

Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses

Natalja Fjodorova · Marjan Vračko · Marjan Tušar ·
Aneta Jezierska · Marjana Novič · Ralph Kühne ·
Gerrit Schüürmann

Received: 29 April 2009 / Accepted: 26 July 2009 / Published online: 15 August 2009
© Springer Science+Business Media B.V. 2009

Abstract The new European chemicals regulation Registration, Evaluation, Authorization and Restriction of Chemicals entered into force in June 2007 and accelerated the development of quantitative structure–activity relationship (QSAR) models for a variety of endpoints, including carcinogenicity. Here, we would like to present quantitative (continuous) and qualitative (categorical) models for non-congeneric chemicals for prediction of carcinogenic potency. A dataset of 805 substances was obtained after a preliminary screening of findings of rodent carcinogenicity for 1,481 chemicals accessible via Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network originated from the Lois Gold Carcinogenic Potency Database (CPDB). Twenty seven two-dimensional MDL descriptors were selected using Kohonen mapping and principal component analysis. The counter propagation artificial neural network (CP ANN) technique was applied. Quantitative models were developed

exploring the relationship between the experimental and predicted carcinogenic potency expressed as a tumorigenic dose TD_{50} for rats. The obtained models showed low prediction power with correlation coefficient less than 0.5 for the test set. In the next step, qualitative models were developed. We found that the qualitative models exhibit good accuracy for the training set (92%). The model demonstrated good predicted performance for the test set. It was obtained accuracy (68%), sensitivity (73%), and specificity (63%). We believe that CP ANN method is a good *in silico* approach for modeling and predicting rodent carcinogenicity for non-congeneric chemicals and may find application for other toxicological endpoints.

Keywords Counter propagation artificial neural network · *In silico* · Quantitative structure–activity relationship · Qualitative (categorical) models · Quantitative (continuous) models · Rodent carcinogenicity · Tumorigenic dose TD_{50}

Electronic supplementary material The online version of this article (doi:10.1007/s11030-009-9190-4) contains supplementary material, which is available to authorized users.

N. Fjodorova (✉) · M. Vračko · M. Tušar · A. Jezierska · M. Novič
National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana,
Slovenia
e-mail: natalja.fjodorova@ki.si

A. Jezierska
Faculty of Chemistry, University of Wrocław, 14 F. Joliot-Curie,
50-383 Wrocław, Poland

R. Kühne · G. Schüürmann
UFZ Department of Ecological Chemistry, Helmholtz Centre
for Environmental Research, Permoserstr. 15, 04318 Leipzig,
Germany

G. Schüürmann
Institute for Organic Chemistry, Technical University Bergakademie
Freiberg, Leipziger Strasse 29, 09596 Freiberg, Germany

Abbreviations

REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
QSAR	Quantitative structure–activity relationship
DSSTox	Distributed Structure-Searchable Toxicity
CPDB	Lois Gold Carcinogenic Potency Database
PCA	Principal component analysis
CP ANN	Counter propagation artificial neural network
TD_{50}	Tumorigenic dose

Introduction

Carcinogenicity and mutagenicity of chemical substances are widely discussed issues with respect to their potential negative effects on human health [1]. According to the newly implemented European's Registration, Evaluation,

Authorization and Registration of Chemicals directive (REACH), the toxicological information shall be submitted for the registration or authorization of chemicals. The identification of chemical mutagens and carcinogens is of high priority within the EU and other countries. More details about the carcinogenicity hazard assessment can be found in OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 35 [2].

Chemicals are defined as carcinogenic if they induce tumors, increase tumor incidence, or shorten the time to tumor occurrence. Carcinogenicity is a very complex biological process that includes at the first stage, a possible metabolic transformation of a chemical followed by interaction with biomolecules like DNA or proteins which may end in the malfunctioning of cells. Basically, carcinogens may be divided into two categories: genotoxic and non-genotoxic. Genotoxic carcinogens (GCs) initiate carcinogenesis by direct interaction with DNA, resulting in DNA damage or chromosomal aberrations that can be detected by genotoxicity tests. In contrast, non-genotoxic carcinogens (NGCs) exert their carcinogenic effects through a variety of so-called epigenetic mechanisms [3,4].

The data for carcinogen identification include human epidemiologic studies, long-term bioassays in experimental animals, and other relevant data on toxicokinetics and cancer mechanisms including in vitro methods. Transgenic assays might be used as additional source of information. Each source of data has a role in the overall assessment. An integrated testing strategy for carcinogenicity in details is provided in report of Defra REACH project [5]. The use of in vitro data is described in more detail in Ref. [6] and mini review [7].

It should be highlighted that a discussion is raised on whether animal data are adequate to estimate hazards in humans. To answer questions about the similarity of response between animals and humans, studies of toxicokinetics and mechanisms of actions have been employed [8–12]. However, it has become a regulatory principle that, without human data, only data from animal bioassays are acceptable as definitive evidence of carcinogenicity. Substances that induce tumors in animals are considered as presumed or suspected human carcinogens until convincing evidence to the contrary is presented [13].

At the current stage of research, the majority of our knowledge on carcinogenicity relies on the data generated from rodent's carcinogenicity assays. Toxicity and carcinogenicity studies should be conducted according to OECD Guidelines for testing of chemicals, so-called Test Guidelines (TGs). Consideration should be given in carrying out a combined chronic toxicity and carcinogenicity study (TG 453), rather than separate execution of a chronic toxicity study (TG 452) and carcinogenicity study (TG 451) [14–16]. The above-mentioned animal experiments count to the most expensive tests. First,

the duration of a single test takes a life span of animals, i.e., 18–36 months. Second, to get statistically significant results, a large number of animals must be killed that rises the ethical concerns. Increasing the social, economic pressures to reduce the use of animal testing and existence of large databases on rodent carcinogenicity appear as important reasons to develop “in silico” carcinogenicity models [17,18]. Thus, in silico approaches are considered to be an alternative solution to animal testing and widely used not only in risk assessment by different regulation bodies, but also, e.g. in drug discovery [19–25].

In silico models are divided into three groups: quantitative structure–activity relationship models (QSARs), qualitative structure–activity relationship models (SARs) and expert systems. All the above mentioned approaches can be applied for carcinogenicity prediction. SARs belong to qualitative approaches, often occurring in the form of structural alerts, which are related to the presence or absence of biological activity. Under computer supported alternative methods, SARs play an important role [1,26].

Quantitative structure–activity relationships are quantitative in nature, producing categorical or continuous prediction scale. The evaluation of non-commercial QSARs for mutagenicity and carcinogenicity is reported by Benigni et al. [27]. Many in silico models for prediction of carcinogenicity were developed (for review articles see [18,28–30]). Some QSARs models were developed for distinct chemical classes (models for congeneric chemicals), for example, the models for prediction of amines, nitro compounds, polycyclic aromatic hydrocarbons [26,31–36]. Generally, the topic of QSARs for individual classes of compounds was assessed by Passerini [37] and reported by Benigni [38]. The advantage of models based on chemical class is that they have a strong mechanistic basis that leads to better interpretation of predictions. Their disadvantage is that they are models with only a limited applicability domain which is uncomfortable for regulatory uses as many complementary models should be supplied to cover all chemical classes. Models for non-congeneric chemicals based on heterogeneous databases are developed and described in the literature as well [39–41]. A large number of expert systems such as MultiCASE, OncoLogic, and PASS exist to predict carcinogenicity. Some approaches dedicated for prediction only carcinogenicity and others integrate different endpoints [42–46]. Different issues for prediction of carcinogenicity are presented by regulatory bodies, research centers, and laboratories [47–55]. In carcinogenicity study, the endpoint must be well-defined. Some models use well-defined in vitro measurements of carcinogenicity potency and mutagenicity [56–61]. The biggest rodent carcinogenicity Gold data base gives the tumourigenic dose (TD₅₀) as the continuous variable to measure the carcinogenic potency [62]. In some cases, the *No Observed Effect Concentration* (NOEC) can be used as alternative measure for carcinogenic potency [63].

In this article, we present models built with Counter Propagation Artificial Neural Network (CP ANN). The CP ANN is described in many textbooks [64–69]. In the section below, some basic ideas are presented. We have developed carcinogenic potency prediction models for 805 non-congeneric chemicals using CP ANN technique in the scope of European project Computer Assisted Evaluation of industrial chemical Substances According to Regulations (CAESAR) [70].

It is to emphasize that qualitative and quantitative data can be used in the development of models depending on response. Usually, one talks about quantitative models when the potency is expressed as continuous variable like tumorigenic dose TD_{50} . Alternatively, the potency can be expressed only as ranking information. For example, the compounds can be classified as positive and negative. Models built with such data are usually addressed as qualitative ones. Both kinds of models are considered in this article. In our first trials, we have applied the CP ANN for modeling of continuous data using TD_{50} . Some findings concerned the log TD_{50} distribution are discussed in this article. Numerous quantitative models were built and optimized; however, none answers the validation criteria suitable for regulation. Using the categorical data (yes/no), we have built classification models with good prediction power. Our final model should serve as a useful tool for the preliminary ranking and prioritization of chemicals for carcinogenicity, as required by REACH.

Data

Database

The investigated carcinogenicity dataset contains 805 chemicals which were extracted from the database of 1,481 substances taken from Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network (http://www.epa.gov/nccet/dsstox/sdf_cpdbas.html) derived from the Lois Gold Carcinogenic Database (CPDB) (<http://potency.berkeley.edu/>)—the biggest rodent carcinogenicity studies collection. Here, the carcinogenic potency is expressed as tumorigenic dose TD_{50} which was measured according to standard protocol. We took into consideration only rat-derived data (1,186 compounds) because such data are often considered to be more suitable for human carcinogenicity prediction. (The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) agreed that, in certain circumstances, only one long-term rodent bioassay is required. This should be preferably be in the rat as all known human carcinogen are also positive in rats.)

We focused only on well-defined organic compounds therefore mixtures, polymers, inorganic compounds, metalloorganic compounds, salts, complexes and compounds with

out well-defined structure were excluded. Furthermore, the composed database was cross checked by at least one of the partners involved in the CAESAR project and then submitted to another group for descriptors calculation and mathematical modeling. The list of 805 chemicals used for carcinogenic potency modeling along with original rat-derived rodent carcinogenic potency expressed as discrete endpoint and with prediction results is given in Supporting Information (Table 1SI). Discrete endpoint in Table 1SI is given as target value in column named “target”, where 1 corresponds to positive (P) or active and 0 stands for negative/not positive (NP) or inactive. Prediction value is given in column named “prediction”.

Division of the database into the training and test sets

The analyzed dataset consisting of 805 organic compounds was subdivided into multiple training (644 chemicals)/test (161 chemicals) set pairs. The training set contains 327 positive and 317 negative compounds and the test set 95 positive and 66 negative ones. It is evident that positive and negative compounds are evenly distributed over both sets. This part of the study has been done at the Helmholtz Centre for Environmental Research—UFZ in Germany by one of the groups involved in the CAESAR project.

Considerable number of papers concerning the descriptor (and target) domain of training and prediction sets was published (recent examples: [71, 72]).

In the study, it is presented a new algorithm aimed to split the organic compounds into training/prediction sets with respect to the chemical space. First, the compounds are sorted according to a hierarchical system of compound classes. The top level of this hierarchy consists of hydrocarbons, compounds with O, N, N and O (not bonded), NO (bonded), S, SO (bonded), P, and other elements. The order of these classes determines a priority to ensure a proper treatment of compounds with multiple different functional groups. The priority increases within this order, i.e., in the absence of rare elements, P is assigned the highest priority. For example, a compound containing P-, S-, and O-groups together would primarily be sorted into the P class. Within a class, the same sorting according to classes of lower priority is applied. In the example just mentioned, there would be an S subclass, and an O subclass would be within the S subclass.

Next, the subclasses are divided according to their halogen content, with the subclasses none, F, Cl, Br, I, mixed halogens, and the compounds are then sorted with respect to the number of halogen atoms within each of these classes. Then, the next step is to further discriminate these subclasses according to functional groups. For each of the main classes, a set of the 5–15 most relevant functional groups with respect to organic chemicals has been defined. For example, the O class contains groups as OH, carbonyls, carboxyls, etc.

In some cases, there is a further sub-grouping, as p-OH, s-OH, t-OH, aro-OH, or dividing the carbonyls into aldehydes and ketones. Unidentified functional groups containing the respective atom (in the example: O) are treated as a separate subgroup. As with halogens, the compounds are at last sorted according to the number of the respective functional group. In these steps, the treatment of mixed occurrences of functional groups differs from the procedure for the primary classes and subclasses. There is no priority for the functional group classes, but there is an additional mixed group's class. Within the mixed class, there is again a stratified sorting according to priorities.

After finishing the sub-sorting according to functional groups, several steps follow to distinguish between connectivity aspects; compound are distinguished between aromatics and non-aromatics, aromatics are further divided into simple and fused. Regardless of aromatic or not, a further separation follows for the non-aromatic part of the molecule (if any) into saturated or containing double or triple bonds. Saturated non-aromatic substructures are divided into straight chains only or containing branches or rings. Finally, within the deepest leave of that classification tree, compounds remaining within the same subclass are sorted with respect to the number of C atoms.

The sorting of the compounds according to the procedure described here is implemented to run fully automated with the software ChemProp [73,74].

From the sorted list, the prediction sets will be separated by simple counting with respect to the desired ratios between the subsets. For the 80:20 splitting ratio in this study, each fifth compound from the sorted list has been put into the second list. For a splitting into 10 samples of equal size, compound one would go to set one, compound two to set two, etc. This way, the procedure keeps the relations between compound classes in the resulting sets as close as possible to the relations in the total set. It is to emphasize that the splitting to training and test set was performed independently from modeling procedure, i.e., is neither a part of descriptor selection nor the model optimization.

Description of the carcinogenic potency

In toxicology, the acute lethal dose LD₅₀ of a chemical plays a central role. It is defined as the dose of the chemical that kills 50% of the test animals. A large value for the LD₅₀ indicates a substance of low acute toxicity, while a small LD₅₀ indicates a potent poison. Although the LD₅₀ varies with strain, species, and experimental conditions, it has been proved to be a useful and practical measure of acute toxicity, and it is widely used and well understood. In order to adopt some roughly analogous measure for the tumorigenicity of a particular agent, the tumorigenic dose has been employed in the study. The TD₅₀ is defined as the tumorigenic dose rate where

50% of the test animals got any kind of cancer. Using other words, the TD₅₀ is that chronic dose rate [in mg/kg body weight per day (mg/kg-bw/day) or mmol/kg body weight per day (mmol/kg-bw/day)] which would give half of animal tumors within some standard experiment time—the “standard lifespan” for the species. In our study, we consider the dose for rats expressed as logarithm of (mmol/kg bw/day). The dose was adopted as the output value in quantitative modeling.

We have accepted an assignment of carcinogenic categorical activity based on evidence for or against activity within the species group in target sites of rats (male, female or both) as provided in the CPDB Summary Table. Hence, “active” or P or carcinogen was assigned for compound if one or more TD₅₀ and tumor site are listed for one or more rat carcinogenicity sex/species cell (rat male, rat female, rat both) and “inactive” or NP or non-carcinogens was assigned for compound if no TD₅₀ or tumor site are listed and one or more “no positive results” entry for one or more rat carcinogenicity sex/species cell, i.e. one or more experiments are reported in the CPDB for species, but none are positive. In other words, chemicals from the studied dataset were classified as not carcinogenic when the results obtained during animal tests on rats were assigned as NP (or not active). Compounds were classified as P (or active) when the in vivo assays gave defined TD₅₀ value. In the studied dataset (805 compounds), 421 chemicals were classified as carcinogenic and remaining 384 as non-carcinogens. This classification scheme was adopted for quantitative modeling.

Descriptors used in the modeling

The descriptors space for the current study was generated using MDL[®]QSAR program (MDL-QSAR version 2.2.2.0.7, MDL Information Systems, 14600 Catalina St. San Leonardo). MDL QSAR software calculates over 240 physical-chemical, electro topological E-state, connectivity, and other descriptors. E-state indices are a combination of electronic, topological, and valence state information. These indices incorporate information related to atom types and electron accessibility, hydrogen atom E-states, and connectivities that are influenced by all of the sub-structural features of a molecule [75–77]. Hence, we have got the molecular structure information using topological descriptors, including atom-type and group-type, E-State and hydrogen E-state indices, molecular connectivity, chi indices, topological polarity, and counts of molecular features. The set of 254 MDL descriptors was obtained for each of the 805 compounds by the group of BioChemics Consulting SAS (BCX, France) involved in the CAESAR project.

Before the descriptors space was reduced, all variables were normalized into $-1+1$ range. The main goal of the descriptors space reduction is the extraction of significant

variables for the future QSAR modeling. In present study, Kohonen Artificial Neural Network (KANN) and Principle Component Analysis (PCA) were used to reduce the number of descriptors as described in the following section.

Methods

Artificial neural networks

In the current study, we focused on three tasks: first, reduction of number of descriptors using KANN and PCA technique, second, construction of quantitative models [trying to model the continuous value (tumorigenic dose TD_{50})], and third, development and evaluation of categorical models using as a response binary characteristics of carcinogenic potency (active/inactive). Obtained models have been discussed from regulator's point of view.

Kohonen Artificial Neural Networks and CP ANNs have been used in our study. KANNs have been applied for selection of descriptors and CP ANNs have been employed to build predictive models. KANN [64,78–81] represents a basic type of neural network. Mathematically, it is a mapping from multi-dimensional descriptor space onto two-dimensional network of neurons, which are vectors of weights. The dimension of the vectors is equal to the dimension of descriptor space. The mapping runs via non-linear algorithm also known as training. A result is two-dimensional network where the neurons are occupied with objects. The arrangement of objects shows the similarity relationships among them, i.e., the similar objects are located close to each other in the network. It is to emphasize that the training of KANN is unsupervised, i.e., the property values are not a part of the training. The KANN method was successfully applied for descriptors selection, e.g. in structure mutagenicity study of aromatic amines [82] or classification models [83,84]. In the procedure aimed for reduction of descriptors space the Kohonen network is trained with transpose matrix, i.e., the matrix where the roles of objects and descriptors are exchanged. A result is a map where the neurons are occupied with descriptors. Two selection criteria were applied. If a neuron is occupied with a single descriptor, it is placed into descriptor set, and, if a neuron is occupied with more descriptors those with the shortest and larger distance were selected. The distance is defined as the Euclidian distance between the descriptor vector and the neuron weight vector [78–81].

The CP ANN is a generalization of KANN [64,85,86]. The architecture of CP ANN is presented in Fig. 1. The CP ANN is built up from two layers of neurons arranged in two-dimensional rectangular matrix. The input (or Kohonen) layer contains information about the descriptors (input values, see vector representing structure in Fig. 1) while the output layer represents the output values, the carcino-

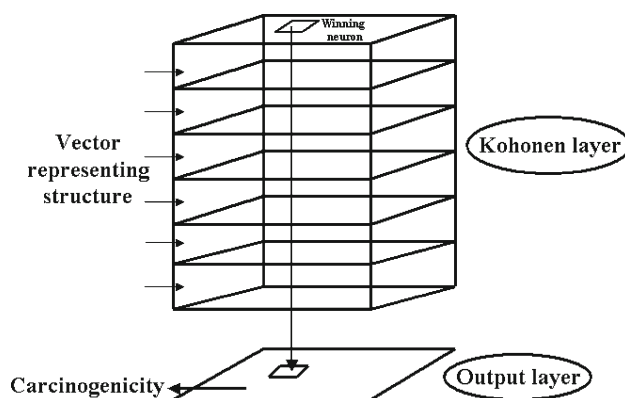


Fig. 1 The architecture of counter propagation artificial neural network (CP ANN)

genic potency in our case. We have used the CP ANN to build the models for prediction of carcinogenic potency doses (quantitative models) and for classification of compounds into two classes (qualitative models). In the first case, the output layer consists of a single layer associated to a single output variable. In the second case, the output layer is two-dimensional, one dimension corresponds to P and the second one to NP response.

Special program TRACEANN in Matlab (<http://www.mathworks.com/products/>) was developed for CP ANN modeling in frame of Sixth Framework Program under the Food Quality and Safety Priority (TRACE project no. 006942). The ANN Toolbox performs the classification of multivariate data using the Kohonen mapping method, and predictive modeling using Counter-propagation neural network including the visualization (contour plots, 3D visualization, and colored neurons) of Kohonen levels. It was created by Marjan Tušar (National Institute of Chemistry Slovenia, Laboratory of Chemometrics, Hajdrihova 19, 1001, Ljubljana, Slovenia). Further applications and program details are reported by Groselj et al. [87].

Validation procedures

Robustness, prediction ability, and validation are crucial issues in developing QSAR models. The partners of CAESAR project agreed to accept statistical parameters described in detail below for the development of quantitative and qualitative models.

Quantitative models validation criteria can be described as follows. For the training set, it was proposed to use: the square correlation coefficient R^2 between predicted and experimental values; the cross-validated values, Q^2 by leave-one-out (LOO) and leave-several-out (LSO); the root mean square error (RMSE). Similar set of parameters was proposed for the test set: square correlation coefficient R^2 and RMSE scores like for the training set.

Moreover, the following conditions by Golbraikh et al. [88] should be considered to confirm that proposed QSAR models have both, robustness and predictive power.

$$Q^2 > 0.5; \quad (1)$$

$$R^2 > 0.6, \quad (2)$$

where R^2 =square correlation coefficient and Q^2 =cross-validated correlation coefficient. More detailed requirements for validation of QSAR models used for regulatory uses are described in the following papers [89–92].

The evaluation of the performance of a categorical or classification model can be assessed using the Cooper statistics [93,94]. First, it should be calculated the number of true and false positive and negative [true positive (TP), true negative (TN), false positive (FP), false negative (FN)]. Cooper statistics express the ability of classification model to detect known active compounds (sensitivity), non-active compounds (specificity), and all chemicals in general (accuracy). The statistical standard binary measures used for categorical models in the study are described in follow-up.

Sensitivity is defined as the percentage of correctly classified carcinogens among the total number of carcinogens. Sensitivity (SE) = TP/(TP + FN).

Specificity is the percentage of correctly classified non-carcinogens among the total number of non-carcinogens. Specificity (SP) = TN/(TN + FP).

Accuracy (Acc) (or concordance) is defined as the total number of carcinogens and non-carcinogens correctly predicted among the total number of tested compound. Acc = (TN + TP)/(TN + TP + FN + FP).

The statistical characterization of the classification (categorical) models is based on the “confusion matrix”. The positive and negative classification rates focused more on the effects of individual chemicals, since they are conditional probabilities. Thus, the positive classification rate is a probability that a chemical classified as active is really active, while negative classification rate gives the probability that chemical classified as inactive chemical is really inactive.

Results and discussion

Descriptors space reduction

Initial set of 254 descriptors was reduced in four steps procedure. First step was dedicated to a visual inspection of the descriptors set and elimination of four descriptors with zero values for all objects (see Table 2SI). As a result, the number of descriptors was reduced from 254 to 250. Sec-

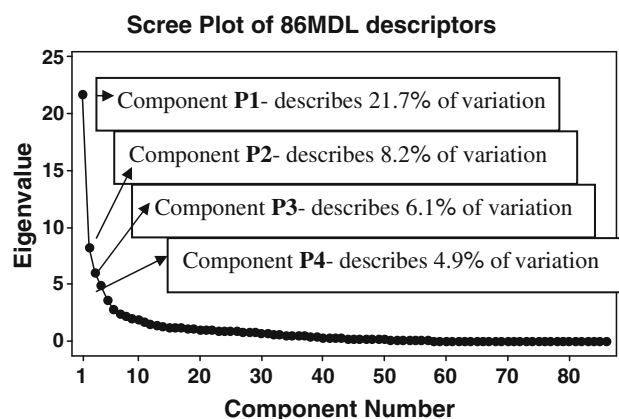


Fig. 2 The graphical representation of Principle Component Analysis (PCA) results

ond step includes the implementation of KANN technique where the number of descriptors was reduced from 250 to 94 MDL descriptors listed in Table 3SI. It should be highlighted that the most homogenous distribution of descriptors over the network was obtained for dimension of Kohonen map 7×7 and number of learning epochs equal to 100. Initial variables mapping or distribution of descriptors in 7×7 top map of Kohonen neural network is shown in the Table 4SI where the number of descriptors occupying an individual neuron is pointed in each square. After selection of pairs of descriptors (with shortest and largest distance) from each of square, we have got distribution of descriptors in 7×7 top map shown in Table 5SI. Third step in descriptors space reduction involves elimination of eight descriptors with average value close to zero and standard deviation close to zero (see Table 6SI). This way the number of descriptors was reduced from 94 to 86. Finally, step four includes application of the PCA [95] where the number of descriptors was reduced to 27. PCA has been performed using Minitab program (Minitab statistical software, Minitab Inc., Quality Plaza, 1829 Pine Hall Rd, State College, PA, USA (www.minitab.com)). The main goal of PCA was to display multidimensional data in the space of lower dimensionality with minimum loss of information and to extract basic features behind the data. The graphical representation of PCA results is shown in Fig. 2.

According to data in correlation matrix presented in Table 7SI and in Fig. 2, the first principle component P1 has variance 21.742 (equal to the largest eigenvalue) and accounts for 0.253 (25.3%) of total variation in data, the second principle component P2 (variance 8.242) accounts for 0.096 (9.6%) of total variation in data, the third one (P3) (variance 6.077) accounts for 0.071 (7.1%) of total variation in data, the fourth one (P4) (variance 4.909) accounts for 0.057 (5.7%) of total variation in data and the fifth one (P5) (variance 3.668) accounts for 0.043 (4.3%) of total variation in the data. The remaining components from P6 to P86 are less informative. Summarizing, we can conclude

Table 1 The set of 27 MDL descriptors obtained after descriptors space reduction using PCA

MDL_ID	Index	Definition	Descriptors categories
MDL001	SsCH3	Sum of all (–CH3) E-State values in molecule	Atom-Type E-State
MDL006	SaaCH	Sum of all (CH) E-State values in molecule	Atom-Type E-State
MDL007	SsssCH	Sum of all (> CH–) E-State values in molecule	Atom-Type E-State
MDL042	SsCH3_acnt	Count of all (–CH3) groups in molecule	Atom-Type E-State Acnt
MDL052	SaaC_acnt	Count of all (CH) groups in molecule	Atom-Type E-State Acnt
MDL083	x0	Simple 0 order chi indices	Connectivities simple
MDL088	xp5	Simple fifth order path chi indices	Connectivities simple
MDL105	dx0	Difference simple zero order chi indices	Connectivities simple
MDL124	nxc3	Number of three way clusters	Connectivities subgraph counts
MDL131	nxch7	Number of seven-membered rings	Connectivities subgraph counts
MDL148	xvpc4	Valence fourth order path/cluster chi index	Connectivities valence
MDL160	dxvp3	Difference valence third order path chi indices	Connectivities valence
MDL165	dxvp8	Difference valence eighth order path chi indices	Connectivities valence
MDL176	SHsOH	Sum of all [–OH] E-State values in molecule	HE-State categories
MDL186	Hmin	Smallest atom hydrogen E-State value in molecule	HE-State categories
MDL187	Gmin	Smallest atom E-State value in molecule	HE-State categories
MDL193	SHarom	sum of hydrogen E-State on aromatic CH	HE-State for groups
MDL226	LogP	Calculated value of LogP	Log P
MDL229	nelem	Number of chemical elements	Molecular properties
MDL231	ncirc	Number of graph circuits	Molecular properties
MDL235	numHBa	Number of hydrogen bond acceptors	Molecular properties
MDL240	SHHBa	Sum of Atom-type E-State indices for hydrogen bond Acceptors	Molecular properties
MDL243	Qsv	Average polarity	Molecular properties
MDL248	sumI	Total of simple topological indices	Total topological indices
MDL249	TTs(4) Simple	Total of valence topological indices	Total topological indices
MDL252	totop	Total topological index based on the molecular connectivity formalism	Total topological indices
MDL253	Wt	Total Wiener number	Total topological indices

that five principle components describe 52% of total data variation. Another way for determination of principle components used in our study is the analysis of the scree plot (see Fig. 2) where the component numbers are displayed versus the corresponding eigenvalues. The eigenvalues of correlation matrix are equal to the variation of principle component. The number of components was chosen based on the size of eigenvalues. The selected five principal components contain 27 MDL descriptors used for further modeling and listed in Table 1.

These 27 descriptors belong mostly to topological structure descriptors, including atom-type and group-type E-state indices (three Atom-Type E-State, two Atom-Type E-State Acnt); molecular connectivity chi indices (eight descriptors), hydrogen E-state indices (four descriptors), four total topological indices, counts of molecular features (five molecular properties descriptors), and log P. (See MDL QSAR Users guide for specific interpretation of topological descriptors.) For references see also [96–98].

Quantitative CP ANN models development and evaluation

Quantitative models aimed for prediction continues values such as tumourigenic dose TD₅₀ in case of prediction of carcinogenic potency. They can be used as support to carcinogenicity assessment and carcinogenic potency evaluation, for instance to evaluate relative risk of different compounds, or of metabolite or parent compound.

In this part of study, quantitative models for prediction carcinogenic potency expressed as continuous value log TD₅₀ are presented. Some findings concerned statistical evaluation of log TD₅₀ distribution are discussed. We have investigated the effect of NP data and molecular weight (MW) on statistical performance of models.

First of all, statistical evaluation of distribution of carcinogenic potency expressed as log TD₅₀ in mmol/kg bw per day for dataset of 805 compounds was performed in our study. Minitab program version 15 was used for graphical representation and evaluation of response log TD₅₀.

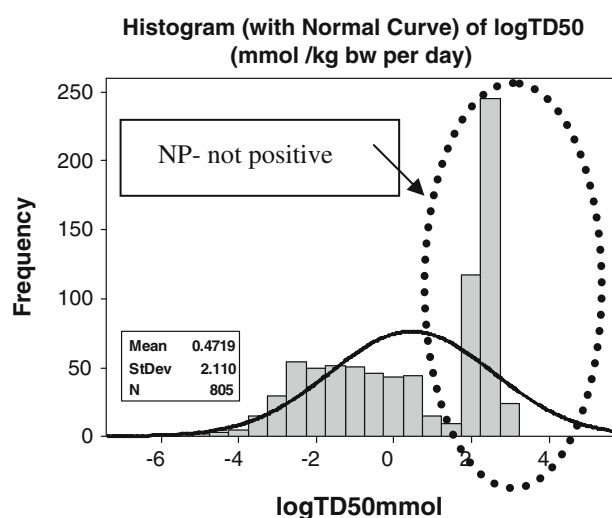


Fig. 3 Histogram of log TD₅₀ (mmol/kg-bw/day)

It should be noticed that NP results was transformed as a number. The tumorigenic dose of 5,000 mg was assumed for not positive tests, as the biggest dose in our dataset appeared to be 3,000 mg. To transform mg/kg-bw/day into mmol/kg bw/day one must divide TD₅₀ expressed in mg/kg bw/day on MW. We used a histogram to assess the shape and spread of continuous data log TD₅₀ (mmol/kg-bw/day) (see Fig. 3).

To draw a histogram, Minitab divides sample values into many intervals called bins. By default, bars represent the number of observations falling within each bin (its frequency). Y-axis corresponds to frequency and X-axis corresponds to log TD₅₀. It is visible on histogram that we deal with two distributions. In Fig. 3, values of log TD₅₀ approximately greater than 1.5 belong to NP results (non-carcinogens) and the area for values less than 1.5 belongs to positive ones (carcinogens). Thus, we have got clusters in the histogram for compounds belonging to carcinogens (P) and non-carcinogens (NP).

Second, a number of quantitative models with different number of MDL descriptors were developed and evaluated. The following sets of descriptors were employed in the study:

- (i) 250 MDL descriptors from initial set calculated with MDL software;

- (ii) 86 MDL descriptors obtained after descriptors selection using Kohonen Network technique;
- (iii) 27 MDL descriptors obtained after variable reduction using PCA.

During modeling, the optimal dimension of neural network (35 × 35) for dataset of 805 compounds was found. CP ANN was trained from 100 to 1,600 learning epochs. Summary statistical performance of CP ANN models using 250, 86, and 27 MDL descriptors network dimension 35 × 35 and number of learning epochs 100 presented in Table 2.

Correlation coefficient “*R*” for training set for best models is in range 0.72–0.74. As for validation (test) set, the correlation coefficient do not exceed 0.47, which is low for reliable model. The obtained results for the training and the test sets showed that the number of descriptors does not influence the statistical quality of models. The correlation coefficients for test sets are not high enough to fulfill criteria for the predictive power of QSAR models concerning prediction of carcinogenic potency.

Fitted line plot [predicted versus experimental log TD₅₀ (mmol/kg-bw/day)] was built for model containing 27 MDL descriptors (dimension 40 × 40, 100 learning epochs) (see Fig. 4).

It should be highlighted that clusters for P and NP results are visible in Fig. 4, where P are marked as rectangles and NP are marked as triangles. Statistical performance of training set (644 compounds) in this model was (*r* = 0.730, RMSE=1.545) (see Fig. 4a) and prediction power of test set (161 compounds) was (*r* = 0.460, RMSE=1.794) (see Fig. 4b).

First, additional modeling was done using only P results. We eliminated all 384 NP compounds (non-carcinogens). Then the rest 421 P compounds (carcinogens) were divided into training (327) and test (95) sets. The performance of the model was not improved; the correlation coefficient for test set was *r* = 0.44.

Second, additional modeling was dedicated to dataset with MW less than 380. In the statistical analysis of descriptors, the MW is one of the most important descriptors. It is speculated that large molecules behave differently in the transport through the cell membranes and that they interact

Table 2 Statistical performance of quantitative CP ANN models using sets of 250, 86, and 27 MDL descriptors (network dimension 35 × 35 and number of learning epochs 100)

Number of MDL descriptors	Method used for the descriptors space reduction	<i>R</i> _{training}	RMSE _{training}	<i>R</i> _{test}	RMSE _{test}
250	–	0.74	1.51	0.47	1.78
86	Kohonen map	0.72	1.54	0.42	1.90
27	PCA	0.74	1.52	0.45	1.80

RMSE root mean square error, *R* correlation coefficient

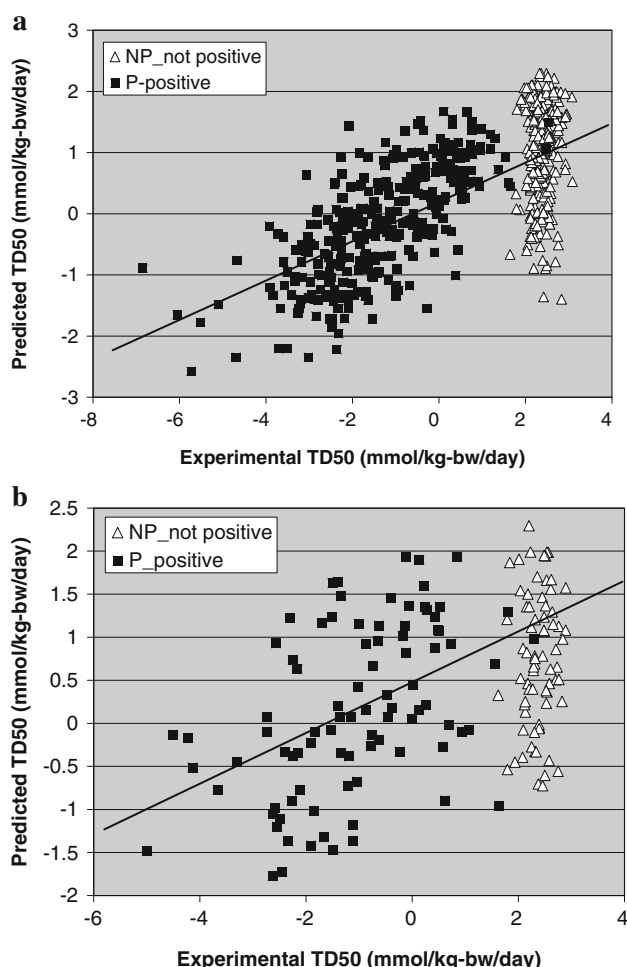


Fig. 4 Fitted line plot [predicted versus experimental logTD₅₀ (mmol/kg-bw/day)] for model containing 27 MDL descriptors (dimension 40 × 40, 100 learning epochs) a-for training set (644 compounds) ($r = 0.730$, RMSE=1.545); b-for test set (161 compounds) ($r = 0.460$, RMSE=1.794). Positives (P) and not positives (NP) are marked as rectangles and triangles, respectively; r correlation coefficient, RMSE root mean square error

differently with biomolecules. The PCA analysis was performed on set of 805 compounds described with 86 MDL descriptors where 39 compounds with the MW larger than 380 appeared as outliers outside the dotted lined square (see Fig. 5).

These 39 compounds are also listed in Table 8SI. Correlation coefficient for test sets for examined models using only compounds with MW less than 380 did not exceed 0.5.

Summarizing the results of our study of quantitative models using CP ANN method, we would like to highlight that we have got poor statistical performance of quantitative models. Even for the best models, the correlation coefficient of test sets was less than 0.5. The low prediction power of the obtained models is probably due to big variation of TD₅₀ values which can considerably vary across the results of

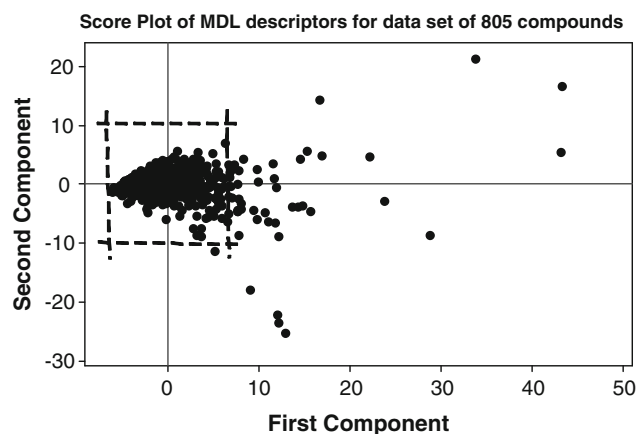


Fig. 5 Score plot of MDL descriptors for data set of 805 compounds some standard experiments for certain chemical. To overcome these circumstances we have chosen categorical model for future modeling.

Qualitative CP ANN models development and assessment

Qualitative models were developed for classification (categorization) of molecules. This approach aims at deriving classifiers in terms of molecular descriptors to describe the distribution of compounds over the respective classes. Often classification method was employed if biological measurements are not precise enough to present biological potency on a continuous scale. In our study, a case with two classes [e.g. biologically active (carcinogen) versus biologically inactive compounds (non-carcinogens)] was considered. As it was mentioned above, we transformed the tumorigenic dose TD₅₀ into two categories: P or active (carcinogens) and NP or inactive (non-carcinogens). Then categorical models were developed and optimized. The dataset of 805 chemicals divided into training (644 chemicals) and test (161 chemicals) was used in modeling as it was described above. Training set examples with class were used for learning and test set examples with class value were used for evaluation. Twenty-seven MDL descriptors previously selected were employed in the modeling.

In order to optimize the computational parameters of CP ANN, first, the dimension of network was examined over the range 25 × 25 to 45 × 45 and number of learning epochs from 100 to 1,800. The quality of the obtained models was assessed using the validation (test) set. This way we have found the optimal dimension of network equal to 35 × 35. Second, we examined statistical performance of neural network with dimension 35 × 35 being trained in the range from 100 to 1,800 epochs. The graphical representation of Acc for training and test sets versus number of learning epochs is shown in Fig. 6. Where the dotted lines indicate the location of maximal Acc for test set 0.68 corresponding to 200

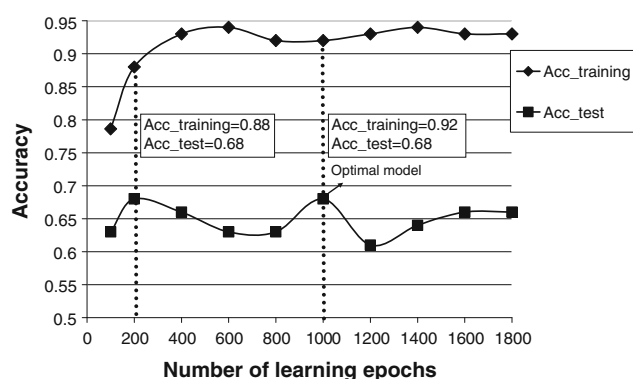


Fig. 6 The accuracy of the training and test sets versus number of learning epochs obtained for models containing 27 MDL descriptors. The dimension of CP ANN network is 35×35 . Acc accuracy. Dotted lines indicate the location of maximal accuracy for test set corresponding to 200 and 1,000 learning epochs, respectively

and 1,000 learning epochs, respectively. It should be highlighted that the training set has notably higher Acc for model trained 1,000 epochs (equal to 0.92) than for one trained 200 epochs (equal to 0.88). Therefore, optimal model corresponds to model with 1,000 learning epochs. Analyzing data plotted in Fig. 6, we can conclude that categorical models (containing 27 MDL descriptors) being trained from 100 to 1,600 learning epochs give Acc for training set from 0.79 to 0.94 and for test set in the range 0.61–0.68.

On the one hand, Acc shows how well a model classifies. From another side, higher Acc does not necessarily imply better performance on target task. It is important to know the number of carcinogens and non-carcinogen predicted correctly. SE and SP are very important characteristics used in categorical method for determination of TP and negative rate of model. Therefore, the next part of study is dedicated to relationships between Acc, SE, and SP and number of learning epochs. We considered here statistical performance of validation (test) set and focused on prediction power of studied models. The Fig. 7 shows Acc, SE, and SP plotted versus number of learning epochs obtained for test set. It is seen that statistical performance of considered models can be modulated by number of epochs. The lowest value of Acc (0.61) corresponds to 1,200 epochs, while the highest one (0.68) to 200 and 1,000 learning epochs, respectively. The highest SE (0.75 and 0.78) was obtained in case of 800 and 1,800 learning epochs, respectively. It should be noted that models with highest sensitivities (0.75 and 0.78) exhibit the lowest values of Acc (0.63 and 0.66) and SP (0.56 and 0.58). The

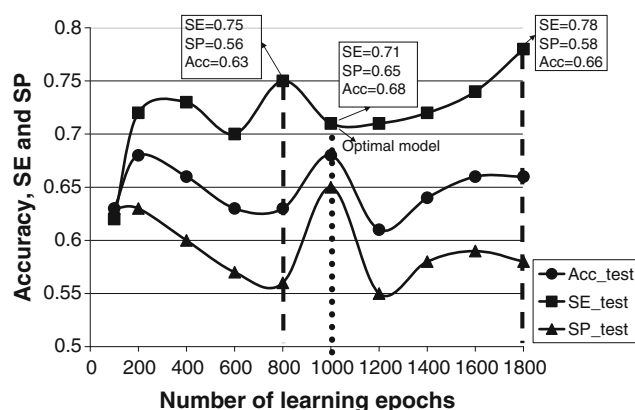


Fig. 7 The accuracy, sensitivity, and specificity versus number of learning epochs obtained for the test set. Acc Accuracy, SE Sensitivity, SP specificity. Dotted line indicates model with the maximal accuracy (0.68) of test set and corresponding SE (0.71) and SP (0.65) in case of optimal model and stroked lines indicate 2 models with maximal sensitivity (SE) equal to 0.75 and 0.78 and corresponding Acc (0.63 and 0.66) and specificity (SP) (0.56 and 0.58), respectively

optimal model was set at 1,000 epochs showing the SE, SP, and Acc at 0.74, 0.65 and 0.68, respectively.

Summary statistical characterization of optimal model (35×35 , 1,000 epochs) for total, training, and test sets is presented in Table 3. Acc (% of compounds with correct model prediction) for internal (training set) is equal to 92.2%, while for external (test) set is 68.3% and for total set is 87.5. SP values for pointed sets are equal to 98.9, 65.2, 92.0 and SE 87.5, 70.7, 84.2%, respectively.

Confusion matrices were composed for the training (see Table 4a) and test (see Table 4b) sets. Of the 644 training compounds, 265 are classified as TN, 329 are classified as TP and only three are classified as FN and 47 as FP. As for test set we have got the following distribution of correct and not correct predicted chemicals: of the 161 test compounds, 45 are classified as TN, 65 are classified as TP, 24 are classified as FN and 27 as FP.

As described in the section “Methods”, the prediction of classification model is given as two-dimensional vector of real numbers. The numbers express the affiliation of compound to a particular class. A crucial question in evaluation of predicted results is the threshold value above which a compound is classified as positive and below as negative. The threshold determines the Acc, SP, and SE of model, however, higher SP means lower SE, and vice versa. Figure 8 shows the dependence of three parameters (SE, SP, and Acc)

Table 3 Summary statistical performance of optimal model for total (805), training (644), and test (161) sets

Data sets	Specificity (%)	Sensitivity (%)	Accuracy (%)
Training set (644 compounds)	98.9	87.5	92.2
Test set (161 compounds)	65.2	70.7	68.3
Total set (805 compounds)	92.0	84.2	87.5

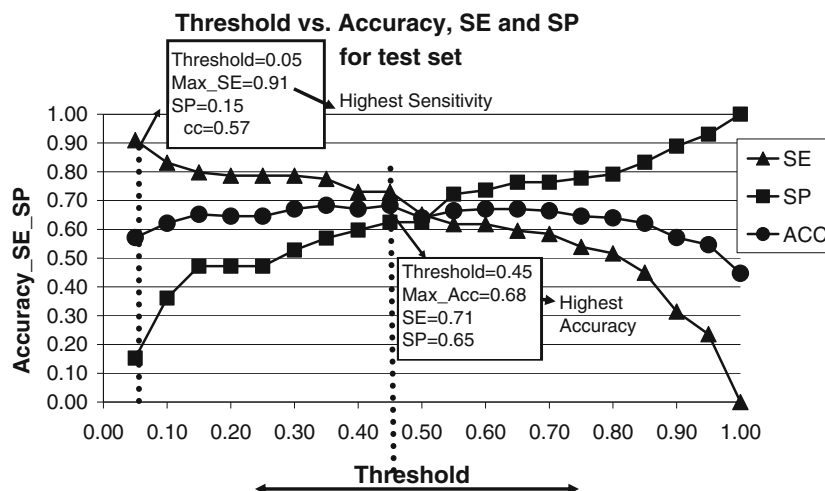
Table 4 Confusion matrix obtained for the training set of 644 chemicals (a) and for the test set of 161 chemicals (b)

	Predicted class	
	NP	P
<i>a</i>		
Observed class		
NP	TN	FP
312 chemicals	265	47
P	FN	TP
332 chemicals	3	329
<i>b</i>		
Observed class		
NP	TN	FP
72 chemicals	45	27
P	FN	TP
89 chemicals	24	65

NP not positive, P positive, TN True Negative, FP False Positive, FN False Negative, TP True Positive

versus threshold for test set of 161 chemicals. From regulatory point of view Acc prediction of carcinogens (SE) is more important for human health safety than prediction Acc of non-carcinogens (SP). Depending on requirements for models used for regulatory decision support and their application we are able to create models capable to predict carcinogens rather than non-carcinogens and vice versa. Figure 8 demonstrates how Acc, SE and SP vary depending on threshold of categorical model for test set of 161 chemicals. Changing of threshold leads to revision of SE and SP. The highest Acc (0.68) corresponds to threshold 0.45. SE and SP in this point are equal to 0.71 and 0.65, respectively. This data addressed to optimal model which was described above and also shown in previous Figs. 6 and 7 from another point of view. Maximal sensitivity (Max_SE=0.91) corresponds to SP (0.15), Acc (0.57), and threshold equal to 0.05 (Fig. 8).

Fig. 8 The threshold value obtained for test set (161 chemicals). Dotted lines pointed to the threshold values 0.05 and 0.45 correspond to models with highest sensitivity (SE=0.91) and highest accuracy (Max_Acc=0.68), respectively. Acc accuracy, SE sensitivity, SP specificity



Referring to Fig. 8, we can see that changing the threshold from 0 for 1 leads consequently to reducing SE and increasing SP. Thus, we can increase the number of correctly predicted carcinogens or non-carcinogens. Selection of proper model is the matter of making decision.

Moreover, the validation set of 931 compounds was composed by CAESAR project partner (*Istituto di Ricerche Farmacologiche “Mario Negri” (IRFMN), Milano, Italy*) and implemented for our optimal model. As a result, we have got for validation set the following statistical performance: Acc 0.66, SE 0.61, and SP 0.69. We have also performed leave 20% out procedure. As a result, we have obtained the predictive power of model equal to 66%.

Another presentation of quality of model, which considers SE and SP, is Receiver Operating Characteristic (ROC) curve. In ideal case, the area under the curve is equal to 1. We have applied Receiving Operation Characteristics (ROC) curve to show the threshold (cut-off or trade-off) that exists between performance on positive and negative cases. The ROC curves are plotted for training (644 chemicals) and test (161 chemicals) sets and are shown in Fig. 9.

A measure of model performance is provided by the area under the ROC curve. For training and test sets, we have got an area of 0.988 and 0.699, respectively, which is promising initial result in modeling carcinogenicity. Using categorical yes/no response, we managed to escape some noises and variations that presented in continuous parameter of tumorigenic dose TD₅₀ and to build reliable model for database contained noncongeneric chemicals.

Conclusions

The main goal of our study was to develop QSAR models applying the CP ANN technique, which can be used for regulatory purposes. The models were built on a dataset of 805 non-congeneric organic compounds, which were described

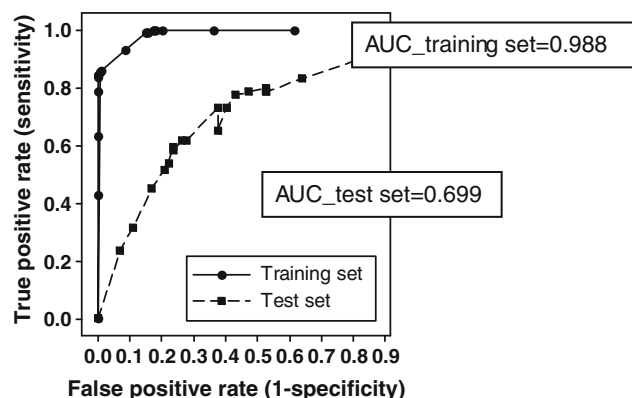


Fig. 9 Receiver operating characteristic (ROC) curve for the optimal model. The area under the curve is 0.988 and 0.699 in the training and test sets, respectively

with physico chemical and wholistic structural descriptors. The 27 MDL descriptors for the final models were selected in an independent procedure with implementation of KANN and PCA methods. In the first part, we examined the quantitative models where the carcinogenic potency was expressed as logarithm of tumorigenic dose. The models show poor prediction abilities even though the data set was reduced to only positive compounds or to compounds with MW lower than 380. More speculative conclusions can be done. Indeed, the tumorigenic dose is associated with the diversity of mechanism and mode of action.

Furthermore, the dose is measured with *in vivo* assays, which probably contains a large variation in the measured values. As for categorical models, based on YES/NO response, satisfactory prediction results were obtained. The CP ANN classification models developed in this study are in accordance with current expectations for carcinogenic predictions with QSAR approach.

Acknowledgments The authors thank the European Commission for the financial support under project CAESAR (SSPI-022674), the Slovenian Ministry of Higher Education, Science and Technology (grant P1-017), and integrated project OSIRIS (contract no. 037017). The authors also like to thank Ralf-Uwe Ebert (UFZ) for his technical support in running the training/prediction set splitting.

References

- Benigni R (2005) Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Rev* 105:1767–1800. doi:10.1021/cr030049y
- OECD (2002) Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies. OECD Environment, Health and Safety Publications Series on Testing and Assessment No. 35, Paris, France

- Combes RD (2000) The use of structure-activity relationships and markers of cell toxicity to detect non-genotoxic carcinogens. *Toxicol In Vitro* 14:387–399. doi:10.1016/S0887-2333(00)00026-6
- Choy WN (2001) Genotoxic and non-genotoxic mechanisms of carcinogenesis. In: Choy WN (ed) Genetic toxicology and cancer risk assessment. Marcel Dekker, New York pp 47–72
- Combes R, Grindon C, Cronin MTD, Roberts DW, Garrod JF (2007) Proposed integrated decision-tree testing strategies for mutagenicity and carcinogenicity in relation to the EU REACH Legislation. *ATLA* 35:267–287
- Fenech M (2000) The *in vitro* micronucleus technique. *Mutat Res Fundam Mol Mech Mutagen* 455:81–95. doi:10.1016/S0027-5107(00)00065-8
- Bernauer U, Oberemm A, Madle S, Gundert-Remy U (2005) The use of *in vitro* data in risk assessment. *Basic Clin Pharmacol Toxicol* 96:176–181. doi:10.1111/j.1742-7843.2005.pto960306.x
- Knight A, Bailey J, Balcombe J (2006) Animal carcinogenicity studies: implications for the REACH systems. *ATLA* 34(Suppl 1):139–147
- Omenn GS (1995) Assessing the risk assessment paradigm. *Toxicology* 102:23–28. doi:10.1016/0300-483X(95)03034-D
- Long ME (2007) Predicting carcinogenicity in humans: the need to supplement animal-based toxicology. *AATEX* 14:553–559
- Helma C (ed) (2005) Predictive toxicology. Marcel Dekker, New York
- Kuschner M (1995) The relevance of rodent tumors in assessing carcinogenicity in human beings. *Regul Toxicol Pharmacol* 21:250–251. doi:10.1006/rtp.1995.1037
- IARC (2006) Preamble, IARC monographs on the evaluation of carcinogenic risks to humans
- OECD (1981) Guidelines for the testing of chemicals test no. 451: carcinogenicity studies
- OECD (1981) Guidelines for the testing of chemicals test no. 452: chronic toxicity studies
- OECD (1981) Guidelines for the testing of chemicals test no. 453: combined chronic toxicity/carcinogenicity studies
- Benigni R, Giuliani A (2003) Putting the predictive toxicology challenge into perspective: reflections on the results. *Bioinformatics* 19:1194–1200. doi:10.1093/bioinformatics/btg099
- Richard AM, Benigni R (2002) AI and SAR approaches for predicting chemical carcinogenicity: survey and status report. *SAR QSAR Environ Res* 13:1–19. doi:10.1080/10629360290002055
- Benigni R, Bossa C (2008) Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat Res* 659:248–261. doi:10.1016/j.mrrev.2008.05.003
- Lill MA (2007) Multi-dimensional QSAR in drug discovery. *Drug Discov Today* 12:1013–1017. doi:10.1016/j.drudis.2007.08.004
- Kruhlak NL, Contrera JF, Benz RD, Matthews EJ (2007) Progress in QSAR toxicity screening of pharmaceutical impurities and other FDA regulated products. *Adv Drug Deliv Rev* 59:43–55. doi:10.1016/j.addr.2006.10.008
- Mager DE (2006) Quantitative structure—pharmacokinetic/pharmacodynamic relationships. *Adv Drug Deliv Rev* 58:1326–1356. doi:10.1016/j.addr.2006.08.002
- Oprea TI, Matter H (2004) Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* 8:349–358. doi:10.1016/j.cbpa.2004.06.008
- Fjodorova N, Novich M, Vrachko M, Smirnov V, Kharchevnikova N, Zholdakova Z, Novikov S, Skvortsova N, Filimonov D, Poroiakov V, Benfenati E (2008) Directions in QSAR modeling for regulatory uses in OECD member countries EU and in Russia. *J Environ Sci Health C* 26:201–236. doi:10.1080/10590500802135578
- Fjodorova N, Novich M, Vrachko M, Kharchevnikova N, Zholdakova Z, Sinitsyna O, Benfenati E (2008) Regulatory assessment

- of chemicals within OECD member countries, EU and in Russia. *J Environ Sci Health C* 26:40–88. doi:[10.1080/10590500801907365](https://doi.org/10.1080/10590500801907365)
26. Benigni R, Giuliani A, Franke R, Gruska A (2000) Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chem Rev* 100:3697–3714. doi:[10.1021/cr9901079](https://doi.org/10.1021/cr9901079)
 27. Benigni R, Bossa C, Netzeva T, Worth A (2007) Collection and evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity. European Commission Directorate General Joint Research Centre 2007 EUR 22772EN © European Communities
 28. Patlewicz G, Rodford R, Walker JD (2003) Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity. *Environ Toxicol Chem* 22:1885–1893. doi:[10.1897/01-461](https://doi.org/10.1897/01-461)
 29. Helguera AM, Perez MCA, Combes RD, González MP (2005) The prediction of carcinogenicity from molecular structure. *Curr Comput Aided Drug Des* 1:237–255. doi:[10.2174/1573409054367655](https://doi.org/10.2174/1573409054367655)
 30. Morales Helguera A, Cabrera Perez MA, Perez González M, Molina Ruiz R, Gonzalez-Diaz H (2005) A topological substructural approach applied to the computational prediction of rodent carcinogenicity. *Bioorg Med Chem* 13:2477–2488. doi:[10.1016/j.bmc.2005.01.035](https://doi.org/10.1016/j.bmc.2005.01.035)
 31. Franke R, Gruska A, Giuliani A, Benigni R (2001) Prediction of rodent carcinogenicity of aromatic amines: a quantitative structure-activity relationships model. *Carcinogenesis* 22:1561–1571. doi:[10.1093/carcin/22.9.1561](https://doi.org/10.1093/carcin/22.9.1561)
 32. Gini G, Lorenzini M, Benfenati E, Grasso P, Bruschi M (1999) Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. *J Chem Inf Comput Sci* 39:1076–1080. doi:[10.1021/ci9903096](https://doi.org/10.1021/ci9903096)
 33. Helguera Morales A, Cabrera Pérez MA, Combes RD, Pérez González M (2006) Quantitative structure-activity relationships for the computational prediction of nitrocompounds carcinogenicity. *Toxicology* 220:51–62. doi:[10.1016/j.tox.2005.11.024](https://doi.org/10.1016/j.tox.2005.11.024)
 34. Zhang L, Sannes K, Shusterman AJ, Hansch C (1992) The structure-activity relationships of skin carcinogenicity of aromatic hydrocarbons and heterocycles. *Chem Biol Interact* 81:149–180. doi:[10.1016/0009-2797\(92\)90032-G](https://doi.org/10.1016/0009-2797(92)90032-G)
 35. Villemain D, Cherqaoui D, Mesbah A (1994) Predicting carcinogenicity of polycyclic aromatic hydrocarbons from back-propagation neural network. *J Chem Inf Comput Sci* 34:1288–1293. doi:[10.1021/ci00022a010](https://doi.org/10.1021/ci00022a010)
 36. Richard AM, Woo YT (1990) A CASE-SAR analysis of polycyclic aromatic hydrocarbon carcinogenicity. *Mutat Res* 242:285–303. doi:[10.1016/0165-1218\(90\)90047-6](https://doi.org/10.1016/0165-1218(90)90047-6)
 37. Passerini L (2003) QSARs for individual classis of chemical mutagens and carcinogens. In: Benigni R (ed) Quantitative structure-activity relationship (QSARs). Models of mutagens and carcinogens. CRC Press, Boca Raton, pp 81–123
 38. Benigni R (ed) (2003) Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens. CRC Press, Boca Raton, pp 88–144
 39. Contrera JF, Matthews EJ, Benz RD (2003) Prediction the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul Toxicol Pharmacol* 38:243–259. doi:[10.1016/S0273-2300\(03\)00071-0](https://doi.org/10.1016/S0273-2300(03)00071-0)
 40. Loew GH, Poulsen M, Kirkjian E, Ferrell J, Sudhindra BS, Rebagliati M (1985) Computer-assisted mechanistic structure-activity studies: application to diverse classes of chemical carcinogens. *Environ Health Perspect* 61:69–96. doi:[10.2307/3430063](https://doi.org/10.2307/3430063)
 41. Vračko M (1997) A study of structure-carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures. *J Chem Inf Comput Sci* 37:1037–1043. doi:[10.1021/ci970231y](https://doi.org/10.1021/ci970231y)
 42. Klopman G, Chakravarti SK, Zhu H, Ivanov JM, Saiakhov RD (2004) ESP: a method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases. *J Chem Inf Comput Sci* 44:704–715. doi:[10.1021/ci030298n](https://doi.org/10.1021/ci030298n)
 43. Klopman G, Ivanov J, Saiakhov R, Chakravarti S (2005) MC4PC-An artificial intelligence approach to the discovery of quantitative structure-toxic activity relationship. In: Helma C (ed) Predictive toxicology. CRC Press, Boca Raton, pp 423–457
 44. Matthews EJ, Contrera JF (1998) A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASEQSAR-ES software. *Regul Toxicol Pharmacol* 28:242–264. doi:[10.1006/rtph.1998.1259](https://doi.org/10.1006/rtph.1998.1259)
 45. Woo Y-T, Lai DY (2005) OncoLogic: a mechanism-based expert system for predicting the carcinogenic potential of chemicals. In: Helma C Predictive toxicology. CRC Press, Boca Raton, pp 385–413
 46. Lagunin AA, Dearden JC, Filimonov DA, Poroikov VV (2005) Computer-aided rodent carcinogenicity prediction. *Mutat Res* 586:138–146. doi:[10.1016/j.mrgentox.2005.06.005](https://doi.org/10.1016/j.mrgentox.2005.06.005)
 47. Benfenati E, Gini G (1997) Computational predictive programs (expert systems) in toxicology. *Toxicology* 119:213–225. doi:[10.1016/S0300-483X\(97\)03631-7](https://doi.org/10.1016/S0300-483X(97)03631-7)
 48. Benigni R, Richard AM (1998) Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity. *Methods* 14:264–276. doi:[10.1006/meth.1998.0583](https://doi.org/10.1006/meth.1998.0583)
 49. Richard AM (1998) Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet?. *Mutat Res* 400:493–507. doi:[10.1016/S0027-5107\(98\)00068-2](https://doi.org/10.1016/S0027-5107(98)00068-2)
 50. Dearden JC, Barratt MD, Benigni R, Bristol DW, Combes RD, Cronin MTD, Judson PN, Payne MP, Richard AM, Tichy M, Worth AP, Yourick JJ (1997) The development and validation of expert systems for predicting toxicity. The report and recommendations of an ECVAM/ECB workshop (ECVAM Workshop 24). *ATLA* 25:223–252
 51. Klopman G (1992) MULTICASE 1. A hierarchical computer automated structure evaluation program. *Quant Struct Act Relat* 11:176–184. doi:[10.1002/qsar.19920110208](https://doi.org/10.1002/qsar.19920110208)
 52. Cunningham AR, Rosenkranz HS, Zhang YP, Klopman G (1998) Identification of “genotoxic” and “non-genotoxic” alerts for cancer in mice: the carcinogenicity potency data base. *Mutat Res* 398:1–17. doi:[10.1016/S0027-5107\(97\)00202-9](https://doi.org/10.1016/S0027-5107(97)00202-9)
 53. Rosenkranz HS, Cunningham AR, Zhang YP, Claycamp HG, Macina OT, Sussman NB, Grant SG, Klopman G (1999) Development, characterization and application of predictive-toxicology models. *SAR QSAR Environ Res* 10:277–298. doi:[10.1080/10629369908039181](https://doi.org/10.1080/10629369908039181)
 54. King RD, Srinivasan A (1996) A prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environ Health Perspect* 104:1031–1040
 55. Perrotta A, Malacarne D, Taningher M, Pesenti R, Paolucci M, Parodi S (1996) A computerized connectivity approach for analyzing the structural basis of mutagenicity in *Salmonella* and its relationship with rodent carcinogenicity. *Environ Mol Mutagen* 28:31–50. doi:[10.1002/\(SICI\)1098-2280\(1996\)28:1<31::AID-EM7>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1098-2280(1996)28:1<31::AID-EM7>3.0.CO;2-H)
 56. Clare MG, Lorenzon G, Akhurst LC, Marzin D, van Delft J, Montero R, Botta A, Bertens A, Cinelli S, Thybaud V, Lorge E (2006) SFTG international collaborative study on in vitro micronucleus test II. Using human lymphocytes. *Mutat Res* 607:37–60. doi:[10.1016/j.mrgentox.2006.04.001](https://doi.org/10.1016/j.mrgentox.2006.04.001)
 57. DiPaolo JA, Nelson RL, Donovan PJ (1971) Morphological, monogenic and carpological characteristics of Syrian hamster embryo cells transformed in vitro by carcinogenic polycyclic hydrocarbons. *Cancer Res* 31:1118–1127
 58. Kirkland D, Aardema M, Henderson L, Müller L (2005) Evaluation of the ability of a battery of 3 *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitiv-

- ity, specificity and relative predictability. *Mutat Res* 584:1–256. doi:[10.1016/j.mrgentox.2005.10.002](https://doi.org/10.1016/j.mrgentox.2005.10.002)
59. Kirsch-Volders M (2003) Report from the in vitro micronucleus assay working group. *Mutat Res* 540:153–163. doi:[10.1016/j.mrgentox.2003.07.005](https://doi.org/10.1016/j.mrgentox.2003.07.005)
 60. National Research Council (2006) Animal and in vitro toxicity testing, in toxicity testing for assessment of environmental agents: interim report. National Academies Press, Washington, 26–70
 61. ESAC (2006) Statement on the scientific validity of the in vitro micronucleus test as an alternative to the in vitro chromosome aberration assay for genotoxicity testing. Joint Research Centre, European Commission, 2 pp
 62. Gold LS, Zeiger E (eds) (1997) Handbook of carcinogenic potency and genotoxicity databases. CRC Press, Boca Raton
 63. Hoshi M, Morimura K, Wanibuchi H, Wei M, Okochi E, Ushijima T, Takaoka K, Fukushima S (2004) No-observed effect levels for carcinogenicity and for in vivo mutagenicity of a genotoxic carcinogen. *Toxicol Sci* 81:273–279. doi:[10.1093/toxsci/kfh241](https://doi.org/10.1093/toxsci/kfh241)
 64. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design, 2nd edn. Wiley-VCH Verlag GmbH, Weinheim
 65. Hecht-Nielsen R (1987) Counter propagation networks. In: Caudill M, Butler C (eds) Proceedings of the IEEE first international conference on neural networks, vol 2. SOS Printing, San Diego, pp 19–32
 66. Zupan J, Novič M, Gasteiger J (1995) Neural networks with counter-propagation learning strategy used for modeling. *Chemom Intell Lab Syst* 27:175–187. doi:[10.1016/7439\(94\)00016-C](https://doi.org/10.1016/7439(94)00016-C)
 67. Eric S, Solmajer T, Zupan J, Novič M, Oblak M, Agbaba D (2004) Prediction of selectivity of α_1 -adrenergic antagonists by counter propagation neural network (CP-ANN). *Il Farmaco* 59:389–395. doi:[10.1016/j.farmac.2003.12.009](https://doi.org/10.1016/j.farmac.2003.12.009)
 68. Novič M, Vračko M (2001) Comparison of spectrum-like representation of 3D chemical structure with other representations when used for modeling biological activity. *Chemom Intell Lab Syst* 59:33–44. doi:[10.1016/S0169-7439\(01\)00142-3](https://doi.org/10.1016/S0169-7439(01)00142-3)
 69. Vračko M, Novič M, Zupan J (1999) Study of structure-toxicity relationship by a counter propagation neural network. *Anal Chim Acta* 384:319–332. doi:[10.1016/S0003-2670\(98\)00782-X](https://doi.org/10.1016/S0003-2670(98)00782-X)
 70. Computer Assisted Evaluation of industrial chemical Substances According to Regulations (CAESAR) project, <http://www.caesar-project.eu> Accessed 22 April 2009
 71. Yeap SK, Walley RJ, Snarey M, van Hoorn WP, Mason JS (2007) Designing compound subsets: comparison of random and rational approaches using statistical simulation. *J Chem Inf Model* 47:2149–2158. doi:[10.1021/ci600382m](https://doi.org/10.1021/ci600382m)
 72. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb Sci* 25:235–251. doi:[10.1002/qsar.200510161](https://doi.org/10.1002/qsar.200510161)
 73. Schüürmann G, Kühne R, Kleint F, Ebert R-U, Rothenbacher C, Herth P (1997) A software system for automatic chemical property estimation from molecular structure. In: Chen F, Schüürmann G Quantitative structure-activity relationships in environmental sciences—VII. SETAC Press, Pensacola, pp 93–114
 74. Schüürmann G, Ebert R-U, Nendza M, Dearden JC, Paschke A, Kühne R (2007) Prediction of fate-related compound properties. In: van Leeuwen K, Vermeire T (eds) Risk assessment of chemicals. An introduction. Springer Science, Dordrecht, pp 375–426
 75. Kier LB, Hall LH (1999) Molecular structure description: the electro topological state. Academic Press, New York
 76. Kier LB, Hall LH (2001) Database organization and searching with E-state indices. *SAR QSAR Environ Res* 12:55–74. doi:[10.1080/10629360108035371](https://doi.org/10.1080/10629360108035371)
 77. Kier LB, Hall LH (1999) The electro topological state: structure modeling for QSAR and database analysis. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Reading, pp 491–562
 78. Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern* 43:59–69. doi:[10.1007/BF00337288](https://doi.org/10.1007/BF00337288)
 79. Kohonen T (1984) Self-organization and associative memory. Springer, Berlin
 80. Kohonen T (2001) Self-organizing maps. Springer, Berlin
 81. Vračko M (2005) Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. *Curr Comput Aided Drug Des* 1:73–78. doi:[10.2174/1573409052952224](https://doi.org/10.2174/1573409052952224)
 82. Jezierska A, Vračko M, Basak SC (2004) Counter-propagation artificial neural network as a tool for the independent variable selection: structure-mutagenicity study on aromatic amines. *Mol Divers* 8:371–377. doi:[10.1023/B:MODI.0000047502.66802.3d](https://doi.org/10.1023/B:MODI.0000047502.66802.3d)
 83. Roncaglioni A, Novič M, Vračko M, Benfenati E (2004) Classification of potential endocrine disrupters on the basis of molecular structure using a non-linear modeling method. *J Chem Inf Comput Sci* 44:300–309. doi:[10.1021/ci030421a](https://doi.org/10.1021/ci030421a)
 84. Kirew DB, Chretien JR, Bernard P, Ros F (1998) Application of Kohonen Neural Networks in classification of biologically active compounds. *SAR QSAR Environ Res* 8:93–107. doi:[10.1080/10629369808033262](https://doi.org/10.1080/10629369808033262)
 85. Dayhoff JE (1990) Neural network architectures: an introduction. Van Nostrand-Reinhold, New York
 86. Cherkassky V, Friedman JH, Wechsler H (1994) From statistics to neural networks-theory and pattern recognition applications. Springer, Berlin
 87. Grošelj N, van der Veer G, Tušar M, Vračko M, Novič M (2008) Verification of the geological origin of bottled mineral water using artificial neural networks. *Food Chem.* doi:[10.1016/j.foodchem.2008.11.085](https://doi.org/10.1016/j.foodchem.2008.11.085)
 88. Golbraikh A, Tropsha A (2002) Beware of q²! *J Mol Graph Model* 20:269–276. doi:[10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
 89. OECD (2004) Report from the expert group on (quantitative) structure-activity relationships ((Q)SARs) on the principles for the validation of (Q)SARs. OECD Environment Health and Safety Publications, Series on Testing and Assessment No. 49. Environment Directorate, OECD, Paris, France
 90. Jaworska JS, Comber M, Auer C, Van Leeuwen CJ (2003) Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ Health Perspect* 111:1358–1360. doi:[10.1289/ehp.5757](https://doi.org/10.1289/ehp.5757)
 91. Benfenati E (ed) (2007) Quantitative structure-activity relationship (QSAR) of pesticide for regulatory purposes. Elsevier, The Netherlands
 92. Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, Worth AP (2003) Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environ Health Perspect* 111:1391–1401. doi:[10.1289/ehp.5759](https://doi.org/10.1289/ehp.5759)
 93. Cooper JA, Saracci R, Cole P (1979) Describing the validity of carcinogen screening test. *Br J Cancer* 39:87–89
 94. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Belmont
 95. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer-Verlag, New York
 96. Kier LB (1986) Molecular connectivity in structure-activity analysis. Wiley, New York
 97. Hall LH, Kier LB (1991) The molecular connectivity chi indices and kappa shape indexes in structure-property relations. In: Boyd D, Lipkowitz K (eds) Reviews of computational chemistry. VCH Publishers, Inc., pp 367–422
 98. Kellogg GE, Kier LB, Gaillard P, Hall LH (1996) E-state fields: applications to 3D QSAR. *J Comput Aided Mol Des* 10:513–515. doi:[10.1007/BF00134175](https://doi.org/10.1007/BF00134175)