SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

Prediction of estrogenicity: validation of a classification model

A. Gallegos Saliner [a]; T. I. Netzeva [a];A. P. Worth [a]
[a] †European Chemicals Bureau (ECB), Institute for Health and Consumer Protection, 21020 Ispra (VA), Italy

To cite this Article Saliner, A. Gallegos , Netzeva, T. I. andWorth, A. P.(2006) 'Prediction of estrogenicity: validation of a classification model', SAR and QSAR in Environmental Research, 17: 2, 195 — 223
To link to this Article: DOI: 10.1080/10659360600636022
URL: http://dx.doi.org/10.1080/10659360600636022

SAR and QSAR in Environmental Research
Vol. 17, No. 2, April 2006, 195–223

Prediction of estrogenicity: validation of a classification model†

A. GALLEGOS SALINER*, T. I. NETZEVA and A. P. WORTH

†European Chemicals Bureau (ECB), Institute for Health and Consumer Protection, Joint Research Centre, European Commission, 21020 Ispra (VA), Italy

(*Received 13 October 2005; in final form 31 January 2006*)

(Q)SAR models can be used to reduce animal testing as well as to minimise the testing costs. In particular, classification models have been widely used for estimating endpoints with binary activity. The aim of the present study was to develop and validate a classification-based quantitative structure-activity relationship (QSAR) model for endocrine disruption, based on interpretable mechanistic descriptors related to estrogenic gene activation. The model predicts the presence or absence of estrogenic activity according to a pre-defined cut-off in activity as determined in a recombinant yeast assay. The experimental data was obtained from the literature. A two-descriptor classification model was developed that has the form of a decision tree. The predictivity of the model was evaluated by using an external test set and by taking into account the limitations associated with the applicability domain (AD) of the model. The AD was determined as coverage of the model descriptor space. After removing the compounds present in the training set and the compounds outside of the AD, the overall accuracy of classification of the test chemicals was used to assess the predictivity of the model. In addition, the model was shown to meet the OECD Principles for (Q)SAR Validation, making it potentially useful for regulatory purposes.

Keywords: Quantitative structure-activity relationship (QSAR); Applicability domain (AD); Relative estrogenic gene activation; Endocrine disruptor (ED); QSAR validation; Classification model (CM)

1. Introduction

Endocrine disrupters (EDs) are chemicals that cause adverse environmental or human health effects in healthy organisms or their offspring by altering the function of the endocrine system. They include a large number of substances such as natural hormones, environmental pollutants, antioxidants and metabolites, synthetically produced chemicals, and industrial chemicals or their by-products [1]. Some EDs have the ability to bind to estrogen receptors (ER), and are therefore thought to cause their effects

*Corresponding author. Email: ana.gallegos@jrc.it
†Presented at CMTPI 2005: Computational Methods in Toxicology and Pharmacology Integrating Internet resources (Shanghai, China, October 29–November 1 2005).

SAR and QSAR in Environmental Research
ISSN 1062-936X print: ISSN 1029-046X online
DOI: 10.1080/10659360600636022

in this way. In wildlife, EDs cause abnormalities and impaired reproductive performance in some species, and they are also associated with changes in immunity, behaviour, and skeletal deformities. In humans, EDs are suspected to be responsible for changes in health patterns over recent decades, including declining sperm counts in some geographical regions, increased numbers of male children born with genital malformations, and certain types of cancer [2].

The environmental and human health impacts of endocrine disruption have stimulated worldwide initiatives to understand and assess ED effects. In the early 90s, the US Environmental Protection Agency (EPA) initiated research in this field [3–5]. In particular, the US EPA screening program uses a tiered approach for determining whether a substance may have an effect in humans that is similar to an effect produced by naturally occurring estrogen, androgen, or thyroid hormones [6].

In the European Union (EU), the European Commission adopted a community strategy for endocrine disrupters in 1999 [6, 7]. This strategy established a number of actions related to identification of substances, monitoring, research, international co-ordination and communication to the public [8, 9]. Emphasis was placed on the need to develop quick and effective risk management strategies, and the need for consistency with the overall chemicals policy and legislation. The first progress report on the implementation of this strategy was presented in 2001 [10]. The actions of the strategy have contributed to the gathering of scientific data and to the identification of substances for further assessment, to research and monitoring efforts, and to the identification of specific exposure groups [11].

The European Commission and EU Member States participate in the OECD (Organisation for Economic Co-operation and Development) Endocrine Disrupter Testing and Assessment Task Force (EDTA TF), which was set up in 1998 with the goal of developing agreed test methods for EDs for some environmental and human health effects. In 2003, the OECD EDTA TF developed a conceptual framework for the screening, testing and assessment of EDs [12], foreseeing the use of quantitative structure-activity relationships (QSARs) and *in vitro* tests before using *in vivo* tests. Subsequently, a QSAR Task Group of the Validation Management Group for Non-Animal testing (VMG NA) was established [13]. This group has recognised the need to define the physicochemical domains of chemical inventories of regulatory interest within OECD Member Countries and to review the state-of-the-art in QSARs for endocrine disruption, to provide the basis for evaluating the applicability of *in silico* models for screening purposes [14].

Thus, the issue of endocrine disruption is being addressed in multiple research and policy initiatives. In the Commission's legislative proposal for a new regulatory system called REACH (Registration, Evaluation, and Authorisation of Chemicals), adopted by the Commission in 2003 [15], EDs are covered by the authorisation procedure for substances of very high concern, according to the precautionary principle. The precautionary principle states that where there is uncertainty as to the existence or extent of risks to human health, protective measures should be taken without having to wait until the reality and the seriousness of those risks become apparent. EDs are also being considered in discussions on data requirements and principles for risk assessment of plant protection products [16]. While the current and proposed EU chemicals legislation accounts for detrimental endocrine related effects on reproduction or specific diseases, it does not use endocrine disruption as a required endpoint *per se*, but treats endocrine disruption as a possible underlying mechanism of other effects of concern.

Given the status of the science in this area, there is a need to establish the most relevant endpoints for predicting endocrine disrupting effects, and a need to develop consensus on how to use the data within regulatory frameworks [17].

Guidance on the use of (Q)SARs is provided in Annex IX of the proposed REACH legislation. It states that (Q)SARs may be used to indicate the presence or absence of a dangerous property if results are derived from a (Q)SAR model whose scientific validity has been established, which is adequate for the purpose of classification and labelling and risk assessment, and if adequate and reliable documentation of the method is provided [18, 19].

Traditionally, the use of (Q)SARs for regulatory purposes has often been limited, partly due to the lack of an accepted approach for the evaluation of scientific validity. To overcome this barrier to acceptance, a number of initiatives were organised, including a workshop organised by CEFIC/ICCA in Setubal (Portugal) in 2002 [20], during which a number of requirements for the validity of (Q)SARs were proposed. Based on this proposal and following an in-depth analysis, five principles for (Q)SAR validation were subsequently adopted by the OECD Member Countries and the European Commission in 2004 [21]. According to these principles, to facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with: (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictivity; and (5) a mechanistic interpretation, if possible. These principles provide a useful conceptual framework for assessing the validity of (Q)SARs, but practical guidance on how to apply the principles is also needed [22]. Preliminary guidance has been proposed by the European Chemicals Bureau [23], and is being developed by the OECD *ad hoc* Group on (Q)SARs.

The mechanisms might include receptor binding, altered post-receptor activation, altered steroidgenesis, perturbation of hormone storage, release and elimination [24, 25]. The receptor binding assays quantify directly (or as a result of competitive interaction) the ability of chemicals to bind to estrogen and androgen receptors, while the transcriptional activation (TA) assays as used in this study detect a stimulation of a reporter gene as a result of receptor activation. Thus, the relationship between chemical structure and biological activity determined in a TA assay appears less obvious compared to the binding assay. However, it is easier to trace in comparison to the steroidgenesis assays or *in vivo* studies, where multiple mechanisms can be involved in the affection of the complex hormonal system of mammals. A number of approaches have been used to develop (Q)SARs for estrogenicity, including Automatic Docking models (ADAM) [26], Comparative Molecular Field Analysis (CoMFA) [27, 28], classical QSAR and Hologram QSAR (HQSAR) [29], the Common Reactivity Pattern (COREPA) approach [30], Molecular Quantum Similarity Analysis (MQSA) [31], and Decision Forests (DF) [32].

In a previous study, we developed a classification model (CM) for estrogenic activity and evaluated its compliance with the OECD Principles for the validation of (Q)SARs [14]. To assess predictivity with chemicals that had not been used for model development, we split the whole data set of 117 chemicals in two subsets. The larger subset (approximately 90% of all chemicals) was used for model development, whereas the smaller subset (the remaining 10%) was used for external validation. Although this strategy for splitting the data is widely used in predictivity assessments by external

validation, it reduces the number of chemicals used in the modelling which may lead to a loss of information.

The aim of the current study was to develop a classification model for estrogenic activity by using the training set of 117 chemicals that was used in the previous study [14] and to validate it with a real external test set obtained recently from the literature.

2. Methods and results

2.1 *Datasets and estrogenic activity*

2.1.1 Training set. The training set was derived from an experimentally measured set of an assorted group of 117 aromatic compounds published by Schultz *et al*. [33], including bisphenols, benzophenones, flavonoids, biphenyls, phenols, and other aromatic chemicals.

For these chemicals, the *in vitro* Glaxo Wellcome-derived *Saccharomyces cerevisiae* recombinant yeast assay [34] had been performed following the protocol of Schultz *et al*. [35]. This assay is used to express the α-human Estrogen Receptor (hERα) gene, and has a 0.08 nM detection limit for 17β-estradiol and a 10000-fold responsiveness.

The estrogenic compounds interacting with the receptor transcribe the genes and produce β-galactosidase, the activity which is measured colorimetrically. The endpoint determined for each compound is the concentration eliciting 50% of the activity of the positive control 17β-estradiol (EC50), expressed in molar units. The relative gene expression of each test chemical was calculated by dividing the EC50 value of 17β-estradiol by the EC50 of the test chemical and multiplied by 100. For modelling purposes, experimental data were classified as active (A) or inactive (I), as reported previously [31]. Those compounds that reached LC50 (lethal concentration) without triggering release of β-galactosidase, detected as a no colour change compared to the untreated control, are called inactive. The chemical names, CAS numbers, binary-converted experimentally observed activities, and activities predicted by the classification model are given in table 1.

2.1.2 Test set. An independent external test set was taken from a dataset published by Nishihara *et al*. [36]. The original test set consisted of more than 500 commercially available chemicals obtained from different sources, of which 64 compounds were evaluated as positive. The dataset consists of natural substances, medicines, pesticides, and industrial chemicals (including PCBs, PCDFs, PAHs, phenols, benzenes, phthalates, adipates), and thus covers a wide range of chemical classes, representing a broad degree of structural diversity.

The compounds had been evaluated by using an *in vitro* reporter gene expression assay in yeast cells with high repeatability and sensitivity, based on the ligand-dependent interaction of two proteins, a hormone receptor (ERα) and a coactivator (TIF2), where the hormonal activity is detected by the β-galactosidase activity [37].

The results were expressed as the 10% of the relative effective concentration (REC10), that is, the concentration of the test chemical showing 10% of the agonist activity of the optimum concentration of 17β-estradiol. When the activity of the test

Table 1. The chemical names, Chemical Abstract Service (CAS) number (if available), logarithm of the octanol-water partition coefficient (log K_{OW}), number of hydrogen bond donor groups (n_{Hdon}), and observed and predicted binary classification activity (1 if considered active, 0 if considered inactive) for the training set (117 compounds).

ID	*Chemical name*	*CAS*	*log K_{OW}*	*n_{Hdon}*	*Binary activity (observed)*	*Binary activity (predicted)*
1	17-β-Estradiol*	50-28-2	4.01	2	1	1
2	4,4′-Diethylethylene bisphenol	6898-97-1	4.79	2	1	1
3	4,4′-Cyclohexylidene bisphenol	843-55-0	4.90	2	1	1
4	4,4′-Thiodiphenol*	2664-63-3	3.25	2	1	1
5	*bis* (4-Hydroxyphenyl)methane*	620-92-8	3.56	2	1	1
6	4,4′-Ethylidene bisphenol	2081-08-5	3.89	2	1	1
7	Bisphenol A*	80-05-7	4.32	2	1	1
8	4,4′-(1,3-Adamantanediyl)bisphenol	37677-93-3	5.42	2	1	1
9	4,4′-Dihydroxybenzophenone*	611-99-4	2.70	2	1	1
10	*Bis* (4-Hydroxyphenyl)sulphone**	80-09-1	2.30	2	1	1
11	1,1,1-*tris*(4-Hydroxyphenyl)ethane	27955-94-8	5.26	3	1	1
12	4,4′-Dimethoxybiphenyl	2132-80-1	3.23	0	0	0
13	4,4′-Dipyridyl	553-26-4	2.04	0	0	0
14	2,4-Dihydroxybenzophenone	131-56-6	2.70	2	1	1
15	2,2′,4,4′-Tetrahydroxyl benzophenone	131-55-5	2.13	4	1	1
16	4-Chloro-4′-hydroxy benzophenone	42019-78-3	3.50	1	1	1
17	3-Hydroxybenzophenone	13020-57-0	2.98	1	1	1
18	4-Hydroxybenzophenone	1137-42-4	2.98	1	1	1
19	2,3,4-Trihydroxybenzophenone	1143-72-2	2.41	3	1	1
20	2,4,4′-Trihydroxybenzophenone	1470-79-7	2.41	3	1	1
21	2,2′-Dihydroxybenzophenone	835-11-0	2.70	2	0	1
22	4,4′-Dichlorobenzophenone	90-98-2	4.30	0	0	0
23	2-Hydroxybenzophenone	117-99-7	2.98	1	0	1
24	4-Methoxybenzophenone	611-94-9	3.02	0	0	0
25	4-Chlorobenzophenone	134-85-0	3.79	0	0	0
26	4-Methylbenzophenone	134-84-9	3.74	0	0	0
27	4-Nitrobenzophenone	1144-74-7	3.22	0	0	0
28	Benzophenone*	119-61-9	3.27	0	0	0
29	Genistein (4′,5,7-Trihydroxyisoflavone)*	446-72-0	1.96	3	1	1
30	Biochanin A (5,7-Dihydroxy-4′-methoxy isoflavone)**	491-80-5	1.99	2	1	1
31	Naringenin (4′,5,7-Trihydroxyflavanone)*	480-41-1	1.99	3	1	1
32	Morin hydrate (3,3′,5,5′,7-Pentahydroxyflavone)	480-16-0	0.28	5	1	0

(*continued*)

Table 1. Continued.

ID	Chemical name	CAS	log K_{OW}	n_{Hdon}	Binary activity (observed)	Binary activity (predicted)
33	Daidzein (4′,7-Dihydroxyisoflavone)*	486-66-8	2.25	2	1	1
34	Phloretin (2′,4,4′,6′-Tetrahydroxychalcone)**	60-82-2	2.46	4	1	1
35	4′-Hydroxychalcone	2657-25-2	3.39	1	1	1
36	Galangin (3,5,7-Trihydroxyflavone)	548-83-4	0.85	3	0	0
37	Baicalein (5,6,7-Trihydroxyflavone)	491-67-8	1.46	3	0	0
38	Chrysin (5,7-Dihydroxyflavone)	480-40-0	1.75	2	0	0
39	Flavone*	525-82-6	2.32	0	0	0
40	Flavanone	487-26-3	2.84	0	0	0
41	*trans*-Chalcone	614-47-1	3.68	0	0	0
42	2′,4′,6′-Trichloro-4-biphenylol	–	5.00	1	1	1
43	2′,3′,4′,5′-Tetrachloro-4-biphenylol	–	5.52	1	1	1
44	2′,5′-Dichloro-4-biphenylol	–	4.48	1	1	1
45	4′-Chloro-4-biphenylol	–	3.96	1	1	1
46	2′,3′,4′,5′-Tetrachloro-3-biphenylol	–	5.52	1	1	1
47	2,2′,5′-Trichloro-4-biphenylol	–	5.00	1	1	1
48	2′,5′-Dichloro-3-biphenylol	–	4.48	1	1	1
49	4,4′-Biphenyldiol*	92-88-6	3.16	2	1	1
50	4-(1-Hydroxyethyl)biphenyl	3562-73-0	3.61	1	1	1
51	3-Hydroxybiphenyl	580-51-8	3.45	1	1	1
52	4-Hydroxybiphenyl*	92-69-3	3.45	1	1	1
53	4-(2-Hydroxypropyl)biphenyl	34352-74-4	3.69	1	1	1
54	4-Biphenylmethanol	3597-91-9	3.20	1	1	1
55	3-Chloro-4-biphenylol	–	3.96	1	1	1
56	2-Chloro-4-biphenylol	–	3.96	1	1	1
57	2-Hydroxybiphenyl**	90-43-7	3.45	1	1	1
58	4-Methoxybiphenyl	613-37-6	3.48	0	1	0
59	2′,5′-Dichloro-2-biphenylol	–	4.48	1	1	1
60	3,4′,5-Trichloro-4-biphenylol	–	5.00	1	0	1
61	3,3′,5,5′-Tetrachloro-4,4′-biphenyldiol	–	5.23	2	0	1
62	Biphenyl*	92-52-4	3.73	0	0	0
63	4-(1-Adamantyl)phenol	29799-07-3	4.06	1	1	1
64	4-(4-Bromophenyl)phenol	29558-77-8	4.24	1	1	1
65	Ethyl-4′-hydroxy-4-biphenyl carboxylate	50670-76-3	3.52	1	1	1
66	Benzyl-4-hydroxybenzoate	94-18-8	3.27	1	1	1
67	Isoamyl-4-hydroxybenzoate	6521-30-8	3.03	1	1	1
68	2-Ethylhexyl-4′-hydroxy benzoate	5153-25-3	4.29	1	1	1

69	4-Cyclohexylphenol	1131-60-8	3.64	1	1	1
70	Nonyl-4-hydroxybenzoate	38713-56-3	4.68	1	1	1
71	4-(*tert*-Octyl)phenol*	140-66-9	4.95	1	1	1
72	Phenyl-4-hydroxybenzoate	17696-62-7	3.17	1	1	1
73	4-Phenoxyphenol	831-82-3	3.19	1	1	1
74	*N*-(4-Hyroxyphenyl)-2-naphthylamine	93-45-8	3.88	2	1	1
75	4-(Benzyloxy)phenol	103-16-2	3.29	1	1	1
76	4-Hydroxyoctanophenone	2589-73-3	3.68	1	1	1
77	Benzyl-4-hydroxyphenyl ketone	2491-32-9	2.92	1	1	1
78	4-Hexanoyl resorcinol	3144-54-5	2.60	2	1	1
79	4-Heptyloxyphenol	13037-86-0	3.91	1	1	1
80	4-Octylphenol**	1806-26-4	5.00	1	1	1
81	Resorcinol monobenzoate	136-36-7	3.17	1	1	1
82	Butyl-4-hydroxybenzoate*	94-26-8	2.70	1	1	1
83	4-Hydroxydiphenylmethane	101-53-1	3.84	1	1	1
84	2-Hydroxydiphenylmethane	28994-41-4	3.84	1	1	1
85	4-Cyclopentyl phenol	1518-83-8	3.24	1	1	1
86	4-Hexyloxyphenol	18979-55-0	3.51	1	1	1
87	3-Hydroxydiphenylamine	101-18-8	2.87	2	1	1
88	4-(*tert*-Pentyl)phenol*	80-46-6	3.79	1	1	1
89	4-*n*-Pentylphenol*	14938-35-3	3.81	1	1	1
90	4-Pentyloxyphenol	18979-53-8	3.11	1	1	1
91	4-Butoxyphenol	122-94-1	2.72	1	1	1
92	*N*-benzyl-4-hydroxyaniline	103-14-0	2.97	2	1	1
93	Ethyl-4-hydroxybenzoate*	120-47-8	1.83	1	1	1
94	4-Hydroxypropiophenone	70-70-2	1.70	1	1	0
95	2-(4-Hydroxyphenyl)-5-pyrimidinol	142172-97-2	1.87	2	1	1
96	4-Propoxyphenol	18979-50-5	2.32	1	1	1
97	4-Propylphenol*	645-56-7	3.02	1	1	1
98	2-(Benzyloxy)phenol	6272-38-4	3.29	1	1	1
99	4-(Imidazol-1-yl)phenol	10041-02-8	2.09	1	1	1
100	4-(4-Hydroxyphenyl)-2-butanone	5471-51-2	2.34	1	1	1
101	4-Methylphenol**	106-44-5	2.23	1	0	1
102	Phenol*	108-95-2	1.76	1	0	0
103	5-Pentylresorcinol	500-66-3	3.53	2	0	1
104	Homovanillyl alcohol	2380-78-1	1.23	2	0	0
105	1-(4-Hydroxyphenyl)-1H-tetrazole-5-thiol	52431-78-4	2.22	2	0	1
106	2-Phenylhydroquinone	1079-21-6	3.16	2	0	1
107	1-Benzyl-4-hydroxypiperidine	4727-72-4	1.49	1	0	0

(*continued*)

Table 1. Continued.

ID	*Chemical name*	*CAS*	*log K_{OW}*	n_{Hdon}	*Binary activity (observed)*	*Binary activity (predicted)*
108	4-Phenylpyridine	939-23-1	2.88	0	0	0
109	2-(4-Hydroxyphenyl)ethanol	501-94-0	1.48	2	0	0
110	2-(2-Hydroxyethyl)resorcinol	49650-88-6	1.07	2	0	0
111	1-Benzyl-3-pyrrolidinol	775-15-5	1.44	1	0	0
112	Toluene*	108-88-3	2.51	0	0	0
113	Chlorobenzene*	108-90-7	2.56	0	0	0
114	Benzyl benzoate	120-51-4	3.55	0	0	0
115	Isoamyl benzoate	94-46-2	3.31	0	0	0
116	Methyl benzoate	93-58-3	1.78	0	0	0
117	Benzene	71-43-2	2.05	0	0	0

*Compounds present in the training and test set with the same binary activity (22 compounds).
**Compounds present in the training and test set with different activity (6 compounds).

substance was higher than REC10 within the concentrations tested, the chemical was classified as positive (active), and when it was classified as negative, the highest tested dose was indicated (inactive).

2.2 *Data screening*

Prior to the analysis, data screening techniques were applied to the original test set to assess the quality of the input. After removing redundant chemicals present twice with different synonyms, compounds for which it was not possible to retrieve the structure, and compounds for which it was not possible to calculate the descriptors, 470 compounds were left. In addition, any chemicals present both in the training and the test set (28 chemicals) were removed. After extracting the compounds out of the applicability domain of the classification model, as illustrated in a following section, the final test set consisted of 343 compounds, shown in table 2.

The 28 overlapping compounds in the training set and the original test set were used to compare the classifications obtained by the two tests. In this comparison, 22 compounds showed a coherent activity, i.e. they were consistently classified by both assays, namely 17-β-estradiol, 4,4′-thiodiphenol, *bis* (4-hydroxyphenyl)methane, bisphenol A, 4,4′-dihydroxybenzophenone, benzophenone, genistein, naringenin, daidzein, flavone, 4,4′-biphenyldiol, 4-hydroxybiphenyl, biphenyl, 4-(*tert*-octyl)phenol, butyl-4-hydroxybenzoate, 4-(*tert*-pentyl)phenol, 4-*n*-pentylphenol, ethyl-4-hydroxybenzoate, 4-propylphenol, phenol, toluene, and chlorobenzene. In contrast, six compounds (*bis*(4-hydroxyphenyl)sulphone, biochanin A, phloretin, 2-hydroxybiphenyl, 4-octylphenol, and 4-methylphenol) were classified as active in the training set, and as inactive in the test set. This is due to the higher sensitivity of the method performed by Schultz *et al*. [35] compared to that of Nishihara *et al*. [36]. The correlation between experimental binding affinity and gene expression for the overlapping chemicals in the training and test set was about 80%. Thus, it was assumed that the factors (model parameters) driving ER binding and gene expression are the same or strongly related.

2.3 *Computational methods*

2.3.1 Molecular structures. For the training set, molecular structure details can be obtained from the literature [31]. For the test set, CAS numbers were retrieved from the published chemical names [36], and then SMILES were obtained by using EPISUITE [38]. In cases were no CAS numbers were available, the SMILES were compiled from the molecular structures drawn in TSAR [39].

2.3.2 Calculation of descriptors. The same descriptors found to be useful in the modelling part of the training set [14] were also calculated for the test set. Both the logarithm of the octanol-water partition coefficient ($\log K_{OW}$), and the number of hydrogen bond donor groups (n_{Hdon}) were calculated by using the TSAR software package for Windows [39]. However, for some molecules in the test set it was not possible to obtain estimated $\log K_{OW}$ values. In such cases, the EPISUITE software was used to fill in the missing values [38]. Since these programs are based on different algorithms for the $\log K_{OW}$ calculation, where calculated values from both programs were available, estimated values from TSAR were compared to those obtained

Table 2. The chemical names, Chemical Abstract Service (CAS) number (if available), logarithm of the octanol-water partition coefficient (log K_{OW}), number of hydrogen bond donor groups (n_{Hdon}), and observed and predicted binary classification activity (1 if considered active, 0 if considered inactive) for the test set (343 compounds). Only chemicals that were in the AD of the training set are shown.

ID	*Chemical name*	*CAS*	*log K_{OW}*	*n_{Hdon}*	*Binary activity (observed)*	*Binary activity (predicted)*
1	Dibutyl adipate	105-99-7	2.89	0	0	0
2	2-Butoxyethyl phthalate	117-83-9	3.59	0	0	0
3	Dimethyl phthalate	131-11-3	1.51	0	0	0
4	Di-*n*-pentyl phthalate	131-18-0	4.71	0	0	0
5	Di-*iso*-butyl adipate	141-04-8	2.91	0	0	0
6	Diethyl adipate	141-28-6	1.17	0	0	0
7	Di-*iso*-propyl adipate	6938-94-9	1.99	0	0	0
8	Dicyclohexyl phthalate	84-61-7	4.68	0	0	0
9	Diethyl phthalate	84-66-2	2.19	0	0	0
10	Di-*iso*-butyl phthalate (DIBP)	84-69-5	3.93	0	0	0
11	Di-*n*-Butyl phthalate (DBuP)	84-74-2	3.92	0	0	0
12	Dibutyl phthalate	84-74-2	3.92	0	0	0
13	Di-*n*-hexyl phthalate	84-75-3	5.51	0	0	0
14	4-Chloronitrobenzene	100-00-5	2.52	0	0	0
15	Terephtalic acid (TPA)	100-21-0	1.44	2	0	0
16	Ethyl benzene	100-41-4	2.91	0	0	0
17	Styrene	100-42-5	2.70	0	0	0
18	Benzylalcohol	100-51-6	1.51	1	0	0
19	Benzaldehyde	100-52-7	1.72	0	0	0
20	*N*-methylaniline	100-61-8	1.48	1	0	0
21	Phenylhydrazine	100-63-0	1.36	2	0	0
22	Diphenylmethane	101-81-5	4.13	0	0	0
23	trans-Stilbene	103-30-0	4.26	0	0	0
24	Dibenzyl ether	103-50-4	3.57	0	0	0
25	*N*-Ethylaniline	103-69-5	1.82	1	0	0
26	*n*-Butyl benzene	104-51-8	3.70	0	0	0
27	1,4-Diethylbenzene	105-05-5	3.77	0	0	0
28	4-Chlorotoluene	106-43-4	3.03	0	0	0
29	4-Chloroaniline	106-47-8	1.78	1	0	0
30	1,3-Diphenylpropane	1081-75-0	4.92	0	0	0
31	1,3,5-Trimethylbenzene	108-67-8	3.45	0	0	0
32	2-Methylpyridine	109-06-8	1.38	0	0	0
33	*o*-Tolidine	119-93-7	3.10	2	0	1
34	1,2,4-Trichlorobenzene	120-82-1	3.60	0	0	0

35	*N,N*-Dimethylaniline	121-69-7	1.84	0	0	0
36	Diphenylamine	122-39-4	3.16	1	0	1
37	1,2-Diethylbenzene	135-01-3	3.77	0	0	0
38	*N*-Phenyl-2-naphthylamine	135-88-6	4.16	1	0	1
39	2-Mercaptobenzothiazole	149-30-4	2.30	1	0	1
40	Neopentyl glycol dimethacrylate	1985-51-9	2.77	0	0	0
41	trans-1,2-Diphenylcyclobutane	20071-09-4	4.68	0	0	0
42	Diethylbenzene	25340-17-4	3.77	0	0	0
43	1,2,3-Trimethylbenzene	526-73-8	3.45	0	0	0
44	1,2-Dinitrobenzene	528-29-0	1.95	0	0	0
45	2,4-Dichloroaniline	554-00-7	2.30	1	0	1
46	Menadione	58-27-5	1.32	0	0	0
47	2-Ethyltoluene	611-14-3	3.38	0	0	0
48	3-Ethyltoluene	620-14-4	3.38	0	0	0
49	Aniline	62-53-3	1.26	1	0	0
50	cis-Stilbene	645-49-8	4.26	0	0	0
51	4-Toluenesulfonamide	70-55-3	1.03	1	0	0
52	*N*-Nitrosodiphenylamine	86-30-6	3.37	0	0	0
53	1,2,3-Trichlorobenzene	87-61-6	3.60	0	0	0
54	1-Chloro-2-nitrobenzene	88-73-3	2.52	0	0	0
55	*N*-Phenyl-1-naphthylamine	90-30-2	4.16	1	0	1
56	Quinoline	91-22-5	2.14	0	0	0
57	4-Amino butylbenzoate	94-25-7	2.20	1	0	1
58	1,2-Dichlorobenzene	95-50-1	3.08	0	0	0
59	2-Aminotoluene	95-53-4	1.73	1	0	0
60	1,2,4-Trimethylbenzene	95-63-6	3.45	0	0	0
61	2,4-Diaminotoluene	95-80-7	0.95	2	0	0
62	2,5-Dichloroaniline	95-82-9	2.30	1	0	1
63	1,2,4,5-Tetramethylbenzene	95-93-2	3.92	0	0	0
64	1,2-Epoxyethylbenzene	96-09-3	1.68	0	0	0
65	1-Chloro-2,4-dinitrobenzene	97-00-7	2.47	0	0	0
66	2,4-Dinitroaniline	97-02-9	1.17	1	0	0
67	4-*tert*-Butylbenzoic acid	98-73-7	3.37	1	0	1
68	Cumene	98-82-8	3.24	0	0	0
69	α-Methylstyrene	98-83-9	2.85	0	0	0
70	Nitrobenzene	98-95-3	2.00	0	0	0
71	4-Nitrotoluene	99-99-0	2.47	0	0	0
72	2,4-Diphenyl-1-butene	no CAS	4.86	0	0	0
73	1,3,5-Triethyltoluene	no CAS	5.10	0	0	0

(*continued*)

Table 2. Continued.

ID	*Chemical name*	*CAS*	*log K_{OW}*	*n_{Hdon}*	*Binary activity (observed)*	*Binary activity (predicted)*
74	4-*iso*-Propyl-3-methylphenol	3228-02-2	3.42	1	0	1
75	4-*n*-Nonylphenol	104-40-5	5.40	1	0	1
76	2,4-Dimethylphenol	105-67-9	2.70	1	0	1
77	Resorcinol	108-46-3	1.48	2	0	0
78	2,4,6-Tribromophenol	118-79-6	4.18	1	0	1
79	Catechol	120-80-9	1.48	2	0	0
80	4-Hydroxybenzaldehyde	123-08-0	1.44	1	0	0
81	Hydroquinone	123-31-9	1.48	2	0	0
82	Bisphenol-A-diglycidyl ether (BPA-GE)	1675-54-3	3.84	0	0	0
83	4-*n*-Heptylphenol	1987-50-4	4.61	1	0	1
84	Bisphenol-A-ethoxylate (BPA-E)	32492-61-8	3.50	2	0	1
85	2,2′-Dihydroxybiphenyl	4225-26-7	1.80	0	0	0
86	2,4-Dinitrophenol	51-28-5	1.67	1	0	0
87	3-Nitrophenol	554-84-7	1.72	1	0	0
88	2,5-Dichlorophenol	583-78-8	2.80	1	0	1
89	3-*tert*-Butylphenol	585-34-2	3.39	1	0	1
90	3-Aminophenol	591-27-5	0.98	2	0	0
91	2,4-Dibromophenol	615-58-7	3.35	1	0	1
92	2,4,6-Trichlorophenol	88-06-2	3.32	1	0	1
93	2-*tert*-Butylphenol	88-18-6	3.39	1	0	1
94	2-Nitrophenol	88-75-5	1.72	1	0	0
95	2-*sec*-Butylphenol	89-72-5	3.35	1	0	1
96	2-Methylphenol (*o*-cresol)	95-48-7	2.23	1	0	1
97	2,2-*Bis*[4-(2-hydroxy-3-methacryloxypropoxy)pheny] propane (*Bis*GMA)	no CAS	5.07	2	0	1
98	4-*tert*-Octylphenol polyethoxylate (5)	no CAS	4.27	1	0	1
99	4-Nonylphenol polyethoxylate (5)	no CAS	4.33	1	0	1
100	Antimony (III) chloride	10025-91-9	1.66	0	0	0
101	Triphenyltin (IV) chloride (TPT)	639-58-7	3.93	0	0	0
102	Tributyltin (IV) chloride (TBT)	98-51-1	4.14	0	0	0
103	Dicyclohexylamine	101-83-7	2.97	1	0	1
104	Diphenyl carbonate	102-09-0	3.85	0	0	0
105	1,2-Dibromoethane	106-93-4	1.71	0	0	0
106	Cyclohexylamine	108-91-8	0.97	1	0	0
107	Cyclohexanol	108-93-0	1.32	1	0	0
108	Cyclohexanone	108-94-1	1.53	0	0	0

109	Triethylene glycol dimethacrylate	109-16-0	1.28	0	0	0
110	*Bis*(2-chloroethyl) ether	111-44-4	1.42	0	0	0
111	*n*-Decyl alcohol	112-30-1	3.32	1	0	1
112	1-Tridecanol	112-70-9	4.51	1	0	1
113	*Tris*(2-chloroethyl) phosphate	115-96-8	2.67	0	0	0
114	Triethylamine	121-44-8	1.18	0	0	0
115	Chlorodibromomethane	124-48-1	1.55	0	0	0
116	Tributyl phosphate	126-73-8	4.18	0	0	0
117	Tetrachloroethylene	127-18-4	2.52	0	0	0
118	*n*-Butyl acrylate	141-32-2	1.82	0	0	0
119	1-Nonanol	143-08-8	2.92	1	0	1
120	Sodium lauryl sulfate (SDS)	151-21-3	1.69	0	0	0
121	Trimethylolpropane triacrylate	15625-89-5	2.49	0	0	0
122	Diethoxyleneglycol dimethacrylate	2358-84-1	1.45	0	0	0
123	2-Chloro-1,1,2-trifluoroethyl ethyl ether	310-71-4	1.91	0	0	0
124	Trimethylpropane trimethacrylate	3290-92-4	3.32	0	0	0
125	Camphorquinone	465-29-2	1.38	0	0	0
126	1,3-Dichloropropene	542-75-6	1.97	0	0	0
127	1,1,1,2-Tetrachloroethane	630-20-6	2.44	0	0	0
128	Diethyl sulfate	64-67-5	1.14	0	0	0
129	1-Butanol	71-36-3	0.94	1	0	0
130	Didecyldimethylammonium chloride	7173-51-5	4.66	0	0	0
131	Urethane dimethacrylate	72869-86-4	3.47	2	0	1
132	Bromoform	75-25-2	1.52	0	0	0
133	Bromodichloromethane	75-27-4	1.58	0	0	0
134	Dicyclopentadiene	77-73-6	2.17	0	0	0
135	*Tris*(butoxyethyl) phosphate	78-51-3	3.68	0	0	0
136	Isophorone	78-59-1	1.59	0	0	0
137	2-Methyl-1-propanol	78-83-1	0.95	1	0	0
138	Methyl methacrylate (MMA)	80-62-6	0.89	0	0	0
139	Hexachloro-1,3-butadiene	87-68-3	2.61	0	0	0
140	Nonoxynol iodide	9016-45-9	4.82	1	0	1
141	Coumaric acid	91-64-5	1.82	0	0	0
142	1,2,3-Trichloropropane	96-18-4	2.36	0	0	0
143	1,3-Dichloro-2-propanol	96-23-1	1.21	1	0	0
144	Monoethoxyleneglycol dimethacrylate	97-90-5	1.61	0	0	0
145	Butylated hydroxytoluene (BHT)	128-37-0	5.48	1	0	1

(*continued*)

Table 2. Continued.

ID	*Chemical name*	*CAS*	*log K_{OW}*	*n_{Hdon}*	*Binary activity (observed)*	*Binary activity (predicted)*
146	Butylated hydroxyanisole (BHA)	25013-16-5	3.14	1	0	1
147	Dexamethasone	50-02-2	1.71	3	0	0
148	Hydroxy-flutamide	52806-53-8	1.92	2	0	1
149	Benzoic acid	65-85-0	1.75	1	0	0
150	4-Hydroxybenzoic acid	99-96-7	1.46	2	0	0
151	Dodecylol polyethoxylate (3)	no CAS	4.58	1	0	1
152	Dodecylol polyethoxylate (5)	no CAS	4.25	1	0	1
153	PhlP	105650-23-5	2.13	1	0	1
154	Ferulic acid	1135-24-6	1.62	2	0	0
155	4-Aminobenzoic acid	150-13-0	0.96	2	0	0
156	Glycitein	40957-83-3	1.99	2	0	1
157	Glycitein	40957-83-3	1.99	2	0	1
158	Hinokitiol	499-44-5	0.87	1	0	0
159	Carvacrol	499-75-2	3.42	1	0	1
160	Tyramine	51-67-2	1.13	2	0	0
161	Testosterone	58-22-0	2.90	1	0	1
162	Tyrosine	60-18-4	0.87	3	0	0
163	Trp-P-2	62450-07-1	1.84	2	0	0
164	MelQx	77500-04-0	1.12	1	0	0
165	Thymol	89-83-8	3.42	1	0	1
166	Acetyleugenol	93-28-7	2.34	0	0	0
167	Eugenol	97-53-0	2.55	1	0	1
168	Isoeugenol	97-54-1	2.51	1	0	1
169	2-Aminoanthraquinone	117-79-3	1.66	1	0	0
170	Anthracene	120-12-7	4.05	0	0	0
171	Pyrene	129-00-0	4.37	0	0	0
172	8-Hydroxy benzo[*a*]pyrene	13345-26-1	5.09	1	0	1
173	9-Hydroxy fluorine	1689-64-1	3.00	1	0	1
174	9-Hydroxy benzo[*a*]pyrene	17573-21-6	5.09	1	0	1
175	Benzo[*e*]pyrene	192-97-2	5.37	0	0	0
176	Benzo[*b*]fluoranthene	205-99-2	5.37	0	0	0
177	Benzo[*k*]fluoranthene	207-08-9	5.37	0	0	0
178	5-Hydroxy benzo[*a*]pyrene	24027-84-7	5.09	1	0	1
179	4-Hydroxy-4′-monochlorobiphenyl	28034-99-3	3.96	1	0	1
180	6-Hydroxy benzo[*a*]pyrene	33953-73-0	5.09	1	0	1
181	4-Hydroxy benzo[*a*]pyrene	37574-48-4	5.09	1	0	1

182	7-Hydroxy benzo[*a*]pyrene	37994-82-4	5.09	1	0	1
183	1,6-Dinitropyrene	42397-64-8	4.28	0	0	0
184	1,8-dinitropyrene	42397-65-9	4.28	0	0	0
185	Benzo[*a*]pyrene	50-32-8	5.37	0	0	0
186	1-Hydroxy pyrene	5315-79-7	4.09	1	0	1
187	1-Nitropyrene	5522-43-0	4.32	0	0	0
188	Benz[*a*]anthracene	56-55-3	5.05	0	0	0
189	10-Hydroxy benzo[*a*]pyrene	56892-31-0	5.09	1	0	1
190	11-Hydroxy benzo[*a*]pyrene	56892-32-1	5.09	1	0	1
191	12-Hydroxy benzo[*a*]pyrene	56892-33-2	5.09	1	0	1
192	1,8-Dimethylnaphthalene	569-41-5	3.98	0	0	0
193	1,2-Dimethylnaphthalene	573-98-8	3.98	0	0	0
194	2,6-Dimethylnaphthalene	581-42-0	3.98	0	0	0
195	2-Nitrofluorene	607-57-8	3.72	0	0	0
196	2-Aminoanthracene	613-13-8	3.27	1	0	1
197	2-Hydroxydibenzofuran	86-77-1	2.48	1	0	1
198	3-Nitrofluoranthene	892-21-7	4.32	0	0	0
199	1-Methylnaphthalene	90-12-0	3.52	0	0	0
200	Naphthalene	91-20-3	3.05	0	0	0
201	4-Hydroxy-2′,3,5,5′-tetrachlorobiphenyl	no CAS	5.52	1	0	1
202	7-Hydroxy-1,2,3,6,8-pentachlorodibenzofuran	no CAS	5.07	1	0	1
203	3-Hydroxy-2,8-dichlorodibenzofuran	no CAS	3.52	1	0	1
204	9-Hydroxy-2,6-dichlorodibenzofuran	no CAS	3.52	1	0	1
205	6-Hydroxy-3,4-dichlorodibenzofuran	no CAS	3.52	1	0	1
206	9-Hydroxy-3,4-dichlorodibenzofuran	no CAS	3.52	1	0	1
207	Xylylcarb	2425-10-7	2.46	1	0	1
208	Chinomethionate	2439-01-2	2.48	0	0	0
209	Fenthoate (PAP)	2597-03-7	3.52	0	0	0
210	Mancozeb	8018-01-7	2.05	4	0	1
211	Chloropropham	101-21-3	2.79	1	0	1
212	Simetryne	1014-70-6	1.26	2	0	0
213	Heptachlor epoxide	1024-57-3	2.83	0	0	0
214	Chlorobenside	103-17-3	4.95	0	0	0
215	Carbendazim	10605-21-7	1.30	2	0	0
216	1,4-Dichlorobenzene	106-46-7	3.08	0	0	0
217	Propoxur	114-26-1	2.02	1	0	1
218	Endosulfan (benzoepin)	115-29-7	3.50	0	0	0

(*continued*)

Table 2. Continued.

ID	*Chemical name*	*CAS*	*log K_{OW}*	*n_{Hdon}*	*Binary activity (observed)*	*Binary activity (predicted)*
219	Fensulfothion	115-90-2	2.61	0	0	0
220	Malathion	121-75-5	1.84	0	0	0
221	Fenitrothion (MEP)	122-14-5	3.39	0	0	0
222	Simazine (CAT)	122-34-9	1.36	2	0	0
223	Maneb	12427-38-2	2.05	4	0	1
224	Kitazin (IBP)	13286-32-3	3.36	0	0	0
225	Captans (1,2,3,6-tetrahydro-*N*-(trichloromethylthio)phthalimide)	133-06-2	3.24	0	0	0
226	Thiram	137-26-8	3.38	0	0	0
227	Ziram	137-30-4	1.14	0	0	0
228	Propazin	139-40-2	2.19	2	0	1
229	Thiabendazole	148-79-8	1.74	1	0	0
230	Carbofuran	1563-66-2	1.57	1	0	0
231	Trifluralin	1582-09-8	4.25	0	0	0
232	Alachlor	15972-60-8	3.66	0	0	0
233	Malaoxon	1634-78-2	1.11	0	0	0
234	Edifenfos (EDDP)	17109-49-8	4.95	0	0	0
235	Nitrofen (NIP)	1836-75-5	4.47	0	0	0
236	Chlornitrofen (CNP)	1836-77-7	4.98	0	0	0
237	Isoxathion	18854-01-8	4.31	0	0	0
238	Chlorothalonil (TPN)	1897-45-6	3.85	0	0	0
239	Atrazine	1912-24-9	1.77	2	0	0
240	EPN (*O*-Ethyl-*O*-(4-nitrophenyl)phenyl-phosphono thioate)	2104-64-5	4.37	0	0	0
241	Metribuzin	21087-64-9	2.31	1	0	1
242	Cyanazine	21725-46-2	1.47	2	0	0
243	Molinate	2212-67-1	1.89	0	0	0
244	Tetrachlorovinphos	22248-79-9	4.05	0	0	0
245	Phosalone	2310-17-0	4.08	0	0	0
246	Thiophanate-methyl	23564-05-8	2.58	4	0	1
247	Propyzamide	23950-58-5	2.93	1	0	1
248	Triforine	26644-46-2	2.34	2	0	1
249	Tetrachlorofthalide	27355-22-2	3.58	0	0	0
250	Thiobencarb	28249-77-6	3.41	0	0	0
251	Primiphos-methyl	29232-93-7	3.34	0	0	0
252	Aldrin	309-00-2	3.55	0	0	0
253	α-Hexachlorocyclohexane (HCH, BHC)	319-84-6	4.65	0	0	0
254	Procymidone	32809-16-8	3.44	0	0	0

255	Linuron	330-55-2	2.42	1	0	1
256	Endosulfan (β-Benzoepin)	33213-65-9	3.50	0	0	0
257	Diazinon	333-41-5	4.27	0	0	0
258	Diflubenzuron	35367-38-5	3.33	2	0	1
259	Iprodione	36734-19-7	2.34	1	0	1
260	2-(1-methylpropyl)phenol methylcarbamate (BPMC)	3766-81-2	3.11	1	0	1
261	Pendimethalin	40487-42-1	4.01	1	0	1
262	Metamitron	41394-05-2	1.40	1	0	0
263	Bifenox	42576-02-3	4.20	0	0	0
264	Triadimefon	43121-43-3	4.55	0	0	0
265	Vinclozolin	50471-44-8	3.40	0	0	0
266	Isoprothiolane (IPT)	50512-35-1	1.99	0	0	0
267	Chlorobenzilate	510-15-6	4.30	1	0	1
268	Pretilachlor	51218-49-6	4.11	0	0	0
269	Triadimenol	55219-65-3	3.81	1	0	1
270	Fenthion (MPP)	55-38-9	3.53	0	0	0
271	Neburon	555-37-3	3.36	1	0	1
272	Chlorpyrifos-methyl	5598-13-0	4.28	0	0	0
273	Ethyl parathion (Parathion)	56-38-2	3.61	0	0	0
274	Dieldrin	60-57-1	2.67	0	0	0
275	γ-hexachlorocyclohexane (HCH, BHC)	608-73-1	4.65	0	0	0
276	Dimepiperate	61432-55-1	3.42	0	0	0
277	Dichlorvos (DDVP)	62-73-7	1.84	0	0	0
278	Carbaryl (NAC)	63-25-2	2.52	1	0	1
279	Cyhalothrin	68085-85-8	5.48	0	0	0
280	Fluazifop-butyl	69806-50-4	5.28	0	0	0
281	Propanil (DCPA)	709-98-8	2.56	1	0	1
282	Endrin	72-20-8	2.67	0	0	0
283	Methoxychlor	72-43-5	4.85	0	0	0
284	Mefenacet	73250-68-7	2.95	0	0	0
285	Bromobutide	74712-19-9	4.05	1	0	1
286	Heptachlor	76-44-8	3.67	0	0	0
287	Tefluthrin	79538-32-2	5.17	0	0	0
288	Toxaphene (camphechlor)	8001-35-2	5.47	0	0	0
289	Pentachloronitrobenzene (PCNB)	82-68-8	4.59	0	0	0
290	Esprocarb	85785-20-2	4.18	0	0	0
291	Pentachlorophenol (PCP)	87-86-5	4.35	1	0	1

(*continued*)

Table 2. Continued.

ID	*Chemical name*	*CAS*	*log K_{OW}*	*n_{Hdon}*	*Binary activity (observed)*	*Binary activity (predicted)*
292	Fenoprop	93-72-1	3.43	1	0	1
293	2,4,5-Trichlorophenoxyacetic acid (2,4,5-T)	93-76-5	2.89	1	0	1
294	2,4-dichlorophenoxyacetic acid (2,4-D)	94-75-7	2.37	1	0	1
295	Methidathion	950-37-8	3.20	0	0	0
296	2,4,5-Trichlorophenol	95-95-4	3.32	1	0	1
297	Endosulfan (α-Benzoepin)	959-98-8	3.50	0	0	0
298	1,2-Dibromo-3-chloropropane	96-12-8	2.49	0	0	0
299	Dicloran (AL-50)	99-30-9	2.25	1	0	1
300	Di-*n*-propyl phthalate	131-16-8	3.13	0	1	0
301	Di-*iso*-propyl phthalate	605-45-8	3.02	0	1	0
302	Benzylbutyl phthalate (BBP)	85-68-7	4.49	0	1	0
303	cis-1,2-Diphenylcyclobutane	7694-30-6	4.68	0	1	0
304	4-Bromophenol	106-41-2	2.55	1	1	1
305	4-Chlorophenol	106-48-9	2.28	1	1	1
306	2,4-Dichlorophenol	120-83-2	2.80	1	1	1
307	4-Ethylphenol	123-07-9	2.63	1	1	1
308	4-*n*-butylphenol	1638-22-8	3.42	1	1	1
309	4-*n*-Hexylphenol	2446-69-7	4.21	1	1	1
310	4-Chloro-3-methylphenol	59-50-7	2.75	1	1	1
311	2,2-*Bis*(4-hydroxy-3-methylphenyl)propane	79-97-0	5.26	2	1	1
312	4-(branched)-Nonylphenol (BNP)	84852-15-3	5.33	1	1	1
313	4-Chloro-3,5-xylenol	88-04-0	3.21	1	1	1
314	3,4-Dichlorophenol	95-77-2	2.80	1	1	1
315	4-*tert*-Butylphenol	98-54-4	3.39	1	1	1
316	4-*sec*-Butylphenol	99-71-8	3.35	1	1	1
317	4-Hydroxyacetophenone	99-93-4	1.07	1	1	0
318	4-*tert*-Octylphenol polyethoxylate (2)	no CAS	4.77	1	1	1
319	2,2-*Bis*(4-hydroxy-phenyl)butane	no CAS	4.72	2	1	1

320	4-Nonylphenol polyethoxylate (2)	no CAS	4.82	1	1	1
321	Diethylstilbestrol (DES)	56-53-1	4.79	2	1	1
322	17α-Ethynylestradiol	57-63-6	4.02	2	1	1
323	β-estradiol-17-acetate	6045-67-6	4.14	1	1	1
324	*n*-Propyl 4-hydroxybenzoate	94-13-3	2.30	1	1	1
325	Methyl 4-hydroxybenzoate	99-76-3	1.49	1	1	0
326	Coumestrol	479-13-0	2.24	2	1	1
327	Estriol	50-27-1	3.24	3	1	1
328	Apigenin	520-36-5	1.46	3	1	0
329	Dihydrotestosterone (DHT)	521-18-6	3.93	1	1	1
330	Estrone	53-16-7	4.54	1	1	1
331	Equol	531-95-3	3.22	2	1	1
332	17α-Estradiol	57-91-0	4.01	2	1	1
333	3-Hydroxy benzo[*a*]pyrene	13345-21-6	5.09	1	1	1
334	4-Hydroxy-2′,4′,6′-trichlorobiphenyl	14962-28-8	5.00	1	1	1
335	2-Hydroxy benzo[*a*]pyrene	56892-30-9	5.09	1	1	1
336	8-Hydroxy-3,4,6-trichlorodibenzofuran	no CAS	4.03	1	1	1
337	2-Hydroxy fluorene	no CAS	3.48	1	1	1
338	7-Hydroxy-3,4-dichlorodibenzofuran	no CAS	3.52	1	1	1
339	8-Hydroxy-2,3,4-trichlorodibenzofuran	no CAS	4.03	1	1	1
340	8-Hydroxy-3,4-dichlorodibenzofuran	no CAS	3.52	1	1	1
341	3,8-Dihydroxy-2-chlorodibenzofuran	no CAS	2.71	2	1	1
342	8-Hydroxy-3-monochlorodibenzofuran	no CAS	3.00	1	1	1
343	8-Hydroxy-2-monochlorodibenzofuran	no CAS	3.00	1	1	1

with EPISUITE. A correlation coefficient of 89% was calculated for the relationship between the $\log K_{OW}$ values generated by the two methods.

2.3.3 Classification model. The classification model was developed by using Formal Inference-based Recursive Modelling (FIRM) analysis as implemented in the TSAR software [39]. FIRM is a form of decision tree analysis able to model nonlinear relationships where a large set of data is split into subgroups based on relevant descriptors [40]. This process continues until there is no further meaningful way to split the nodes. Nodes which mainly contain active compounds point to molecular structures associated with activity, whereas the path leading to a node of inactive cases is indicative of non-active or inhibitory structural features. Although the method itself incorporates a variable selection procedure that allows multivariate splitting, in this study univariate splitting was performed on the basis of our previous experience in modelling a part of the current training set. More sophisticated probabilistic classification methods for modelling complex ligand-receptor interactions are also available. They are based on the simultaneous analysis of different parameters for each chemical using, for example, Bayesian trees [41].

The result of the FIRM analysis is a tree diagram or dendrogram that describes the splitting of the data set into smaller and more homogeneous subsets according to the predictor variables. The structure of the dendrogram shows which variables are important to explain the dependent variable. The decision diagram can also be used as a method of prediction for new chemicals with an unknown dependent variable. Following a path down the tree, the independent variables classify the new chemical into a terminal node. The mean of the response in that node can be used as the prediction.

In the tree diagram, each node is drawn as a box with text detailing the name and values/range of predictor variable, the number of data points in the node, mean, standard deviation and standard error of the response. For nodes which can be split further, the name of the predictor variable by which it is split and the statistical significance (p-value) is reported. A cut-off value of 0.5 was applied. Chemicals falling in a terminal node with mean response higher than 0.5 were classified as positive (active). Chemicals falling in a terminal node with mean response lower than 0.5 were classified as negative (inactive). In this study, a maximum bin number of 10 for continuous splitting was allowed. The p-value for merging was 0.05 and for splitting it was 0.06 (the default TSAR values). Any p-value below 0.05 indicates a statistically significant split.

The binary classification model using mechanistically interpretable descriptors for the training set of 117 chemicals has been developed to predict the presence or absence of activity. The classification tree is depicted in figure 1.

This figure shows that the initial group of 117 chemicals was split in three nodes, according to the value of the more important predictor variable, i.e. n_{Hdon}. The significance level of the split was measured by its p-value ($2.18E10^{-13}$), which is far below 0.05. Whereas the first node, corresponding to inactive compounds, was not split further, the other two nodes were both subdivided in two terminal nodes on the basis of the second predictor variable, $\log K_{OW}$. For the latter two nodes, the compounds falling into the lower range of $\log K_{OW}$ ($1.44 \leq \log K_{OW} \leq 1.76$ for $n_{Hdon} = 1$, and $0.28 \leq \log K_{OW} \leq 1.75$ for $n_{Hdon} \geq 2$) were classified as inactive. The summary statistics are also displayed in figure 1.

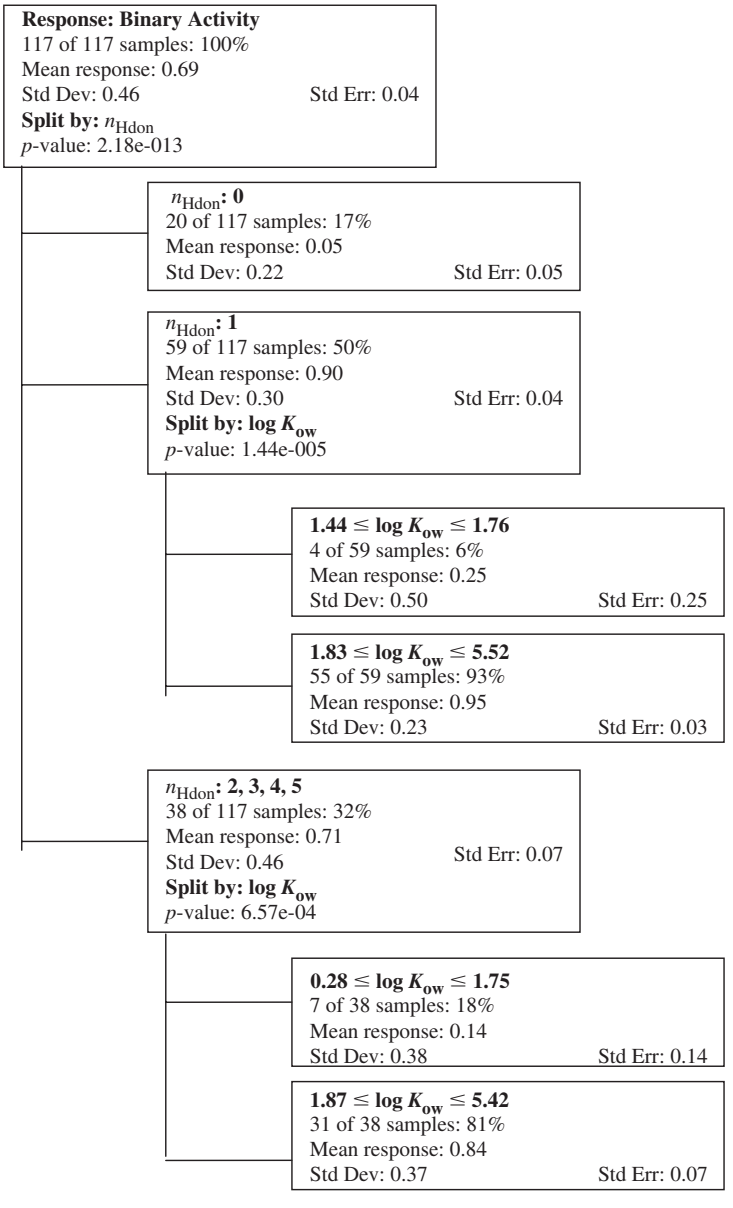

Figure 1. Classification model developed for prediction of estrogenic activity.

2.4 *Applicability domain characterisation*

The applicability domain (AD) was taken into account in order to consider the scope and limitations of the model, i.e. the range of chemical structures for which the model is considered to be applicable [42]. The model domain can be represented by ranges of molecular descriptors for training set compounds, that is, the descriptor space of the training set. More precisely, the AD could be determined as an optimum prediction

space [43]. This approach assesses the population density of chemicals within the training set to determine a probability density function.

The coverage of the training set in the model descriptor space was defined by using ranges of individual descriptors [14]. Within this statistically based method, the data distribution is assumed to be normal. Some limitations of this approach are that interior empty spaces are not detected and there is no correction for correlations between descriptors. It is assumed that interpolated estimates are more reliable than extrapolated ones. An interpolation region in one-dimensional descriptor space is the interval between the minimum and the maximum values of the training data set. In general, in an *n*-dimensional space, the interpolation region is an *n*-dimensional hyper-rectangle with sides parallel to the coordinate axes. In this study, a two-dimensional descriptor space is considered; thus, the minimum and maximum descriptor values define a rectangle in one plane.

In order to apply the range approach, the distributions of descriptors for both the training and test set were examined.

2.4.1 log K_{OW}. It can be observed from figure 2 that the log K_{OW} distribution for both the training and the test set is approximately normal. However, the minimum and maximum values for the chemicals in the test set are different from those of the training set. Thus, a large number of the test compounds lie in the extrapolation region of the model.

2.4.2 n_{Hdon}. The distribution of the number of hydrogen bond (HB) donors is close to normal for the training set but for the test set shows a heavy tailing towards the lower values of this descriptor (figure 3). Thus, for the whole data set, the distribution is heavily left-skewed and sharp. This distribution reflects the fact that many chemicals from the test set were inactive and do not have potential HB donor groups. Besides, it can be noted that a larger number of chemicals with more than four n_{Hdon} are present in the test set compared to the training set.

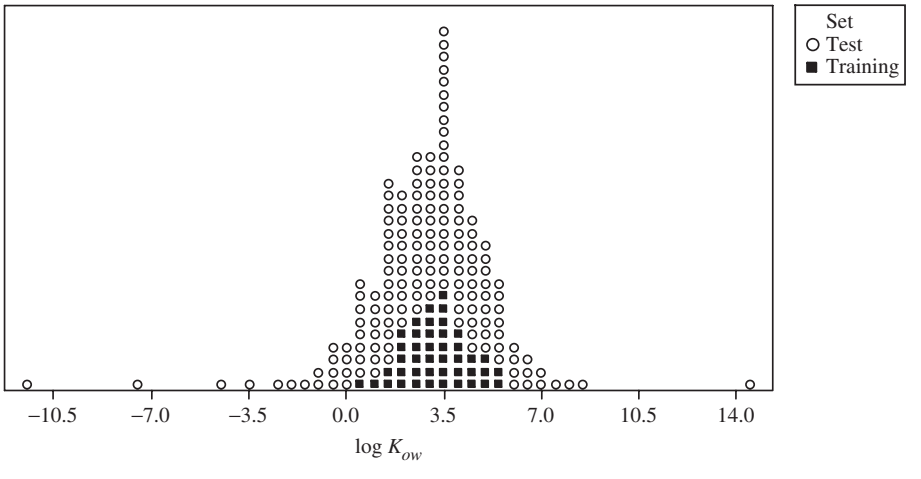

Figure 2. Dotplot of log K_{OW} for both the training and the validation test set.

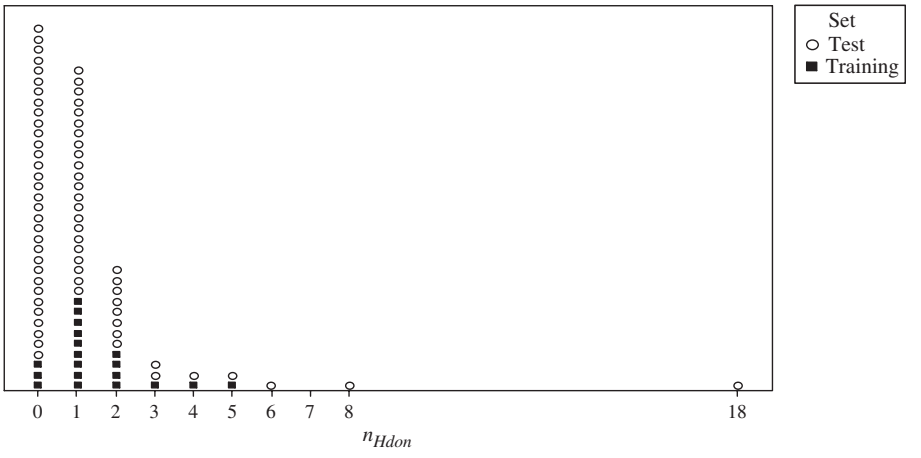

Figure 3. Dotplot of n_{Hdon} for both the training and the validation test set.

Table 3. Ranges of the descriptors used in the classification model for estrogenicity.

	Training set			*Test set*		
Range	*Minimum*	*Maximum*	*# cmpds out AD*	*Minimum*	*Maximum*	*# cmpds out AD*
$\log K_{\mathrm{OW}}$ (TSAR)	0.2796	5.5186		−11.70	14.65	
$\log K_{\mathrm{OW}}$ (TSAR) AD	0.8484	5.5186	1	0.8484	5.5186	96
N. HB Donors	0	5		0	18	
N. HB Donors AD	0	4	1	0	4	7

One chemical in the training set (morin) was considered to be out of the AD due to the outlying values of both descriptors ($\log K_{\mathrm{OW}} = 0.28$, and $n_{\mathrm{Hdon}} = 5$). A test chemical is considered to be out of the domain if at least one descriptor is out of the range defined by the descriptors of the training set. For the test set, 96 chemicals out of the $\log K_{\mathrm{OW}}$ range were eliminated (36 compounds with $\log K_{\mathrm{OW}}$ values between 5.52 and 14.65, and 60 compounds with $\log K_{\mathrm{OW}}$ between 0.85 and −11.70). In addition, 7 compounds with n_{Hdon} values higher than 4 were extracted from the test set. The ranges of the individual descriptors for the training and test set chemicals are indicated in table 3.

The scatterplot in figure 4 shows the coverage of the test set (empty circles) and the training set classification model (empty squares), visualised in the $\log K_{\mathrm{OW}}/n_{\mathrm{Hdon}}$ plane. The solid triangle indicates a chemical excluded from the training set for definition of the AD (morin), whereas the X's represent the test set chemicals out of the AD (99 chemicals). Figure 5 shows the coverage of the training set and part of the test set considered in the domain of the classification model, in the $\log K_{\mathrm{OW}}/n_{\mathrm{Hdon}}$ plane.

2.5 *Performance of the model: Goodness of fit, robustness and predictivity*

The evaluation of the performance of a CM can be assessed in terms of the Cooper statistics [44, 45]. The goodness-of-fit Cooper statistics for a CM can be defined

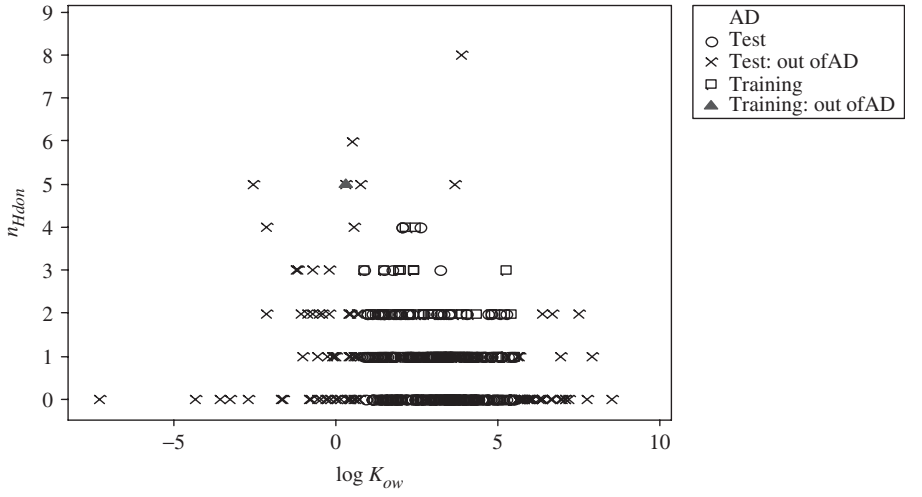

Figure 4. Scatterplot showing the coverage of the external test set (empty circles) and the training set classification model (empty squares), visualised in the log K_{OW}/n_{Hdon} plane. The solid triangle indicates the chemical excluded from the training set for definition of the AD (morin), whereas the ex-es represent the test set chemicals out of the AD. Three compounds (fenbutatin oxide, chlorhexidine gluconate, and EDTA) were excluded from the graph due to their extreme values for either log K_{OW} or n_{Hdon}.

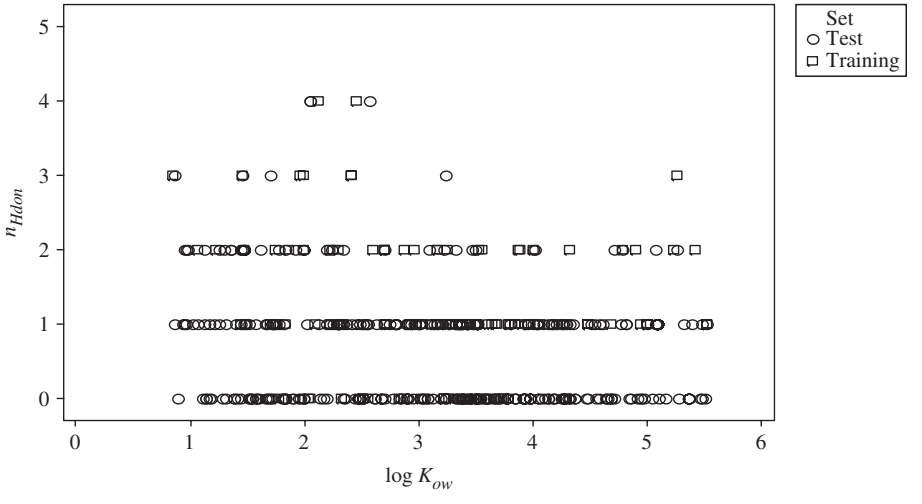

Figure 5. Scatterplot showing the overlap between the training set chemicals and those chemicals from the test set that were considered in the AD of the classification model according to the selected method (ranges in the descriptor space).

according to the results in the contingency matrix [46], where the rows represent the classification according to the reference reporter gene assay, and the columns represent the predicted classes assigned by the CM. The main diagonal represents the chemicals correctly classified into each class, while the non-diagonal cells represent the

Table 4. Cooper statistics of the classification model for the training set chemicals, for the averaged results from the leave-many-out cross-validation, and for the external test set.

Cooper statistic	*Training set classification model*	*Leave-many-out cross-validation*	*Validation test set*
Sensitivity	78/81 (96.30%)	70/73 (95.89%)	37/44 (84.09%)
Specificity	28/36 (77.78%)	25/32 (78.13%)	206/299 (68.90%)
Accuracy or concordance	106/117 (90.60%)	100/105 (90.48%)	243/343 (70.85%)
False positive rate	8/36 (22.22%)	7/32 (21.88%)	93/299 (31.10%)
False negative rate	3/81 (3.70%)	3/73 (4.11%)	7/44 (15.91%)
Positive predictivity	78/86 (90.70%)	70/77 (90.91%)	37/130 (28.46%)
Negative predictivity	28/31 (90.32%)	25/28 (89.29%)	206/213 (96.71%)

misclassifications. The last column reports the number of objects belonging to each class, whereas the last row reports the total number of objects assigned to each class according to the CM.

Cooper statistics express the ability of a CM to detect known active compounds (*sensitivity*), non-active compounds (*specificity*), and all chemicals in general (*concordance or accuracy*). The *false positive* and *false negative* rates can be calculated from the specificity and sensitivity, respectively. The *positive* and *negative classification rates* focus more on the effect of individual chemicals, since they are conditional probabilities. Thus, the positive classification rate is the probability that a chemical classified as active is really active, while the negative classification rate gives the probability that a chemical classified as inactive chemical is really inactive.

The robustness of the model was assessed by introducing a systematic perturbation in the training set. A 10-fold *leave-many-out* (LMO) cross-validation procedure with random selection of 10% chemicals of the training set excluded during each run was performed. The proportion of active ($n = 8$) and inactive chemicals ($n = 4$) was kept proportional to the original training test proportions for each run.

The model predictivity, i.e. ability to predict the response for new chemicals with unknown responses, was evaluated by using compounds that were not used in the model development. The statistical parameters used to evaluate the goodness-of-prediction are the same ones used for evaluating goodness-of-fit, i.e. the model capability to fit the data of the training set. The performance of the model is summarised in table 4.

The Cooper statistics based on the training set indicated an accuracy of 90.6% and a high value of sensitivity (96.3%). However, the specificity (77.8%) was lower than the sensitivity. To assess the robustness, the model was developed again ten times with the same descriptors for the smaller number of chemicals and the activities of the excluded ones were predicted. This trial led to a 95.9% correct prediction of active compounds and a 78.1% correct prediction of inactive compounds, which is comparable to the classification performance of the original model. Finally, the predictive power of the model was evaluated by using the independent external test set. Predictions were done only for those chemicals that were within the model AD, defined as coverage in the model descriptor space. Based on this test set, the accuracy is slightly lower (70.9%) than for the training set, due to decreases in both sensitivity (84.1%) and specificity (68.9%). Regarding the goodness-of-fit parameters, it should be noted that the classification accuracy usually assumes a balance between the positives and negative

cases. If the number of negative cases exceeds the number of positive ones, it is expected that there will be higher percentage of correctly predicted active than correctly predicted inactive chemicals.

3. Discussion

(Q)SARs are being increasingly used as tools for the prediction of environmental and human health endpoints. However, to be acceptable for decision making and regulatory purposes, questions concerning their reliability are likely to be raised. For quantitative regression models, the application of the OECD principles for (Q)SAR validation has been well illustrated [47, 48], while guidance on their application to classification-based models is still scarce. The aim of this work was to develop and validate a classification model, emphasising its internal performance and external predictivity, thereby illustrating the application of the OECD validation principles to a CM.

The CM developed in this study was evaluated according to the five OECD Principles for (Q)SAR validation and the observations are summarised below.

According to the first OECD principle, a (Q)SAR should be associated with a defined endpoint. Although there is no general agreement on the most meaningful endpoint to predict estrogenicity, commonly used approaches consist of the study of ligand-receptor binding interactions or ligand-induced gene activation [25]. Although the use of a different protocol to measure the endpoint could be questioned, the paucity of publicly available data and the high concordance in the binary prediction of estrogenic potential lend support to this approach. It has been noted in comparative studies [49] that these assays show slight differences. The former is more sensitive than the latter. This difference was also observed in the present study and could explain partially the lower accuracy of the CM for the chemicals in the test set compared to the training set.

The unambiguity of the Formal Inference-based Recursive Modelling (FIRM) algorithm used to derive the classification model was addressed by showing the explicit dendrogram which illustrates the decision rules for the splitting of chemicals into active and inactive. This model is transparent and reproducible.

The domain of applicability was defined in terms of the ranges of individual descriptors considered to be relevant for the endpoint. The range method is the simplest method for defining the coverage of the training and the test sets. Even if the chemicals from the both sets have the same coverage, there is no guarantee that all structural features in the test set were known from the training set. In this study, the structural similarity in terms of functional groups and their spatial orientation was omitted but its consideration could improve the assessment of predictivity.

The performance of the model was evaluated in several steps to fulfil the fourth OECD principle. The goodness-of-fit of the model was evaluated by using Cooper statistics derived from the training set. The robustness was assessed by LMO validation performed several times to minimise the error from the random selection of the excluded groups of chemicals. Finally, the predictive power was checked by applying the model to an external test set. According to the defined domain of applicability, the test set chemicals considered to be out of the AD were excluded from the test set to avoid unreliable extrapolations. Predictions were made only for those chemicals within the domain defined by the selected method. In general, the acceptance criteria for a CM should be defined taking into account the quality of the predictor and response data

as well as the purpose of the model. For stand-alone classification models, the Cooper statistics should be significantly greater than 50%. This requirement is clearly met for the CM.

Taking into account the results for the external test set, it was observed that the relatively high sensitivity (84.1%) is associated with a high false positive rate (31.1%). The CM is good at identifying known active compounds, but this is at the expense of over-classifying known inactive compounds. Besides, given a fixed sensitivity and specificity, the positive and negative classification values also vary according to the prevalence or proportion of chemicals in the population. Thus, the positive predictivity (28.5%) is much lower than the negative predictivity (96.7%). The overall accuracy is influenced by the performance on the bigger class.

In order to evaluate the relative importance of the results, they were compared to the scenario characterised by the absence of a model. This means that all the objects are assigned to the class most represented among the ones compared (in the present case, the inactive class). The value for this reference condition, called *no-model*, is unique and independent of the classification method adopted, and in this case was 12.8%. As the goodness-of-fit was far better than the *no-model* scenario, the CM can be regarded as statistically significant

Finally, the mechanistic interpretation of the model corroborates the hypothesis for the gene expression mechanism. Two descriptors with clear physical meaning (the logarithm of the octanol-water partition coefficient – $\log K_{OW}$, and the number of hydrogen bond donor groups – n_{Hdon}) were selected by using the FIRM approach. The $\log K_{OW}$ descriptor accounts for hydrophobicity, which was found to be an important predictor of estrogenic activity in a number of studies [50]. The n_{Hdon} accounts for the number of electron-donating groups, which could be correlated with the number of phenol groups in the molecule. These descriptors used to derive the model clearly corroborate the empirical findings according to which the hydrophobicity and the presence of phenolic groups play a central role in determining the estrogenic activity.

In a further phase, a more sophisticated and context related model could be derived by using three-dimensional descriptors, which are considered to be significant for modelling receptor mediated endpoints. In addition, the validation of the model with an external test set based on the same endpoint would be valuable. It would be also interesting to study the performance of the model on the external validation set at different activity thresholds. However, although this approach could be used to analyse differences in experimental protocols in more depth, it was not considered in this study due to the lack of numerical activity values for some compounds.

4. Conclusions

This paper provides an example of iterative model development and external validation by using additional experimental data. The model was developed by using a data set which was previously split to allow external validation. The new model was also evaluated according to the OECD principles for (Q)SAR validation.

The transparent and interpretable classification model for predicting the presence or absence of estrogenic activity was based on only two descriptors that have a clear physico-chemical meaning. The descriptors were selected empirically based on previous experience. The type of descriptors used and the form of the classification tree allowed

mechanistic interpretation of the model developed in this study. The model demonstrated robustness in LMO validation, and a good predictivity of an external set of chemicals. One drawback of the external validation is the fact that although well defined, the training and test set chemicals are based on slightly different endpoints. Nevertheless, the overall accuracy of prediction for the external test set clearly demonstrated that the model predicts much better than the chance (no-model situation).

Acknowledgements

A. Gallegos Saliner acknowledges receipt of a Joint Research Centre postdoctoral fellowship (contract number 22575-2004-12 P1B30 ISP IT).

References

[1] K.M. Lai, M.D. Scrimshaw, J.N. Lester. *Crit. Rev. Toxicol.*, **32**, 113 (2002).
[2] E. Hood. *Env. Health Persp.*, **113**, 670 (2005).
[3] R.J. Kavlock, G.P. Daston, C. DeRosa, P. Fenner-Crisp, L.E. Gray, S. Kaattari, G. Lucier, M. Luster, M.J. Mac, C. Maczka, R. Miller, J. Moore, R. Rolland, G. Scott, D.M. Sheehan, T. Sinks, H.A. Tilson. *Environ. Health Persp.*, **104**, 715 (1996).
[4] U.S. Code. Food Quality Protection Act: PL 104 (1996) and the U.S. Code. 1996. Safe Water Drinking Act: PL 104 (1996).
[5] http://www.epa.gov/scipoly/oscpendo (last accessed 9 March 2006).
[6] U.S. Environmental Protection Agency. Endocrine Disruptor Priority-Setting Database (EDPSD v.2 Beta) Eastern Research Group.
[7] EC. *Communication from the (European) Commission to the Council and the European Parliament on the Community Strategy for Endocrine Disrupters – a Range of Substances Suspected of Interfering with the Hormone Systems of Humans and Wildlife* (COM(1999)706). Brussels, Belgium (1999).
[8] CSTEE (Scientific Committee for Toxicity, Ecotoxicity and the Environment). *Opinion on Human and Wildlife Health Effects of Endocrine Disrupting Chemicals, with Emphasis on Wildlife and on Ecotoxicology Test Methods* (1999). http://europa.eu.int/comm/health/ph_risk/committees/sct/documents/out37_en.pdf (last accessed 9 March 2006).
[9] http://europa.eu.int/comm/environment/endocrine (last accessed 9 March 2006).
[10] EC. *Communication from the (European) 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). Brussels, Belgium (2001).
[11] EC. *Communication from the (European) Commission on a European Environment and Health Strategy* (COM(2003)338). Brussels, Belgium (2003).
[12] OECD. *Final Report of the Sixth Meeting of the Task Force on Endocrine Disrupters Testing and Assessment (EDTA 6)*, (ENV/JM/TG/EDTA/M(2003)4). Paris, France (2003).
[13] OECD. *Final Report of the 2nd Meeting of the Validation Management Group for Non-Animal Testing (VMG NA) of the Task Force on Endocrine Disrupters Testing and Assessment (EDTA)*, (ENV/JM/TG/EDTA/M(2005)2). Paris, France (2005).
[14] T.I. Netzeva, A. Gallegos Saliner, A.P. Worth. *Environ. Toxicol. Chem.* in press (2006).
[15] EC. *Proposal for a Regulation of the European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency and amending Directive 1999/45/EC and Regulation (EC) {on Persistent Organic Pollutants}*, (COM(2003)0644). Brussels, Belgium (2003).
[16] K.A. Thayer, R. Melnick, K. Burns, D. Davis, J. Huff. *Env. Health Persp.*, **113**, 1271 (2005).
[17] ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods). *Expert Panel Evaluation of the Validation Status of In Vitro Test Methods for Detecting Endocrine Disruptors* (2002).
[18] A.P. Worth, C.J. van Leeuwen, T. Hartung. *SAR QSAR Environ. Res.*, **5–6**, 331 (2004).
[19] http://europa.eu.int/comm/enterprise/reach/overview.htm (last accessed 9 March 2006).

[20] CEFIC (European Chemical Industry Council). *(Q)SARs for Human Health and the Environment – Workshop on Regulatory Acceptance* (2002).
[21] OECD. *Annexes to the Final Report on Principles for Establishing the Status of Development and Validation of (Quantitative) Structure-Activity Relationships [(Q)SARs]*, (ENV/JM/TG(2004)27/ANN). Paris, France (2004).
[22] SPORT (Strategic Partnership on REACH Testing). *Report of the pilot project: The SPORT Report: Making REACH work in practice*. http://www.sport-project.info (last accessed 9 March 2006).
[23] A.P. Worth, A. Bassan, A. Gallegos, T.I. Netzeva, G. Patlewicz, M. Pavan, I. Tsakovska, M. Vracko. *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*. EUR technical report 21866. Ispra, Italy (2005). http://ecb.jrc.it/DOCUMENTS/QSAR/QSAR_characterisation_EUR_21866_EN.pdf (last accessed 9 March 2006).
[24] P. Schmieder, O. Mekenyan, S. Bradbury, G. Veith. *Pure Appl. Chem.*, **75**, 2389 (2003).
[25] H. Fang, W. Tong, W.J. Welsh, D.M. Sheehan. *J. Mol. Struct. (Theochem)*, **622**, 113 (2003).
[26] M.Y. Mizutani, N. Tomioka, A. Itai. *J. Mol. Biol.*, **243**, 310 (1994).
[27] W. Tong, R. Perkins, L.I. Xing, W.J. Welsh, D.M. Sheehan. *Endocrinology*, **138**, 4022 (1997).
[28] S.J. Yu, S.M. Keenan, W. Tong, W.J. Welsh. *Chem. Res. Toxicol*, **15**, 1229 (2002).
[29] 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**, 669 (1998).
[30] P.K. Schmieder, A.O. Aptula, E.J. Routledge, J.P. Sumpter, O.G. Mekenyan. *Environ. Toxicol. Chem.*, **19**, 1727 (2000).
[31] A. Gallegos Saliner, L. Amat, R. Carbó-Dorca, T.W. Schultz, M.T.D. Cronin. *J. Chem. Inf. Comput. Sci.*, **43**, 1166 (2003).
[32] W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang, R. Perkins. *Environ. Health Perspect.*, **112**, 1249 (2004).
[33] T.W. Schultz, G.D. Sinks, M.T.D. Cronin. *Environ. Toxicol.*, **17**, 14 (2002).
[34] E.J. Routledge, J.P. Sumpter. *Environ. Toxicol. Chem.*, **15**, 241 (1996).
[35] T.W. Schultz, G.D. Sinks, M.T.D. Cronin. *Environ. Toxicol. Chem.*, **19**, 2637 (2000).
[36] T. Nishihara, J. Nishikawa, T. Kanayama, F. Dakeyama, K. Saito, M. Imagawa, S. Takatori, Y. Kitagawa, S. Hori, H. Utsumi. *J. Health Sci.*, **46**, 282 (2000).
[37] J. Nishikawa, K. Saito, J. Goto, M. Matsuo, T. Nishihara. *Jpn. J. Toxicol. Environ. Health*, **44**, P–32 (1998).
[38] KOWWIN v1.67. U.S. Environmental Protection Agency (2000).
[39] TSAR v3.3. Oxford Molecular Ltd., Oxford, UK (2000).
[40] D.M. Hawkins, S.S. Young, A. Rusinko III. *Quant. Struct.-Act. Relat.*, **16**, 296 (1997).
[41] O. Mekenyan, N. Nikolova, P. Schmieder, G. Veith. *QSAR Comb. Sci.*, **23**, 5 (2004).
[42] T.I. Netzeva, A.P. Worth, A. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang. *ATLA*, **33**, (2004).
[43] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, O. Mekenyan. *J. Chem. Inf. Model.*, **45**, 839 (2005).
[44] A.P. Worth, M.T.D. Cronin. *ATLA*, **29**, 447 (2001).
[45] J.A. Cooper, R. Saracci, P. Cole. *Brit. J. Cancer*, **39**, 87 (1979).
[46] L. Breiman, J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees Wadsworth International Group, Belmont, CA (1984).
[47] P. Pavan, A.P. Worth, T.I. Netzeva. *Comparative Assessment of QSAR Models for Aquatic Toxicity*. EUR technical report 21750. Ispra, Italy (2005). http://ecb.jrc.it/DOCUMENTS/QSAR/Report_Comparative_assessment_QSAR_models.pdf (last accessed 9 March 2006).
[48] P. Pavan, A.P. Worth, T.I. Netzeva. *Preliminary Analysis of an Aquatic Toxicity Dataset and Assessment of QSAR models for Narcosis*. EUR technical report 21749. Ispra, Italy (2005). http://ecb.jrc.it/DOCUMENTS/QSAR/Report_QSAR_model_for_narcosis.pdf (last accessed 9 March 2006).
[49] M. Nakai. *Receptor Binding Assay and Reporter Gene Assay of Medaka*. Chemical Evaluation and Research Institute, Japan. http://www.env.go.jp/chemi/end/pdfs/e06_chapter2.pdf (last accessed 9 March 2006).
[50] P.S. Danielian, R. White, J.A. Lees, M.G. Parker. *EMBO J.*, **11**, 1025 (1992).