyls |
| 5 | nCIC | Number of rings |
| 6 | ALOGP | Ghose–Crippen octanol–water partition coefficient |
| 7 | nR = Ct | Number of aliphatic tertiary C($sp^2$) atoms |
| 8 | nCb- | Number of substituted benzene atoms |
| 9 | nR#CH/X | Number of terminal C($sp$) atoms |
| 10 | nOHs | Number of secondary OH groups |
| 11 | Hy | Hydrophilic factor |
| 12 | Ui | Unsaturation index |
| 13 | nArCOOR | Number of aromatic esters |
| 14 | nCt | Number of total tertiary C($sp^3$) atoms |

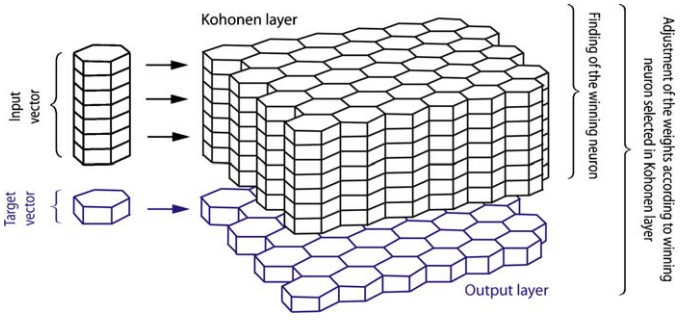


**Fig. 1.** Graphical representation of counter-propagation neural network. The Kohonen layer serves for mapping of the (multidimensional) input data – $x$ into two-dimensional map and finding the winning neurons, while the weights in both layers (Kohonen and output) are adjusted under the same conditions (the learning rate and the neighborhood function) using pairs of input and target vectors ($x$, $y$).

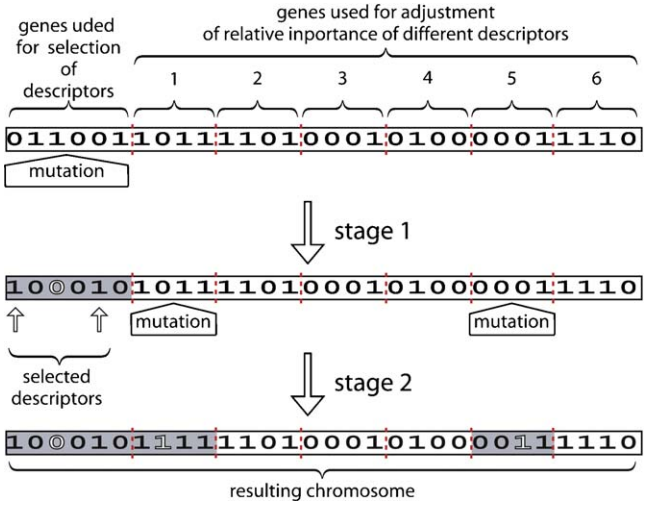


**Fig. 2.** In order not to lose valuable information in some of the genes responsible for relative importance of descriptors not selected in the model defined by the current chromosome the mutation is applied only to those parts of the offspring chromosome that correspond to selected descriptors. The genes in the part responsible for the adjustment of relative importance for the selected descriptors are then transformed into decimal numbers which represent the relative importance of the input variables.

#### *1.1.3. Genetic algorithm*

This algorithm is suitable for automated search for the extremes of the functions which do not posses *nice* properties like differentiability and continuity. Mimicking the main ideas on how the genetic material is progressing in the living organisms (with crossovers, mutations, and survival of the fittest strategy), the genetic algorithms (GA) helps in optimization of different problems. The details of this algorithm and its use in chemistry have been discussed in the chemical literature [44–50] and it has been proven as a valuable tool for optimization of models based on different algorithms.

#### *1.1.4. Automatic adjustment of the relative importance of the input variables*

Whenever we do not have an experience about the relative importance of the input variables, the auto-scaling or range-scaling are the most commonly used procedures for preprocessing [35].

During the training of CPANN there are no interactions between the weight levels that correspond to different input variables. This lack of interactions among the weight levels may be used as an advantage of CPANN, because it could help not only in interpretation of the influence of the input variables on the modeled property, but also it allows us to perform automated adjustment of their relative importance which could simplify the model [35].

The main idea of the algorithm for automatic adjustment of the relative importance of the input variables (illustrated in Fig. 2) is in the use of GA for selection of the suitable input variables (in our case descriptors) and in the use of the selective mutation to the part of the chromosomes responsible for the adjustment of their relative importance. Using this algorithm it is possible to obtain simpler models and at the same time models which are easier for interpretation. More detailed explanation about the algorithm is presented in our recently published paper [35].

#### *1.1.5. Encoding of the chromosomes*

In this work we used GA not only for variable selection and for adjustment of the relative importance of the variables, but also for the selection of the most appropriate size (width and length) of the CPANN. Eight genes were used for this purpose (4 for the width and 4 for the length of the CPANN). Their values were changed in the interval between 4 and 19. The training of the CPANN was performed in two phases: rough and fine training phase. In the rough training phase the number of the epochs was searched in the interval between 10 and 25 (four genes were used). The number of training epochs in the fine training phase was searched in the interval between 1 and 128. However, in order to have always longer fine training phase, the obtained number of epochs by these genes was increased by twice the number of epochs obtained for the rough training phase.

In order to find the relative importance of the input variables for each descriptor five more genes were used for adjustment of their relative importance (in the interval between 1/32 and 32/32). For this purpose additional 70 (=14 × 5) genes were used. In total, each chromosome had 103 genes.

For the GA optimization we used populations of 100 chromosomes. 20% of the chromosomes with the best performances were used as parents for formation of the offspring in the next generation. The whole optimization procedure lasted for 600 generations. The performances of the models during the GA were checked using root mean square error of prediction (*RMSEP*) for the results obtained using cross-validation leave-5%-out performed using the structures which were part of the training set. Additionally, the performances of the final models were evaluated using the independent test set of structures which were not used during the cross-validation.

Variable mutation rate was applied during the optimization. At the beginning the mutation probability was 0.10 in order to help a better initial exploration of the space in which our models were defined. The mutation was kept at this level until generation 100. After that the probability of occurrence of the mutation in the chromosomes linearly decreased down to 0.05 in generation 200. From here on, until the end of the GA optimization, the mutation probability was kept at this level.

The mutation in the part of the chromosomes responsible for adjustment of the relative importance of the input variables was managed in a different way. In this part of the chromosomes we used larger mutation probabilities in order to enforce the search for the most appropriate relative importance of the input variables. So, until the generation 150 mutation probability was 0.25. After that, until the generation 300 the mutation was 0.20. Starting from generation 301 until the end of the optimization the probability for the occurrence of the mutation was 0.15.

The entire optimization using GA was repeated several times. We obtained several models with good performances which at the same time were suitable for interpretation. The performance as well as the data exploratory analysis of one of these models is discussed in details.

## 2. Results and discussion

One of the simplest models that we obtained (presented in Fig. 3) will be discussed here in details. The discussed CPANN model has rectangular shape (13 × 10 neurons). It was trained using 18 epochs in the rough training phase and with 156 epochs in the fine training phase. This model consists of only five descriptors (nC, nArOH, nClC, nR = Ct, nCb-).

The selected descriptors for this model are in agreement with SAR analysis of estrogen active endogenous disruptors, where a few general structural requirements were found to be relevant for a compound to bind to the estrogen receptor: (1) H-bonding ability of the phenolic ring mimicking the 3-OH of estradiol ($E_2$), (2) H-bond donor mimicking the 17$\beta$-OH and O–O distance between 3- and 17$\beta$-OH of $E_2$, (3) precise steric hydrophobic centers mimicking steric 7$\alpha$- and 11$\beta$-substituents of $E_2$, (4) hydrophobicity and (5) a ring structure [51,52]. The SAR analysis for ER binding by individual chemical classes indicate that some features may well represent binding dependencies for all structural classes, while other features

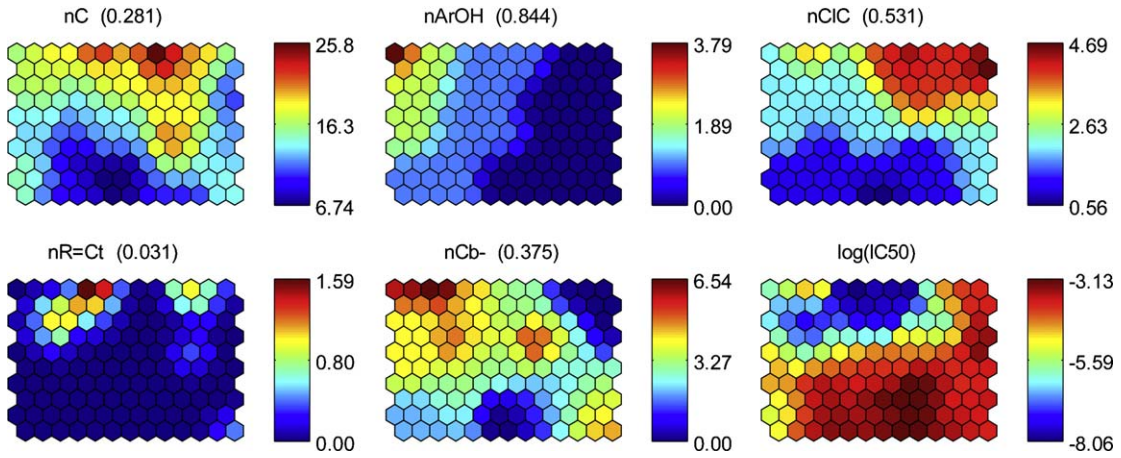


**Fig. 3.** Weight levels for the discussed CPANN model. Next to the labels of the descriptors the relative importance of each of the descriptors is presented in brackets.

may better represent binding dependencies for a specific structural class, which is also shown in our model. However, these structural features are inherently related, suggesting that structural commonality exists among structurally diverse estrogens.

The expected versus found values for the training and for the test set obtained using the discussed model are presented in Fig. 4 ($R^2$ = 0.854 for the training set and $R^2$ = 0.741 for the test set). By careful examination of the weight levels and the relative importance of the descriptors (given in the brackets next to the descriptor labels in Fig. 3), one can notice that the most influential descriptor (with the largest relative importance) on the performed mapping and, at the same time, on the performances of the discussed CPANN model is the number of aromatic hydroxyl groups (nArOH). This feature included in the discussed model is in agreement with the findings that the contribution of the phenolic OH group at position 3 is much more significant for binding than any other structural feature of the estrogen active substances.

The recently reported crystal structure of the estrogen receptor ER–$E_2$ complexes reveals that the 3- and 17$\beta$-OH groups of $E_2$ derivatives primarily serve as H-bond donors or acceptors while interacting with the receptor binding site [53]. The elimination or modification of either of these two groups reduces a chemicals binding affinity for the receptor. This impact is more dramatic at 3-position than at 17$\beta$-position. The grouping of the most of the structures that posses steroid skeleton on the CPANN is in agreement with these findings. Namely, the most of the *steroidal estrogens* which are strong binders have 3- and 17$\beta$-OH groups. In our model, these structures are labeled with the numbers 1, 4, 6 and 7 (Fig. 5a). The small influence of the absence of OH group at position 17 is illustrated with the estra-1,3,5(10)-trien-3-ol (structure number 2 on Table 1) which is the second strongest binder in this group (Figs. 3 and 5a the weight level for log($IC_{50}$)). Further right, outside the region represented with blue color, which corresponds to small values for log($IC_{50}$), moderate and weak binders, that belong to this group of substances, are mapped. A very good example to illustrate the importance of the presence of phenolic ring in the structure, using CPANN, is the comparison of the estriol (which is a strong binder) with three 3-methylestriol (the 11th structure in Table 1). As a result of the substitution of the phenolic OH group with methoxy group, the 3-methylestriol has smaller activity and it belongs to a group of moderate binders, indicating that phenolic OH group interacts as H-bond donor. This is consistent with the findings from the ER–$E_2$ crystal structure that the 3-OH has H-bonding interactions with Glu 353, Arg 394, and a water molecule, whereas 17$\beta$-OH only forms one H-bond with His 524. Studies also showed that the elimination of 3-OH causes a greater reduction in the activity of weak estrogens than strong ones. The H-bond donor ability of phenolic OH depends on the number of

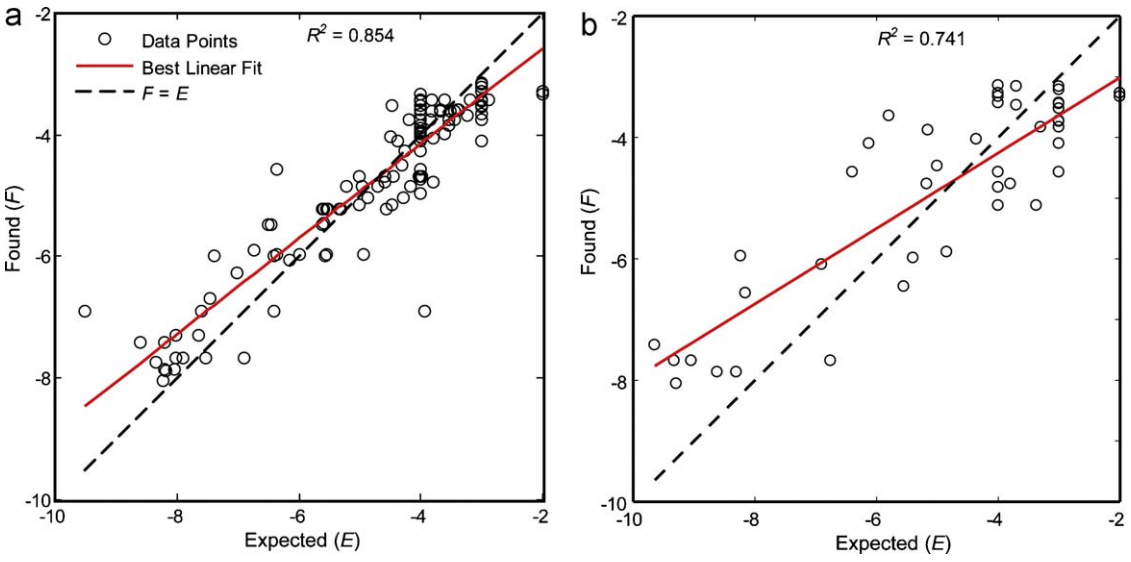


**Fig. 4.** Expected versus found values for log($IC_{50}$) for (a) the training set and (b) the test set.

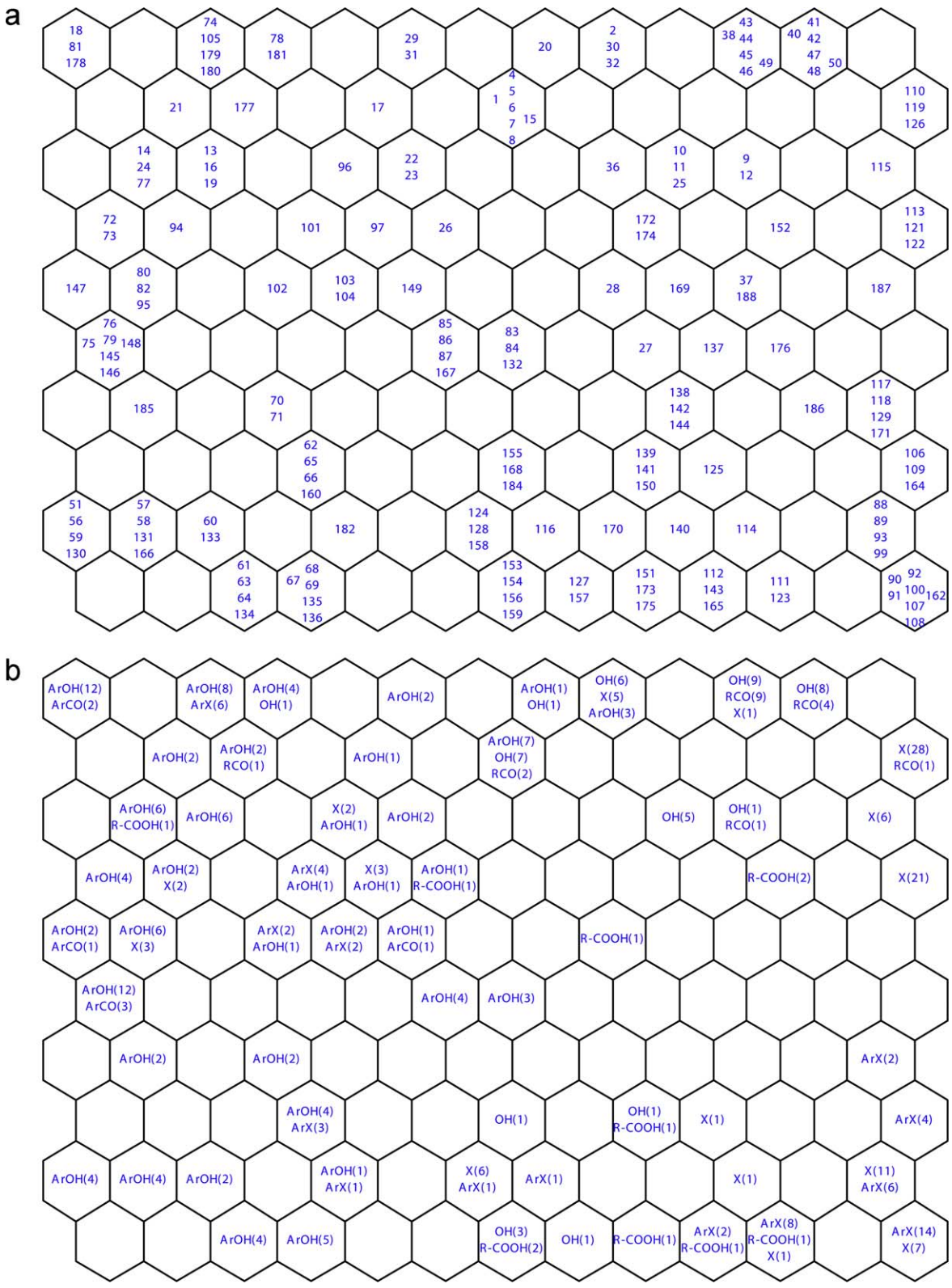


**Fig. 5.** Trained CPANN labeled with (a) labels for the structures taken from Table 1 and (b) the most frequently occurring groups in the data set.

factors, particularly the nature of ortho substituents that affect the OH accessibility.

Another interesting feature emerges when we compare the weight level that corresponds to nArOH descriptor (Fig. 3) with the unified distance matrix (UDM) for the CPANN (Fig. 6). As expected, the examination of the labels presented on the UDM, shows that this descriptor roughly separated the non-binding (NB) structures (in the lower and right part of the CPANN) from the weak (WB), moderate (MB) and strong (SB) binders to the ER (in the upper and upper left part of the network).

The UDM also shows that, as a result of the influence of the nArOH, the CPANN is divided into three regions. We already men-

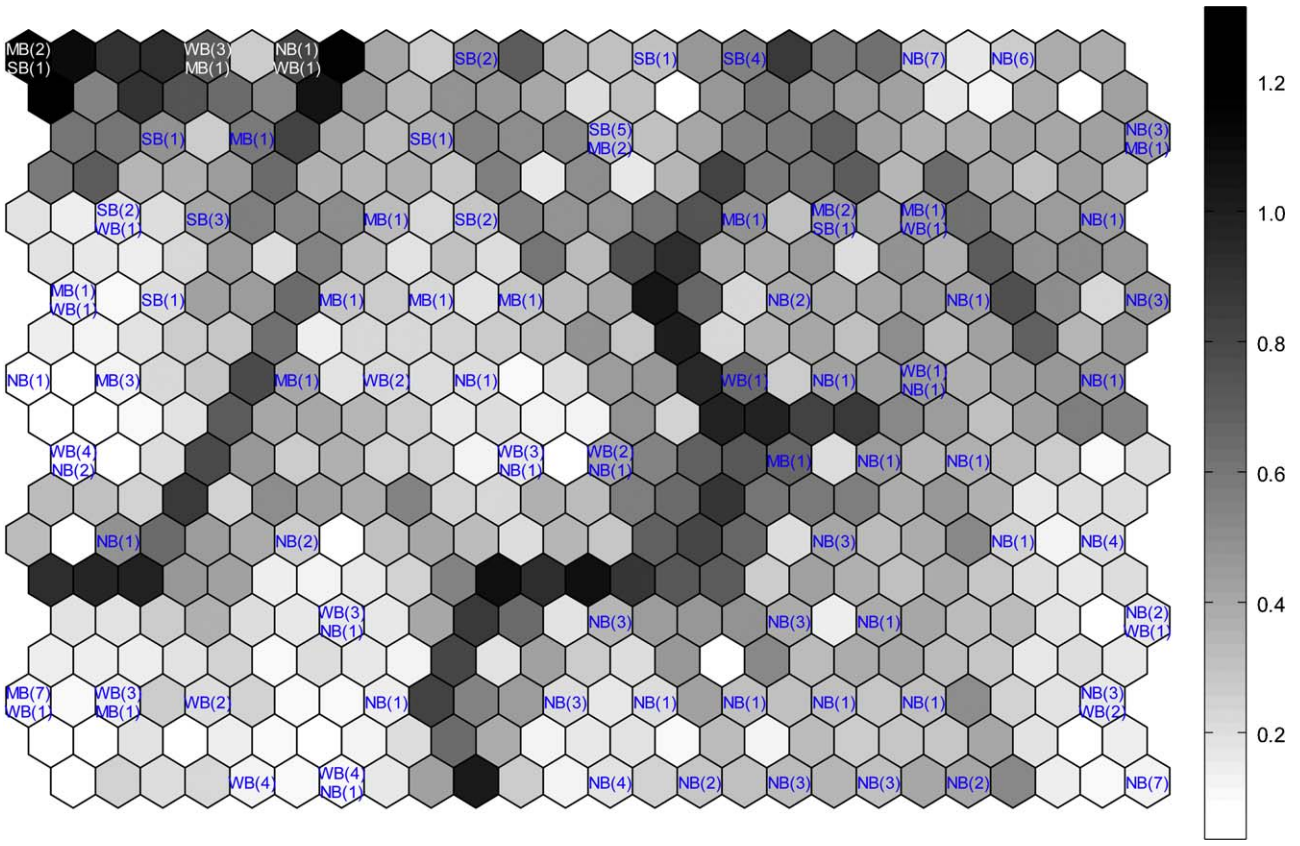


**Fig. 6.** Unified distance matrix for the discussed CPANN labeled with which type of binders the structures represent: SB, strong binders; MB, moderate binders; WB, weak binders; NB, nonbinding substances. On the UDM the distances between the neurons are represented with different colors (the larger the distance – the darker the color).

tioned that most of the NB substances are grouped in the largest region (on the lower and right part of the CPANN). The compounds mapped there do not possess any phenolic group. In the middle part of the CPANN, the structures with only one aromatic phenolic group are mapped (Fig. 6). The remaining structures, with more than one aromatic OH group are mapped in the upper left part. In the latter two regions (Figs. 3 and 6) the most of the strong binders are grouped in their upper part.

The most of the MB are surrounding the region of the CPANN that corresponds to small values of the log($IC_{50}$). The region predominantly occupied by weak binders is the one in the lower left part of the CPANN (Fig. 6), although several WB and MB could be found in the upper part of network.

Many of the effective ER ligands have a ring structure [53]. Although the flat aromatic phenolic ring is more important than other types of rings, the overall ring structure is of a definite importance since their construction increases the rigidity of both the structure and steric centre, which favors ER binding. In our model, the descriptor nCIC represents the importance of the presence of rings for the estrogenic activity.

Using nArOH with additional "assistance" from the descriptor nClC, which shows the second largest relative importance (17/32), the most of the strong binders are grouped in the upper central part of the CPANN (Figs. 3 and 5a – the weight level for log($IC_{50}$)).

Additionally, four of the five substances that have five rings (the largest number of rings in the data set) are grouped in the upper right part of the CPANN. These four substances are pesticides with condensed rings (and large number of chlorine atoms in the structure). These substances are: kepone, aldrin, dieldrin and mirex. The fifth structure with five rings is nafoxidine (label: 36) which belongs to the group of antagonists. Nafoxidine is a moderate binder and probably due to its size it is mapped in the lower left part of the region where the structures with the highest number of rings are grouped.

The volume of the ER binding pocket is about twice that of $E_2$ [54]. The length and breadth of strong binders should be well matched with the receptor, but at the active site of estrogenic receptors there are cavities that allow groups of certain size to fit. Thus we can say that, in this part of the CPANN, nC descriptor, which represents the number of C atoms, helps in better separation of the strongest binders (in the upper central part of the CPANN) with the larger substituents (like: estra-1,3,5(10)-trien-3-ol) or with large side chain (the structures have labels: 30, 32) from the rest of the strong binders.

The strong binders among the *synthetic estrogens* which posses structure of diethylstilbestrol (DES) are not influenced so much by nC and nClC as the strong binders with steroid skeleton. That is the reason why most of these structures (Fig. 5a) are grouped in the left part of the region of the CPANN that corresponds to low values of log($IC_{50}$) (Figs. 3 and 5).

The cavities of the binding pocket are of large importance for other xenoestrogens, including DES-like chemicals, diphenylmethanes, and diphenyls. Due to the size of the cavities, the binding affinities might be influenced by the size of the parts of the molecule that fit with cavities or by the size of substituents at the certain positions, rather than size of the whole molecule. The importance of nC is in accordance with above findings.

The non-binding compounds with steroid skeleton (labels: 38–50), as a result of the absence of ArOH groups, are mapped further right from the part of the CPANN that corresponds to the most active substances (Figs. 3 and 5a – weight level for log($IC_{50}$)). Actually, some of these molecules possess OH group in their structure, but the ring A in the steroid skeleton is not aromatic so that OH group does not have H-bond donor or acceptor abilities to bind with

receptor. Only three of these structures are weak binders (labels: 38–41), while the remaining structures are non-binding.

The nCb- descriptor (the relative importance of this descriptor is 12/32) also appears to be important for ER binding. Careful examination of weight levels that correspond to nCb- with the weight levels of log($IC_{50}$), shows that this descriptor is actually responsible for the grouping of the structures which do not posses aromatic rings in their structure in the lower central part of the CPANN and in the upper right part of the network (Figs. 3 and 5a – weight level for log($IC_{50}$)). The most of these structures (labeled as: 38, 40–50, 110, 113, 115, 116, 119, 121, 122, 124, 126–128, 150, 153, 154, 156–159) are non-binders. Only three of them are weak binders and they belong to the group which, on Table 1, is labeled as *Other miscellaneous steroids* (structures: 38, 40–50). The remaining structures belong to the groups of *Pesticides* and *Miscellaneous compounds*.

Unlike the structures which do not have benzene rings (the most of them are non-binders), the structures which have one or more benzene rings (in the most cases nCb- > 0) as a structural feature are very diverse both in their ER binding affinity and structurally. Among the 23 structures with five or more substituents on the benzene rings present in their molecules there are 4 strong binders (labeled as: 18, 21, 29 and 31), 6 moderate binders (labels: 36, 81, 101, 102, 178 and 179), 5 weak binders (labels: 28, 74, 105, 180 and 181) and the remaining 8 structures are non-binding (labels: 78, 107, 108, 123, 147, 162, 172 and 174). So, having this in mind we can conclude that, for the discussed model, this descriptor is the most responsible for the grouping of the structures that do not possess aromatic ring in the upper right and in the lower central part of the CPANN.

The descriptor with the smallest relative importance (1/32) is nR = Ct. This descriptor represents the number of aliphatic tertiary $sp^2$ hybridized C atoms. Different types of molecules have tertiary aliphatic $sp^2$ hybridized C atoms. The larger fraction of these molecules (10 of 27) have steroid skeleton where the carbonyl group is attached to the ring A in the skeleton. Due to a larger influence of the nCIC and nCb- descriptors (in comparison to nR = Ct) these substances are brought in this part of the CPANN mainly by these descriptors. However, in the left part of the CPANN that corresponds to large values of nR = Ct (Fig. 3) one can notice that this descriptor is responsible for grouping of the structures similar to diethylstilbestrol (labels: 16, 17 and 19) into the region of the CPANN that represents small values for the log($IC_{50}$). Another two strong binders (4-hydroxytamoxifen and 3-hydroxytamoxifen) which are mapped in the same region (with only one aliphatic tertiary $sp^2$ hybridized C atom) are also influenced by this descriptor.

Some other patterns will emerge if we try to examine mapping of the different types of compounds on the CPANN. For example, alkylphenolic compounds (in Fig. 5a these compounds have labels in the interval: 51–71) and parabenes (alkylhydroxybenzoat, labeled with: 130–136) are grouped in the lower left part of the CAPNN due to the similarities in the structure and in their estrogen activity. The activities of the phenols are largely dependent on the alkyl chain. In general, longer the side chain, the greater the binding affinity for the ER. There is however the limit to the side-chain carbons that increases binding to the ER since 4-dodecylphenol (12 carbons) exhibited lower relative binding affinity than 4-nonylphenol (9 carbons). Concerning parabens, the study of their activity is of particular interest due to their commercial importance. All of the parabens present in data set are showing moderate to weak binding affinity. Thus, due to their inherent estrogenicity and their wide range of application, it is apparent that the parabens pose a potential hazard as endocrine disruptors. The parabens, like the alkylphenolic compounds, demonstrated the dependence of binding affinity on the chain length [26].

In the region of the CPANN which corresponds to the compounds with more than one phenolic group (Figs. 5b and 6 – upper left part) together with the synthetic estrogens with DES-like structures, the most of the diphenyl derivatives (structures: 72–82) and benzophenone compounds (145–148) are grouped. The remaining structures of these two types of compounds, as a result of the presence of only one phenolic group, and consequently smaller estrogen binding affinity, are grouped in the central part of the CPANN (structures labeled as: 83–87 and 149). When both OH groups contribute to the binding, a rigid structure is critical for a better fit to the ER. When a substance contains only one phenolic ring, the binding depends on how well the rest of the structure fits into the binding pocket and the binding affinity is more favorable to a chemical with certain flexibility. That is why part of these structures are strong binders and the remaining structures are moderate or weak binders.

Pesticides are another important group suspected of being endocrine disruptor compounds [55]. Among the pesticides investigated in this work, dihydroxymethoxychlor olefin is only one strong binder to ER. The activity of the *organochlorines* (labels: 88–109) varies from moderate binders to non-binders and it strongly depends on their structure. The absence of the aromatic OH group in part of these structure (labels: 88–93, 100, 106–108) helps in grouping of all non-binding compounds and 2 weak binders in the lower right part of the CPANN (Fig. 5a and b). As expected, the presence of the phenolic group(s) in the structure makes the remaining organochlorines compounds, in most of the cases, moderate binders. The remaining pesticides (labels: 100–129) as weak binders are mapped in the left region of the CPANN.

The phthalates (137–144) present in the data set are non-binding and accordingly they are grouped in the lower right part of the CPANN. In addition, most of the substances labeled as other miscellaneous compounds (150–188) are non-binding and according to this, most of them are grouped in the lower right part of the CPANN. The two moderate binders present among these substances and 2 weak binders are grouped in the upper left part of the network.

## 3. Conclusion

Self-organizing maps and CPANN have been predominantly used as mapping algorithms in analysis of endogenous disruptors, serving for clustering or for classification of the data using large number of input variables [37]. CPANN have also been tested for establishing the model with the good predictive ability of receptor binding affinity of EEDCs [50]. Although the classical approach in using of these algorithms is very convenient, using our previously developed algorithm for automatic adjustment of the relative importance of the input variables [35] we developed not only a model with good generalization performance but also an interpretable model.

The mechanistic interpretation of the results, the interpretation on how the input variables are influencing the mapping and the model itself were simplified using our approach. The simplicity of the model enabled us to perform detailed data exploratory analysis of the results obtained using CPANN and to get an insight into the structural features influencing the activity of environmental chemicals on the estrogenic receptor. The examination showed that, as expected, the presence of the aromatic OH group is important for the estrogenic activity. The descriptor nC confirms the significance of the size of the molecule for fitting at certain position of ER, whilst descriptor nCb indicates the importance of substituted aromatic ring for ER binding. Above findings were very helpful in grouping of different types of structures of EEDCs and gave us better insight into mechanism of binding of these substances.

## References


[1] U.A. Boelsterli, Mechanistic Toxicology. The Molecular Basis of How Chemicals Disrupt Biological Targets, Taylor & Francis, New York, 2002.
[2] W.V. Welshons, K.A. Thayer, B.M. Judy, J.A. Taylor, E.M. Curran, F.S. vom Saal, Environ. Health Perspect. 8 (2003) 994–1006.
[3] S.P. Bradbury, R.L. Lipnick, Environ. Health Perspect. 87 (1990) 181–182.
[4] D.W. Singleton, S.A. Khan, Front. Biosci. 8 (2003) 110–118.
[5] B. Hileman, Chem. Eng. News 72 (1994) 19–23.
[6] B. Hileman, Chem. Eng. News 75 (1997) 24–25.
[7] T.H. Hutchinson, D.B. Pickford, Toxicology 181–182 (2002) 383–387.
[8] T. Colborn, F.S.V. Saal, A.M. Soto, Environ. Health Perspect. 101 (1993) 378–384.
[9] B.E. Erickson, Chem. Eng. News 87 (2009) 25–26.
[10] H. Fang, W. Tong, R. Perkins, A. Soto, N. Prechtl, D.M. Sheehan, Environ. Health Perspect. 108 (2000) 723–729.
[11] M. Nakai, Y. Tabira, D. Asai, Y. Yakabe, T. Shimyozu, M. Noguchi, M. Takatsuki, Y. Shimohigashi, Biochem. Biophys. Res. Commun. 254 (1999) 311–314.
[12] W. Tong, D.R. Lowis, R. Perkins, Y. Chen, W.J. Welsh, D.W. Goddette, T.W. Heritage, D.M. Sheehan, J. Chem. Inf. Comput. Sci. 38 (1998) 669–677.
[13] T.W. Shultz, M.T. Cronin, T.I. Netzeva, J. Mol. Struct. (Theochem) 622 (2003) 23–38.
[14] S.P. Bradbury, C.L. Russom, G.T. Ankley, T.W. Schultz, J.D. Walker, Environ. Toxicol. Chem. 22 (2003) 1789–1798.
[15] A.D.P. Worgan, J.C. Dearden, R. Edwards, T.I. Netzeva, M.T.D. Cronin, QSAR Comb. Sci. 22 (2003) 204–209.
[16] M.T.D. Cronin, J.D. Walker, J.S. Jaworska, M.H.I. Comber, C.D. Watts, A.P. Worth, Environ. Health Perspect. 111 (2003) 1376–1390.
[17] M.T.D. Cronin, T.W. Schultz, J. Mol. Struct. (Theochem) 622 (2003) 39–51.
[18] C. Zhao, E. Boriani, A. Chana, A. Roncaglioni, E. Benfenati, Chemosphere 73 (2008) 1701–1707.
[19] M. Novič, M. Vračko, Molecules 15 (2010) 1987–1999.
[20] REACH – in brief, European Commission, Enterprise & Industry Directorate General, February 2007.
[21] J. Ahlers, F. Stock, B. Werschkun, Environ. Sci. Pollut. Res. 15 (2008) 565–572.
[22] Guidance Document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models. OECD Environment, Health and Safety Publications. Series on Testing and Assessment No. 69. ENV/JM/MONO (2007)2, Paris.
[23] H. Fang, W. Tong, L.M. Shi, R. Blair, R. Perkins, W. Branham, B.S. Hass, Q. Xie, S.L. Dial, C.L. Moland, D.M. Sheehan, Chem. Res. Toxicol. 14 (2001) 280–294.
[24] C.L. Waller, T.I. Oprea, K. Chae, H.K. Park, K.S. Korach, S.C. Laws, T.E. Wiese, W.R. Kelce, L.E. Gray Jr., Chem. Res. Toxicol. 9 (1996) 1240–1248.
[25] W. Tong, R. Perkins, L. Xing, W.J. Welsh, D.M. Sheehan, Endocrinology 138 (1997) 4022–4025.
[26] A.G. Saliner, L. Amat, R. Carbo-Dorca, T.W. Schultz, M.T.D. Cronin, J. Chem. Inf. Comput. Sci. 43 (2003) 1166–1176.
[27] R.M. Blair, H. Fang, W.S. Branham, B.S. Hass, S.L. Dial, C.L. Moland, W. Tong, L. Shi, R. Perkins, D.M. Sheehan, Toxicol. Sci. 54 (2000) 138–153.
[28] R. Hecht-Nielsen, Proc. IEEE First Int. Conf. on Neural Networks, 1987, p. 19.
[29] R. Hecht-Nielsen, Appl. Opt. 26 (1987) 4979–4984.
[30] R. Hecht-Nielsen, Neural Netw. 1 (1988) 131–139.
[31] J. Dayhoff, Neural Network Architectures, an Introduction, Van Nostrand Reinhold, New York, 1990, p. 192.
[32] J. Zupan, M. Novič, I. Ruisánchez, Chemometr. Intell. Lab. Syst. 38 (1997) 1–23.
[33] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley, Weinheim, New York, 1999.
[34] I. Kuzmanovski, M. Novič, Chemometr. Intell. Lab. Syst. 90 (2008) 84–91.
[35] I. Kuzmanovski, M. Novič, M. Trpkovska, Anal. Chim. Acta 642 (2009) 142–147.
[36] T.W. Schultz, G.D. Sinks, M.T.D. Cronin, Environ. Toxicol. 17 (2002) 14–23.
[37] F. Marini, A. Roncaglioni, M. Novič, J. Chem. Inf. Model. 45 (2005) 1507–1519.
[38] L.M. Shi, H. Fang, W.D. Tong, S.L. Dial, C.I. Moland, J. Wu, R. Perkins, R.M. Blair, W.S. Branham, D.M. Sheehan, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.
[39] Talete srl, Dragon for Windows (Software for Molecular Descriptor Calculations), Version 5.4, http://www.talete.mi.it/, 2006.
[40] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[41] S. Eric, T. Solmajer, J. Zupan, M. Novic, M. Oblak, D. Agbaba, Il Farmaco 59 (2004) 389–395.
[42] T. Kohonen, Neural Netw. 1 (1988) 3–16.
[43] T. Kohonen, Self-organizing Maps, 3rd edition, Springer, Berlin, 2001.
[44] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. de Noord, Anal. Chem. 67 (1995) 4295–4301.
[45] R. Leardi, A. Lupiáñez Gonzáles, Chemometr. Intell. Lab. Syst. 41 (1998) 195–207.
[46] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci. 37 (1997) 306–310.
[47] B.M. Smith, P.J. Gemperline, Anal. Chim. Acta 423 (2000) 167–177.
[48] H. Handels, T. Rob, J. Kreusch, H.H. Wolff, S.J. Pöppl, Artif. Intell. Med. 16 (1999) 283–297.
[49] S.S. So, M. Karplus, J. Med. Chem. 39 (1996) 5246–5256.
[50] H. Yoshida, R. Leardi, K. Funatsu, K. Varmuza, Anal. Chim. Acta 446 (2001) 485–494.
[51] A.G. Saliner, L. Amat, R. Carbó-Dorca, T. Wayne Schultz, M.T.D. Cronin, J. Chem. Inf. Comput. Sci. 43 (2003) 1166–1176.
[52] H. Hong, W. Tong, H. Fang, L. Shi, Q. Xie, J. Wu, R. Perkins, J.D. Walker, W. Branham, D.M. Sheehan, Environ. Health Perspect. 110 (2002) 29–36.
[53] A.M. Brzozowski, A.C. Pike, Z. Dauter, R.E. Hubbard, T. Bonn, O. Engström, L. Öhman, G.L. Greene, J.A. Gustafsson, M. Carlquist, Nature 389 (1997) 753–758.
[54] H. Fang, W. Tong, W.J. Welsh, D.M. Sheehan, J. Mol. Struct. (Theochem) 622 (2003) 113–125.
[55] G. Lyons, Mixed messages: pesticides that confuse hormones. Pesticide Action Network UK, Briefing 2 (2000) 1–6.