Analytica Chimica Acta 759 (2013) 28–42

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

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

Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study☆

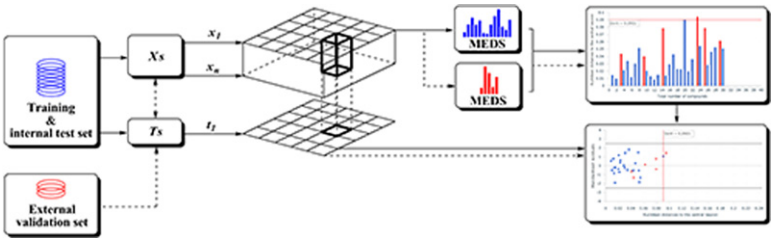
Nikola Minovski**,[1], Špela Župerl[1], Viktor Drgan, Marjana Novič*

National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia

HIGHLIGHTS

- The concept of applicability domain (AD) in QSAR modeling is discussed.
- The AD assessment method for non-linear neural network predictive models is proposed.
- The counter-propagation artificial neural network (CP-ANN) was applied for modeling.
- Minimal Euclidean distance space (MEDS) of CP-ANN model was defined and analyzed.
- The resulting outliers coincide with those from linear models (leverage based AD).

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 9 July 2012
Received in revised form 30 October 2012
Accepted 2 November 2012
Available online 15 November 2012

Keywords:
Quantitative structure–activity relationship
Artificial neural networks
Counter-propagation artificial neural network
Applicability domain
Euclidean distance
Leverage approach

ABSTRACT

Alongside the validation, the concept of applicability domain (AD) is probably one of the most important aspects which determine the quality as well as reliability of the established quantitative structure–activity relationship (QSAR) models. To date, a variety of approaches for AD estimation have been devised which can be applied to particular type of QSAR models and their practical utilization is extensively elaborated in the literature. The present study introduces a novel, simple, and effective distance-based method for estimation of the AD in case of developed and validated predictive counter-propagation artificial neural network (CP ANN) models through a proficient exploitation of the Euclidean distance (ED) metric in the structure-representation vector space. The performance of the method was evaluated and explained in a case study by using a pre-built and validated CP ANN model for prediction of the transport activity of the transmembrane protein bilitranslocase for a diverse set of compounds. The method was tested on two more datasets in order to confirm its performance for evaluation of the applicability domain in CP ANN models. The chemical compounds determined as potential outliers, i.e., outside of the CP ANN model AD, were confirmed in a comparative AD assessment by using the leverage approach. Moreover, the method offers a graphical depiction of the AD for fast and simple determination of the extreme points.

Abbreviations: QSAR, quantitative structure–activity relationship; KANN, Kohonen artificial neural networks; CP ANN, counter-propagation artificial neural network; ED, Euclidean distance; MEDS, minimum Euclidean distance space; AD, applicability domain.

☆ Paper presented at the XIII Conference on Chemometrics in Analytical Chemistry (CAC 2012), Budapest, Hungary, 25–29 June 2012.

* Corresponding author. Tel.: +386 1 4760 253; fax: +386 1 4760 300.
** Corresponding author. Tel.: +386 1 4760 383; fax: +386 1 4760 300.
E-mail addresses: nikola.minovski@ki.si (N. Minovski), spela.zuperl@ki.si (Š. Župerl), viktor.drgan@ki.si (V. Drgan), marjana.novic@ki.si (M. Novič).

[1] These authors contributed equally.

0003-2670/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2012.11.002

1. Introduction

In the past 40 years, the interest in the quantitative structure–activity relationship (QSAR) as a concept is undoubtly increased. Just in the period between 1975 and 2011, an incredible number of approximately 11,000 published articles observing the problem of QSAR can be found (Web of Knowledge), which data clearly indicate the emergence and usefulness of the developed QSAR models for solving a variety of chemical problems [1,2]. Beside these facts, the quality as well as reliability of the QSAR models must be also taken into account [3]. Therefore, the OECD principles for validation of the QSAR models for regulatory purposes, clearly stated that a model should be used within the boundaries of its applicability domain (AD) [4]. According to the Setubal's Workshop guidance and acceptability criteria for application of QSAR models for chemical management purposes, the AD of a QSAR model is defined as a "*physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The AD of a QSAR should be described in terms of the most relevant parameters, i.e., usually those that are descriptors of the model*" [5].

Depending on the modeling strategy utilized, several approaches for assessing the AD were developed [6] that can be applied to a particular type of QSAR model (The Report and Recommendations of ECVAM Workshop 52) [7]. A comprehensive description of these approaches can be found in the review articles by Jaworska and Nikolova-Jeliazkova [8,9]. As described there, the methodologies for assessment of the AD of QSAR models are classified in four major categories: (1) range-based methods, (2) geometric methods, (3) distance-based methods, and (4) probability density distribution methods.

Distance (similarity) metrics are very frequently used in QSAR for solving different chemical problems [10–12] and their practical utilization in the model's AD evaluation is recently described in the literature [13,14]. Among the various distance metrics for model's AD assessment (e.g., Euclidean, Mahalanobis, Manhattan, Hotelling T^2, and Leverage), the Euclidean distance (ED) metric is one of the most commonly utilized. In its simplest form, it can be defined as an "ordinary distance", i.e., a line connecting two points A and B defined by their two-dimensional coordinates $A(x_1, x_2)$ and $B(y_1, y_2)$. This definition is mainly related to two-dimensional space, i.e., Euclidean plane [15]. However, the majority of the statistical methods used in the QSAR modeling (including artificial neural networks) deal with the problem of modeling multidimensional data and therefore the ED between a query point (e.g., a training/test/external validation set compound) and the centroid of the model can be determined within the framework of a more complex high-dimensional Euclidean space [16,17]. The query object n_i (e.g., a compound from the training/test/external validation set) is defined as a vector of independent or input variables (e.g., molecular descriptors), whereas the model's centroid is usually defined as a multidimensional averaged vector obtained by averaging the input vectors for all objects representing the compounds in the dataset.

In non-linear modeling methods such as artificial neural networks (ANNs), there are two general strategies for determining the activation signal of a neuron, by either calculating the dot product (vector scalar product) between the input vector and the weight vector (synaptic efficacy) for each neuron, or by determining similarity (Euclidean distance) between the input vector and the neuron in the Kohonen type of neural networks. In both cases, the ANNs algorithm computes the so-called net input, which procedure runs repeatedly for all the neurons constructing the network. In the learning strategy of the Kohonen or self-organizing type of neural networks, the neuron with the maximal output of the dot product or the minimal ED is chosen as the "*winning*" or "*central*" neuron [18,19]. This iterative procedure (i.e., learning of the network) results in calculation of the minimum EDs for each input object (compound) to the "*central*" neuron and construction of a so-called "*minimum ED space*" (MEDS) [20]. Therefore, the model's coverage, i.e., the boundary to which a model is applicable could be simply defined by selection of the training set object with maximal value for ED within the MEDS (Fig. 1).

Comparing to some recommended distance-based methods widely used for solving the AD problem for linear models (e.g., the well known leverage approach [3]), a very small number of publications observing the problem of distance-based AD estimation for non-linear models can be found in the literature [21–23]. In a comprehensive comparative study published recently, Fjodorova et al. demonstrated a successful utilization of ED metrics for estimation of AD in case of non-linear ANNs-based classification models [24].

This study presents an effective methodology for graphical assessment of AD for non-linear ANNs-based prediction models (descriptor space vs. model response space) – counter-propagation artificial neural network (CP ANN) models by taking into account the so-called "*minimum ED space*" (MEDS) as a function of the total number of objects (compounds from the training/test/external validation set), coupled with a standard residual analysis. We focused on a detailed explanation along with a graphical depiction of the AD for three pre-built and validated CP ANN models, as well as on a thorough in-depth assessment of the detected outliers. Furthermore, the same case studies were approached by the partial least squares (PLS) method and the leverage-based AD assessment of the models was performed for the purpose of comparison of both methods.

2. Data and computational methods

In order to define and assess the practical applicability for a given non-linear ANNs model utilizing the MEDS methodology, a developed and validated ANN model must be available. For these purposes we used our previously published CP ANN predictive models developed for three different datasets. The data, modeling strategies utilized as well as the CP ANN prediction models developed are already elaborated in details [25–27], and therefore we give here only a short abridgement.

2.1. Datasets, experimental details and CP ANN models development

2.1.1. Dataset 1

The biological assay data (a total of 88 compounds) which belong to different chemical classes (nucleobases, nucleosides, nucleotides, and various endogenous molecules, dyes and drugs) were used as an initial data source for development of the in silico CP ANN prediction model. The model was built to estimate the transport activity of the transmembrane protein bilitranslocase for a large set of diverse endogenous compounds and xenobiotics (Table 1) [25].

The biological activity values were known for 81 compounds. For the majority of them (a total of 75 compounds) the inhibition constants (K_I [mmol L^{-1}], expressed in logarithmic units as pK_I) were experimentally determined [25], while for the rest 6 compounds the inhibition constants were obtained from the literature [28–30]. For the remaining 7 compounds (which are part of the total 88 compounds) the experimental assays are not yet finished.

The effect of 75 compounds on bilitranslocase transport activity was evaluated spectrophotometrically in rat hepatical plasma membrane vesicles. The initial rate of the bromosulphophtalein

Table 1
List of all 88 compounds (*Dataset 1*) separated as training, external test, and external validation set, together with their activity classification code (ACC), the experimental/predicted inhibition constants ($pK_{I\text{-}exp}$, $pK_{I\text{-}pred}$), as well as their calculated Euclidean distances (ED) [25].

ID	Compound	ACC	$K_{I\text{-}exp}$ [mmol L^{-1}][a]	$pK_{I\text{-}exp}$	$pK_{I\text{-}pred}$	ED
Training set						
1	Adenine[b]	I	–	−2.000	−1.1137	0.0863
2	Adenosine[b]	I	–	−2.000	−1.5260	0.0211
3	Adenosine 3′-monophoshate[b]	C	0.95	0.022	−0.5844	0.0226
5	Adenosine 3′,5′-cyclic monophosphate[b]	I	–	−2.000	−1.9741	0.0419
6	Adenosine 5′-diphosphate[c]	C	1.42	−0.152	−0.1373	0.0494
7	Adenosine 5′-triphosphate[b]	C	0.385	0.415	−0.0958	0.0482
8	Adenosine-5′-diphosphoglucose[c]	I	–	−2.000	−1.2062	0.1581
9	Adenosine 5′-(α.β-methylene) diphosphate[b]	C	1.31	−0.117	−0.1373	0.0095
10	Adenine 9-β-D-arabinofuranoside[b]	NC	3.76	−0.575	−1.5260	0.0158
11	Adenosine 3′-phosphate 5′-phosphosulfate[b]	C	0.148	0.830	−0.0958	0.0728
12	S-(5′-Adenosyl)-L-homocysteine[c]	C	0.44	0.357	0.3340	0.0658
13	S-(5′-Adenosyl)-L-methionine chloride[b]	C	0.408	0.389	0.3340	0.0127
15	Guanosine[b]	I	–	−2.000	−1.5260	0.0337
16	Guanosine 5′-monophosphate[b]	C	13.92	−1.144	−0.5844	0.0198
17	Guanosine 3′,5′-cyclic monophosphate[b]	I	–	−2.000	−1.9741	0.0396
18	Guanosine 5′-diphosphate[b]	C	4.55	−0.658	−0.0958	0.0459
19	Guanosine 5′-triphosphate[b]	NC	7.66	−0.884	−0.0958	0.0299
20	Uracil[b]	I	–	−2.000	−1.9811	0.0436
21	Uridine[b]	C	2.58	−0.412	−1.4886	0.0288
22	Uridine 5′-monophosphate[b]	C	4.13	−0.616	−0.6220	0.0155
23	Uridine 5′-diphosphate[c]	C	3.10	−0.491	−0.4331	0.0412
24	Uridine 5′-triphosphate[b]	C	1.425	−0.154	−0.1645	0.0501
25	Uridine 5′-diphosphoglucose[b]	I	–	−2.000	−1.2062	0.0125
26	Uridine 5′-diphosphogalactose[b]	C	2.47	−0.393	−1.2062	0.0196
28	Thymine[c]	I	–	−2.000	−1.9811	0.1127
29	Thymidine[b]	I	–	−2.000	−1.4886	0.0543
30	Thymidine 5′-monophosphate[c]	C	3.71	−0.570	−0.6220	0.0879
31	Thymidine 5′-diphosphate[b]	C	2.23	−0.348	−0.4331	0.0062
32	Thymidine 5′-triphosphate[b]	C	1.45	−0.161	−0.1645	0.0473
33	Cytosine[b]	I	–	−2.000	−1.9811	0.0453
34	Cytidine[b]	I	–	−2.000	−1.4886	0.0445
35	Cytidine 2′-monophosphate[c]	I	–	−2.000	−1.8244	0.0398
37	Cytidine 5′-monophosphate[b]	I	–	−2.000	−1.8244	0.0146
42	Uric acid[b]	C	1.50	−0.176	−1.1137	0.0842
43	Ouabain[b]	I	–	−2.000	−1.9872	0.0872
44	Aucubin[b]	I	–	−2.000	−1.9531	0.0460
45	Loganin[c]	I	–	−2.000	−1.9531	0.0718
47	Isovitexin[b]	I	–	−2.000	−1.7429	0.0175
48	Vitexin-2′-O-rhamnoside[b]	I	–	−2.000	−1.9529	0.0467
49	Cibacron Blue F3G-A[b]	C	0.00347	2.460	2.4751	0.0497
50	Digoxin[c]	I	–	−2.000	−1.9872	0.3497
51	Taurocholate[b]	I	–	−2.000	−1.9872	0.0868
52	Sulfobromophtalein[b]	C	0.00532	2.274	1.9787	0.0397
54	Bilirubin[b]	C	0.00111	2.955	2.9023	0.0452
55	Biliverdin[c]	C	0.00111	2.955	2.9023	0.0683
External test set						
4	Adenosine 5′-monophoshate	C	2.63	−0.420	−0.5844	0.0560
14	Guanine	Insoluble	/	/	−1.1137	0.0660
27	Uridine 5′-diphosphoglucuronic acid	I	–	−2.000	−1.2062	0.0686
36	Cytidine 2′:3′-cyclic monophosphate	/	/	/	−1.9741	0.1021
38	Cytidine 5′-diphosphate	/	/	/	−0.4285	0.0742
39	Cytidine 5′-triphosphate	/	/	/	−0.1645	0.0788
40	Cytosine β-D-arabinofuranoside	/	/	/	−1.4886	0.0448
41	Cytosine β-D-arabinofuranoside 5′-monophosphate	/	/	/	−1.8244	0.0336
46	Verbenalin	I	–	−2.000	−1.9531	0.0947
53	Thymol Blue	C	/	/	1.9787	0.3090
External validation set						
56	Nicotinic Acid	I	–	−2.000	−1.9811	0.2600
57	[D-Ala^2]-Deltorphin II	I	–	−2.000	2.5195	0.4062
58	[D-Pen^{2,5}]-Enkephtalin (DPDPE)	I	–	−2.000	−1.9529	0.2716
59	17β-Estradiolo 3-glucuronide	I	–	−2.000	−1.9872	0.1440
60	5-Fluorouracil	I	–	−2.000	−1.9811	0.1051
61	Acetazolamide	I	–	−2.000	−1.1137	0.1988
62	Acycloguanosine (Acyclovir)	I	–	−2.000	−1.5260	0.1231
63	Etacrynic Acid	I	–	−2.000	1.9787	0.3078
64	Mefenamic Acid	I	–	−2.000	1.9787	0.4074
65	Dehydroisoandrosterone 3-sulfate	I	–	−2.000	−1.9531	0.3027
66	Diclofenac	I	–	−2.000	1.9787	0.3533
67	Dichlorofluorescein (DCFH)	I	–	−2.000	1.9787	0.2660
68	Enalapril maleate	I	–	−2.000	−1.9531	0.2721
69	Estrone 3-sulfate	I	–	−2.000	−0.6692	0.3385
70	Phenacetin	I	–	−2.000	−1.4796	0.4971
71	Furosemide	I	–	−2.000	−1.3612	0.2661

Table 1 (*Continued*)

ID	Compound	ACC	$K_{I\text{-}exp}$ [mmol L^{-1}][a]	$pK_{I\text{-}exp}$	$pK_{I\text{-}pred}$	ED
72	Glycocholate	I	–	−2.000	−1.9872	0.0919
73	Ibuprofen	I	–	−2.000	1.9787	0.4966
74	Hydrochlorothiazide	C	0.056	1.251	−1.4796	0.3436
75	Hydrocortisone	I	–	−2.000	−1.9531	0.1745
76	Indomethacin	I	–	−2.000	1.9787	0.2714
77	Ketoprofen	I	–	−2.000	1.9787	0.3769
78	L-Thyroxine (T4)	I	–	−2.000	−1.4886	0.2759
79	Methotrexate	I	–	−2.000	0.8018	0.1852
80	Naproxene	I	–	−2.000	1.9787	0.4501
81	Piroxicam	I	–	−2.000	−1.1137	0.2969
82	Pravastatin	I	–	−2.000	−1.9872	0.1840
83	Probenecid	I	–	−2.000	1.9787	0.3771
84	Progesterone	I	–	−2.000	−1.9531	0.3072
85	Prostaglandin E_2	I	–	−2.000	−1.9531	0.2164
86	Sulindac	C	0.062	1.209	1.9787	0.3018
87	Triiodo-L-thyronine (T3)	NC	0.174	0.759	1.9787	0.2880
88	β-Estradiol	I	–	−2.000	−1.9531	0.4088

[a] Inactive compounds were considered for modeling purposes at hypothetical concentration 100 mmol L^{-1}, with $pK_{I\text{-}exp} = -2.000$.
[b] Internal training set compounds (a total of 35 molecules).
[c] Internal test set compounds (a total of 10 molecules).

(BSP) passage into vesicles (called electrogenic BSP uptake) was used as the measure of bilitranslocase transport activity [28,31–34]. The experimental assays performed, successfully identified 28 compounds as active (of which 25 compounds as competitive ("**C**") and 3 compounds as non-competitive inhibitors ("**NC**")), and 53 compounds as inactive ("**I**"). For all compounds determined as inactive ("**I**"), a $pK_{I\text{-}exp}$ value of ($pK_{I\text{-}exp} = -2.000$) that corresponds to a hypothetical concentration of 100 mmol L^{-1} was assigned (Table 1) [25].

The geometry of all 88 chemical structures was initially optimized by using AM1 semi-empirical procedure [35], whereas Kohonen artificial neural network (KANN) [36,37] was employed for dataset division (Table 1) into training set (45 compounds), external test set (10 compounds), and external validation set (33 compounds). The training set was additionally split into an internal training set (Table 1; 35 compounds indexed by b) and internal test set (Table 1; 10 compounds indexed by c) for the optimization of the ANN parameters [25]. Counter-propagation artificial neural network (CP ANN) [38,39] which can be considered as an extension of KANN, was exploited for development of the 14-parameters CP ANN predictive model ($R^2 = 0.89$). The model was built using the internal training set, optimized with the internal test set ($Q^2_{te/F3} = 0.89$), as well as externally validated using the external validation set ($Q^2_{ext/F3} = -1.83$) which was not used during the model development [25]. The model's predictive ability quantificators, were calculated according to the following equations (Eqs. (1) and (2)) [40,41]:

$$R^2 = 1 - \frac{\left[\sum_{i=1}^{n_{tr}} (\hat{y}_i - y_i)^2\right] / n_{tr}}{\left[\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2\right] / n_{tr}} \quad (1)$$

$$Q^2_{(te/ext)/F3} = 1 - \frac{\left[\sum_{i=1}^{n_{(te/ext)}} (\hat{y}_i - y_i)^2\right] / n_{(te/ext)}}{\left[\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2\right] / n_{tr}} \quad (2)$$

where n_{tr} designates the number of training set objects, whereas $n_{(te/ext)}$ is the number of test set, i.e., number of external validation set objects, respectively. Here we would like to stress that the Q^2 parameter as defined by (Eq. (2)) might be negative in case of large error in prediction of the external validation set compared to the range of the target values in the training set [40]. For the purpose of this study, an additional PLS model ($R^2 = 0.78$, $Q^2_{te/F3} = 0.47$, $Q^2_{ext/F3} = 0.15$) was developed by using the same dataset division as well as the number of molecular descriptors, as described previously.

2.1.2. Dataset 2

The second dataset consists of 59 trypsin inhibitors, initially selected from the Brookhaven PDB database, that were experimentally tested for their inhibitory potency (pK_I) against the trypsin

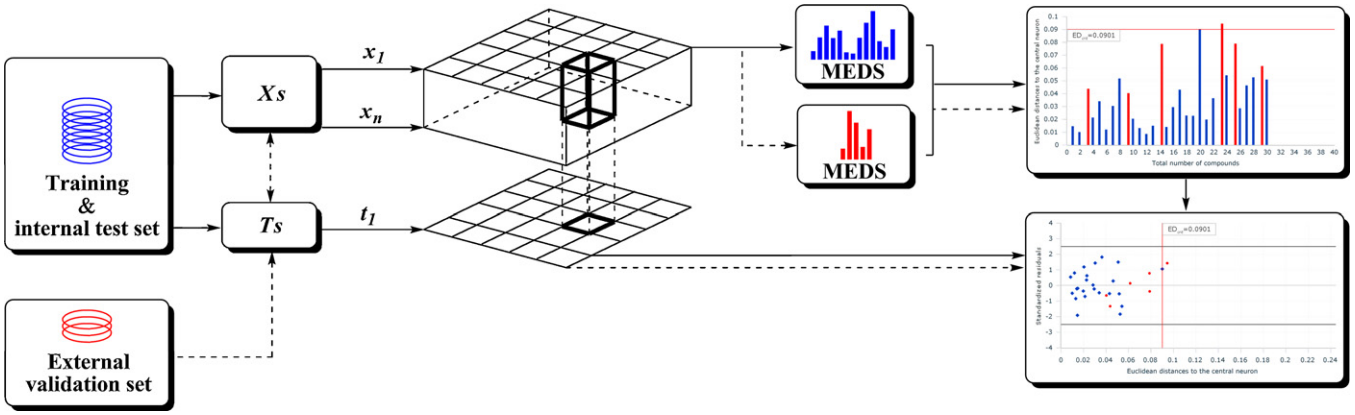


Fig. 1. Schematic representation of the overall methodology workflow, depicting different stages, where: X_s, is multidimensional input vector (e.g., calculated molecular descriptors); T_s, is multidimensional target vector (e.g., experimentally-determined endpoint values); MEDS for the training & internal test set (model) is represented in blue, while MEDS for the external validation set is represented in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

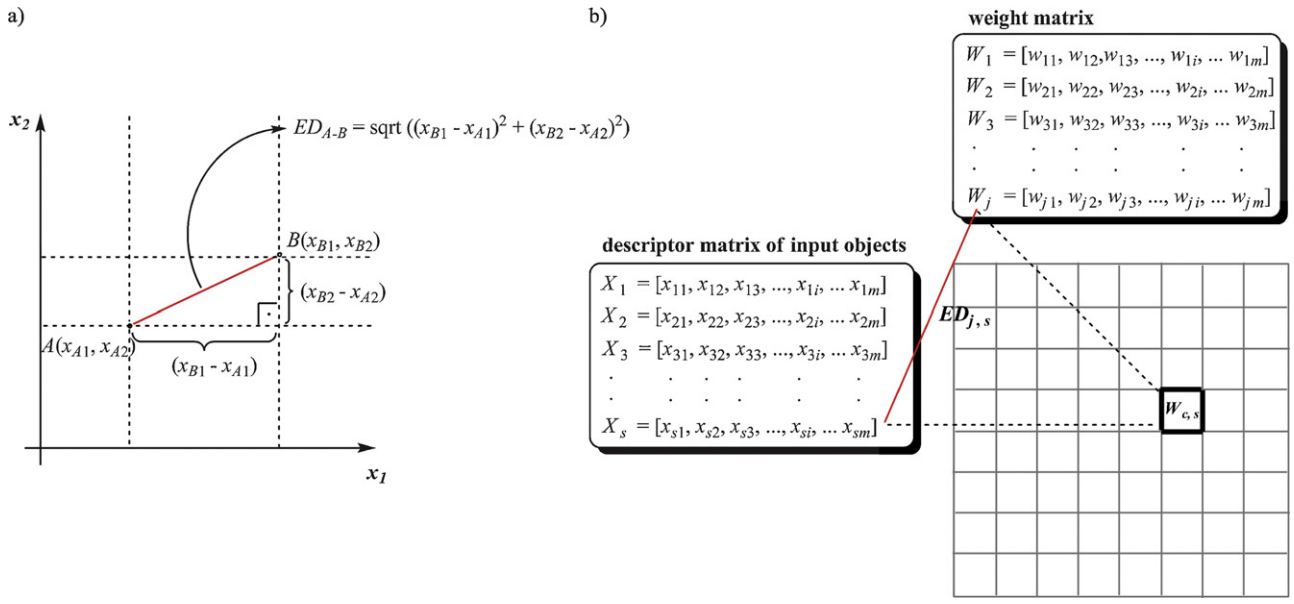


Fig. 2. Euclidean geometry and calculation of EDs between objects [47]. (a) Determination of ED between two points *A* and *B* in a simple two-dimensional Cartesian (Euclidean) space, defined by their two-dimensional Cartesian coordinates. (b) Mapping of multidimensional input data (e.g., compounds) encoded as multidimensional descriptor vectors onto two-dimensional grid of neurons (Kohonen map) and determination of the minimal ED (the red line, $\min(W_j - X_s) \rightarrow ED_{j,s}$) for each input object to the "*central*" neuron ($W_{c,s}$) [38]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

enzyme. All inhibitors are heterogeneous, non-covalently bound to the same site on the enzyme surface. The dataset, modeling strategy applied, as well as the CP ANN prediction model are described in details in our previous work [26]. The best model which was selected ($R^2_{[CP\,ANN-97-p]} = 0.89$, $Q^2_{ext/F3[CP\,ANN-97-p]} = 0.85$) was developed by using 97 molecular descriptors, while 26, 15 and 18 compounds in the training, internal test, and external validation set, respectively, were used for its modeling and validation. Furthermore, the same dataset division was used for building two additional QSAR models taking into account a reduced number of molecular descriptors (7 out of total pool of 97 molecular descriptors were selected) for the purpose of this study; CP ANN ($R^2_{[CP\,ANN-7-p]} = 0.82$, $Q^2_{ext/F3[CP\,ANN-7-p]} = 0.45$) and PLS ($R^2_{[PLS-7-p]} = 0.42$, $Q^2_{ext/F3[PLS-7-p]} = 0.33$). The descriptors reduction here was performed explicitly to avoid obtaining an ill-conditioned diagonalized matrix [42] constructed of extremely large diagonal elements (such as the leverage values, h_{ii}) which are usually impractical for work and further interpretation (e.g., identification of structurally-influential outliers).

2.1.3. Dataset 3

The third dataset includes the retention factors (k) of ionized inorganic and organic acids measured on ion-exchange analytical column (IonPac AS18, Dionex) at 25 mM eluent (KOH) [27]. CP ANN and PLS models were built to predict retention factor of ions from molecular structures of their corresponding acids. Starting with a pre-selected pool of total 64 molecular descriptors, the best CP ANN model was build ($R^2_{[CP\,ANN-64-p]} = 0.99$, $Q^2_{ext/F3[CP\,ANN-64-p]} = 0.96$) with 15 training and 5 internal test compounds, while 5 compounds were used as the external validation set [27]. Similarly to *Dataset 2*, two more QSAR models (CP ANN and PLS) were additionally developed by using a subset of 13 descriptors from the starting pool of total 64 descriptors; CP ANN ($R^2_{[CP\,ANN-13-p]} = 0.99$, $Q^2_{ext/F3[CP\,ANN-13-p]} = 0.86$) and PLS ($R^2_{[PLS-13-p]} = 0.71$, $Q^2_{ext/F3[PLS-13-p]} = 0.91$) which were used to assess the ADs in a comparative manner.

2.2. From trained network to definition of the "minimum Euclidean distance space"

No matter which modeling strategy (linear or non-linear) will be utilized for developing of a QSAR model for chemical management purposes, two substantial things must be taken into account: (1) chemical structures must be represented numerically in a form of molecular descriptors, and (2) a defined experimentally-determined endpoint (e.g., pK_I for *Dataset 1 and 2*, or retention factor k for *Dataset 3*) must be available for each compound. Since the molecular descriptors are unique numerical values obtained by quantification of various structural and/or physico-chemical properties of the molecule, they could be successfully used not only for prediction of the endpoint values of unknown compounds, but also for performing similarity measurements between two or more molecular structures. For those purposes, different distance (similarity) metrics are available (e.g., Euclidean distance (ED), Mahalanobis distance, Manhattan (City-block) distance, etc.) [13,14,43], which can be exploited in several different ways [44–47]. Among them, the Euclidean distance (ED) metric is one of the most widely used metrics for the determination of similarity. The most simple form of ED could be defined as a distance between two points in a two-dimensional Cartesian space (known as Cartesian or Euclidean plane) [15]. Therefore, if *A* and *B* are two points defined by their two-dimensional Cartesian coordinates $A(x_{A1}, x_{A2})$ and $B(x_{B1}, x_{B2})$, respectively (Fig. 2a), then the ED between them can be simply calculated using the following equation (Eq. (3)) [48]:

$$ED_{A-B} = \sqrt{(x_{B1} - x_{A1})^2 + (x_{B2} - x_{A2})^2} \quad (3)$$

The smaller the ED between the points *A* and *B* is, the higher the similarity between them is. Although, this similarity-measurement concept is very simple, it can be effectively used for solving more complex problems. The QSAR modeling methods are constructed for handling large amount of data describing a so-called multidimensional data space, where each chemical structure is usually encoded in a form of a multidimensional descriptor vector. Consequently, the similarity measures between the objects must be

calculated in such complex space according to the dimensionality of the descriptor vector.

In order to simplify the problem of complexity, a data compression (mapping) from the multidimensional space to a space of lower complexity is required, which is the basis of the Kohonen artificial neural networks (KANN) [36,49,50]. KANN provides a so-called self-organizing maps (SOMs) which are not comparable to the Cartesian space mentioned previously, since the metrics are not conserved. It is conserving the topology of the multidimensional input data (usually multidimensional descriptor vectors) within a two-dimensional network of neurons (as a result of unsupervised competitive learning), which are captivated with objects in a way where similar objects are situated on same or close neighbor neurons [38].

The mapping of the objects progresses iteratively over a non-linear algorithm (known as training of the network), which is mainly based on object-similarity determination through utilization of ED metrics. It can be literally described as "*winner takes all strategy*", which means that for each multidimensional input vector, only the most excited neuron (or the so-called "*winning*" or "*central*" neuron) is selected and all the corrections are performed around it. The learning algorithm selects the "*central*" neuron according to the minimal ED calculated between the multidimensional input vector X_s ($x_{s1}, x_{s2}, x_{s3}, \ldots, x_{si}, \ldots, x_{sm}$) for each input object (e.g., a compound from the training set) and all the weight vectors W_j ($w_{j1}, w_{j2}, w_{j3}, \ldots, w_{ji}, \ldots, w_{jm}$), using the following equation (Eq. (4)) [38]:

$$ED_{j,s} = d_{j,s}^{Eucl} = \sqrt{\sum_{i=1}^{m}(W_{j,i} - X_{s,i})^2},$$

$$\min\{d_{j,s}^{Eucl}, j = 1, \ldots, N_{net}\} \rightarrow W_{c,s} \quad (4)$$

where N_{net} describes the total number of neurons constructing the network, while $W_{c,s}$ designates the selected "*central*" neuron to which the object s is the most similar (Fig. 2b). At the end of the learning process (developing of the model), every object is assigned to its "*central*" or closest neuron; there are as much minimum EDs calculated, as is the total number of input (training) objects, which together define a so-called "*minimum ED space*" (MEDS) [20]. Since, each training compound used for building of the ANN model is characterized with its unique ED, the MEDS concept could be effectively used for assessing of the model's applicability domain (AD) which boundaries can be simply defined by selection of the maximal ED (hereafter named as ED_{crit}) within the MEDS (Fig. 1). For the purpose of assessing the AD, the MEDS boundaries should be extended to the internal test compounds involved in the model parameters optimization (internal test set distances). Consequently, each object (e.g., a compound from the external test/external validation set) entering the trained network (model) with calculated ED greater than ED_{crit}, can be considered as it is outside of the model's AD, and vice versa. The distances in MEDS depend on the dimension of the descriptor vectors entering the network. In order to unify their scale, they should be normalized by division with the total number of molecular descriptors used for modeling. Hereafter, normalized values are considered for ED and ED_{crit}.

2.3. Descriptor space versus model response space

While the MEDS concept is carrying only the structural information (encoded in a form of molecular descriptors) extracted from the trained Kohonen map for the compounds used in the QSAR study, it can be applied only for assessment of the possible outliers within the compound's descriptors space. Although, these data could be extremely helpful for detection of the structurally-influential outliers that could affect the quality of the QSAR model (e.g., a two dimensional plot of the EDs vs. total number of compounds), they do not tell anything about the accuracy of predictions, and therefore the AD of the model would be only partially defined. To alleviate this problem as well as to define as much as informative domain of model's applicability, the model's predictability information (model response space) must be additionally incorporated into the AD definition.

Contrary to KANNs which are constructed of only one layer of neurons, the counter-propagation artificial neural networks (CP ANNs) contain an additional layer located exactly below the Kohonen layer, and therefore it can be regarded as an extension of KANN [37,38,50]. This additional layer is trained in exactly the same manner as the Kohonen layer, i.e., it receives the vector of target (dependent) variables T_s (e.g., experimentally-determined values of K_I), and as a result of the learning procedure controlled by the Kohonen network above, outputs their predicted (response) values which reflect the model's response space (Fig. 1) [51]. Taking these additional information into account, the AD of the CP ANN predictive model could be defined in a more informative way through expression of the MEDS of the model as a function of the standardized residuals.

2.4. Partial least squares model development

In order to assess the performances of the MEDS-based AD method developed for non-linear CP ANN models with a well known leverage approach [3] used in multiple linear regression or partial least squares (PLS) regression modeling, the linear models were built with the same data as non-linear ones. PLS models were developed taking into account the same settings (number and type of molecular descriptors as well as the same training/internal test/external validation set division) as described previously for the neural network models (see Sections 2.1.1–2.1.3).

2.5. The hat matrix and influential observations – leverage approach

Analogously to the MEDS-based AD approach defined previously, the leverage approach which is usually applied for regression diagnostics (outliers identification) in case of linear models (e.g., MLR, PLS) [52,53], was additionally included in this study. Although, both methods are substantially diverse with regards to the difference between the distance metrics on which are based, they were jointly utilized just for comparative purposes.

In the linear regression it is often very useful to determine the influence of a given y_i value (e.g., experimentally-determined endpoint values such as K_I) over each predicted $\hat{y}_i$ value. Contrary to the relationship between y_i and $\hat{y}_i$, which interpretation can be easily performed through implementation of a simple residual analysis ($y_i - \hat{y}_i$), the influence of the independent variables x_i (e.g., calculated molecular descriptors) on the model might be difficult to determine. The solution to this problem lies into the so-called hat matrix H [54,55], which maps the vector of observed values to the vector of fitted values. It is an $n \times n$ symmetric matrix which diagonal elements h_{ii} (known as leverage values) directly reflect the structural influence of a compound to the values predicted by the model (i.e., a distance metric which shows how far a compound is from the model experimental space) [53].

Therefore, for a given set of compounds encoded in a form of calculated molecular descriptors, one can solve the hat matrix H through an implementation of a simple set of matrix algebra rules [54,55]. Let us X_{tr} be a multidimensional $n \times m$ matrix carrying the structural information m_j (calculated molecular descriptors) for each training set compound n_i separately. The calculation of the

hat matrix H of the original X_{tr} matrix requires several consecutive steps:

(1) generation of a transposed version ($X_{tr}^T = m \times n$) of the original descriptor matrix X_{tr}.
(2) multiplication of the transposed matrix X_{tr}^T and the original matrix X_{tr}: ($X_{tr}^T X_{tr}$).
(3) inversion of the product matrix ($X_{tr}^T X_{tr}$) obtained by multiplication in (2): $(X_{tr}^T X_{tr})^{-1}$.
(4) multiplication of the original X_{tr}, the inverse $(X_{tr}^T X_{tr})^{-1}$, and the transposed matrix X_{tr}^T.

The obtained result is an $n \times n$ symmetric matrix (H, hat matrix) [54] which maps the $y_{i,tr}$ values into $\hat{y}_{i,tr}$ (Eq. (5)):

$$H_{tr} = X_{tr}(X_{tr}^T X_{tr})^{-1} X_{tr}^T \rightarrow \hat{y}_{tr} = H_{tr} y_{tr} \leftrightarrow \hat{y}_{tr} = X_{tr}(X_{tr}^T X_{tr})^{-1} X_{tr}^T y_{tr} \quad (5)$$

where $\hat{y}_{tr}$ and y_{tr} are the predicted and experimentally-determined endpoint values for the training set compounds, respectively, whereas H_{tr} designates the hat matrix for the training set which diagonal elements ($h_{ii,tr}$) describe the distance of each training set compound to the structural centroid of the model [53].

The calculation of the hat matrix as well as the extraction of the diagonal elements (leverages) for any test/external validation set compound entering the model, can be performed as for the training set described above, but slightly modified (Eq. (6)):

(1) generation of a transposed version ($X_{te/ext}^T = m \times n$) of the original descriptor matrix of test data $X_{te/ext}$.
(2) generation of a clone version of the inverse matrix solved above for the training data: $(X_{tr}^T X_{tr})^{-1}$.
(3) multiplication of the original $X_{te/ext}$, the inverse $(X_{tr}^T X_{tr})^{-1}$, and the transposed matrix $X_{te/ext}^T$.

$$H_{te/ext} = X_{te/ext}(X_{tr}^T X_{tr})^{-1} X_{te/ext}^T \rightarrow \hat{y}_{te/ext} = H_{te/ext} y_{te/ext} \leftrightarrow \hat{y}_{te/ext} = X_{te/ext}(X_{tr}^T X_{tr})^{-1} X_{te/ext}^T y_{te/ext} \quad (6)$$

where $\hat{y}_{te/ext}$ and $y_{te/ext}$ are the predicted and experimentally-determined endpoint values for the test/external validation set compounds, respectively, whereas $H_{te/ext}$ designates the hat matrix for the test/external validation set which diagonal elements ($h_{ii,te/ext}$) describe the distance of each test/external validation set compound to the structural centroid of the training set (model) [53].

One of the recommended hat-based methods for AD investigation in case of linear QSAR models is the widely known leverage approach [3,53,56]. The method offers a graphical assessment of the leverage values (h_{ii}), as a function of the standardized cross-validated residuals (Williams plot) [53,57] and it is suitable not only for detection of the structurally-influential outliers, but also for determination of the response outliers. The model predictions should be referred as unreliable for those compounds for which h_{ii} diagonal elements are greater than the cut-off leverage value (h^*). These compounds are located far from the structural centroid of the model, and therefore could be referred as structurally-influential outliers, too. The cut-off leverage value (h^*) is usually defined as (Eq. (7)):

$$h^* = \frac{3(p+1)}{n} \quad (7)$$

where p is the total number of descriptors used for developing of the QSAR model, while n is the total number of the training set compounds. Moreover, the compounds for which the calculated standardized residual values are greater than three standard deviation units ($>\pm 3\sigma$), could be considered as response outliers.

3. Results and discussion

No matter how good (robust, meaningful, and validated) a developed QSAR model could be, its inherent limitation is data-driven correlation pursuit: it cannot be expected to give reliable predictions for the property under investigation for the compounds out of the domain of chemical space defined by the training data. In order to be used for chemical management (screening) purposes, the QSAR model should have a clearly defined domain of applicability, and therefore the property estimations for only those compounds situated within the model's AD boundaries can be considered reliable [3–5]. Among the various methods available today for AD estimation of established QSAR models, the distance-based approaches (e.g., ED-based or leverage-based) proved to be particularly useful not only in the case of linear models [53], but also for non-linear models such as artificial neural networks [24]. The following study involves a simple and effective approach for AD estimation of a non-linear QSAR models [25–27] developed using CP ANN modeling method through the utilization of the MEDS concept elaborated in Section 2.2.

3.1. MEDS-based applicability domain assessment for CP ANN predictive models

3.1.1. Dataset 1

In order to demonstrate the practical significance of the MEDS concept for graphical assessment of the AD (Fig. 1), a developed and validated predictive CP ANN model was used. As described previously, the MEDS concept for AD assessment is simply a distance-based method (based on ED metrics), and therefore the availability of calculated ED data is crucial. For these purposes, as a result of the network training which process runs iteratively (i.e., learning of the network), the minimum EDs (Table 1) between each input object (e.g., a compound expressed as a multidimensional descriptor vector) entering the network and the so-called "*central*" neuron are calculated (Fig. 2b) according to the (Eq. (4)) [38]. Once calculated, the AD of the CP ANN model could be easily defined as a two-dimensional column plot (Fig. 3) which expresses the calculated minimum EDs (MEDS) for each compound (e.g., training/external test/external validation set) as a function of the total number of objects (88 compounds).

As shown in Fig. 3, the AD coverage of our CP ANN prediction model is simply defined by the threshold value ($ED_{crit} = 0.350$), which corresponds to the training set object with maximal value for ED (Table 1; ID = 50, Digoxin). Therefore, all other objects (e.g., external test or external validation set compounds) for which the calculated ED is greater than ED_{crit}, could be considered as being outside of the model AD. According to Fig. 3, a total of 9 compounds (ID = 57, 64, 66, 70, 73, 77, 80, 83, and 88) belonging to the external validation set (33 compounds), could be distinguished as potential outliers (Table 1) with ($ED > ED_{crit} = 0.350$). Except two compounds (ID = 57 (Deltorphin; a heptapeptide of exogenic origin) and 88 (β-Estradiol; steroidal hormone)) which belong to different therapeutic classes, i.e., selective δ-opioid agonist (analgesic) and selective estrogen receptor modulator (contraceptive), respectively, the rest of the identified outliers are mainly non-steroidal antiinflammatory agents. These compounds are structurally very different comparing to the compounds used for building of the CP ANN model which are mainly purine and pyrimidine derivatives, and therefore their calculated EDs are much larger than the selected ED threshold ($ED_{crit} = 0.350$). Consequently, the identified outlying

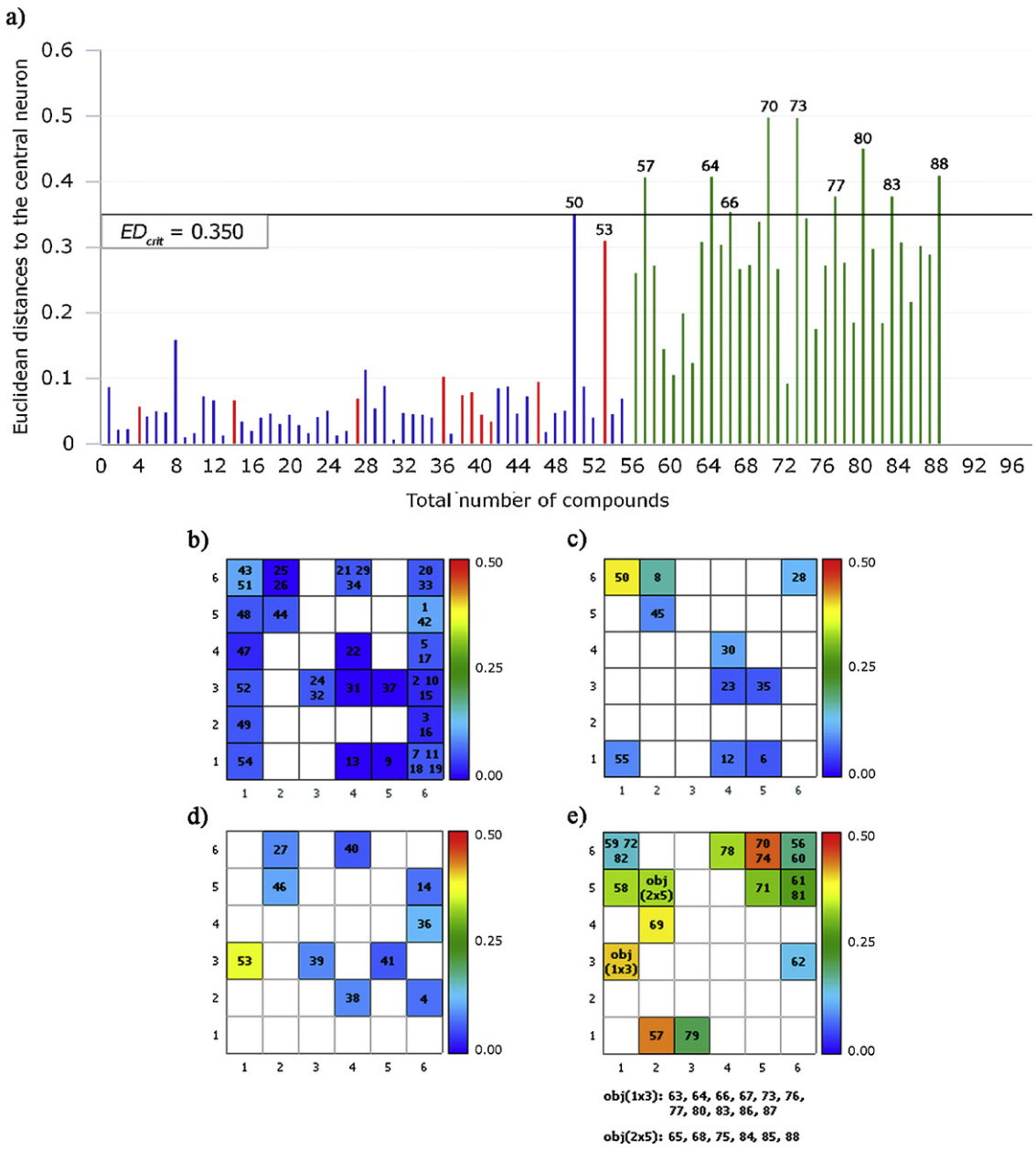


Fig. 3. Two-dimensional column plot of the AD for the CP ANN prediction model for bilitranslocase transport activity together with the corresponding CP ANN's obtained output layers (*Dataset 1*) [25]: (a) the training & internal test set objects (45 compounds) are represented as solid blue columns, the external test set objects (10 compounds) as solid red columns, while the external validation set objects (33 compounds) as solid green columns. The domain coverage (AD threshold, ED_{crit} = 0.350) is depicted as a horizontal black line. The identified outlying objects (green columns with ED > ED_{crit}) are signed by their ID numbers (Table 1); below the plot (a), EDs-based output layers (MEDS) for each set separately are given: (b) internal training set, (c) internal test set, (d) external test set, and (e) external validation set; all 88 compounds are mapped by their corresponding ID number (Table 1); the central neurons are colored according to the calculated ED values in the range between dark blue (minimal ED value) and strong red (maximal ED value), while the intact (empty) neurons were colored in white.

objects could be considered as pure structurally-influential outliers, i.e., outside of the CP ANN model's AD. All the objects identified as outliers (Fig. 3), were experimentally-determined as inactive ("**I**") compounds (Table 1).

The MEDS analysis of the CP ANN model (Fig. 3) makes possible to assess the AD regarding the compound's structural space. Thus it is capable to detect only those objects (compounds) which are structurally different comparing to the compounds used for developing of the model, taking into account only the structural information (i.e., compounds descriptor space) stored in the framework of the trained Kohonen map of the CP ANN architecture (Fig. 1). In order to define the domain of applicability regarding the response data (the molecular property or activity values), the information about the predictivity of the CP ANN model (response space) must be additionally included into the AD definition. As a result, the AD of our CP ANN predictive model could now be defined as a two-dimensional dot plot representing the model response space (depicted in a form of calculated standardized residuals) as a function of the previously defined MEDS (Fig. 4a). In this way, one can identify those compounds for which the predicted activity values are questionable [53,56].

Similarly to the column plot represented in Fig. 3, the structurally-influential outliers (signed by the corresponding ID numbers in Fig. 4a) can be easily (visually) determined

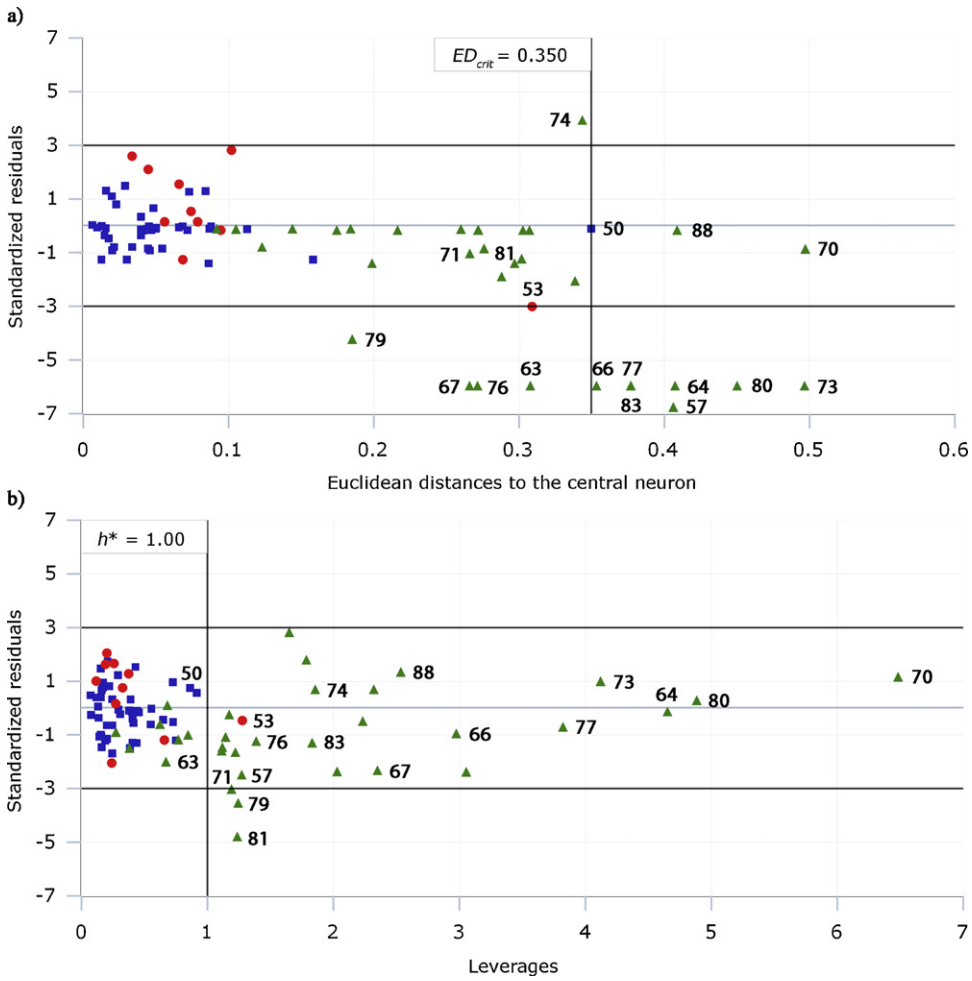


Fig. 4. Graphical representation (dot plots) of the (a) MEDS-based versus (b) leverage-based method for AD estimation for our CP ANN prediction model and its linear counterpart (PLS model), respectively, for prediction of the bilitranslocase transmembrane protein transport activity (*Dataset 1*) [25]. Training set objects (45 compounds) are depicted as solid blue rectangles, external test set object (10 compounds) as solid red circles, while the external validation set objects (33 compounds) as solid green triangles. The critical hat value ((b) $h^* = 1.00$) is calculated according to (Eq. (7)). The compounds identified as outliers (outside of the model's AD) by both methods are signed by their corresponding ID numbers ((a) Table 1; (b) Supporting information in Table S1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(ED > ED_{crit} = 0.350), however only two of them (ID = 70 and 88) can be recognized as correctly predicted by the model, with calculated standardized residual values in the boundaries between $\pm 3\sigma$ units [53,56]. The rest 7 outlying compounds regarding structural domain (ID = 57, 64, 66, 73, 77, 80 and 83) are also out of the 3σ boundaries (with large prediction errors, see Table 1 and Ref. [25]) and therefore they could be recognized as pure response outliers, as well. A closer look at Fig. 4a indicates five additional response outliers from the external validation set (ID = 63, 67, 74, 76, and 79). According to the calculated ED values, these external validation set compounds are located within the domain boundaries as defined by ED_{crit} = 0.350, however their calculated standardized residual values greater than $\pm 3\sigma$ units pinpoint to objects wrongly predicted by the model (Table 1). In summary, 9 compounds from the external validation set were found out of structural AD and thus the predictions are considered unreliable, and 7 of them were really poorly predicted. On the other hand, 23 compounds were found within the structural AD and thus expected as reliably predicted by the model, and 18 of them were predicted correctly, only 5 of them were response outliers. Thus, by assessing the structural AD outliers only, we got rid of 7 out of 12 response outliers. This is important because the assessing of structural AD is affordable also in case of unknown experimental values of properties of new active/toxic compounds.

3.2. MEDS-based versus leverage-based applicability domain assessment

3.2.1. Dataset 1

As demonstrated in Figs. 3 and 4a, the MEDS-based AD assessment of our CP ANN prediction model [25] performed so far, showed reasonable and in some instance satisfactory results. Namely, all compounds which were identified as potential structurally-influential outliers (Table 1) are indeed structurally very different than the compounds used for developing of the model, and therefore they could be expected to be situated far from the structural centroid of the model, i.e., outside of the model's AD. Comparing to the well known leverage approach [3] and particularly its graphical representation for AD estimation (Williams plot; standardized residuals vs. leverages) [53], a certain degree of similarity between both methods is evident (distance-based methods) [8,9]. On the other hand, both methods are considerably different in regard to the distance metrics used (Euclidean distances vs. leverages) and consequently limited to the nature of the developed QSAR model (non-linear, i.e., linear model, respectively) to which can be applied, however, for comparison purposes we decided to evaluate and demonstrate both of them.

Analogously to the MEDS-based AD approach (Fig. 4a) for which definition the calculated EDs for each investigated object are

required, the leverage-based method for AD estimation (Williams plot) [53] requires the leverage values (h_{ii}), which are the diagonal elements of the hat matrix (H; see Eqs. (5) and (6)) [52,53]. In order to compare both methods as well as to assess the performances of the MEDS-based AD approach applied to our CP ANN prediction model (Fig. 4a) [25], a linear model (14-parameters PLS model) was additionally developed (see Section 2.1.1) using the same data (same number and type of molecular descriptors and same training/external test/external validation set division were exploited) to which the leverage-based method was applied (Fig. 4b). The experimental and predicted activity values ($pK_{I\text{-}exp}$, $pK_{I\text{-}pred}$) and calculated leverage values (h_{ii}) for each investigated compound are available as Supporting information in Table S1.

As shown in Fig. 4, both methods give comparable distribution of compounds in the AD plots, however, the vertical border (cut-off value) of the AD showing the limit of the model with regard to the structural space is much more restrictive for the PLS model (Fig. 4b). Consequently, the external validation set compounds (ID = 57, 64, 66, 70, 73, 77, 80, 83, and 88) which were identified as structurally-influential outliers by MEDS-based approach in the CP ANN model (Fig. 4a), were identified as extreme points ($h > h^* = 1.00$) in the Williams plot of the PLS model (Fig. 4b), too. While the majority of these compounds (more precisely the objects with ID = 57, 64, 66, 73, 77, 80 and 83) were determined as response outliers (points with calculated standardized residual values greater than $\pm 3\sigma$) on the MEDS plot (Fig. 4a), the same compounds could be identified as acceptably predicted by the PLS model (see Supporting information, Table S1; Fig. 4b). Comparing to the MEDS-based approach (Fig. 4a), the leverage-based approach (Fig. 4b) identified an external test set compound thymol blue (thymolsulphonephthalein, ID = 53) as a structurally-influential outlier ($h_{ID=53} = 1.277$). Indeed, this compound is structurally very different comparing to the training/internal test set compounds, and therefore it is located far from the structural centroid of the PLS model (Fig. 4b). It is also far from the training/internal test set ensembles ($ED_{ID=53} = 0.309$; Fig. 4a), nevertheless, it is situated within the CP ANN model AD boundaries since the cut-off value (ED_{crit}) is determined as 0.350 by the training compound Digoxin (ID = 50). Surprisingly, of total 33 compounds from the external validation set, the leverage-based approach identified only 7 compounds (ID = 62, 63, 72, 75, 78, 82, and 85) situated within the PLS model AD (Fig. 4b), while the majority of them are located out of the model's domain ($h > h^*$), although they are structurally-similar to the compounds used for developing of the model. On the other hand, these compounds were correctly identified by the MEDS-based approach as non-outlying objects (inside the model's applicability domain) as shown in Fig. 4a.

3.2.2. Dataset 2

The *Dataset 2* of total 59 trypsin inhibitors was used to build three models, as described in Section 2.1.2. The first one was our previously published CP ANN model constructed using 97 molecular descriptors ($R^2_{[CP\,ANN\text{-}Dataset\,2;\,97-p]} = 0.89$, $Q^2_{ext/F3[CP\,ANN\text{-}Dataset\,2;\,97-p]} = 0.85$) [26], while the other two models were built additionally utilizing a reduced number of molecular descriptors. In order to avoid the obtaining an ill-conditioned hat matrix [42] (constructed of extremely large elements) and consequently extremely large leverage values h_{ii} which are usually not very useful for construction of the Williams plot, we decided to reduce the number of molecular descriptors in a way where the number of molecular descriptors not exceed the total number of training set objects. Taking this limitation into account and for the purpose of this study, the additional CP ANN and PLS models were built by using only 7 out of 97 molecular descriptors through elimination of those descriptors for which the intercorrelation coefficients were more than or equal than 0.40 [58]. Although the performance of the obtained models with 7 descriptors was worse comparing to the original one (CP ANN model: $R^2_{[CP\,ANN-Dataset\,2;\,7-p]} = 0.82$, $Q^2_{ext/F3[CP\,ANN\text{-}Dataset\,2/7\text{-}p]} = 0.5$, PLS model: $R^2_{[PLS\text{-}Dataset\,2;\,7\text{-}p]} = 0.42$, $Q^2_{ext/F3[PLS\text{-}Dataset\,2;\,7\text{-}p]} = 0.33$), they are jointly elaborated here to compare the novel MEDS-based AD method with the leverage one (Fig. 5).

As demonstrated in Fig. 5a, the CP ANN model constructed by using 97 molecular descriptors covers very well the structural domain of the compounds from the external validation set – all the objects are situated within the MEDS boundary with calculated EDs significantly smaller than the critical ED value ($ED_{crit} = 0.153$) that corresponds to the training set object (ID = 17). However, two external validation set compounds (ID = 42 and 59) could be determined as pure response outliers (see Supporting information, Table S2). On the other hand, the CP ANN model constructed by using 7 molecular descriptors assigns three external validation set compounds (ID = 34, 42, and 56 with EDs > $ED_{crit} = 0.279$) to be situated out of the MEDS (Fig. 5b). They are also determined outside of the AD of the PLS model ($h > h^* = 0.59$) as represented on its Williams plot (Fig. 5c). Chemical structure of the compounds with ID = 34 and 56 do not contain SO_2 group as most of the compounds in the training set, while the outlier ID = 42 contains three fluor atoms, unlike the rest of the compounds in the *Dataset 2*. While no response outliers could be identified in the PLS model's AD depicted on the Williams plot (Fig. 5c), only one external validation set compound (ID = 42) is identified as a common response outlier on the MEDS plots (Fig. 5a and b).

3.2.3. Dataset 3

Similarly to the *Dataset 2*, the first model of the retention factor (k) was taken from previous work [27] – a CP ANN model built with 15 training and 5 internal test set compounds based on 64 molecular descriptors (see Section 2.1.3). The model was validated by using 5 external validation set compounds ($R^2_{[CP\,ANN-Dataset\,3;\,64-p]} = 0.99$, $Q^2_{ext/F3[CP\,ANN-Dataset\,3;\,64-p]} = 0.96$). Analogously to the *Dataset 2*, two more additional models were constructed by using a reduced number of molecular descriptors (13 out of 64 descriptors)–CP ANN ($R^2_{[CP\,ANN\text{-}Dataset\,3;\,13-p]} = 0.99$, $Q^2_{ext/F3[CP\,ANN\text{-}Dataset\,3;\,13-p]} = 0.86$) and PLS ($R^2_{[PLS\text{-}Dataset\,3;\,13\text{-}p]} = 0.71$, $Q^2_{ext/F3[PLS\text{-}Dataset\,3;\,13-p]} = 0.91$). The applicability domain assessment for all three models is shown in Fig. 6.

As demonstrated in Fig. 6, one can observe satisfactory performances in all three cases (almost all objects are situated within the models AD). In the AD plot for the CP ANN model constructed by using 64 molecular descriptors (Fig. 6a) only one object from the external validation set is evidently structurally-influential outlier, ID = 20; This is sulfurous acid ant its 64-dimensional descriptor vector shows significant difference from the most similar structure in the training set, which is sulfuric acid that has a different number of oxygen atoms bound to sulfur atom with a different oxidation state. No outlying objects could be identified in the 13-parameters CP ANN and PLS models AD as to be located outside of their structural centroids (Fig. 6b and c). On the other hand, regarding the models response space, the AD for the 13-parameters CP ANN model (Fig. 6b) identified an external validation set object (ID = 9) as a response outlier, hence it is acceptably predicted by 64-parameters CP ANN and 13-parameters PLS models as depicted in Fig. 6a and c (see Supporting information, Table S3).

3.3. Is the MEDS-based AD estimation affected by the neural network architecture?

All the MEDS-based AD estimations performed so far, were established by using a single-architecture CP ANN prediction

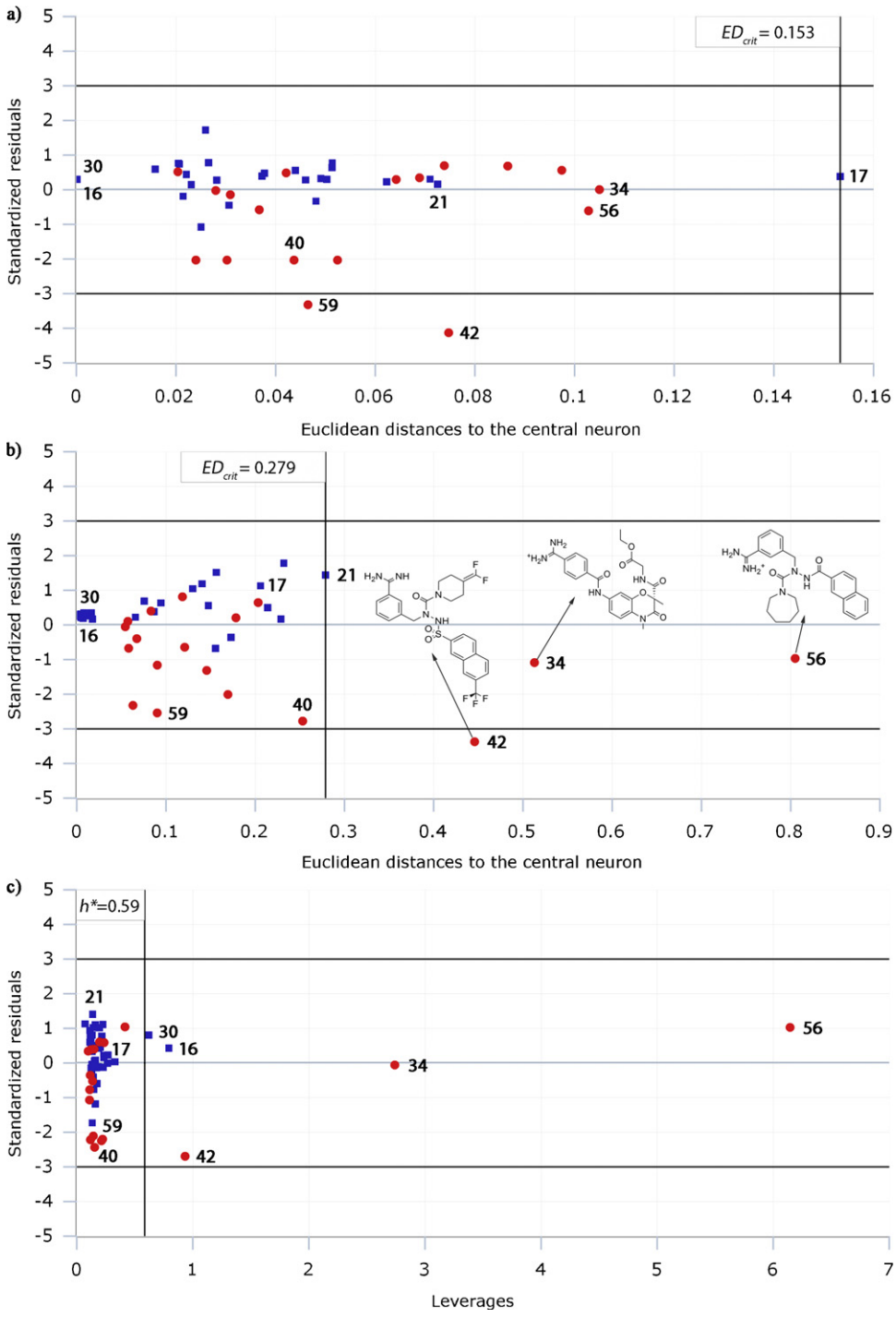


Fig. 5. Graphical representation of the (a, b) MEDS-based versus (c) leverage-based method for AD estimation for our (a) 97-parameters CP ANN prediction model [26] and the additionally developed 7-parameters (b) CP ANN and (c) PLS models, for prediction of the inhibitory potency for a set of 59 trypsin inhibitors (*Dataset 2*) [26]. Training/internal test set objects (41 compounds) are depicted as solid blue rectangles, while the external validation set object (18 compounds) as solid red circles. The critical hat value ((c) $h^* = 0.59$) is calculated according to (Eq. (7)). The compounds identified as outliers (outside of the model's AD) by both methods ((a, b) MEDS vs. (c) leverage) are signed by their corresponding ID numbers ((a–c) Supporting information, Table S2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model. On the other hand, when working with non-linear QSAR modeling methods such as ANN, the colloquial strategy is to train different ANN architectures before to make a final decision which ANN model will be accepted (usually according to the highest value for the cross-validated coefficient of correlation for the model) [51]. According to this, the probability to find more than one acceptable model (trained under conditions different than those used to develop the best one) must be also taken into account, and therefore a crucial question arises: *Is the MEDS-based AD estimation really affected by the ANN architecture employed?* In order to give an answer to this question, a side-by-side comparative evaluation of the ADs for different ANN models of *Dataset 1* developed under various network architectures must be performed as demonstrated in Fig. 7.

Fig. 7 shows the MEDS-based AD plots for three CP ANN models for prediction of the bilitranslocase transport activity for 88

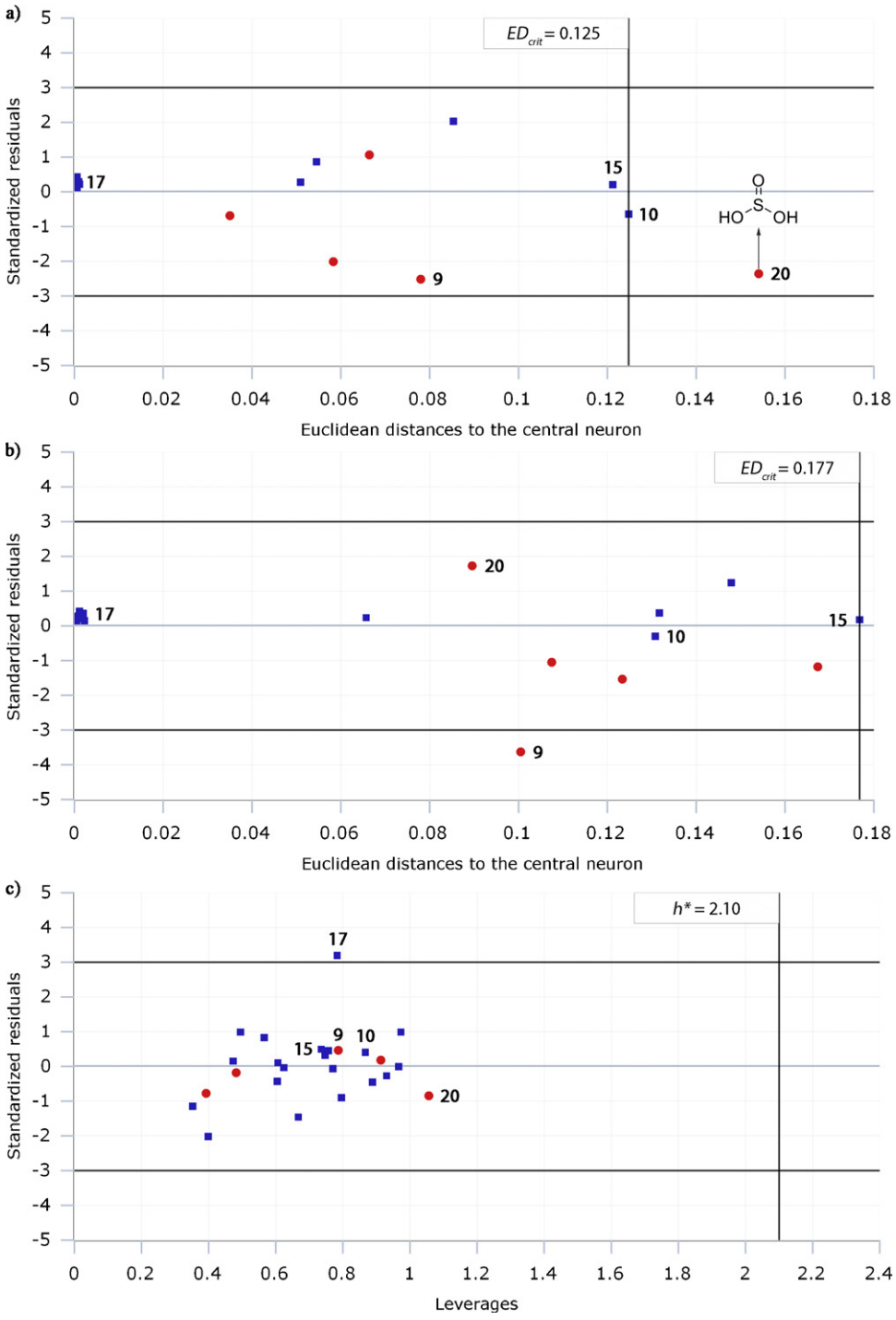


Fig. 6. Graphical representation of the (a, b) MEDS-based versus (c) leverage-based method for AD estimation for our (a) 64-parameters CP ANN prediction model [27] and the additionally developed 13-parameters (b) CP ANN and (c) PLS models, for prediction of the retention factors (k) for a set of 25 ions (*Dataset 3*) [27]. Training/internal test set objects (20 compounds) are depicted as solid blue rectangles, while the external validation set object (18 compounds) as solid red circles. The critical hat value ((c) $h^* = 2.10$) is calculated according to (Eq. (7)). The objects identified as outliers (outside of the model's AD) by both methods ((a, b) MEDS vs. (c) leverage) are signed by their corresponding ID numbers ((a–c) Supporting information, Table S3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

compounds (*Dataset 1*) developed under different conditions (network architectures). Namely, alongside our published CP ANN model [25] which AD plot is represented in Fig. 7b (6 × 6 network), for comparative purposes we selected two more models developed under two vicinal network architectures, i.e., 5 × 5 network and 7 × 7 network, which AD plots are depicted in Fig. 7a and c, respectively. Comparing to the 6 × 6 model's AD (Fig. 7b), no significant differences were determined between all three cases in regard to the identified outliers. The external validation set compounds (ID = 57, 64, 66, 70, 73, 77, 80, 83, and 88) which were previously identified as extreme points, i.e., structurally-influential outliers (ED > ED_{crit}) as well as the response outliers (ID = 63, 67, 74, 76, and 79), were found to be outside of the AD in all three cases. The only questionable compound predicted by 5 × 5 (Fig. 7a) and 7 × 7 (Fig. 7c) models as structurally-influential outlier is the compound hydrochlorothiazide (ID = 74; $ED_{ID=74[5\times5]}$ = 0.357

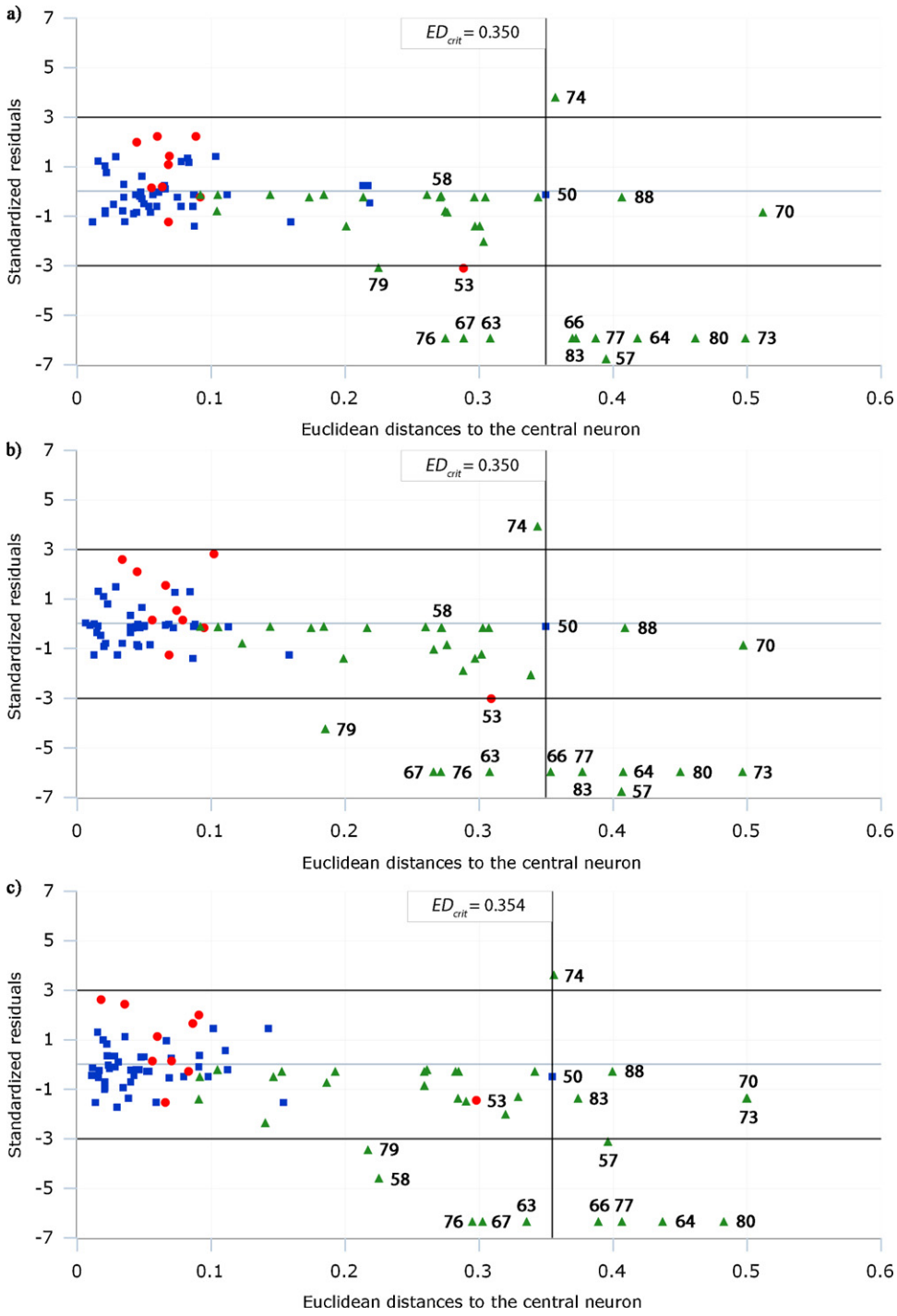


Fig. 7. Side-by-side comparative assessment of the MEDS-based AD for three CP ANN prediction models for estimation of the bilitranslocase transport activity developed using different network architectures: (a) 5 × 5, (b) 6 × 6 (our published model, *Dataset 1*) [25], and (c) 7 × 7. The normalized ED threshold values (a) ED_{crit} = 0.350, (b) ED_{crit} = 0.350, and (c) ED_{crit} = 0.354, respectively, correspond to the training set object with maximal ED value. The training set objects (45 compounds) are depicted as solid blue rectangles, external test set objects (10 compounds) as solid red circles, while the external validation set objects (33 compounds) are represented as solid green triangles. The objects identified as potential outliers (ED > ED_{crit}) are signed with their corresponding ID number ((a, c) Supporting information Table S4; (b) Table 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and $ED_{ID=74[7\times7]}$ = 0.356) that belongs to the thiazide diuretic's therapeutic class. Indeed, hydrochlorothiazide is structurally very different than the compounds used for developing of the CP ANN models (no similar compound can be found) and therefore should be expected to be outside of the model's AD, even it was experimentally determined as bilitranslocase competitive inhibitor (Table 1). Hence, the same compound is situated within the 6 × 6 model's AD as shown in Fig. 7b.

4. Conclusion

In this study a novel, simple and effective distance-based methodology for AD estimation in case of established prediction-like ANN models was introduced taking into account the so-called "*minimum Euclidean distance space*" (MEDS) concept. The method offers a graphical depiction of the ANNs model AD for fast and accurate visual determination of the detected structurally-influential

outliers. The performances of our MEDS-based AD estimation concept were thoroughly evaluated through three case studies utilizing a pre-built and validated CP ANN prediction models. In the model for the estimation of the transport activity of the transmembrane protein bilitranslocase for a diverse set of compounds, the method identified 9 compounds out of total 33 compounds (used for external validation of the CP ANN model) as pure structurally-influential outliers with calculated EDs greater than the critical threshold value (ED_{crit}), of which the majority were also determined as response outliers (objects with standardized residual values greater than $\pm 3\sigma$ units). The structural analysis of these compounds clearly confirmed that no similarity exist between them and the chemical structures used for developing of the CP ANN model, and therefore it is very likely to be situated far from the structural centroid of the model. Furthermore, the PLS modeling methodology was applied on the same data and the leverage-based AD was estimated and compared with the MEDS-based AD results. Although, both AD estimation methods are essentially different with regards to the distance metrics utilized (EDs vs. leverages), they gave comparable outcome. The same outlying objects identified by using the MEDS-based AD approach were again identified as extreme points on the assessed Williams plot ($h > h^*$), together with a much larger ensemble of external validation set objects. These results explicitly show that the calculated cut-off leverage value (h^*) is much more restrictive for the PLS model, comparing to the ED_{crit} for the CP ANN model. Finally, the side-by-side comparative assessment of the AD plots generated for different CP ANN models developed under various ANN architectures close to the optimal one showed no significant differences in the identified outliers, allowing us to conclude that the MEDS-based AD assessment is not significantly affected by the neural network size.

From pharmacological point of view, the majority of the compounds identified as outliers throughout this case study belong to the class of non-steroidal antiinflammatory drugs (NSAIDs) which are totally different than the training set compounds. Therefore we believe that their permanent appearance out of the model AD pinpoint to some degree of weakness of the model for prediction of NSAIDs. Additionally, the same methodology was applied on two more case studies and similar outcomes were observed.

In conclusion, our MEDS-based approach demonstrated satisfactory performances for proficient assessment of the AD in case of non-linear predictive distance-based ANN models such as KANN and CP ANN, as has been already shown for classification CP ANN models [24]. This method is supposed to improve the choice of standard methods needed in providing the high quality validations in the QSAR world.

5. Notes

All the AD plots provided in this study were generated and analyzed by using the CPNN-AD Builder and MLR-AD Builder available upon request.

Acknowledgement

Authors thank Agency of Research of R. Slovenia (ARRS) for the financial support through the Grants P1-0017.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aca.2012.11.002.

References

[1] M.T.D. Cronin, in: T. Puzyn, J. Leszczynski, M.T.D. Cronin (Eds.), Recent Advances in QSAR Studies: Methods and Applications, Series: Challenges and Advances in Computational Chemistry and Physics, vol. 8, Quantitative Structure-Activity Relationships (QSARs) – Applications and Methodology, Springer, Dordrecht, 2010, pp. 3–11.
[2] A.Z. Dudek, T. Arodz, J. Gálvez, Comb. Chem. High Throughput Screen. 9 (2006) 213–228.
[3] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.
[4] OECD, Principles for the Validation for Regulatory Purposes of (Quantitative) Structure–Activity Relationship Models, OECD, Paris, France, 2004.
[5] J. Jaworska, M. Comber, C. Van Leeuwen, C. Auer, Environ. Health Perspect. 111 (2003) 1358–1360.
[6] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Molecules 17 (2012) 4791–4810.
[7] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang, ATLA 33 (2005) 1–19.
[8] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, ATLA 33 (2005) 445–459.
[9] N. Nikolova-Jeliazkova, J. Jaworska, ATLA 33 (2005) 461–470.
[10] J.M. Barnard, G.M. Downs, P. Willett, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[11] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, S.K. Kearsley, J. Chem. Inf. Comput. Sci. 44 (2004) 1912–1928.
[12] J. Polanski, A. Bak, R. Gieleciak, T. Magdziarz, Molecules 9 (2004) 1148–1159.
[13] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, J. Chem. Inf. Comput. Sci. 48 (2008) 1733–1746.
[14] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, V.V. Kovalishyn, V.V. Prokopenko, I.V. Tetko, J. Chemom. 24 (2010) 202–208.
[15] M.M. Deza, E. Deza, Encyclopedia of Distances, Springer-Verlag, Heidelberg, Berlin, 2009.
[16] S. Wold, W.J. Dunn, J. Chem. Inf. Comput. Sci. 23 (1983) 6–13.
[17] S. Haykin, Neural Networks, A Comprehensive Foundation, second ed., Prentice Hall, New Jersey, 1999.
[18] T. Kohonen, Self-Organizing and Associative Memory, third ed., Springer-Verlag, New York, 1988.
[19] R. Bharath, J. Drosen, Neural Network Computing, Windcrest/McGraw-Hill, New York, 1994.
[20] X. Yong-jin, G. Hua, QSAR Comb. Sci. 22 (2003) 422–429.
[21] T.S. Schroeter, A. Schwaighofer, S. Mika, A.T. Laak, D. Suelzle, U. Ganzer, N. Heinrich, K.-R. Müller, J. Comput. Aided Mol. Des. 21 (2007) 651–664.
[22] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, P.A. Koutentis, G. Kollias, Chem. Biol. Drug Des. 76 (2010) 397–406.
[23] A. Afantitis, G. Melagraki, P.A. Koutentis, H. Sarimveis, G. Kollias, Eur. J. Med. Chem. 46 (2011) 497–508.
[24] N. Fjodorova, M. Novič, A. Roncaglioni, E. Benfenati, J. Comput. Aided Mol. Des. 25 (2011) 1147–1158.
[25] Š. Župerl, S. Fornasaro, M. Novič, S. Passamonti, Anal. Chim. Acta 705 (2011) 322–333.
[26] Š. Župerl, G. Mlinšek, T. Šolmajer, M. Novič, J. Chemom. 21 (2007) 346–356.
[27] V. Drgan, Modeling of retention processes and structure-retention relationships in ion chromatography, doctoral dissertation, Ljubljana, 2010.
[28] G.L. Sottocasa, S. Passamonti, L. Battiston, L. Pascolo, C. Tiribelli, J. Hepatol. 24 (1996) 36–41.
[29] S. Passamonti, M. Terdoslavich, A. Margon, A. Cocolo, N. Medic, F. Micali, G. Decorti, M. Franko, FEBS J. 272 (2005) 5522–5535.
[30] M. Terdoslavich, Trasporto epatocellulare del Cibacron Blue. Ruolo della bilitranslocasi, Dipartimento di Biochimica Biofisica e Chimica delle Macromolecole, Università degli Studi, Trieste, 2002.
[31] S. Passamonti, U. Vrhovsek, F. Mattivi, Biochem. Biophys. Res. Commun. 296 (2002) 631–636.
[32] S. Passamonti, F. Tramer, E. Petrussa, E. Braidot, A. Vianello, in: A.G. Fett-Neto (Ed.), Plant Secondary Metabolism Engineering, Methods in Molecular Biology, Humana Press Inc., Totowa, NJ, 2010, p. 307.
[33] G. Baldini, S. Passamonti, G.C. Lunazzi, C. Tiribelli, G.L. Sottocasa, Biochim. Biophys. Acta 856 (1986) 1–10.
[34] S. Passamonti, G.L. Sottocasa, Biochim. Biophys. Acta 943 (1988) 119–125.
[35] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902–3909.
[36] M. Novič, J. Zupan, J. Chem. Inf. Comput. Sci. 35 (1995) 454–466.
[37] P. Gramatica, P. Pilutti, E. Papa, J. Chem. Inf. Comput. Sci. 44 (2004) 1794–1802.
[38] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, Germany, 1999.
[39] J. Dayhof, Neural Network Architectures: An Introduction, Van Nostrand Reinhold, New York, 1990.
[40] V. Consonni, D. Ballabio, R. Todeschini, J. Chem. Inf. Model. 49 (2009) 1669–1678.
[41] V. Consonni, D. Ballabio, R. Todeschini, J. Chemom. 24 (2010) 194–201.
[42] T.H. Black, Derivations of Applied Mathematics, The Debian Project (http://www.debian.org), 1967.
[43] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 1–18.
[44] N. Nikolova, J. Jaworska, QSAR Comb. Sci. 22 (2003) 1006–1026.

[45] M.V. Putz, Chem. Cent. J. 5 (2011) 1–11.
[46] S.L. Rodgers, A.M. Davis, N.P. Tomkinson, J. Chem. Inf. Model. 47 (2007) 2401–2407.
[47] F.J. Prado-Prado, O.M. de la Vega, E. Uriarte, F.M. Ubeira, K.-C. Chou, H. González-Diaz, Bioorg. Med. Chem. 17 (2009) 569–575.
[48] J.C. Gower, Math. Sci. 7 (1982) 1–14.
[49] T. Kohonen, Biol. Cybern. 43 (1982) 59–69.
[50] T. Kohonen, Overture, Self-Organizing Neural Networks: Recent Advances and Applications, Springer-Verlag, Inc., New York, 2001.
[51] J. Zupan, M. Novič, I. Ruisanchez, Chemom. Intell. Lab. Syst. 38 (1997) 1–23.
[52] N. Minovski, A. Jezierska-Mazzarello, M. Vračko, T. Šolmajer, Cent. Eur. J. Chem. 9 (2011) 855–866.
[53] P. Gramatica, QSAR Comb. Sci. 26 (2007) 694–701.
[54] F.J. Anscombe, J.W. Tukey, Technometrics 5 (1963) 141–160.
[55] D.C. Hoaglin, R.E. Welsch, Am. Stat. 32 (1978) 17–22.
[56] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[57] M. Meloun, S. Bordovská, K. Kupka, J. Math. Chem. 47 (2010) 891–909.
[58] J.G. Topliss, R.P. Edwards, J. Med. Chem. 22 (1979) 1238–1244.