J Comput Aided Mol Des (2011) 25:1147–1158
DOI 10.1007/s10822-011-9499-9

Evaluating the applicability domain in the case of classification predictive models for carcinogenicity based on the counter propagation artificial neural network

Natalja Fjodorova · Marjana Novič · Alessandra Roncaglioni · Emilio Benfenati

Received: 22 June 2011 / Accepted: 21 November 2011 / Published online: 3 December 2011

Abstract The applicability domain (AD) of models developed for regulatory use has attached great attention recently. The AD of quantitative structure–activity relationship (QSAR) models is the response and chemical structure space in which the model makes predictions with a given reliability. The evaluation of AD of regressions QSAR models for congeneric sets of chemicals can be find in many papers and books while the issue about metrics for the evaluation of an AD for the non-linear models (like neural networks) for the diverse set of chemicals represents the new field of investigations in QSAR studies. The scientific society is standing before the challenge to find out reliable way for the evaluation of an AD of non linear models. The new metrics for the evaluation of the AD of the counter propagation artificial neural network (CP ANN) models are discussed in the article: the Euclidean distances between an object (molecule) and the corresponding excited neuron of the neural network and between an object (molecule) and the representative object (vector of average values of descriptors). The investigation of the training and test sets chemicals coverage in the descriptors space was made with the respect to false predicted chemicals. The leverage approach was used to compare non linear (CP ANN) models with linear ones.

Keywords Counter propagation artificial neural network (CP ANN) · (Quantitative) structure–activity relationship ((Q)SAR) · Rodent carcinogenicity · Applicability domain (AD) · Leverage approach

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9499-9) contains supplementary material, which is available to authorized users.

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

M. Novič
e-mail: marjana.novic@ki.si

A. Roncaglioni · E. Benfenati
Institute for Pharmacological Research "Mario Negri",
Via La Masa 19, 20156 Milan, Italy
e-mail: aroncaglioni@marionegri.it

E. Benfenati
e-mail: benfenati@marionegri.it

Introduction

According to the Organization for Economic Co-operation and Development (OECD) guidance document on the validation of quantitative structure–activity relationship (QSAR) models used for regulatory purposes [1] published in 2007 the prediction models should be accompanied by a definition of the applicability domain (AD). The AD was defined as *the response and chemicals structure space in which the model makes predictions with a given reliability* [1, 2]. The general definition of the AD of a (Q)SAR model is given in several papers [2–4]. The authors [5] proposed a stepwise approach for determining the model applicability domain based on the general requirements for variation of the physicochemical properties of chemicals, the structural similarity between chemicals that are correctly predicted, and on a mechanistic understanding of the modelled endpoint based on the metabolism simulation. Netzeva et al. [2] represented approaches for describing the AD of a model: methods based on ranges of molecular descriptors, geometrical methods, distance-based methods, and probability distribution-based methods. Three distance based approaches are widely used in QSAR research: Euclidean,

Mahalanobis and city-block distance [2]. The Euclidean distance (ED) is commonly used for the determination of the distance between two points which are taken as a similarity measure between two compounds in an n-dimensional descriptor space. In the study we selected ED metrics for the counter propagation artificial neural network (CP ANN) models to explore the training and test sets chemicals coverage and descriptors space in relation to false predictions.

Classification models were applied in this study. Therefore, we focused on the AD for classification problem. Different classification approaches have been developed over the years. The pattern classification methods are reported in the book [6]. Authors compared different classification techniques. The main goal of classification methods was to establish boundaries between groups in an n-dimensional space. These decision boundaries using EDs should be linear if we use regression methods. It means that the groups of points should be linearly separable: active and inactive compounds in the descriptor space can be separated by a line, a plane or a hyper plane. In contrast, the artificial neural networks (ANNs) can generate complex decision boundaries and therefore provide low error classification results being implemented for complex data [7]. It was noted [8] that although neural networks are equivalent to statistical classifiers, they offer more effective computational algorithm.

Carcinogenicity is a very complex endpoint that is extremely difficult to model [9]. We deal with non-liner relationships. The application of exploratory statistical tools like Principal Component Analysis (PCA) is appropriate if the data really does form a line or a plane in the input space, but if the data forms a curved line or a surface and etc., the linear PCA is not suitable. In such a case Kohonen self organizing maps (SOMs) will overcome the approximation problem due to their topological ordering property. The SOM provides a discrete approximation of finding so-called *principal curves* or *principal surfaces*, and may therefore be viewed as a non-linear generalization of PCA [8].

A very important issue that should be taken into account during the determination of the AD of a model is uncertainty. In the literature [10] authors discussed the two main different types of uncertainty: input uncertainty and variability and structural (model) uncertainties that derives from simplifications of the reality due to limited systematic knowledge. Carcinogenicity is a complex endpoint contained uncertainty and noises. The specific of the ANN algorithm enables to get reliable results treating the data contained uncertainty and noises. However, researchers should keep in the mind that due to the uncertainty associated with individual (Q)SAR predictions, some predictions may fall within the defined AD of the model, but be unreliable due to properties and features not accounted by the model. In contrast, a chemical falling outside the defined AD, may still exhibit the response being modelled, because it elicits this response by a mechanism not accounted by the model in question. This problem was discussed in the article [2].

The ADs of models in our study were analyzed using the ED between objects (molecules) and central neurons of neural networks as well as the ED between objects (molecules) and the vector of average values of the descriptors. The investigation of the training and test sets chemicals coverage in models and descriptors space was made with the respect to false predicted chemicals. Additionally, we demonstrated the results of leverage approach [11, 12] applied for the evaluation of descriptors space in CP ANN models for comparison with linear models. Here we considered the AD of models for prediction of carcinogenicity class based on our research work within the CAESAR project [13–15].

There is not a clear consensus about the determination of thresholds in AD for non-linear classification models. As we deal with non-linear models we did not fix a "warning" threshold, but we rather investigated the prediction accuracy of model in the chemical and descriptors space and tried to find out the space where models gave reliable predictions. We evaluated how different metrics relevant to CP ANN correlated with each other and how they can be used in the interpretation of the AD of classification CP ANN models.

Data

In the study we used models that were built using the dataset of 805 chemicals extracted from CPDBAS: Carcinogenic Potency Database Summary Tables located at Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html [16]. The carcinogenic class (carcinogens (P-positive) and non-carcinogens (NP-not positive)) was assigned by the tests results on rats (male and female).

The eight (8) MDL descriptors were used in the ***model A*** and the twelve (12) Dragon descriptors in the ***model B.*** We have got a good statistical prediction of carcinogenic class: an accuracy for the training set (644 compounds) was 91% and 89% for ***models A*** and ***B***, respectively while accuracy of the test set (161 compounds) was equal to 73 and 69% for ***model A*** and ***model B***, correspondingly. The accuracy of the external test set (738 compounds) was equal to 61.4 and 60.0% for ***models A*** and ***B***, respectively. These models were described in our previous study [15].

Methods

CP ANN network and interpretation of input data

CP ANN is one of the most popular algorithms suitable for modeling classes separated with non-linear boundaries. The brief survey of the method is given in the supplementary material section "online resource" (The CPANN network method) to better understand some parameters of CP ANN models which could be used for characterisation of the AD of QSAR models. The description of the CPANN algorithm for classification is included in the supplementary material section "online resource" (CPANN algorithm for classification).

The ED between a molecule (an input pattern) and a central node

The ED is the common way of measuring the distance between vectors and can be calculated as the distance between an input pattern (molecules) and a central (winning) node in the Kohonen layer.

The ED between the input pattern (X) and a central node "***ci***" can be expressed using the following equation:

$$\boldsymbol{ED}(\mathrm{X}, w_{ci}) = \mathrm{sqrt}\left(\left(x_1^T - w_{ci1}\right)^2 + \left(x_2^T - w_{ci2}\right)^2 + \cdots + \left(x_m^T - w_{cim}\right)^2\right), \quad (1)$$

where $w_{ci1}, w_{ci2},\ldots, w_{cim}$ are the weights to the neuron "***ci***" corresponding to particular descriptor, and m is the number of descriptors or levels corresponding to a particular descriptor in the Kohonen layer. Input patterns for each level are illustrated in Table 1 as input vectors. They can be expressed as transposed matrixes. In the equation we denoted the input vector 1, 2,…, m as $x_1^T, x_2^T,\ldots, x_m^T$. Each transpose matrix $(x_1^T, x_2^T,\ldots, x_m^T)$ includes the values of descriptors $D_1, D_2,\ldots, D_m$ correspondingly, calculated for each molecule.

The Kohonen neural network or self organizing map (SOM) is an unsupervised learning algorithm. It does not give us direct information on the actual classification (carcinogens and non-carcinogens) of the data samples. It merely clusters and classifies the data set based on the set attributes (features of the descriptors) used [17]. Authors [18] showed that structural similarity of individual molecules in the same cluster is controlled and directed by the descriptors introduced into the model.

In this study the ED to the central neuron has been used as a "distance to model" metric with respect to true and false predictions, to describe the coverage of chemicals in the training and test sets.

The ED between a molecule (an object) and the vector of average values of descriptors

It should be noted that the ED between molecules and central neurons in Kohonen layers indicates the chemicals space coverage, while the distance between objects (molecules expressed as vectors of values of descriptors) and the representative object (a vector of average values of descriptors) illustrates the descriptors space. The descriptors space in CPANN models was also considered in this study. As was pointed above, the structure of s-th compound represented by m-dimensional vector of descriptors can be expressed as $x_s = (x_{s1}, x_{s2},\ldots x_{sm})$. Therefore, the ED between an object "***s***" (molecule"***s***" represented as a vector of descriptors (1,2,…, m)) and the vector of average values of particular descriptors (1, 2,…, m) related to all chemicals "***k***"in the dataset have been calculated using the following equation (see Table 2):

$$ED_s(x, ave\_d) = \mathrm{sqrt}\left((x_{s1} - ave\_d_1)^2 + (x_{s2} - ave\_d_2)^2 + \ldots + (x_{sm} - ave\_d_m)^2\right) \quad (2)$$

Table 1 The input data matrix using in CP ANN algorithm

Molecules (objects)	***Descriptors***						***Targets***
	D_1	D_2	…	D_i	…	D_m	T
X_1	X_{11}	X_{12}	…	X_{1i}	…	X_{1m}	T_1
X_2	X_{21}	X_{22}	…	X_2	…	X_{2m}	T_2
…	…	…	…	…	…	…	…
X_s	X_{s1}	X_{s2}	…	X_{si}	…	X_{sm}	T_s
…	…	…	…	…	…	…	…
X_k	X_{k1}	X_{k2}	…	X_{ki}	…	X_{km}	T_k
	Input vector_1 (x^T_1)	***Input vector_2*** (x^T_2)	…	***Input vector_i*** (x^T_i)	…	***Input vector_m*** (x^T_m)	***T-inputs during trainig***

where

$$\text{ave\_}d_1 = (x_{11} + x_{21} + \ldots x_{s1} + \cdots + x_{k1})/\text{k}$$
$$\text{ave\_}d_2 = (x_{12} + x_{22} + \ldots x_{s2} + \cdots + x_{k2})/\text{k}$$
$$\text{ave\_}d_\text{m} = (x_{1m} + x_{2m} + \ldots x_{sm} + \cdots + x_{km})/\text{k}$$

In the Eq. 2 the average value of 1st, 2d,…, mth descriptor for dataset containing "**k**" compounds was labelled as ave_d_1, ave_d_2, …, ave_d_m, respectively. The ED between all objects (all molecules represented as a vector of descriptors) and the vector of average values of particular descriptors can be plotted versus predicted values (*Y_predicted*) or difference between target and predicted values (*Y_target–Y_predicted*). As a result, one can get a graphical representation of how the descriptor values of chemicals in the training set are distributed in relation to the class predicted by the model or according to the prediction error.

The leverage approach

The distance-based method, specifically the leverage approach [11, 12], was applied in our study for the comparison between linear and non linear models approaches. Leverages are measures of the distance between the x values for an observation and the mean of x values for all observations. In terms of parameters used in our study this approach provides a measure of the distance between the descriptors values for a chemical and the mean of descriptor values for all chemicals. A large leverage value indicates that the x values of an observation are far from the center of x values for all observations. The leverage h of a compound measures its influence on the model. The leverage of a compound in the original variable space is defined as:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i (i = 1, \ldots, n) \quad (3)$$

where x_i is the descriptor vector of the considered compound and X is the model matrix formed with descriptors values from the training set.

Williams plots can be applied to visualize the AD of QSAR models, where leverage values (or hat values) are plotted versus standardized residuals for each compounds [19, 20]. As we deal with classification non linear CP ANN model we plotted the difference between target and predicted values (*Y_target−Y_predicted*) for our response instead of residues which could be calculated in case of regression model. Then we have investigated the distribution of chemicals along the x axis.

The warning leverage (h^*) of leverage threshold is generally fixed at 3p/n, where n is the number of training chemicals, and p the number of model variables (descriptors) plus one. In Williams plot, chemicals influential on the structural domain of the model, are characterized by leverage (hat value) exceeding warning leverage threshold and they should be carefully examined. Therefore, we considered chemicals outside the limits keeping in mind that these thresholds are suitable only in the case of linear models while we are dealing with non-linear ones.

Results and discussions

The reliability prediction analysis depending on the thresholds of classification models

The goal of an AD is to set up boundaries whereby the obtained predicted values can be trusted with confidence. Hence, it follows that the essential problem of the AD definition is to find out the uncertainty areas where less reliable predictions fall. It should be emphasized that the accuracy (ACC) of models indicates their level of

Table 2 The input data matrix complemented with calculation of Euclidean distance between molecules represented as a vectors of descriptors (1, 2,…, m) and the vector of the average values of particular descriptors (1, 2,…, m) related to all chemicals "***k***" in the dataset

Mol	***Descriptors***				**ED (*x*,*ave_d*)**- Eucledian distance between an object **(x)** and vector of the average values of descriptors **(ave_d)**
	D_1	D_2	…	D_m	
X_1	X_{11}	X_{12}	…	X_{1m}	sqrt $((x_{11} - ave\_d_1)^2 + (x_{12} - ave\_d_2)^2 + \ldots + (x_{km} - ave\_d_m)^2)$
X_2	X_{21}	X_{22}	…	X_{2m}	sqrt $((x_{21} - ave\_d_1)^2 + (x_{22} - ave\_d_2)^2 + \ldots + (x_{km} - ave\_d_m)^2)$
…	…	…	…	…	…
X_s	X_{s1}	X_{s2}	…	X_{sm}	sqrt $((x_{s1} - ave\_d_1)^2 + (x_{s2} - ave\_d_2)^2 + \ldots + (x_{sm} - ave\_d_m)^2)$
…	…	…	…	…	…
X_k	X_{k1}	X_{k2}	…	X_{km}	sqrt $((x_{k1} - ave\_d_1)^2 + (x_{k2} - ave\_d_2)^2 + \ldots + (x_{km} - ave\_d_m)^2)$

$ave\_d_m = (x_{1m} + x_{2m} + \ldots x_{sm} + \ldots + x_{km})/k$

$ave\_d_2 = (x_{12} + x_{22} + \ldots x_{s2} + \ldots + x_{k2})/k$

$ave\_d_1 = (x_{11} + x_{21} + \ldots x_{s1} + \ldots + \ldots + x_{k1})/k$

reliability. Therefore, uncertainty areas in our models depend on ACC of models which is proportional to the correctly predicted compounds (carcinogens and non-carcinogens). The ACC of CPANN classification models, in turn, depends on the selected threshold. (see "CPANN algorithm for classification" in supplementary material section "online resource").

In the modeling we accepted target (T = Y_target) as one (1) for carcinogen class and as zero (0) for non-carcinogen (discrete numbers in our classification models). The predicted values of response (Y = Y_predicted) are expressed as continues values in the interval from 0 to 1. The thresholds of models serve for the separation of carcinogens and non-carcinogens (see Fig. 1). As was reported in the article [15] the threshold of ***model A*** was established as 0.45 while the threshold of ***model B*** was equal to 0.5. If S is a compound to be predicted and $y(S)$ is a continuous prediction value, calculated by the model, then the predicted class for the given compound S ($c(S)$) is identified by the following requirements:

$$c(S)\left\{1 = \text{carcinogen}, y(S) > 0.45;\ 0 = \text{non carcinogen}, y(S) \le 0.45\right\} \text{ for } \textbf{\textit{model A}}$$

$$c(S)\left\{1 = \text{carcinogen}, y(S) > 0.5;\ 0 = \text{non carcinogen}, y(S) \le 0.5\right\} \text{ for } \textbf{\textit{model B}}.$$

It is obvious that for carcinogens with Y_target equal to 1 the predicted values from 0 to 0.45 (***model A***) or from 0 to 0.5 (***model B***) are false predicted while the values closer to 1 will be more reliable and visa versa, in the case of non-carcinogens with Y_target equal to 0 the predicted values from 1 to 0.45 (***model A***) or from 1 to 0.5 (***model B***) are false predicted while the values closer to 0 will be more reliable. The data closer to the threshold can be determined by our algorithm as correctly predicted but they are less reliable. Thus, we can conclude that the area closer to the threshold is the uncertainty area because here the results of a model contain very high uncertainty in prediction.

A measure of the prediction uncertainty can be also expressed as the absolute value of the difference $d(S)$

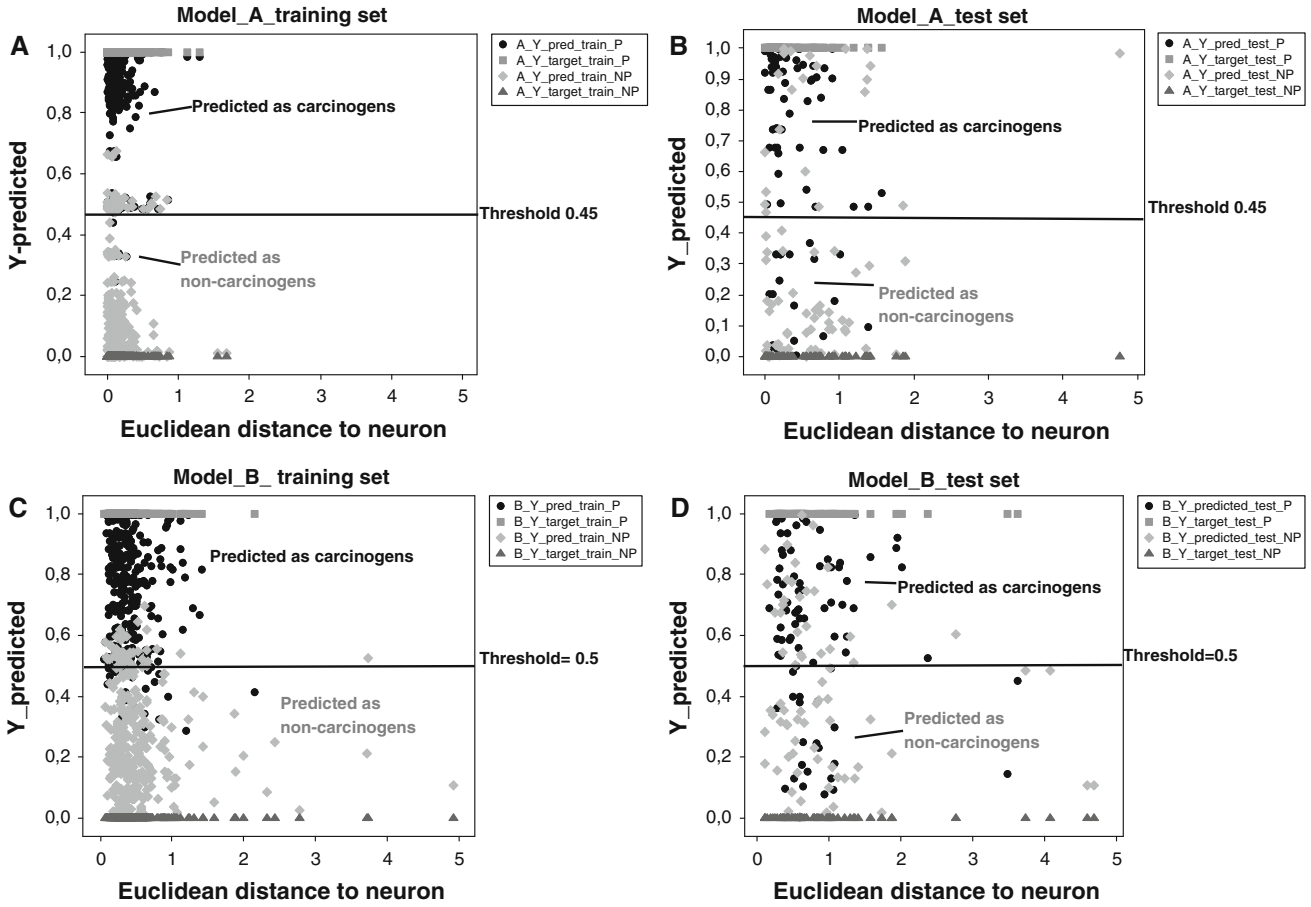

Fig. 1 Plots of *ED to the central neuron* (with indication of carcinogens (**P**) and non-carcinogens (**NP**)) versus *Y_predicted* (and *Y_target*) for ***model A*** (**a** for training set and **b** for test set) and for ***model B*** (**c** for training set and **d** for test set). *Notes*: positive (**P**) target data are specified as squares (*filled square*); non-positive (**NP**) target data marked as triangles (*filled triangle*); positive (**P**) predicted data identified as circles (*filled circles*); non positive (**NP**) predicted data represented as diamonds (*filled diamonds*)

between the prediction value and the nearest label of target values (0 for non-carcinogens and 1 for carcinogens).

$$d(S) = \min\{|0 - y(S)|, |1 - y(S)| \quad (4)$$

Figure 2 plots the *ED to the central neuron* versus (*Y_target-Y_predicted*) and shows the distribution of true predicted (true positive (TP) + true negative (TN)) and false predicted (false positive (FP) + false negative (FN)) chemicals in relation to the threshold. Thus, true predicted chemicals in the ***model A*** (Fig. 2a) are located between −0.45 and 0.55 while in the ***model B*** (Fig. 2b) the true predicted chemicals are distributed between −0.5 and 0.5. All predicted values below −0.45 and above 0.55 are false predicted in the ***model A*** while in the ***model B*** false predicted chemicals are located below −0.5 and above 0.5.

The analysis of the AD metric expressed as the ED between molecules (objects) and the central neuron

The ED between objects (molecules) and central neuron in Kohonen layer of CP ANN models is the essential characteristic of neural network. We used this metrics to compare the training and test sets chemical coverage of models with respect to false predicted chemical space.

The ED represents the interval between a node in the Kohonen layer and an input pattern. The distances are unitless, because all descriptors have been autoscaled. It was noted in the literature [18] that in fact, the sum of distances of all molecules obtained after training of the network during one epoch is equal to the cumulative training error associated with the Kohonen layer. Thus, the ED depends on the input values of descriptors from one side and is connected with the errors from another one.

The distribution of the *ED to the central neuron* versus *Y_predicted* (and *Y_target*) is presented in Fig. 1a–d for ***model A*** and ***B***, respectively, with indication of carcinogens (P) and non-carcinogens (NP). Figure 1a, c correspond to the training set of ***models A*** and ***B*** while Fig. 1b and d are related to the test set of ***model A*** and ***B*** respectively. In Fig. 1 positive (P-carcinogens) target data are specified as squares (■) located on the level Y = 1, non-positive (NP—non-carcinogens) target are marked as triangles (▲) located on the level Y = 0, P predicted data identified as circles (●) while NP predicted data represented as diamonds (◆).

As was pointed above Fig. 1a, c illustrate the discrimination between carcinogens and non-carcinogens on each side of the thresholds 0.45 and 0.5 in case of ***model A*** and **B** correspondingly. The area close to threshold represents the uncertainty zone of the prediction. The false predicted or predicted with high level of uncertainty chemicals are aggregated here. Thus, the predicted carcinogens (P) (●), which are closer to the edge of class P (Y = 1;(■)) and the predicted non-carcinogens (NP) (◆) located close to the edge of class NP (Y = 0;(▲)), are assumed to have better prediction accuracy than dots, located in the middle (uncertainty area) between the classes near the value 0.45 for ***model A*** and 0.5 for ***model B***.

The ***model A*** with higher accuracy (91% (training set) and 73% (test set)) results in less sparse distribution of chemicals in comparison with ***model B*** (accuracy 89% (training set) and 69% (test set) in the plot of the *ED* versus *Y_predicted* (see Fig. 1). The similar observation was obtained comparing the training and test sets. It should be noted that the distribution of chemicals from the training set (Fig. 1a, c) is less sparse in comparison with the test set (Fig. 1b, d) because the ACC of the training set is greater (91% ***model A*** and 89% ***model B***) then the test one (73% ***model A*** and 69% ***model B***).

In the ***model B*** we obtained a less sparse distribution in the case of carcinogens (the majority of chemicals distributed in the narror interval ED < 1.5). The distribution

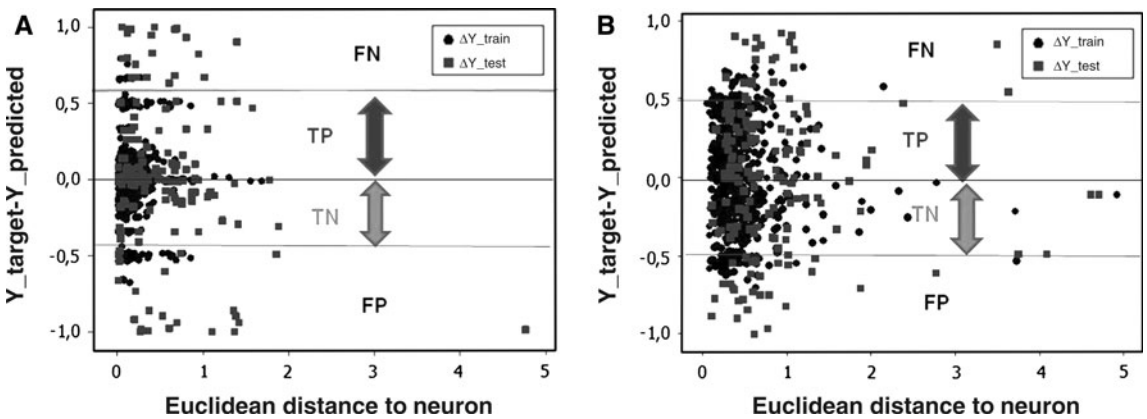

Fig. 2 Plots of *ED to the central neuron* versus (*Y_target—Y_predicted*) for the ***model A*** (**a**) and for the ***model B*** (**b**) with indication of true and false predicted chemicals. *Notes*: *TP* true positive, *TN* true negative, *FP* false positive, *FN* false negative. Test data are specified as squares (*filled square*) while training data are identified as circles (*filled circle*)

of non-carcinogens in turn was observed in the wider interval of ED < 5.

Figure 1 shows that errors of prediction and uncertainty areas don't depend on value of the ED. True and false predicted chemicals are spread more or less evenly throughout the studied space (0 < ED < 5). We can conclude that coverage estimation should be used only as a warning, and not as a final decision of a "model applicability". We have considered chemical space in the interval of the ED from 0 till 5 here. The studied models are characterized by the ED less than 5.

The more detail characterization of chemicals with the largest ED is given below.

Comparison of AD metrics for non linear model and linear one

The comparison of three different metrics: the ED between objects molecules (objects) and the central neuron (*metric 1*), the ED between vectors of descriptors and vectors of average values of the descriptors (*metric 2*) as well as the leverage (*metric 3*) was described here. The first two metrics were aimed for the characterization of non linear models while the leverage approach usually intent for the characterization of linear models.

It was noted in the Joint Research Center (JRS) report [21] that the definition of similarity for molecules depends on the representation of the molecules under consideration in the descriptor space. It should be highlighted that the selected 2D descriptors represent the arrangement of atoms using the topology of molecule and the connectivity of constituting atoms. Thus, the ED here demonstrates the similarity or dissimilarity between the compared objects (molecules). The ED between vectors of descriptors for each molecule in the dataset and the vector of average values of particular descriptors (*metric 2*) and the leverage (*metric 3*) were used to explore the descriptor space.

To compare different AD metrics we have plotted the *ED to central neuron* (Fig. 3a, b) (related to the chemicals space coverage), the *ED to vector of average descriptors* (Fig. 3c, d) and the *leverage* (Fig. 3e, f) (related to the descriptors space) versus (*Y_target*−*Y_predited*). Figure 3a, c, e correspond to the ***model A*** while Fig. 3b, d, f relate to the ***model B***. In Fig. 3 we have marked only the more distant chemicals while in Figure 2SI of the Online Resource one can see all chemicals labelled. The detail information about chemicals labelled in Fig. 3 is shown in the supplementary material "Online Resource" Table 1SI including their name and the CAS registration number (CAS_RN), the molecular weight (MW), and the indication of maximum or minimum value of descriptors intrinsic for studied compounds. Information about carcinogens (P) and non-carcinogens (NP) is presented in the table as well as the indication of model (A, B), if the chemical was found in one or both models approximately at the same level of the ED. Chemicals from the training and the test set are represented separately.

The following findings were obtained analyzing the chemicals with the largest ED to the central neuron or to the vector of average values of particular descriptors or the leverage (see chemicals located from the right side in Fig. 3). First of all, the majority of compounds from the training set have MW greater than 500 or close to 500. Thus, the compounds numbers 151, 19, 107, 635, 521, 550, 226 and 642 have the MW equal to 1,135, 1,255, 536, 811, 666, 823, 457, and 629 respectively. Some of chemicals correspond to the largest (max) or smallest (min) value of descriptors (see Table 1SI of the "Online Resource"). The majority of chemicals are NP-non carcinogens. The pointed features are intrinsic for both models A and B.

Comparing all three metrics one can notice the similarity in the distribution of the most distant chemicals in both models (***model A*** and ***B***). Thus, the compounds marked with number 19, 107, 151, 550, 635, and 642 have the largest value of the ED or the leverage.

The most distant chemical from the test set can be found under the number 37 (see Fig. 3a). This compound is Cyclosporin A (CAS59865-13-3) with formula C62H111N 11O12, MW = 1,202. Cyclosporin A has the Y_target value = 0. It is NP chemicals while it was predicted with score 0.9838 as carcinogen (Y_target_carcinogen = 1). The ED was very high = 4.7606 in the ***model A***. We suggested that due to the complex chemical structure the model could not be accurate in estimating this compound. From the other hand, we have found that this chemical has the smallest value of molecular connectivity descriptor D4 (*dxp9*- difference simple 9th order path chi indices) in comparison with values for the all others chemicals in the dataset (see Table 1SI of the "Online Resource").

It should be pointed that for linear models the Williams plot (the plots of standard residuals on the *y-axis* versus leverage values on the *x-axis*) is typically used as was described in the section methods. From these plots, the applicability domain is established on the left from the leverage threshold (or limit) ($h^* = 3p/n$, where p is number of descriptors plus 1 and n is the number of compounds in the training set). According to this equation we set the leverage threshold at 0.04 for ***model A*** and 0.06 for ***model B***. According to these pointed limits (see Fig. 3e (***model A***) and F (***model B***)) a lot of chemicals appeared to be outside domain. Practically we cannot use such thresholds for the determination of applicability domain of our models as the chemicals outside these limits are correctly predicted and the model is reliable not only inside the pointed domain.

The maximal value of the ED characterizes the boundaries of studied models. The *metric 1* can be used for

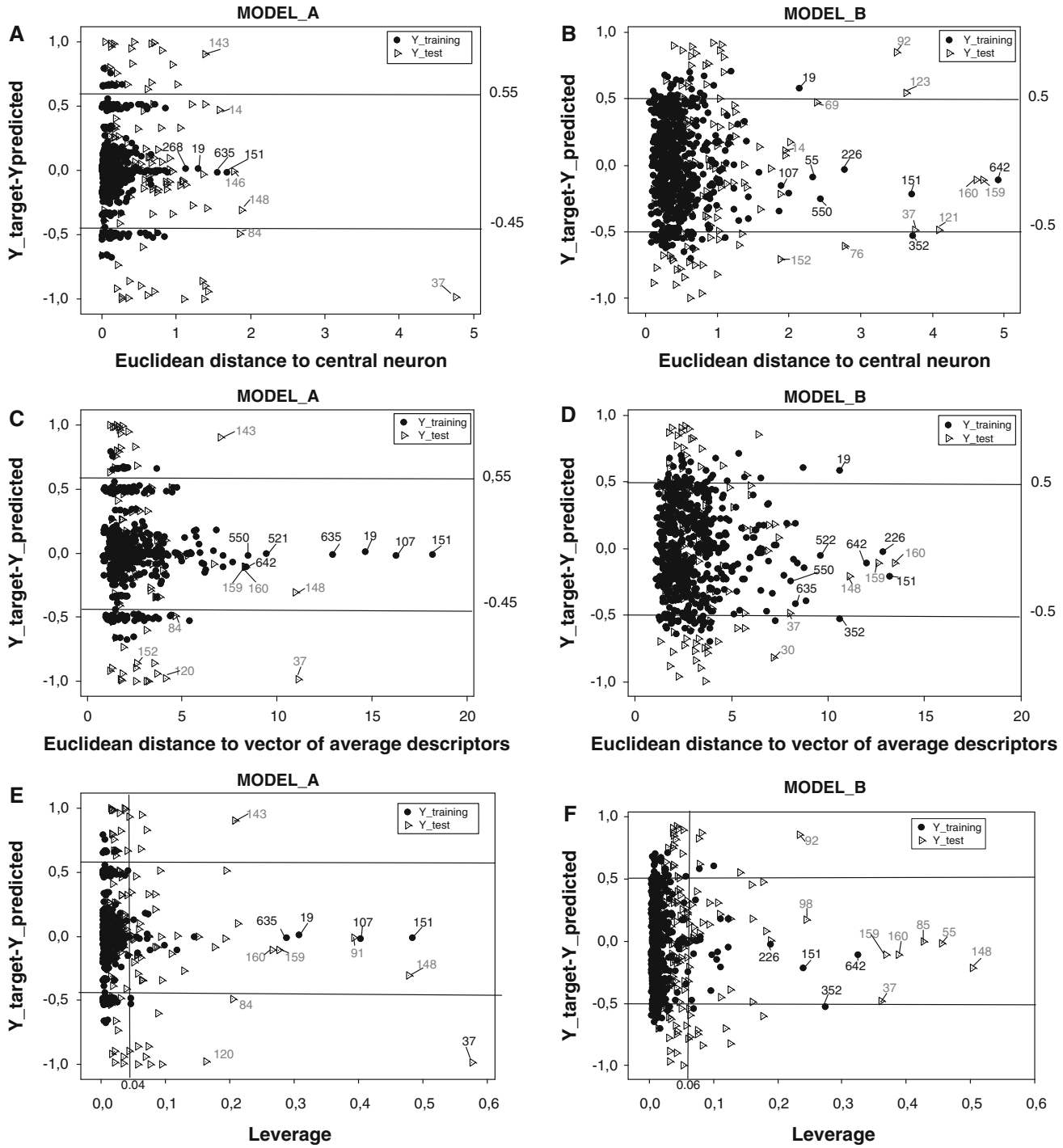

Fig. 3 Plots of *ED to the central neuron* versus (*Y_target−Y_predicted*) (**a, b**); *ED to vector of average descriptors* versus (*Y_target−Y_predicted*) (**c, d**) and t*he leverage* versus (*Y_target−Y_predicted*) (**e, f**) for ***models A*** and ***B*** correspondingly. *Note*: The training data are identified as circles (*filled circle*) while the test data are specified as triangles (*open triangle*)

characterization of chemical space, the *metric 2* characterise the descriptors space.

The similarities in studied metrics were found in respect to distribution of chemicals which have the MW greater 500. In the following section we explored the relationship between the MW and the ED to the central neuron.

How MW affect the distribution of chemicals depending on the ED to the central neuron in respect to the false prediction

The studied chemicals were divided into three groups: compounds with MW greater than 500 (Fig. 4a), compounds

with MW between 250 and 500 (Fig. 4b) and compounds with MW less than 250 (Fig. 4c). Figure 4 shows the plots of the *ED to central neuron* versus *Y_predicted* for pointed groups of chemicals for ***model A***.

In Fig. 4 positive (P) target data are specified as squares (■) located on the level Y = 1, non-positive (NP) target marked as diamonds (◆) located on the level Y = 0, P predicted data identified as circles (○) while NP predicted data represented as triangles (Δ). As shown in Fig. 4a the compounds with largest MW have the largest ED values in comparison with chemicals from other groups (Fig. 4b, c). Interestingly, all chemicals with MW bigger than 500 are correctly predicted while chemicals with MW smaller than 250 (Fig. 4c) have false predicted results (see P predicted (○) and NP predicted (Δ) chemicals located outside the threshold 0.45). For the group of chemicals with MW between 250 and 500 (Fig. 4b) we have got all true predicted carcinogens (P) and a few wrongly predicted non-carcinogens (NP) which are located above the threshold 0.45.

The following conclusions can be done. Chemicals with the MW less than 250 have the smallest ED (see Fig. 4c) in comparison with chemicals from other groups (Fig. 4a, b).

The majority of chemicals with large ED are correctly predicted. CP ANN algorithm is able to correctly predict complex molecules with large MW. On the contrary, the majority of wrongly predicted chemicals have MW less than 250.

How different SAs affect the distribution of chemicals depending on the ED to the central neuron in respect to false prediction

In this part of the study the compounds with specific carcinogenicity structural alerts (SA) were investigated. We have extracted alerts from the Toxtree program for the dataset of 805 compounds. More detailed information related to carcinogenic structural alerts (SA) and description of their symbols is given in the literature [22].

The largest groups of chemicals in our study have the following SAs: SA_8: Aliphatic halogens (47 compounds); SA_21: alkyl and aryl N-nitroso groups (107 compounds); SA_28: primary aromatic amine, hydroxyl amine and its derived esters (52 compounds); SA_X denotes others SAs in the dataset; NA denotes compounds which have no alerts. We also included chemicals containing two different SAs—(*SA27* + *SA28*) (14chemicals) where SA_27 belongs to nitro-aromatic compounds.

Figure 5 represents *ED to central neuron* versus *Y_predicted* (and *Y target*) for each pointed above group of chemicals. Y target data are specified as squares (■) while predicted data represented as circles (○).

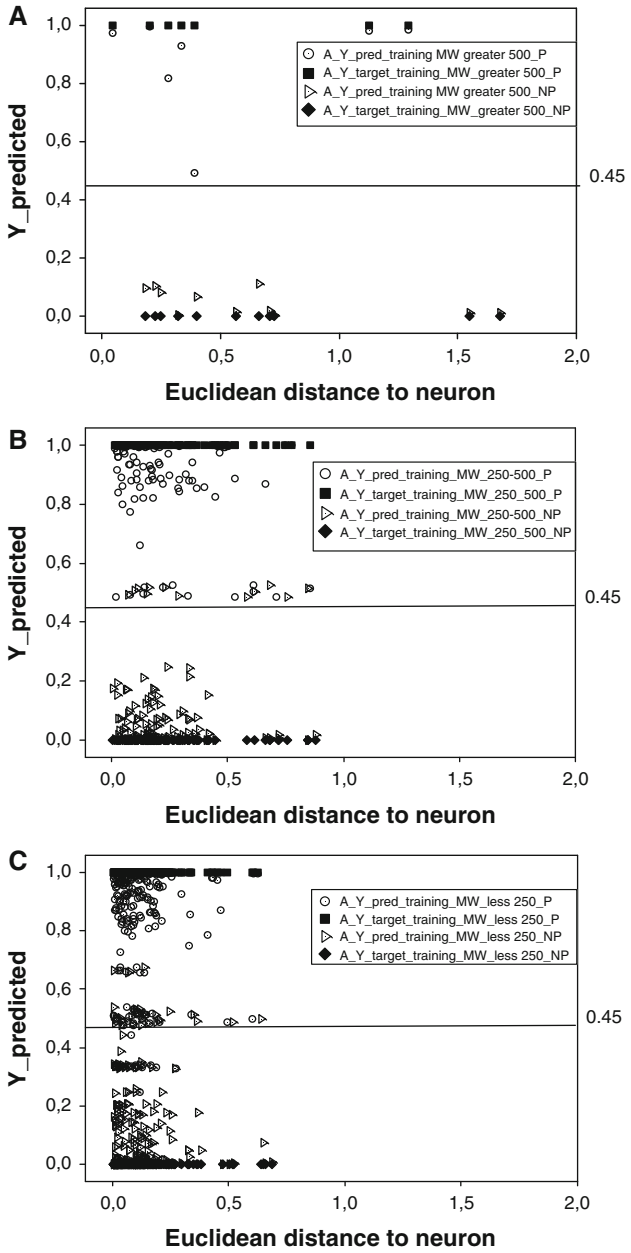

Fig. 4 Plots of *ED to central neuron* versus *Y_predicted* for compounds with MW greater than 500 (**a**), for compounds with MW between 250 and 500 (**b**), and for compounds with MW less than 250 (**c**). *Notes* positive (P) target data are specified as squares (*filled square*), non-positive (NP) target marked as diamonds (*filled diamond*), P predicted data identified as circles (*open circle*) while NP predicted data represented as triangles (*open triangle*)

The groups of chemicals containing pointed above carcinogenic SAs do not have large dispersion by ED value (Fig. 5a–d). The greatest scattering of chemicals was obtained in the case of compounds without alerts (NA chemicals) (Fig. 5f). Non congeneric chemicals with different SAs marked as SA_X gave us also wide data scattering (Fig. 5e).

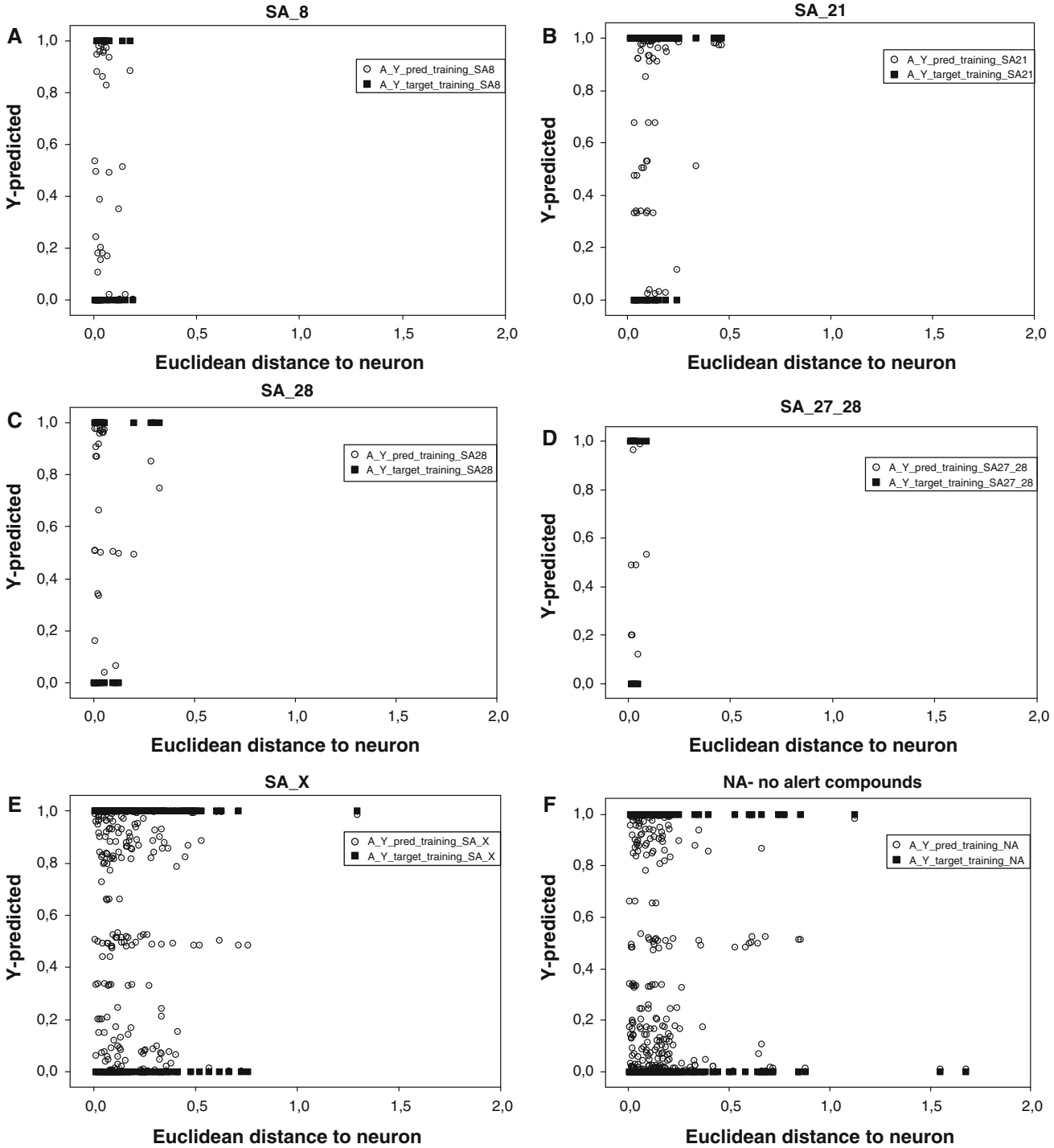

Fig. 5 Plots of *ED to central neuron* versus *Y_predicted* (*and Y_target*) for compounds with SA_8 (**a**), SA_21 (**b**), SA_28 (**c**), SA_27_28 (**d**), SA_X (**e**) and *NA*- no alert compounds (**f**). *Notes* SA_8: Aliphatic halogens (47 compounds); SA_21: alkyl and aryl N-nitroso groups (107 compounds); SA_28: primary aromatic amine, hydroxyl amine and its derived esters (52 compounds); SA_27_28-compounds containing both SAs 27 and 28 (14 compounds) where SA_27: Nitro-aromatic compounds; SA_X- others SAs in dataset; *NA*- compounds with no alerts. Y target data are specified as squares (*filled square*) while Y predicted data represented as circles (*open circle*)

Conclusion

The ED was proposed for characterisation of non-linear CPANN models. Three metrics were compared to describe the domain of models: ED between objects (molecules) and central neurons of neural networks, ED between objects (molecules) and the representative object (vector of average values of descriptors) (as a characteristics of non linear models) as well as and leverage (hat value) (as a characteristic of linear models).

The ED between objects (molecules) and central neuron in Kohonen lyer of CPANN models gave us ability to compare the training and test sets chemical coverage of models with respect to false predicted chemical space. The ED between vectors of real values of descriptors and the vector of average values of descriptors was used to explore the coverage of the descriptor space for the training and the test set chemicals in the models with respect to the space of wrongly predicted chemicals. Additionally, we showed the results of the leverage approach applied for evaluation of descriptors space for comparison with linear models. The threshold of leverage approach is not suitable for non linear method like CP ANN.

Chemicals with the maximal value of ED to central neuron were investigated. The majority of compounds with the largest ED has the largest MW (greater than 500), and/or are NP-non-carcinogens, and/or corresponds to chemicals with largest or smallest value of descriptors. The large value of the ED in model is not evidence of wrong prediction.

In the study we did not fix a "warning" threshold but, rather investigated the prediction accuracy of the model in chemical and descriptors space and try to find out the space where models give reliable predictions.

The ED in the non linear models demonstrates a boundaries where the model was built and is applicable with the determined reliability.

Acknowledgments Authors thank for the European Commission for the financial support under project CAESAR (SSPI-022674), the Slovenian Ministry of Higher Education, Science and Technology (grant P1-017).

References

1. OECD (2007) Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] MODELS. OECD Environment Health and Safety Publications, Series on Testing and Assessment No. 69, [http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)2]
2. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS et al (2005) Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52. ATLA 33:155–173
3. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim 33(5):445–459, ISSN: 0261-1929
4. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A (2008) Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model 48(9): 1733–1746. doi:10.1021/ci800151m
5. Dimitrov SD, Dimitrova GD, Pavlov TS, Dimitrova N, Patlewicz GY, Niemela J, Mekenyan OG (2005) A stepwise approach for defining the applicability domain of SAR and QSAR models. J Chem Inf Model 45:839–849. doi:10.1021/ci0500381
6. Duda RO, Hart PE, and Stork DG (2001) Pattern Classification. Wiley, New York: 654. doi: 10.1007/s00357-007-0015-9
7. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity—a review. QSAR Comb Sci 22(9–10):1006–1026. doi: 10.1002/qsar.200330831
8. Haykin S (1998) Neural networks: a comprehensive foundation, 2nd edn. Prentice Hall. Inc, Englewood Cliffs
9. Benfenati E, Benigni R, DeMarini D, Helma C, Kirkland D, Martin TM, Mazzatorta P, Ouedrago-Arras G, Richard AM, Schilter B, Schoonen WG, Snyder RD, Yang C (2009) Predictive models for carcinogenicity: frameworks, state-of-the-art, and perspectives. J Environ Sci Health C 27:57–90. doi:10.1080/10590500902885593
10. Walker JD, Lars Carlsen L, Jaworska J (2003) Improving opportunities for regulatory acceptance of QSARS: the importance of model domain, uncertainty, validity and predictability. QSAR Comb Sci 22(3):346–350. doi:10.1002/qsar.200390024
11. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375. doi:10.1289/ehp.5758
12. Gramatica P, Pilutti P, Papa E (2003) Predicting the NO″3 radical tropospheric degradability of organic pollutants by theoretical molecular descriptors. Atmos Environ 37:3115–3124. doi: 10.1016/S1352-2310(03)00293-0
13. Caesar project web page http://www.caesar-project.eu/software
14. Fjodorova N, Vračko M, Tušar M, Jezierska A, Novič M, Kühne R, Schüürmann G (2010) Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses. Mol Divers 14(3):581–594. doi:10.1007/s11030-009-9190-4
15. Fjodorova N, Vračko M, Novič M, Roncaglioni A, Benfenati E (2010) New public QSAR model for carcinogenicity. Chem Cent J 4(Suppl 1):S3, http://www.journal.chemistrycentral.com/content/4/S1/S3. doi:10.1186/1752-153X-4-S1-S3
16. Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html
17. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley-VCH Verlag GmbH, Weinheim
18. Maran E, Novic M, Barbieri P, Zupan J (2004) Application of counterpropagation artificial neural network for modelling properties of fish antibiotics. SAR QSAR Environ Res 15(5–6): 469–480. doi:10.1080/10629360412331297461
19. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77. doi:10.1002/qsar.200390007
20. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26(5):694–701. doi: 10.1002/qsar.20061015

21. Saliner AG, Patlewicz G, Worth AP (2005) A similarity based approach for chemical category classification. JRS report EUR 21867 EN:1–44

22. Benigni R, Bossa C, Jeliazkova N, Netzeva TI, Worth AP (2008) The Benigni/Bossa rulebase for mutagenicity and carcinogenicity—a module of Toxtree. EUR 23241 EN:1–70