Classification of Potential Endocrine Disrupters on the Basis of Molecular Structure Using a Nonlinear Modeling Method[†]

Alessandra Roncaglioni,[‡,§] Marjana Novič,*[‡] Marjan Vračko,[‡] and Emilio Benfenati[§]

Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, SLO-1000 Ljubljana,
Slovenia, and Institute Mario Negri, Milano, Italy

Received October 31, 2003

A methodology for the classification of endocrine disruption chemicals is proposed. It is based on a data set of 106 substances extracted from the list of 553 chemicals that were inspected by the European Union Commission for the scientific evidence of their endocrine disruption activity. The substances belong to different categories defined in the EU Commission report: (i) literature evidence for certainly active as endocrine disrupters, (ii) for potentially active, (iii) for less probable active − lacking evidence, and (iv) for certainty nonactive. 3D molecular coordinates were calculated using the AM1or the PM3 optimization method. From 3D coordinates an extensive set of molecular descriptors was calculated. The classification model based on the counterpropagation neural network was constructed and evaluated. This is the first time that the counterpropagation neural network is applied for the classification of compounds regarding their literature evidence for the endocrine disruption activity. The developed classification model is proposed as a tool for a preliminary assessment of potential endocrine disrupters, which would help the assessors to make the priority list for a large amount of chemicals that have to be tested with more expensive in vitro and in vivo methods.

INTRODUCTION

The variety of chemicals released every day in the environment has a large impact on human health and wildlife. Negative effects are encompassed in term toxicity, a broad definition of biochemical property.[1,2] There are different mechanisms of action and different consequences of toxic effects of a given chemical.[2−4] Recent studies showed that many toxic effects are grounded on the malfunctioning and disruption of the endocrine system.[5−8] Chemicals that interfere with the endocrine systems and may lead to adverse effects are recognized as endocrine disrupters (EDs). The most frequent recognition of EDs is their ability to bind to receptors, such as the estrogen and androgen receptors. Most in vitro and in vivo data available in the literature are derived from assays that measure estrogenic or, less frequently, androgen activity.

The number of chemicals is increasing rapidly which results in a greater need for alternative, computational, or in silico methods[9−14] for the prediction of their negative biological activity. The European Union Commission reported the candidate list of 553 man-made substances that are potential endocrine disrupters and 9 synthetic/natural hormones.[15] In this report the substances suspected to act as endocrine disrupters are studied. Numerous information about data available in the literature on several effects related to the endocrine disruption potency are extracted and grouped. A set of 146 compounds was chosen for a deeper investigation having considered two parameters as indicators of exposure probability: the production volume (High Production Volume compounds were selected) and/or their persistency. According to the White paper on strategy, for a further chemical policy High Production Volume chemicals are placed on the EU market in volume exceeding 1000 tones per year. The initial set was further pruned to the final 106 compounds, which were retained for the modeling purpose; for the rest of the compounds it was not possible to calculate the structural descriptors needed for handling the chemical structures. The substances had been categorized into 3 stages of their endocrine disruption potency. It has to be stressed here that the categorization was based on the literature evidence for the endocrine disruption activity, not on the compounds' actual potency. Accordingly, the individual category gives information about the probability of the compound's endocrine disruption potency rather than the potency itself. Chemicals in the first category were confirmed to be endocrine disrupters in an intact organism by at least one study found in the literature. The second category characterizes substances that are potentially active according to in vitro studies, while the in vivo data do not sufficiently prove the ED activity. For the third category there was either no data available or data found for no scientific basis for inclusion into the list. An additional 244 substances were studied; however, the data from the literature about their ED activity was less extensive or convincing.

From the literature sources provided by the report[15] it is obvious that there is a lack of homogeneity in these data. The studies refer to experiments made during tens of years, evaluating several species and a great variety of effects. It means that it is not possible to select a specific feature to be

* Corresponding author phone: 386-1-4760-200; fax: 386-1-7460-300; e-mail: marjana.novic@ki.si.

[†] Dedicated to Professor Bill Milne, our good friend, who always encouraged the scientific development of our research group and has resigned from a long-term editorship of the *Journal of Chemical Information and Computer Sciences.*

[‡] National Institute of Chemistry.

[§] Institute Mario Negri.

10.1021/ci030421a CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/18/2004

Table 1. Categories of Substances, Suspected Endocrine Disrupting Chemicals, Studied by the EU Commission

category	labeling	description
1	endocrine disrupter	At least one study was found providing evidence of endocrine disruption in an intact organism—not a formal weight of evidence approach.
2	potential endocrine disrupter	in vitro data indicating potential for endocrine disruption in intact organisms; also includes effects in vivo that may, or may not, be ED-mediated; may include structural analyses and metabolic considerations
3	nonendocrine disrupter	There is no scientific basis for inclusion in the list of endocrine disrupters.
3A	no certain evidence for non-ED	No data are available on wildlife relevant and/or mammal relevant endocrine effects.
3B	some evidence for non-ED	Some data are available, but the evidence is insufficient for identification.
3C	certain evidence for non-ED	Data are available indicating no scientific basis for inclusion into the list of active ED chemicals.

analyzed with the chemometrics techniques, i.e., to identify a numerical output to be modeled. Consequently, we decided to start with the classification rather than modeling a certain biological endpoint. The classification model developed in the present study on the basis of the counterpropagation neural network is aimed to automatically determine to which class, defined as the degree of evidence for the endocrine disruption activity, a compound belongs. The considered endpoint, i.e., degree of evidence for endocrine disrupting/nondisrupting potency, is a weakly defined quantity. To be able to make the predictions more rigorous, which would satisfy the QSAR Development Principles,[16] more precisely defined endpoints should be included in the modeling. On the other hand, the proposed technique grounds on the concept of similarity in the descriptors' space. If a new compound is presented to the model, it will excite the neurons (model's construction units) containing structurally similar compounds from the training set. Thus the predicted class will be closely related to the classes of structurally similar compounds. This transparency in prediction and the interpretation of individual descriptors which remain unaltered during the modeling procedure are the main advantages of the counterpropagation neural network models. With appropriate data sets the proposed methods satisfy the Setubal principles.[16]

DATA

The initial pool of substances taken from the report of the EU Commission[15] was 553 man-made chemicals suspected to act as endocrine disrupters. Different amounts of information were found in the literature for individual chemicals. The substances were grouped by the EU Commission according to the data available in the literature on several effects related to the endocrine disruption potency. The three categories are described in Table 1.

A subset of 146 compounds was chosen and investigated deeply. First, High Production Volume compounds exceeding 1000 tones per year on the EU market, and second, compounds with high persistency (only for those compounds for which this parameter was reachable) were selected. For this set of compounds a categorization into 3 categories (reported in Table 1) was given on their literature evidence[15] for their endocrine disruption potency. From the subset of 146 compounds, 39 of them were additionally excluded because they were not uniquely defined compounds, such as polymers, salts, mixtures of isomers, or of different substances. These compounds do not satisfy the requirements for applying chemometrics techniques, and it was not possible to obtain structural descriptors. All the remaining 107 structures (see Table 2) were carefully designed and then optimized with the MOPAC program (MOPAC 93 for Windows IMB-PC compatible), using the AM1 or the PM3 semiempirical method.

For a small group of compounds, denoted by a footnote label *a* in Table 2, we encountered problems in the structure optimization procedure. The most important limit of the software is that it was not possible to process this calculation for tin (Sn) compounds with the AM1 method. However, this is a very important category of compounds because their endocrine activity is well-known. The studies focused on aromatic compounds and their interaction with receptors usually do not include the tin compounds because the tin compounds do not act via the same receptor binding mechanism. In our study we decided to include the tin compounds, applying the PM3 semiempirical method, which provides parametrization also for Sn.

In Table 2 the compound 523 is labeled as uncertain, because the structure optimization process was not successfully completed, so it was excluded from further processing. The final data set consisted of 106 compounds.

METHODS

Descriptors. To calculate descriptors of molecular structures, the following methods were applied:

- 3D structure optimization (the AM1 and the PM3 semiempirical methods for the minimization of total molecular energy) to obtain atom coordinates.
- From CODESSA[17] five classes of structural descriptors were obtained: constitutional, geometrical, topological, electrostatic, and quantum-chemical.
- LogP of all 106 compounds was either obtained from the experimental values database[18] or from the Hansch's manual[19] or estimated by the KowWin program.[20]

Classification Model. The counterpropagation neural network was employed as a classification model. We used the software developed in our home laboratory, written in FORTRAN for IBM-compatible PCs and a Windows operating system. The description of the network can be found in the literature.[21−23] Only a brief survey of the method will be given here with an emphasis on the modification which provides the possibility to predict discrete classes of compounds. The counterpropagation neural network is based on a supervised learning method; only one part of the learning process (initial mapping of inputs) involves elements of the unsupervised learning. For the learning procedure a set of input-target pairs $\{\mathbf{X}_s,\mathbf{T}_s\}$ is required. In the classification problem the input $\mathbf{X}_s = (x_{s1}, x_{s2}, ...x_{si} ...x_{sm})$ is a structure representation of the *s*th compound represented by *m* structural descriptors or "variables". The corresponding target $\mathbf{T}_s = (t_{s1}, t_{s2}, ...t_{sj} ...t_{sp})$ is a *p*-component vector of zeros and

Table 2. List of Compounds Optimized with MOPAC, Using the AM1 or the PM3 Semiempirical Method[a]

no.	class	label	CAS no.	name
2	**2**	**P**	10605-21-7	carbendazim
10	**2**	**P**	309-00-2	aldrin
11	**1**	**E**	12789-03-6 (57-74-9)	chlordane
13	**3 B**	**U**	3734-48-3	chlordene
15	**2**	**P**	60-57-1	dieldrin
16	**2**	**P**	115-29-7 (959-98-8 or 33213-65-9)	endosulfan (also alfa and beta)
19	**2**	**P**	72-20-8	endrin
20	**1**	**E**	143-50-0	kepone (chlordecone)
21	**1**	**E**	2385-85-5	mirex
22	**2**	**P**	27304-13-8	oxychlordane
25	**3 B**	**U**	39765-80-5	trans-nonachlor
27	**2**	**P**	94-75-7	2,4-dichlorophenoxy acetic acid (2,4-D)
29	**2**	**P**	67747-09-5	prochloraz
42	**1**	**E**	50-29-3	DDT (technical) = clofenotane = p,p′-DDT
44	**2**	**P**	115-32-2	dicofol = kelthane
57	**1**	**E**	3563-45-9	tetrachloro DDT = 1,1,1,2-tetrachloro-2,2-bis(4-chlorophenyl)ethane
60	**2**	**P**	36734-19-7	iprodione
63	**1**	**E**	50471-44-8	vinclozolin
73	**1**	**E**	137-26-8	thiram
78	**1**	**E**	58-89-9	gamma-HCH (lindane)
85	**2**	**P**	330-54-1	diuron
87	**1**	**E**	330-55-2	linuron (lorox)
104	**2**	**P**	333-41-5	diazinon
106	**2**	**P**	60-51-5	dimethoate
109	**3 C**	**N**	55-38-9	fenthion
113	**2**	**P**	121-75-5	malathion
115	**2**	**P**	298-00-0	methylparathion
119	**2**	**P**	56-38-2	parathion = parathion(-ethyl)
141	**1**	**E**	61-82-5	amitrol = aminotriazol
142	**1**	**E**	1912-24-9	atrazine
156	**2**	**P**	122-34-9	simazine
159	**2**	**P**	43121-43-3	triadimefon
163	**1**	**E**	34256-82-1	acetochlor
164	**1**	**E**	15972-60-8	alachlor
169	**3 A**	**U**	106-93-4	dibromoethane (EDB)
176	**2**	**P**	76-44-8	heptachlor
177	**3 B**	**U**	1024-57-3	heptachlor-epoxide
179	**2**	**P**	74-83-9	methylbromide (bromomethane)
182	**1**	**E**	1836-75-5	nitrofen
183	**3 B**	**U**	4685-14-7	paraquat = 1,1′-dimethyl-4,4′-bipyridinium
187	**2**	**P**	709-98-8	propanil
190	**3 A**	**U**	29082-74-4	octachlorostyrene
191	**1**	**E**	100-42-5	styrene
194	**2**	**P**	120-83-2	2,4 dichlorophenol
195	**2**	**P**	1570-64-5	4-chloro-2-methylphenol
196	**2**	**P**	59-50-7	4-chloro-3-methylphenol
198	**1**	**E**	118-74-1	hexachlorobenzene (HCB)
215	**2**	**P**	98-54-4	4-*tert*-butylphenol
216	**1**	**E**	140-66-9	4-tert-octylphenol=1,1,3,3-tetramethyl-4-butylphenol
277	**3 B**	**U**	103-23-1	bis(2-ethylhexyl)adipate
278	**1**	**E**	85-68-7	butylbenzylphthalate (BBP)
279	**1**	**E**	117-81-7	di-(2-ethylhexyl)phthalate (DEHP)
280	**3 B**	**U**	84-61-7	dicyclohexyl phthalate (DCHP)
281	**3 B**	**U**	84-66-2	diethyl phthalate (DEP)
283	**2**	**P**	26761-40-0	diisodecyl phthalate
284	**2**	**P**	28553-12-0	diisononyl phthalate = 1,2-benzenedicarboxylic acid, diisononyl ester (DINP)
286	**1**	**E**	84-74-2	di-*n*-butylphthalate (DBP)
318	**2**	**P**	1675-54-3	2,2′-bis(4-(2,3-epoxypropoxy)phenyl)propane = 2,2′-[(1-methylethylidene)-bis(4,1-phenyleneoxymethylene)]bisoxirane
326	**1**	**E**	80-05-7	2,2-bis(4-hydroxyphenyl)propan = 4,4′-isopropylidenediphenol = bisphenol A
348	**3 A**	**U**	106-89-8	epichlorohydrin (1-chloro-2,3-epoxypropane)
370	**3 B**	**U**	92-52-4	diphenyl
371	**2**	**P**	90-43-7	*o*-phenylphenol
405	**3 B**	**U**	38380-07-3	PCB 128 (2,2′,3,3′,4,4′-hexachlorobiphenyl)
406	**2**	**P**	38411-22-2	PCB 136 (2,2′,3,3′,6,6′-hexachlorobiphenyl)
408	**1**	**E**	35065-27-1	PCB 153 (2,2′,4,4′,5,5′-hexachlorobiphenyl)
409	**2**	**P**	38380-08-4	PCB 156 (2,3,3′,4,4′,5-hexachlorobiphenyl)
410	**1**	**E**	32774-16-6	PCB 169 (3,3′,4,4′,5,5′-hexachlorobiphenyl)
417	**1**	**E**	2437-79-8	PCB 47 (2,2′,4,4′-tetrachlorobiphenyl)
418	**2**	**P**	70362-47-9	PCB 48 (2,2′,4,5-tetrachlorobiphenyl)
419	**3 A**	**U**	35693-99-3	PCB 52 (2,2′,5,5′-tetrachlorobiphenyl)

Table 2. (Continued)

no.	class	label	CAS no.	name
420	**2**	**P**	33284-53-6	PCB 61 (2,3,4,5-tetrachlorobiphenyl)
421	**2**	**P**	32598-12-2	PCB 75 (2,4,4′,6-tetrachlorobiphenyl)
422	**1**	**E**	32598-13-3	PCB 77 (3,3′,4,4′-tetrachlorobiphenyl)
435	**2**	**P**	No CAS 046	2,2′,4,4′-tetrabrominated diphenyl ether (2,2′,4,4′-tetraBDE)
436	**2**	**P**	No CAS 044	decabrominated diphenyl ether (decaBDE)
444	**3 B**	**U**	135-19-3	2-naphthol
467	**1**	**E**	40321-76-4	1,2,3,7,8-pentachlorodibenzodioxin
472	**1**	**E**	1746-01-6	2,3,7,8-tetrachlorodibenzo-p-dioxin (2,3,7,8-TCDD)
483	**2**	**P**	57117-41-6	1,2,3,7,8-pentachlorodibenzofuran
484	**2**	**P**	83704-53-4	1,2,3,7,9-pentachlorodibenzofuran
485	**2**	**P**	58802-20-3	1,2,7,8-tetrachlorodibenzofuran
486	**2**	**P**	71998-72-6	1,3,6,8-tetrachlorodibenzofuran
487	**1**	**E**	57117-31-4	2,3,4,7,8-pentachlorodibenzofuran (2,3,4,7,8-PeCDF)
488	**2**	**P**	67733-57-7	2,3,7,8-tetrabromodibenzofuran
489	**2**	**P**	51207-31-9	2,3,7,8-tetrachlorodibenzofuran
512[a]	**1**	**E**	688-73-3	tributyltin hydride
513[a]	**1**	**E**	56-35-9	tributyltin oxide = bis(tributyltin) oxide
516[a]	**1**	**E**	4342-30-7	phenol, 2-[[(tributylstannyl)oxy]carbonyl]
517[a]	**1**	**E**	4342-36-3	stannane, (benzoyloxy)tributyl-
518[a]	**1**	**E**	4782-29-0	stannane, [1,2-phenylenebis(carbonyloxy)]bis(tributyl-
521[a]	**1**	**E**	24124-25-2	stannane, tributyl[(1-oxo-9,12-octadecadienyl)oxy]-
522[a]	**1**	**E**	3090-35-5	stannane, tributyl[(1-oxo-9-octadecenyl)oxy]-
523[a,b]	**1**	**E**	26239-64-5	stannane, tributyltin abietate
524[a]	**1**	**E**	1983-10-4	stannane, tributylfluoro-
525[a]	**1**	**E**	2155-70-6	tributyl[(2-methyl-1-oxo-2-propenyl)oxy]stannane
530[a]	**1**	**E**	1461-25-2	tetrabutyltin (TTBT)
531[a]	**1**	**E**	668-34-8	triphenyltin
532[a]	**1**	**E**	900-95-8	fentin acetate = triphenyltin acetate
536	**1**	**E**	95-76-1	3,4-dichloroaniline
538	**1**	**E**	99-99-0	4-nitrotoluene
541	**3 A**	**U**	119-61-9	benzophenone
545	**3 A**	**U**	68-12-2	dimethylformamide (DMFA)
548	**3 C**	**N**	107-21-1	ethylene glycol (ethane-1,2-diol)
557	**2**	**P**	127-18-4	perchloroethylene
558	**3 C**	**N**	108-95-2	phenol
560	**1**	**E**	108-46-3	resorcinol
564	**3 B**	**U**	108-05-4	vinyl acetate

[a] Only 13 compounds were optimized by the PM3 semiempirical method because they contain Sn atoms. The enumeration is taken from the complete set of compounds reported by the EU Commission.[15] [b] The optimized structure was judged unreliable by the software and thus removed from the final data set.

ones. The value $t_{sj} = 1$ or $t_{sj} = 0$ indicates that the sth compound is or is not in the jth class, respectively. The neural network is trained to respond for each input structure representation $\mathbf{X}_s$ from the training set with the output vector $\mathbf{Out}_s$ identical to the target (class vector) $\mathbf{T}_s$. A schematic representation of the counterpropagation neural network architecture with neurons as construction units is shown in Figure 1. The input layer (also called the Kohonen layer) of the counterpropagation neural network consists of $n_x \times n_y$ neurons. The molecules represented as sets of descriptors are placed into the input layer in such a way that clusters of molecules are formed in the top-map.[23]

The clusters are formed in a nonsupervised manner. The structural similarity of individual molecules in the same cluster is controlled and directed by choosing certain descriptors in a reduced set of molecular descriptors introduced into the model. For this step no knowledge about the target vector is needed. Once the position of the input vector is defined, the weights of the neurons in both input and output layers are corrected according to the particular element from the training set, $\{\mathbf{X}_s,\mathbf{T}_s\}$ pair (training object). The trained output layer consists of $n_x \times n_y$ output neurons arranged in a squared neighborhood. The levels of the output layer represent p response surfaces for the p classes. The points of the response surfaces correspond to the weights of the output neurons $\mathbf{Out} = (out_1, out_2, ...out_j ...out_p)$. After the training, each weight out_j is a real number between 0.0 and 1.0. For the final prediction of classes the response surface values must be again transformed into discrete values, zeros and ones. The threshold value between 0.01 and 0.99 must be determined for each of the p classes. Below the threshold all predictions are negative and denoted by a zero, which means that the sth compound does not belong to the jth class, while the predictions above the threshold are positive and denoted by one. The threshold is determined according to the number of correct/wrong class predictions if the trained network is tested by the same objects as it was trained with, i.e., $\{\mathbf{X}_s,\mathbf{T}_s\}$ pairs from the training set.

RESULTS AND DISCUSSION

Set of Descriptors. The data set used to develop and test the classification model contains 106 compounds ($N_{\mathrm{mol}} = 106$). The molecular structures were described by constitutional, topological, geometrical, electrostatic, and quantum-chemical descriptors calculated with CODESSA. Altogether 766 descriptors were provided by this software; however, 484 of them were available only for a limited number of molecules (so-called incomplete descriptors), while 16 descriptors were equal for all molecules and thus neglected. We decided to keep only the remaining 266 descriptors of each molecular structure ($m = 266$), which were descriptive and available for all compounds. An additional 389 incom-

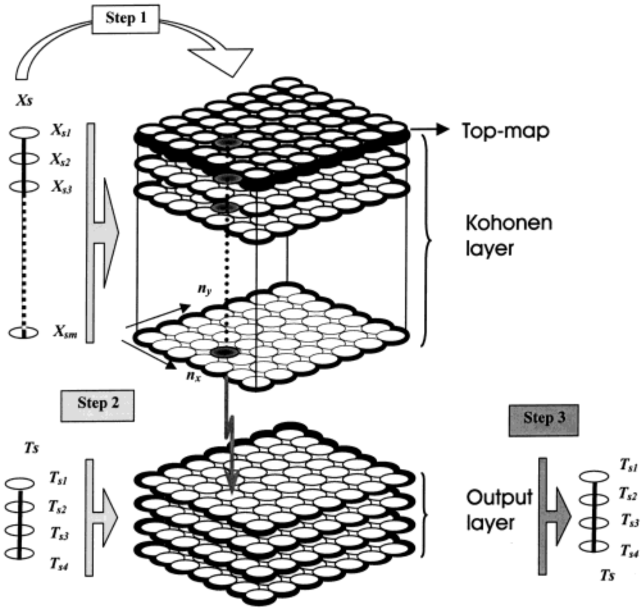


Figure 1. Counterpropagation neural network architecture. The circles represent the weights of individual neurons in the Kohonen and Output layer as well as the components of the molecular structure representation vector and target vector shown vertically on the left side of the plot. Step 1, mapping of the molecule $\mathbf{X}_s$ into the Kohonen layer; Step 2, correction of the weights in both, the Kohonen and the Output layer; Step 3, prediction of the four-dimensional target $\mathbf{T}_s$. The position of the molecule after Step 1 ($n_x = 4$, $n_y = 1$) is visualized by a tiny dashed vertical line and gray circles within the Kohonen layer and in the top-map.

Table 3. ED Categories Associated with 106 Compounds from the Data Set

category	label and description	no. of compounds
1	E (evidently active)	43
2	P (potentially active)	43
3A and 3B	U (uncertain evidence)	17
3C	N (nonactive)	3

plete descriptors available for at least 3 molecules were included into an extended set to make a comparative classification study.

Besides the descriptors calculated by CODESSA from the molecular 3D coordinates, an experimentally obtained parameter LogP was added as an additional descriptor, which reflects the compounds' property usually playing an important role in the mechanism of action of particular biological activity.[24] LogP, the logarithm of the octanol−water partition coefficient, describes the equilibrium partitioning of a chemical between the octanol and the water phases. The experimental values for a great part of the chemicals were available from the literature.[7] A few values were added from Hansch's book,[19] while an estimation method with the KowWin program[20] was used when experimental values were not available from the literature. All descriptors were auto scaled (i.e. normalized with mean = 0 and standard deviation = 1).

Distribution of Compounds between Categories. The ED categories associated with the 106 compounds from the data set are demonstrated in Table 3; for the purposes of our study, categories 3A and 3B were joined into one class, while 3C was taken as the fourth class ($p = 4$). We decided to split the third category into two classes because the uncertain evidence of 3A and 3B is not strong enough for such an important decision, which our predictive model is trained for, that would classify a chemical to be harmless regarding the endocrine disrupting activity. Only for the category 3C there is no doubt about nonactivity. In Table 3 the categories, associated classes, labels, and number of compounds in an individual class are given.

See Table 1 for a detailed description of individual categories. The risk for a chemical to be an endocrine disrupter decreases from the first toward the fourth class.

Training−Test Set Division. We split our data into the training and the test set. The data are not ideally distributed between the classes. It was obligatory to make the selection in a way which would keep a constant amount (approximately two-thirds) of compounds in each particular class for training, while one-third was for testing and validating the constructed classification model. We performed our selection on the basis of the Kohonen neural network.[23,25] The Kohonen neural network of dimension 5 × 5 was applied, which enables one to map objects into 25 positions. Similar objects were mapped into the same position (x, y coordinate in a Kohonen map). Only one part of a representative object from each position in the Kohonen map was chosen for the training set, respecting the original proportion among the different classes and the predefined 2:1 ratio between the training and the test objects. The rest were put into the test set. The following 71 compounds were assigned to the training set: 11, 13, 15, 16, 21, 22, 27, 44, 57, 60, 73, 78, 87, 104, 106, 109, 115, 119, 142, 156, 159, 164, 176, 182, 187, 190, 191, 194, 196, 198, 215, 216, 277, 278, 279, 280, 281, 284, 318, 370, 371, 405, 409, 410, 418, 420, 422, 435, 436, 444, 472, 483, 486, 487, 488, 512, 513, 517, 521, 524, 525, 531, 532, 536, 538, 541, 545, 557, 558, 560, 564. The rest of the 35 compounds were assigned to the test set: 2, 10, 19, 20, 25, 29, 42, 63, 85, 113, 141, 163, 169, 177, 179, 183, 195, 283, 286, 326, 348, 406, 408, 417, 419, 421, 467, 484, 485, 489, 516, 518, 522, 530, 548.

Selection of the Descriptors. The method for selection of the descriptors used in this study was a modified method for selecting objects for the training and the test set described in the previous paragraph. The main difference is in the way the matrix of input data is represented. The transposed data matrix is used instead of the original data matrix. Originally, the rows and columns ($N_{\mathrm{mol}} = 106$ rows and $m = 266$ columns) correspond to molecules and descriptors, respectively. Each molecule is represented as a vector of m components, where m is equal to the number of descriptors. In the transposed matrix, the descriptors are stored in rows, which means that each row represents an N_{mol}-dimensional vector of one of the m descriptors. For example, the first row contains N_{mol} values of the first descriptor with the vector components corresponding to the values of this first descriptor in each of the N_{mol} molecules. The transposed matrix is normalized by columns and then introduced to the Kohonen network which is trained until a limiting error is reached. Instead of mapping **molecules**, we make a map of **descriptors** that are placed onto the $n_x \times n_y$ positions (neurons) of the Kohonen map. We used a small Kohonen network with $5 \times 5 = 25$ neurons producing a map with 25 positions. All 266 descriptors were placed onto these 25 positions (neurons). This means that each neuron was occupied on average by 11 descriptors. In Figure 2 it is demonstrated how the descriptors were distributed.

a

Nx Ny	1	2	3	4	5
1	*12*	*10*	*4*	*9*	*11*
2	*11*	*9*	*3*	*10*	*6*
3	*4*	*8*	*14*	*3*	*6*
4	*22*	*20*	*3*	*8*	*10*
5	*10*	*7*	*4*	*25*	*37*

b

Nx Ny	1	2	3
1	**No. of occupied electronic levels / # of atoms** PNSA-1 Partial negative surface area [Zefirov's PC] FNSA-1 Fractional PNSA (PNSA-1/TMSA) [Zefirov's PC] Tot molecular 1-center E-E repulsion / # of atoms Number of Cl atoms Relative number of Cl atoms Relative molecular weight DPSA-3 Difference in CPSAs (PPSA3-PNSA3) [Zefirov's PC] WNSA-1 Weighted PNSA (PNSA1*TMSA/1000) [Zefirov's PC] Relative number of rings Min partial charge for a C atom [Zefirov's PC] **Number of rings**	*WNSA-3 Weighted PNSA (PNSA3*TMSA/1000) [Semi-MO PC]* PNSA-3 Atomic charge weighted PNSA [Semi-MO PC] PNSA-2 Total charge weighted PNSA [Semi-MO PC] Min net atomic charge FNSA-2 Fractional PNSA (PNSA-2/TMSA) [Semi-MO PC] WNSA-2 Weighted PNSA (PNSA2*TMSA/1000) [Semi-MO PC] FNSA-3 Fractional PNSA (PNSA-3/TMSA) [Semi-MO PC] Final heat of formation Final heat of formation / # of atoms **Min partial charge (Qmin)**	**Number of Br atoms** Relative number of Br atoms Avg bond order of a C atom **Min (>0.1) bond order of a C atom**
2	**Min e-n attraction for a C atom** Min e-e repulsion for a C atom Tot molecular 2-center exchange energy Tot molecular 2-center resonance energy Molecular volume / XYZ Box Min atomic orbital electronic population YZ Shadow / YZ Rectangle Min net atomic charge for a C atom XY Shadow / XY Rectangle Min valency of a C atom **ZX Shadow / ZX Rectangle**	**Relative number of aromatic bonds** Number of aromatic bonds Relative number of benzene rings Number of benzene rings Relative number of C atoms HOMO-1 energy Tot molecular electrostatic interaction / # of atoms Max bonding contribution of a MO **HOMO energy**	**Max 1-electron react. index for a C atom** Max electroph. react. index for a C atom **Max SIGMA-PI bond order**

c

Nx Ny	1	2	3	4	5
1	No. of occupied electronic levels / # of atoms	**WNSA-3 Weighted PNSA (PNSA3*TMSA /1000) [Semi-MO PC]**	Number of Br atoms	Avg electroph. react. index for a C atom	***Principal moment of inertia B / # of atoms***
	Number of rings	**Min partial charge (Qmin)**	Min (>0.1) bond order of a C atom	RNCG Relative negative charge QMNEG/QTMINUSZefirov's PC	***RNCG Relative negative charge (QMNEG/QTMINUSSemi-MO PC***
2	Min e-n attraction for a C atom	Relative number of aromatic bonds	Max 1-electron react. index for a C atom	FNSA-2 Fractional PNSA (PNSA-2/TMSA) [Zefirov's PC]	***Tot molecular 2-center exchange energy / # of atoms***
	ZX Shadow / ZX Rectangle	HOMO energy	Max SIGMA-PI bond order	Tot molecular 1-center E-N attraction	***LUMO energy***
3	Max atomic state energy for a C atom	**Average Structural Information content (order 1)**	DPSA-3 Difference in CPSAs (PPSA3-PNSA3) [Semi-MO PC]	1X BETA polarizability (DIP)	Polarity parameter (Qmax-Qmin)
	Max atomic orbital electronic population	**Average Bonding Information content (order 0)**	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Semi-MO PC]	Avg 1-electron react. index for a C atom	Polarity parameter / square distance
4	HASA-2/SQRT(TMSA) [Zefirov's PC]	***HA dependent HDSA-2/SQRT(TMSA) [Semi-MO PC]***	***PNSA-1 Partial negative surface area [Semi-MO PC]***	Max PI-PI bond order	***Average Complementary Information content (order 0)***
	HACA-1/TMSA [Zefirov's PC]	***Relative number of N atoms***	***FNSA-1 Fractional PNSA (PNSA-1/TMSA) [Semi-MO PC]***	Max nucleoph. react. index for a C atom	***Min 1-electron react. index for a C atom***
5	***HA dependent HDCA-1/TMSA [Semi-MO PC]***	HBCA H-bonding charged surface area [Semi-MO PC]	count of H-donors sites [Zefirov's PC]	Molecular volume	***PPSA-2 Total charge weighted PPSA [Zefirov's PC]***
	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]	***HACA H-acceptors charged surface area [Semi-MO PC]***	Max e-e repulsion for a C atom	Randic index (order 3)	***Max SIGMA-SIGMA bond order***

Figure 2. The distributions of descriptors in the 5 × 5 top-map of the Kohonen neural network: (a) the number of descriptors occupying an individual neuron; (b) the top left section of the top-map with a list of descriptors on the neurons shown; and (c) two descriptors from each neuron chosen on the basis of the smallest and the largest distance between the neuron and the descriptor's vector.

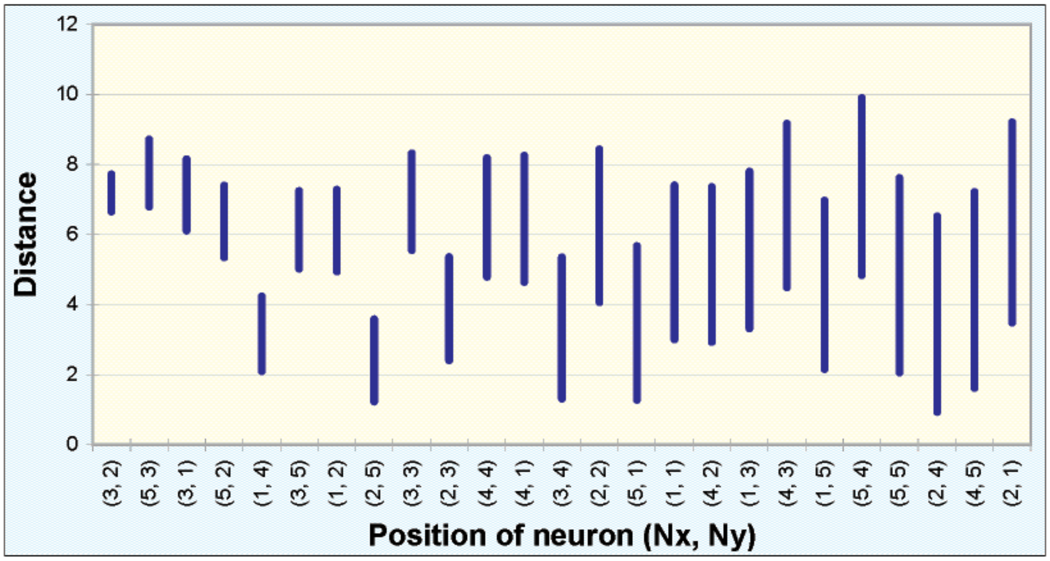


Figure 3. The distances (eq 1) between the neuron at the position (n_x, n_y) and the descriptors' vectors clustered on a particular neuron. The bars represent the differences between the minimal and maximal distances within individual clusters. The positions on the abscissa are ordered by increasing differences (bar's size).

The neurons and descriptors are labeled by the indices *i* and *s*, respectively. In the procedure for selecting descriptors, the dimension of neurons is equal to the number of molecules in the data set (N_{mol} =106); this is also the number of components of the descriptors' representation vectors obtained as rows in the transposed data matrix (X^T_{js}, j=1, N_{mol}). The *i*th neuron is represented as a vector of weights (W_{ji}, j=1, N_{mol}). In the training procedure of the Kohonen neural network, similar descriptors are falling onto the neighboring neurons. If the number of training objects significantly exceeds the number of neurons, many objects occupy the same neuron. The criterion for the selection of descriptors assembled on the same neuron was the Euclidean distance between the descriptor $\mathbf{X}_s^{\mathbf{T}}$ and neuron $\mathbf{W}_i$:

$$d_{s,i} = \sqrt{\sum_{j=1}^{N_{mol}} (X_{j,s}{}^{T} - W_{j,i})^2} \qquad (1)$$

Only two descriptors from each neuron were chosen for the final representation of the molecular structure, the one with the smallest and the one with the largest distance from the excited neuron. See Figure 3 for distances within clusters at all 25 neurons.

The network with dimensions 5 × 5 × 106 was trained for 50, 100, 300, 500, and 1000 epochs. The distribution of objects (descriptors) in the 5 × 5 top-map and the distances of all objects on one neuron was examined. The network trained for 300 epochs was chosen for the final descriptor selection procedure because of the most even distribution of objects and small differences between the maximal and minimal distances calculated at each neuron. In Figure 3 the bars representing the ranges of distances calculated at all 25 neurons are displayed. The reduced set contained 50 descriptors, two from each neuron: the most similar one and the most different one regarding the distance from the particular neuron (eq 1).

A close analysis showed that there were eight neurons (the first eight points in Figure 3) at which the clusters of descriptors were rather homogeneous with calculated differences between the maximal and minimal distances below 2.4 (1.0 < d_{max}−d_{min} < 2.4). The largest difference (the bar at the last point in Figure 3) was more than 2 times larger (d_{max}−d_{min} = 5.7). Accordingly, a larger reduction of the descriptor set was proposed to be tested, which took into account only one descriptor from each of the neuron given at the first eight positions in Figure 3, while two descriptors were depicted from the rest of the 17 neurons, which gives altogether 8 + 34 = 42 descriptors.

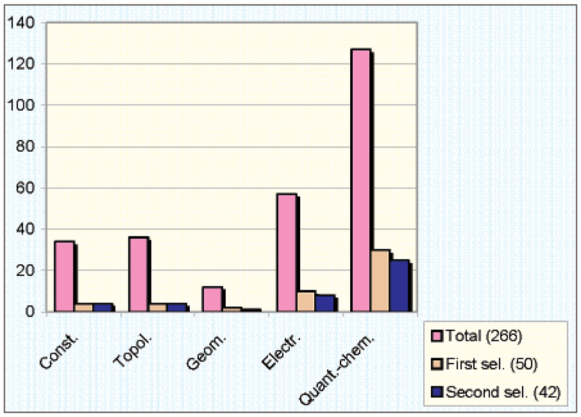


Figure 4. Distribution of descriptors obtained for the three sets of descriptors.

Three different sets of descriptors were prepared: the nonreduced set of 266 descriptors, the reduced set of 50 descriptors, and the reduced set of 42 descriptors, chosen on the basis of the above-described selection procedure. In Figure 4 it is shown how the reduced sets of descriptors are distributed among the four types of descriptors, namely constitutional, topological, geometrical, electrostatic, and quantum-chemical. LogP was added to each set of descriptors. LogP is a descriptor of a different origin (property-descriptor) as the rest of the 266 calculated descriptors, yet important for modeling biological properties.[24] According to the three different descriptors selections described above, the structures of all 106 molecules were encoded, and three data sets were obtained:

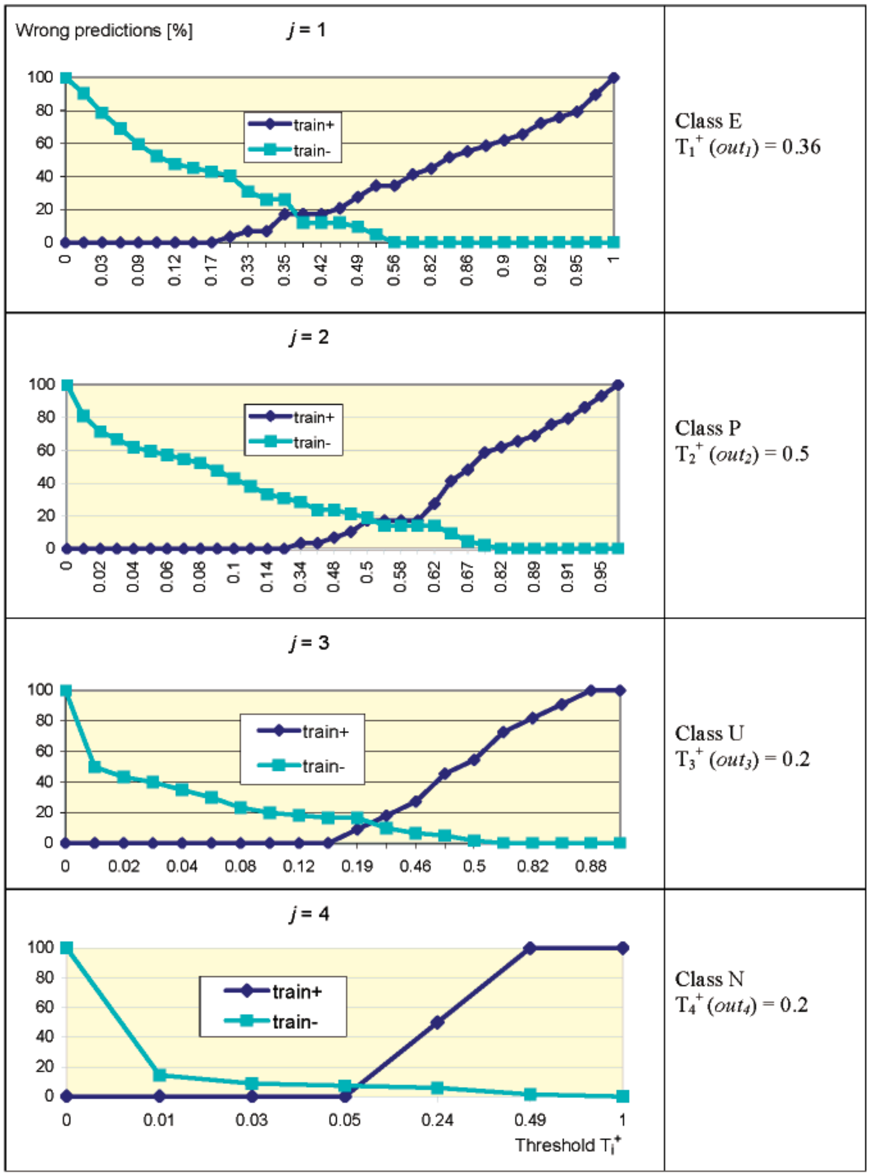


Figure 5. The thresholds T_j^+ determined for the class predictions in the model from the counterpropagation neural network of 9 × 9 neurons, trained for 100 epochs. The diamonds and squares stand for positive (confirmative) and negative (rejecting) predictions, respectively.

♦data set 1 (DS1) → 106 molecular structures represented by 267 descriptors,

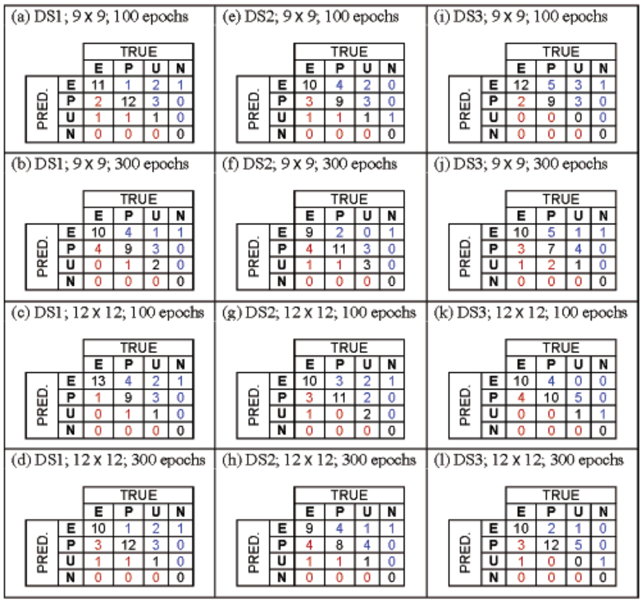
♦data set 2 (DS2) → 106 molecular structures represented by 51 descriptors,

♦data set 3 (DS3) → 106 molecular structures represented by 43 descriptors.

Model Construction and Validation. With three data sets, DS1, DS2, and DS3, each divided into a training (71 molecules) and a test set (35 molecules), different models were built. The CP NN contained from 81 to 144 neurons. The NN architectures varied, concerning the dimensionality, from 9 × 9 × 41 to 12 × 12 × 267 weights per network.[23] The networks were trained with the molecules from the individual training sets for not less than 100 and not more than 1000 epochs.

The predictions of the class of probability for the endocrine disruption activity are provided in the output layer of individual models. As described in the section **Classification Model**, the responses in the output layer, which consists of four levels, are real numbers between zero and one (**Out** = (out_1, out_2, out_3, out_4), $0.00 \leq \mathbf{Out}_j \leq 1.00$). It is necessary to determine the threshold value (T^+), above which the prediction for a *j*th class is positive (confirmative). T^+ enables the transformation of the model output values to discrete class predictions, i.e., one for a confirmative and zero for a rejecting answer. There are four classes, so we need four threshold values for each of the constructed models (T_j^+, $j = 1, 4$). They are determined according to the number of correct/wrong class predictions if the trained network is tested by the same objects as it was trained with, i.e., molecules from the training set. Below the threshold all predictions are rejecting and denoted by a zero, which means that the compound does not belong to the *j*th class, while the predictions above the threshold are positive and denoted by one (the compound belongs to the *j*th class). In Figure 5 an example of the determination of T_j^+ for one of the constructed models is shown.

As can be seen from Figure 5, the threshold is positioned where the cumulative error of the training set (sum of false positive and false negative predictions) is the lowest. If the



(a) DS1; 9 x 9; 100 epochs

		TRUE			
		E	P	U	N
PRED.	E	11	1	2	1
	P	2	12	3	0
	U	1	1	1	0
	N	0	0	0	0

(b) DS1; 9 x 9; 300 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	4	1	1
	P	4	9	3	0
	U	0	1	2	0
	N	0	0	0	0

(c) DS1; 12 x 12; 100 epochs

		TRUE			
		E	P	U	N
PRED.	E	13	4	2	1
	P	1	9	3	0
	U	0	1	1	0
	N	0	0	0	0

(d) DS1; 12 x 12; 300 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	1	2	1
	P	3	12	3	0
	U	1	1	1	0
	N	0	0	0	0

(e) DS2; 9 x 9; 100 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	4	2	0
	P	3	9	3	0
	U	1	1	1	1
	N	0	0	0	0

(f) DS2; 9 x 9; 300 epochs

		TRUE			
		E	P	U	N
PRED.	E	9	2	0	1
	P	4	11	3	0
	U	1	1	3	0
	N	0	0	0	0

(g) DS2; 12 x 12; 100 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	3	2	1
	P	3	11	2	0
	U	1	0	2	0
	N	0	0	0	0

(h) DS2; 12 x 12; 300 epochs

		TRUE			
		E	P	U	N
PRED.	E	9	4	1	1
	P	4	8	4	0
	U	1	1	1	0
	N	0	0	0	0

(i) DS3; 9 x 9; 100 epochs

		TRUE			
		E	P	U	N
PRED.	E	12	5	3	1
	P	2	9	3	0
	U	0	0	0	0
	N	0	0	0	0

(j) DS3; 9 x 9; 300 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	5	1	1
	P	3	7	4	0
	U	1	2	1	0
	N	0	0	0	0

(k) DS3; 12 x 12; 100 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	4	0	0
	P	4	10	5	0
	U	0	0	1	1
	N	0	0	0	0

(l) DS3; 12 x 12; 300 epochs

		TRUE			
		E	P	U	N
PRED.	E	10	2	1	0
	P	3	12	5	0
	U	1	0	0	1
	N	0	0	0	0

Figure 6. Classification tables with the number of correct (diagonal elements), false positive (upper triangle), and false negative predictions (lower triangle). The predictions are acquired from 12 models (from (a) to (l)), constructed on the basis of three different spectral representations (DS1, DS2, and DS3), using two different neural network architectures (9 × 9 and 12 × 12 neurons), while the training time was 100 or 300 epochs.

T_j^+ is close to zero, the predictions of the *j*th class for most of the molecules from the training set will be confirmative ($\mathbf{Out}_j > 0$). So the molecules that are really in the *j*th class will be correctly predicted, while the predictions for those that are not will be wrong, so-called false positives. On the other hand, if the T_j^+ is close to one, the majority of the predictions will be rejecting for class *j*. Now the predictions of molecules that are actually in the *j*th class will be wrong, so-called false negatives. The thresholds were determined for all the models in order to be able to obtain class predictions when testing the predictive ability. The resulting models were validated by checking the class predictions for 35 test molecules. The misclassification tables obtained by a comparison of actual and predicted classes of test compounds are shown in Figure 6.

The resulting predictions from all models were inspected to choose the optimal model. The prediction performance of 12 models demonstrated in Figure 6 are the examples with the largest number of correct predictions (the sum of the diagonal elements). However, to choose the optimal model we have to consider additional criteria, such as the number of false negative predictions, which are more severe errors than false positives, because they would classify a harmful compound as a nontoxic one. An important indicator about the quality of a model is also the sum of the predictions that are wrong for more than one category. For example, the element at the position (1, 4) in the model (a) of Figure 6 is equal to 1. It stands for a prediction of one molecule as being the class E, while in fact it is N (nonactive), which is three categories lower. Taking into account all listed possible criteria, it was found out that the lowest total number of false predictions was 11 (Figure 6, model (a) DS1; 9 × 9; 100 epochs), the lowest number of false negative predictions was 2 (Figure 6, models (c) DS1; 12 × 12; 100 epochs and (i) DS3; 9 × 9; 100 epochs), and the model with zero predictions erroneous for more than one class was model (k) DS3; 12 × 12; 100 epochs. It is not a straightforward decision which model to propose to be the best among the four listed above. 69% of the correct predictions in model (a) is the best result at first glance. It suggests that the descriptors' space covered by the training compounds contains significant information to obtain the model which makes the relationship between the structure and the toxicity class general enough to reasonably predict the classes of the test compounds. However, if the QSAR model is used to make the priority list of compounds that have to be tested by more assured in vivo methods, model (k) is better, because it makes the range-list of tested chemicals from most to least harmful less erroneous (a mistake is never larger than for one class). To recapitulate, the predictions of all the above-described models refer to the test set compounds, which were not considered during the determination of the model parameters and the training procedure. Considering model (a) from Figure 6, the information about the structure−activity relationship contained in 71 compounds from the training set was enough to generalize this relationship to a degree which enabled the correct predictions for more than two-thirds of the test compounds.

Table 4. Situation on the Neuron (1,4) of Model (f)

chem no.	CAS no.	name	class
405	38380-07-3	PCB 128 (2,2′,3,3′,4,4′-hexachlorobiphenyl)	U
406	**38411-22-2**	**PCB 136 (2,2′,3,3′,6,6′-hexachlorobiphenyl)**	**P**
408	**35065-27-1**	**PCB 153 (2,2′,4,4′,5,5′-hexachlorobiphenyl)**	**E**
409	38380-08-4	PCB 156 (2,3,3′,4,4′,5-hexachlorobiphenyl)	P
410	32774-16-6	PCB 169 (3,3′,4,4′,5,5′-hexachlorobiphenyl)	E

There were some errors that were often repeated in different models. There are two sources of problems: (i) to discriminate structurally similar compounds belonging to different classes and (ii) to handle very dissimilar compounds belonging to the same class. In Table 4 there is an example taken from model (f) (see Figure 6), which shows the situation (i). The five PCB derivatives represented with 51 descriptors (DS2) shown in Table 4 are structurally so similar that they occupy the same neuron. Two compounds from the test set are boldface, the other three are from the training set, yet occupying the same neuron. The network was not capable of adapting during the training process to separate those three compounds, so the resulting model did not distinguish between those three PCBs, which generated severe errors, because each of the three PCBs from the training set belonged to a different class. This is one of the so-called conflicting situations at the neuron (1,4), which have to be kept low.

One of the largest difficulty in the present research is the lack of compounds declared as nontoxic (class N). This is a consequence of circumstances that, for many reasons, from economical to ecological ones, it was more urgent to obtain the toxicity tests of the compounds that are likely to be harmful (pesticides, herbicides, fito-estrogenes, drugs). It is not unusual that the experimental data of very toxic compounds is available in a larger extent than for less harmful compounds. With a data set of a more even distribution of compounds between categories the prediction results obtained by the QSAR models proposed in this paper would certainly improve.

CONCLUSIONS

We propose neural network techniques to construct computational models for the classification of endocrine disrupting chemicals on the basis of chemical structure. The Kohonen network technique was applied to reduce the number of descriptors. The extended counterpropagation neural network was chosen as a classification method. A module for the threshold determination was developed in order to transform the predictions obtained from the prediction layer of the counterpropagation neural network into the classification tool. The data set contains structurally diverse chemicals, nevertheless, the two-step modeling principle of the counterpropagation neural network enables one to build a classification model capable of treating all chemicals together. The models were tested with the test set, which was not used for the reduction of descriptors or modeling. The class predictive power of constructed models shows that the method is promising, despite the weakly defined endpoint, which is the degree of evidence for the endocrine disrupting activity/nonactivity. The classification model can be easily adapted to be trained as a two-class (yes/no) model once a list of chemicals is available with precise information about their endocrine disrupting activity. Further development will be directed toward improving the method of the descriptors selection developed during the research work presented in this paper.

ACKNOWLEDGMENT

The authors acknowledge partial financial support of the Ministry of Education, Science, and Sport of Slovenia for financing the research through the project grants P104-507 as well as the European Union 5-th framework program scheme for partial financial support through the program Marie Curie Host Training Site no. HPMT-CT-2001-00240 and IMAGETOX no: HPRN-CT-1999-00015.

REFERENCES AND NOTES

(1) Purchase, I. F. H. Risk assessment. Principles and consequences. *Pure Appl. Chem.* **2000**, *72*, 1051−1056.

(2) Boelsterli, U. A. *Mechanistic Toxicology, The Molecular Basis of How Chemicals Disrupt Biological Targets;* Taylor & Francis: New York, 2002.

(3) Verhaar, H. J. M.; Ramos, E. U.; Hermens, J. L. M. Classifying environmental pollutants. 2. Separation of class 1 (baseline toxicity) and class 2 ('polar narcosis') type compounds based on chemical descriptors. *J. Chemom.* **1996**, *10*, 149−162.

(4) Bradbury, S. P.; Lipnick, R. L. Structural-Properties For Determining Mechanisms of Toxic Action − Introduction. *Environ. Health Perspect.* **1990**, *87*, 181−182.

(5) Colborn, T.; Saal F. S. V.; Soto, A. M. Developmental Effects of Endocrine-Disrupting Chemicals in Wildlife and Humans. *Environ. Health Perspect.* **1993**, *101*, 378−384.

(6) Hutchinson, T. H.; Pickford, D. B. Ecological risk assessment and testing for endocrine disruption in the aquatic environment. *Toxicology* **2002**, *181*, 383−387.

(7) Brunstrom, B.; Axelsson, J.; Halldin, K. Effects of endocrine modulators on sex differentiation in birds. *Ecotoxicology* **2003**, *12*, 287−295.

(8) Singleton, D. W.; Khan, S. A. Xenoestrogen exposure and mechanisms of endocrine disruption. *Front. Biosci.* **2003**, *8*, 110−118.

(9) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I.; The present status of QSAR in toxicology. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 23−38.

(10) Bradbury, S. P.; Russom, C. L.; Ankley, G. T.; Schultz, T. W.; Walker, J. D. Overview of data and conceptual approaches for derivation of quantitative structure−activity relationships for ecotoxicological effects of organic chemicals. *Environ. Toxicol. Chem.* **2003**, *22*, 1789−1798.

(11) Worgan, A. D. P.; Dearden, J. C.; Edwards, R.; Netzeva, T. I.; Cronin, M. T. D.; Evaluation of a novel short-term algal toxicity assay by the development of QSARs and inter-species relationships for narcotic chemicals. *QSAR Comb. Sci.* **2003**, *22*, 204−209.

(12) Cronin, M. T. D.; Walker, J. D.; Jaworska, J. S.; Comber, M. H. I.; Watts, C. D.; Worth, A. P. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environ. Health Perspect.* **2003**, *111*, 1376−1390.

(13) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct. (THEOCHEM*) **2003**, *622*, 39−51.

(14) Moore, D. R. J.; Breton, R. L.; MacDonald, D. B. A comparison of model performance for six quantitative structure−activity relationship packages that predict acute toxicity to fish. *Environ. Toxicol. Chem.* **2003**, *22*, 1799−1809.

(15) European Commission. *Communication from the Commission to the Council and the European Parliament on the Implementation of the Community Strategy for Endocrine Disrupters − a Range of Substances Suspected of Interfering with the Hormone Systems of Humans and Wildlife*, COM (2001) 262 final, Brussels, 14 June 2001. http://europa.eu.int/comm/environment/docum/01262_en.htm#bkh.

(16) QSAR Development Principles: http://ecb.jrc.it/Documents/QSAR/QSAR_development_principles.doc.

(17) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA 2.0, Comprehensive Descriptors for Structural and Statistical Analysis*; Copyright 1994−1996 University of Florida, U.S.A.

(18) Physical Properties Database−PHYSPROP-http://esc.syrres.com/interkow/PhysProp.htm.

(19) Hansch, L.; Leo, A. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(20) KowWin program v1.66, on-line demo: http://esc.syrres.com/interkow/kowdemo.htm.

(21) Hecht-Nielsen, R. Counterpropagation networks, *Appl. Optics* **1987**, *26*, 4979−4984.

(22) Dayhof, J. Neural Network Architectures, An Introduction; Van Nostrand Reinhold: New York, 1990.

(23) Zupan, J.; Novič, M.; Ruisanchez, I. Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 1−23.

(24) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. Comparative QSAR at the interface between chemistry and biology. *Chem. Rev.* **2002**, *102*, 783−812.

(25) Novič, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counter-Propagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.

CI030421A