Prediction of antiprion activity of therapeutic agents with structure–activity models

ARTICLE *in* MOLECULAR DIVERSITY · SEPTEMBER 2013

Impact Factor: 1.9 · DOI: 10.1007/s11030-013-9477-3 · Source: PubMed

3 AUTHORS:

Katja Venko

National Institute of Chemistry

6 PUBLICATIONS 42 CITATIONS

SEE PROFILE

Spela Zuperl

National Institute of Chemistry

6 PUBLICATIONS 17 CITATIONS

SEE PROFILE

Marjana Novič

National Institute of Chemistry

151 PUBLICATIONS 2,141 CITATIONS

SEE PROFILE

Mol Divers
DOI 10.1007/s11030-013-9477-3

FULL-LENGTH PAPER

Prediction of antiprion activity of therapeutic agents with structure–activity models

Katja Venko · Špela Župerl · Marjana Novič

Received: 23 May 2013 / Accepted: 31 August 2013

Abstract We have developed computational structure–activity models for the prediction of antiprion activity of compounds with known molecular structure. The aim is to apply the developed classification and predictive models in further drug design of antiprion therapeutics. The neural network models developed on the counter-propagation reinforcement learning strategy performed better than the linear regression models. The initial data set was composed of 461 compounds representing diverse groups of chemicals (derivatives of acridine, quinolone, Congo red, 2-aminopyridine-3,5-dicarbonitrile, styrylbenzoazole, 2,5-diamino-benzoquinone), which have been tested in comparable cell-screening assay studies for their activity against prion accumulation. Initially, we have designed a classification model for preliminary sorting of compounds into highly active, active, and inactive groups. Further, only the active compounds with IC_{50} less or equal to 10 μM were considered as the initial source of data. Altogether, 158 compounds were used to train the artificial neural network model for the estimation of the antiprion activity. The predictive ability of the model was significantly improved after selection of influential variables with genetic algorithm. The root- mean-squared error of the predicted pIC_{50} values for the external validation set (RMS_{EV}) was slightly above 0.50 log units. A linear regression model, developed for the reasons of comparison, performed with a lower predictive ability (RMS_{EV} 0.92 log units). The applicability domain of the models was assessed by a leverage and distance approach. The set of selected influential structural variables was further studied with the aim to get a better insight into the structural features of compounds potentially involved in disturbing of the prion–prion interactions.

Keywords Prions · Antiprion activity · QSAR models · MLR · Artificial neural networks · Therapeutic compound

Introduction

Abnormal isoforms of the native prion proteins cause prion diseases. These are rare, rapidly progressive and fatal neurodegenerative illnesses resulting in transmissible spongiform encephalopathy, which occurs as a consequence of the self-association and deposition of the pathogenic prion proteins (PrP^{Sc}) in the central nervous system of humans and animals. The prion disease pathology has three modes of initiation: sporadic, genetic, and acquired. In humans, the most prevalent prion diseases are the Creutzfeldt–Jakob disease, Gerstmann–Sträussler–Sheinker disease, fatal familial insomnia and kuru. On the other hand, in animals the most prevalent are the scrapie of sheep, bovine spongiform encephalopathy, and chronic wasting disease of deer [1].

The normal cellular prion protein (PrP^{C}) is highly conserved among vertebrates [2]. It is a cell-surface localized glycoprotein with the glycosyl-phosphatidylinositol (GPI) anchor, which is associated with cholesterol- and glycosphingolipid-rich lipid rafts [3]. The molecular mechanism of the post-translational conversion into the misfolded,

Electronic supplementary material The online version of this article (doi:10.1007/s11030-013-9477-3) contains supplementary material, which is available to authorized users.

K. Venko · Š. Župerl · M. Novič (✉)
Laboratory of Chemometrics, National Institute of Chemistry,
Hajdrihova 19, 1000 Ljubljana, Slovenia
e-mail: marjana.novic@ki.si

K. Venko
e-mail: katja.venko@ki.si

Š. Župerl
e-mail: spela.zuperl@ki.si

Published online: 20 September 2013

β-sheet-rich isoform PrP^{Sc} is still enigmatic, despite numerous studies [4–8]. Although researchers have suggested various mechanisms, the "protein-only" hypothesis is still largely acceptable. According to it, PrP^{Sc} acts as a template, which enhances the conversion of PrP^{C} into protease-resistant PrP^{Sc} [1]. Due to the strong tendency of PrP^{Sc} to aggregate into insoluble amyloid fibrils, further cell mechanisms are activated, which finally result in neuronal death [9]. Interestingly, the comparison between the normal and the pathological isoforms shows a similar global architecture with only few local structural variations, mostly in the α_2–α_3 inter-helical interface and in the α_2–β_2 loop region (for more details see Fig. 4 in Results), which probably have the highest impact on intermolecular interactions and trigger spontaneous generation of infectious PrP^{Sc} conformation [10,11].

In several pathological events protein–protein surface interactions play one of the crucial roles; consequently, their control may offer therapeutic benefits. Therefore, the development of small organic molecules and their usage as modulators to interfere with specific interactions could be a crucial drug design strategy [12,13]. With molecular simulation-based approach, it is revealed that these interactions are also involved in the fibrillation process of prion diseases [14]. Theoretically, five strategies for drug discovery have been suggested: (i) block PrP^{C} synthesis by antisense oligonucleotides targeted to PrP mRNA, (ii) stabilize PrP^{C}, (iii) enhance PrP^{Sc} clearance, (iv) interfere with binding of PrP^{C} to PrP^{Sc}, and (v) prevent binding of protein X to PrP^{C} [15]. Regarding the last strategy, no auxiliary molecule involved in prion replication (e.g., protein X) has been identified yet; because of that, the role of auxiliary molecules still remains unexplained or needs to be tested [16]. Thus, for the design of ligands that will stick onto prion surfaces to disrupt prion–prion interactions and consequently inhibit prion self-assembly, the prion chaperones are needed to be determined. Characterizing chemical chaperones, which will be coherent with pathogenic conversion, is still in progress, as various regions of prion, like the factor X-binding site [15] and other prion surface hot spots [5,7,14,17], are proposed to be involved.

Although lots of immanent studies have been done for prion diseases, currently, no clinical treatment is available. Only one really controlled clinical trial using a prospective double-blinded approach was carried out for flupertine [18]. Other randomized double blinded, placebo-controlled studies for quinacrine and doxycycline are underway. Problems that should be carefully addressed in future trials and overviews of the past human prion diseases treatment cases are accurately reported by Zerr [19] and Appleby et al. [18]. Firstly, natural products, commercially available chemicals and drugs like analgesics, anti-depressants, anti-microbials, and anti-coagulants have been screened and studied in experimental models for their potential as antiprion therapeutics [15,20,21]. Further, great effort was involved in searching for new efficient therapeutics by immanent de novo synthesized libraries of compounds, which were tested by various screening assays either in vitro in cellular lines, or in vivo in animal models [7,13,15,16,22–39]. However, the mode of action and potential cellular targets for most of these compounds remain largely unknown. In basics, the therapeutic compounds could act in two modes: directly on PrP^{C}/PRP^{Sc} like Congo red, or indirectly by interfering with the activity of cellular factors required for prion propagation like antimalarial drugs Quinacrine and Chloroquine [8].

The aim of the research presented here is to develop a data-driven model for prediction of antiprion activity, which could be applied for further drug design, taking into account that the main obstacles in the development of therapeutic agents are the high research and development costs/risks. Commonly, only a relatively small percentage of total R&D costs are used for the initial step. In contrast, the impact on the efficiency of the whole drug developmental process is high since the initial computer screening for hundreds of compounds is crucial for identifying the most promising candidates and thus reducing the number of compounds for further experimental testing [40].

In this study, we have compiled from literature reports a large set of antiprion compounds tested with cell-screening assays. On the basis of the collected data set we have developed data-driven models for further drug design. The new computational models are designed to be used in the initial drug design stage, specifically for in silico PrP^{Sc} inhibition prediction of small antiprion molecules. The models follow the principles of the agreement of OECD (Organisation for Economic Co-operation and Development) member countries for the development of quantitative structure–activity relationships (QSAR) models. "To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information: (1) a defined endpoint, (2) an unambiguous algorithm, (3) a defined domain of applicability, (4) appropriate measures of goodness-of-fit, robustness, and predictivity, (5) a mechanistic interpretation, if possible" [41,42]. The QSAR models are frequently used in rational drug design for reasons of cost, time, and animal welfare [43].

Among several models developed in this study for the prediction of antiprion activity of molecules, the best performance was obtained for a nonlinear model based on counter-propagation neural network (CP-ANN) with molecular descriptors as structural inputs of chemical compounds. Comparison with other models, including linear ones, was performed. For a preliminary categorization of compounds into highly active, active and inactive ones, the specific classification model was designed. The selection of the most influential structural variables with genetic algorithm was integrated into the modeling methodology. According to the

OECD principles for QSAR models, we have assessed the applicability domain of the constructed models using Euclidean distance-based approach for CP-ANN models. With the interpretation of selected influential variables, we have tried to get an insight into the mechanistic explanations of inhibition of prion misfolding and aggregation. Finally, the in silico models are prepared to be used for virtual screening purposes, along with the molecular modeling and docking, in the initial stage of the discovery of antiprion drugs.

Materials and methods

Data set

From literature, we have collected a set of 507 compounds including their experimental values about activities against PrP^{Sc} accumulation. Prion inhibition concentrations of various small chemical compounds were collected from different comparable in vitro studies. Experimental protocols of antiprion activity assays are explained in detail in the references listed in Table 1; here we describe only general points of the experimental schemes. In short, antiprion activity assays were performed on mouse scrapie-infected neuronal cell lines exposed to specific solutions of the tested compounds for 1 week. After incubation, the cells were lysed and digested with proteinase K. The ability to reduce PrP^{Sc} levels was determined in comparison with untreated control by measuring the density of proteinase K-resistant PrP^{Sc}. The results have been presented as the concentration of a compound required to inhibit 50 % of PrP^{Sc} relative to the control (IC_{50}) or percentage of PrP^{Sc} inhibition at 1 or 10 μM of compound (%PrP^{Sc}). Among the studies considered, some experimental conditions varied in different parameters (e.g., cell model, prion strain, incubation period, analyzed methodology). Nevertheless, the studies are comparable due to the implementation of the positive controls. As the positive controls, well-documented antiprion agents such as quinacrine (compound 49), BiCappa (compound 149), imipramine (compound 175), or Congo red (compound 288) were used [15,16,20–34,44].

In several studies, the cell viability control was also performed prior to antiprion activity assay test [26,28,31–36,39]. Therefore, the compounds, which were determined as potentially toxic, were not included in the antiprion activity assay test. Consequently, the information about their antiprion activity is not available and thus removed from our data set. The remaining 461 structurally diverse compounds cover a broad range of chemical space as well as large IC_{50} concentration range. 2D chemical structures and their antiprion activity values are listed in supplementary information, Table S1. Compounds mainly differ in the level of planarity and structure of substituents, attached functional groups, which particularity vary in size, aromaticity, polarity, and hydrogen-bonding capability. Several compounds have a common bivalent structure with a central core and two linkers connecting two terminal moieties with different mono-, bi-, or tricyclic scaffolds (supplementary information, Fig. S1). Compounds were split according to their structural similarity into nine derivative groups: (I) Congo red, (II) Diamino-benzoquinone, (III) Quinolone, (IV) Aminopyridine-dicarbonitrile, (V) Diketopiperazine, (VI) Acridine, (VII) Styrylbenzoazole derivatives, (VIII) prion chaperone compounds, and (IX) other compounds (Table 1).

We have classified the compounds according to their antiprion activity into three classes: highly active, active, and inactive. It is easy to distinguish between highly active and inactive compounds; however, it is very difficult to determine the threshold that would decently separate compounds with low activity from inactive ones. For this reason, we have introduced the middle active class. The threshold for each class was determined as a consensus according to comments of authors of the experimental studies (references listed in Table 1). The criteria for each class were: (A) highly active—$IC_{50} < 1\,\mu M$ or %$PrP^{Sc} > 50$, (a) active—$IC_{50} = 1-10\,\mu M$ or %$PrP^{Sc} = 5-50$, and (N) inactive—$IC_{50} > 10\,\mu M$ or %$PrP^{Sc} < 5$. Among the 461 compounds, 286 were active (135 highly active, 151 active) and 175 were inactive. All 461 compounds were considered for the classifi-

Table 1 List of nine derivative groups of including 507-tested compounds, which were applied in this study

	Groups of compounds	No. of compounds	References
I	Congo red derivatives	88	[22–24]
II	Diamino-benzoquinone derivatives	26	[13,25]
III	Quinolone derivatives	48	[21,26–28,44]
IV	Aminopyridine-dicarbonitrile derivatives	24	[16]
V	Diketopiperazine derivatives	24	[29]
VI	Acridine derivatives	141	[21,26,30–35]
VII	Styrylbenzoazole derivatives	23	[36]
VIII	Prion chaperone compounds	54	[37]
IX	Other compounds	79	[7,15,20,21,28,32,38,39]

cation model, while for the construction and optimization of the predictive models only 158 active compounds with exact IC_{50} values were selected out of the 286 highly active and active compounds.

Computational methods

Molecular descriptors

Optimization of the 3D chemical structures was performed with the MOPAC computational program using AM1 semi-empirical approach with the minimal energy criterion for the optimal conformation of the compound in vacuum [45]. Molecular descriptors (MDs), which mathematically characterize molecular structures, were calculated with the CODESSA program [46]. Molecular descriptors of various types, including 3D descriptors, were considered in our modeling strategy: (i) constitutional, (ii) topological, (iii) geometrical, (iv) electrostatic, and (v) quantum-chemical descriptors. With the applied software 328 MDs for 461 compounds were calculated to obtain 461 × 328 data matrix. Each descriptor was normalized to zero mean and unit standard deviation.

In order to obtain a suitable set of MDs for further modeling, the reduction of a very large number of calculated descriptors is needed. Initially, we have omitted MDs with zero variance or high correlated ones. The set of descriptors was further reduced according to the similarity criterion in the Kohonen map of transposed data matrix, as explained later in Results. The inputs for classification and prediction models consisted of m-dimensional vectors representing the chemical structure of the molecules, m being the number of descriptors selected, and the targets corresponding to the antiprion property value for each molecule.

Linear and nonlinear regression methods

Multiple linear regression (MLR) [47] is one of the most often applied statistical methods used for prediction purposes. Linear function is used to model the relationship between a scalar-dependent variable (y) and descriptive variables (X) where the model parameters are estimated from the data. When all responses are dependent on a single variable ($m = 1$), the regression is called simple or univariate regression and the obtained model is usually straight line with a slope (b_1) and an intercept (b_0). For $m > 2$, polynomial equation is used to describe a linear relationship between m descriptive variables $X = (x_1, x_2, \ldots x_m)$ and a response y. MLR [47] is therefore described with equation (Eq. 1):

$$y_\mathrm{i} = b_0 + b_1\mathrm{x}_{\mathrm{i},1} + \mathrm{b}_2\mathrm{x}_{\mathrm{i},2} + \cdots + \mathrm{b}_\mathrm{m}\mathrm{x}_{\mathrm{i},\mathrm{m}} + \mathrm{e}_{\mathrm{i}}; \quad i = 1, 2, 3 \ldots, n \qquad (1)$$

where n is the number of observations, b_0 is the regression constant, b_1 to b_m are the regression coefficients relating the m explanatory variables, e_i is the error (difference between the value predicted by the model and the observation). Once the unknown regression parameters (b_0, b_1 to b_m) are estimated, we can calculate the predicted values ($\hat{y}$). The best MLR model fits the set of data points in a plane with minimal values of sum of squares between experimentally observed (y) and predicted ($\hat{y}$) values (Eq. 2):

$$\sum\nolimits_i e_i^2 = \sum\nolimits_i \left(y_i - \hat{y}_i\right)^2 \qquad (2)$$

Nonlinear regression is a more appropriate technique to model biological data. Nonlinear regression, with an iterative computational approach, is able to fit the data to any equation that defines Y as a nonlinear function of X and one (or more) parameters.

Kohonen Neural network (KohNN) [48,49] is a modeling tool, usually used for visualization and classification purposes, which provides us with the conversion of the data from the multidimensional space to a lower (usually two) dimensional space. As the input data (multidimensional objects; molecular descriptors $-\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_m)$) is introduced into KohNN, the output is so-called self-organizing map (SOM) or Kohonen neural network, with a two-dimensional arrangement of objects on the net of neurons. As a result of unsupervised competitive learning in the KohNN, the distribution of objects on the Kohonen top map is influenced only by the structural descriptors used as inputs. Therefore, clusters on the Kohonen top map are formed due to structural similarity of the objects.

Counter-propagation artificial neural network (CP-ANN) [48,49], usually used for clustering, classification, and prediction for unknown compounds, consists of two layers of neurons, an input or a Kohonen layer and an output layer, with one to one correspondence of the neurons in both layers. The neurons in the Kohonen and in the output layer have equal coordinates in the 2D arrangement, but different number of weights. The neurons in the Kohonen layer are m-dimensional vectors having as many weights as there are input variables (independent variables—$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_m)$). The number of weights in the output layer depend on the number of output variables (dependent variables; molecular property (IC_{50})—$\boldsymbol{Y} = (\boldsymbol{y}_1)$). Initially, the objects are positioned in the Kohonen layer according to the unsupervised learning strategy. Then, the correction of weights in the output layer is performed with the supervised learning strategy, which requires a set of input–output pairs ($\boldsymbol{X}, \boldsymbol{Y}$). As a result of the learning process in the CP-ANN, we obtain the distribution of objects in two-dimensional map regarding only structural information. Below the Kohonen map, we find the response surface with property values. The new objects with unknown property is first located in the

Kohonen layer regarding the independent variables, then the position of the neuron is projected to the output layer, which gives us the property prediction.

Genetic algorithm (GA) is an optimization technique frequently used for variable selection [50]. In our study, we have used GA coupled with CP-ANN in order to select the influential variables, thus improving the predictive ability and the robustness of the models. The selection processes appearing in nature, such as crossover, mutation, and selection of the fittest are incorporated in the computational part of genetic algorithm, where the chromosomes are represented as *m* dimensional binary vectors (bites of zeros and ones). The number one in the chromosome indicates the selection of a particular variable out of *m* variables in to the set of influential variables. More details on GA can be found in the literature [50]. For KohNN, CP-ANN and GA we have used the softwares developed in-house [48,49], written in FORTRAN for IBM-compatible PCs and Windows operating system, and QSARINS software [51] was used for the MLR models.

Results and discussion

Selection of the training, internal test, and external validation sets

Following recommended methodology for building QSAR models [52,53] we have initially divided the set of 461 compounds in three data sets (training, internal test, and external validation). The training (TR) set was used for the learning of the network, the internal test (TE) set for defining optimal model parameters, and the external validation (EV) set for the model validation. Compounds with their molecular descriptors served as an input to the Kohonen neural network, and the output, i.e. distribution of compounds on the Kohonen top map, was used for data division. Two separate selections for the three data sets were performed; one with all (461) compounds (active and inactive) used for classification model and another with only active (158) compounds used for predictive models.

According to the distribution of the 461 compounds described with the 245 molecular descriptors on the Kohonen top map, we have selected for the classification model 263, 111, and 87 compounds for the training, internal test, and external validation set, respectively. The ratio of the training, internal test and external validation set division of 461 compounds was 57:24:19. In order to insure that all the information space is included in all the sets, compounds were selected in all the three data sets such that they were distributed equally over the entire Kohonen top map. We have optimized the distribution of compounds by testing different neural network-training parameters (number of neurons and epochs, learning rates). The optimal distribution of objects and the highest occupancy of neurons were achieved with the network with following parameters: 441 neurons, 500 learning epochs, 0.5 maximal learning rate, 0.01 minimal learning rate, non-toroidal boundary conditions, and triangular correction function of the neighborhood. In Fig. 1, we can see the optimal distribution of 461 compounds in the Kohonen top map with 60 % occupied and 40 % empty neurons. Moreover, the compounds with similar structures are positioned at the same or neighboring neurons and the derivative groups are well clustered.

For predictive models, a careful selection of the three sets was carried out on the basis of the distribution of 158 active compounds (defined with 254 molecular descriptors) on the Kohonen top map. Initially, for the external validation set, 32 compounds were evenly chosen from the entire Kohonen top map. Then several different training and internal test sets were selected and tested for their model performances. The ratio of the training, internal test and external validation set division of 158 compounds was 60:20:20. Furthermore, several different random selections for the training, and internal test set were performed (for the comparison purposes we used the same 32 compounds for the external validation set as described above). The best results (models) were obtained with following division of compounds: 94, 32, and 32 compounds for the training, internal test, and external validation

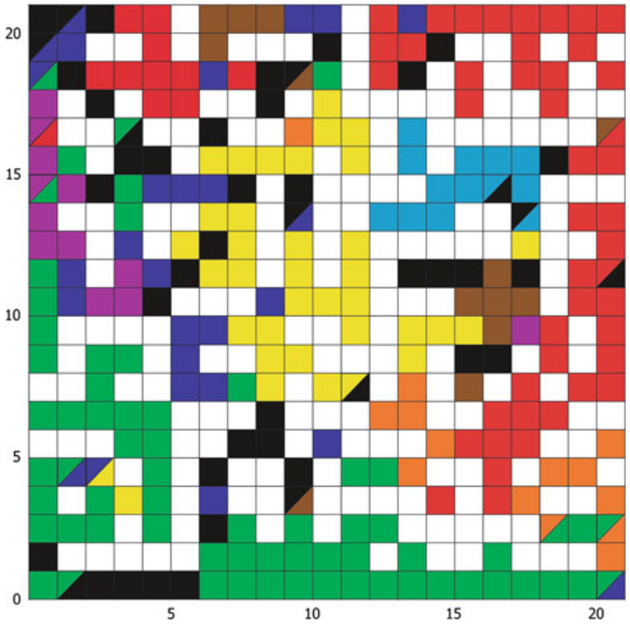

Fig. 1 Distribution of 461 compounds described with 245 MDs on the Kohonen top map with dimensions 21 × 21 neurons. *Different colors* represent groups of derivatives; *red* Congo red derivatives, *orange* diamino-benzoquinone derivatives, *brown* diketopiperazine derivatives, *dark blue* quinolone derivatives, *violet* styrylbenzoazole derivatives, *light blue* aminopyridine-dicarbonitrile derivatives, *green* acridine derivatives, *yellow* prion chaperone compounds, *black* other compounds

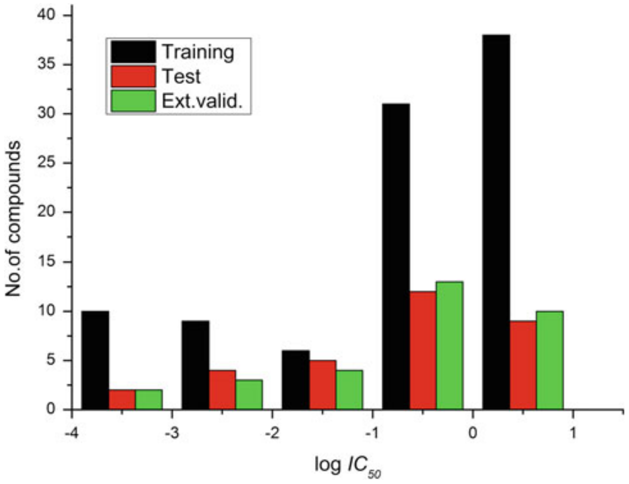

Fig. 2 Distribution of compounds within the antiprion activity range considered for modeling

set, respectively (Table 3). In Fig. 2, the antiprion activity range and the corresponding distribution of compounds for each data set is shown. Finally, the best model was achieved by taking into account the optimal division of compounds and the optimal set of network parameters: 169 neurons, 200 learning epochs, 0.5 maximal learning rate, 0.01 minimal learning rate, non-toroidal boundary conditions, and triangular correction function of the neighborhood.

Classification model

The reduction of variables was the first step in the development of the classification model. From the initial pool of 328 calculated molecular descriptors, we have selected 245 of them with the variance >0.003. Furthermore, this set of descriptors was reduced according to the similarity criterion in the Kohonen top map obtained with the transposed data set (245 × 461 matrix), having the descriptors mapped into 7 × 7 dimensional Kohonen neural network. Each of the 245 descriptors occupied one of the 49 neurons in the 7 × 7 Kohonen neural network. Similar descriptors were mapped onto the same neuron. From each neuron only two descriptors were selected, those with the minimal and maximal Euclidean distances to this particular neuron. In that way, 160 MDs were omitted. For further modeling, the remaining 85 MDs (7 constitutional, 12 topological, 4 geometrical, 18 electrostatic, 44 quantum-chemical descriptors) were used.

For the classification, a set of 461 compounds described with 85 MDs was divided into three subsets according to the protocol described in “Selection of the training, internal test and external validation sets” section. The aim of the classification model was to classify the compounds by their activity into three groups; (A) highly active, (a) active, and (N) inactive (see “Data set” section for classification criteria). After careful consideration to include all representatives of the derivative groups into all three subsets, 263 (77 A, 87 a, 99 N), 111 (28 A, 38 a, 45 N), and 87 (29 A, 30 a, 28N) compounds were selected for the training, internal test, and external validation set, respectively. In order to obtain the optimal classification model, we have tested different network parameters; number of neurons from 16 × 16 to 19 × 19, and number of learning epochs from 100 to 300. The optimal classification model based on 85 MDs was achieved with the following neural network parameters: dimension 19 × 19 neurons, 300 epochs, 0.5/0.01 max/min learning rate, non-toroidal boundary conditions and triangular correction function of the neighborhood. The model was evaluated with the internal test and the external validation set. The best model was chosen on the basis of the classification functions (accuracy, selectivity, sensitivity, Matthews correlation coefficient), which are usually used for evaluation of classification model efficiency [54] and were calculated by equations listed in Table 2. Since the CP-ANN classification model yields a real number between 0.0 and 1.0 as the predicted class for each compound, the threshold of 0.5 was considered to ascribe a class number 0 or 1. Accordingly the number of correctly classified [true positive (TP) and true negative (TN)] and incorrectly classified [false positive (FP) and false negative (FN)] compounds were determined. The performance of the optimized CP-ANN classification model is shown in Table 2. Altogether, the number of true predictions was 398 (267 TP and 131 TN) and the number of false predictions was 63 (39 FP and 24 FN). Cumulative accuracy, sensitivity, and specificity for all compounds of the best classification model were 0.86, 0.92, and 0.77, respectively. The Matthews correlation (MCC) was 0.7, which is sufficiently high as it ranges from −1 to 1.

The main reason for slightly lower accuracy (0.86) and specificity (0.77) was a relatively high number of false negative predictions in the training set (FN = 12) and false positive predictions in the external validation set (FP = 14). Those compounds have similar chemical structure, so they are located on the same or neighboring neurons, but belong to different classes due to larger differences in biological properties. Therefore, a complete separation of active and inactive compounds is not possible. For example, the compounds with ID = 56, 58, 59, 88, 102 have similar chemical structures, therefore are placed on the same neuron, but have different properties (two highly active, two active, one inactive), see Fig. S2 and Table S1 in supplementary information.

Predictive models for active compounds

In order to obtain values for antiprion activity of novel compounds, the predictive model was built with counter-propagation neural network (CP-ANN). Only active compounds were considered for the model development and

Table 2 Confusion matrix for the performance evaluation of the optimized CP-ANN classification model based on 85 molecular descriptors (MD) calculated for total 461 compounds

Classification		Prediction						
		TR		TE		EV		$\sum$
		Active	Inactive	Active	Inactive	Active	Inactive	
TR	Active	**154**(TP)	12(FN)					166
	Inactive	16(FP)	**81**(TN)					97
TE	Active			**61**	5			66
	Inactive			9	**36**			45
EV	Active					**52**	7	66
	Inactive					14	**14**	21
$\sum$		70	93	70	41	66	21	461
Accuracy		0.86						
Sensitivity		0.92						
Specificity		0.77						
Matthews coefficient		0.70						

Accuracy **AC** = (TP + TN) / (TP + TN + FP + FN)
Sensitivity **SE** = TP / (TP + FP)
Specificity **SP** = TN / (TN + FN)
Matthews coefficient **MCC** = ((TP × TN) − (FP × FN))/(SQRT((TP + TN) × (TP + FP) × (TN + FP) × (TN + FN)))
TR training set, *TE* internal test set, *EV* external validation set, *TP* true positive, *FP* false positive, *TN* true negative, *FN* false negative, *bold* true predictions

external validation, and altogether 158 such compounds were used. The reduction of the initial set of descriptors (328 MDs) was performed in the same way as reported for the classification model (see "Classification model" section). 254 MDs with variance >0.003 were considered. The transposed matrix was calculated only for the active compounds (254 × 158), which slightly changed the distribution of the descriptors in the 7 × 7 Kohonen top-map in comparison with the complete dataset of 461 compounds. 87 descriptors were selected for further modeling: 8 constitutional, 10 topological, 3 geometrical, 16 electrostatic, and 50 quantum-chemical.

Different modes of the training, internal test, and external validation set selection were performed and tested. The set of 158 active compounds were divided randomly or according to the Kohonen distribution described in "Classification model" section. The optimal division of compounds according to the internal test results was Kohonen distribution into 94, 32, and 32 compounds for the training, internal test, and external validation set, respectively.

The 87 MDs, together with the corresponding log IC_{50}, were used to train the CP-ANN designed for prediction purposes. The network parameters were optimized to obtain accurate predictions of the CP-ANN model. The following network parameters were varied: number of neurons from 9 × 9 to 20 × 20, number of learning epochs from 1 to 1,000, maximal and minimal learning rate from 0.2 to 0.9, and from 0.01 to 0.1, respectively. The obtained models were evaluated with the internal test set (TE), and the model with a minimal RMS error of the internal test set (RMS_{TE}) was chosen as the optimal model. Minimal RMS_{TE} (0.54) was obtained with the network of 324 neurons (18 × 18) trained with 115 epochs with maximal and minimal learning rates set to 0.5 and 0.01, respectively. In Fig. 3a, the optimal CP ANN model (model_**A**) is shown.

We have included genetic algorithm GA [50] into the CP-ANN modeling procedure in order to select the most influential variables out of the 87 descriptors. With GA coupled to CP-ANN, several combinations of different network and GA parameters [number of neurons, number of learning epochs, number of initial genes, and number of bits in criterion (0 or 1)] were consider in a population of 100 chromosomes (represented as binary vectors) evolving in 500 generations. Besides the varied network parameters, fixed parameters are the following: maximal ($a_{\max} = 0.5$) and minimal ($a_{\min} = 0.01$) learning rate, number of survivals ($N_{\text{surv}} = 23$), and percent of mutations ($N_{\text{mut}} = 0.005$). A large number of GA runs (more than 200), starting from different random origins were performed; the details of the procedure are described in a study by Mlinšek et al. [55]. The criterion for the selection of the best predictive model was the minimal sum of RMS values of the training and the internal test set, calculated in each iteration step, and a low number of selected variables (<15). Top 50 models were kept for the evaluation of variables, while the top-scoring model was proposed for a potential use in drug design.

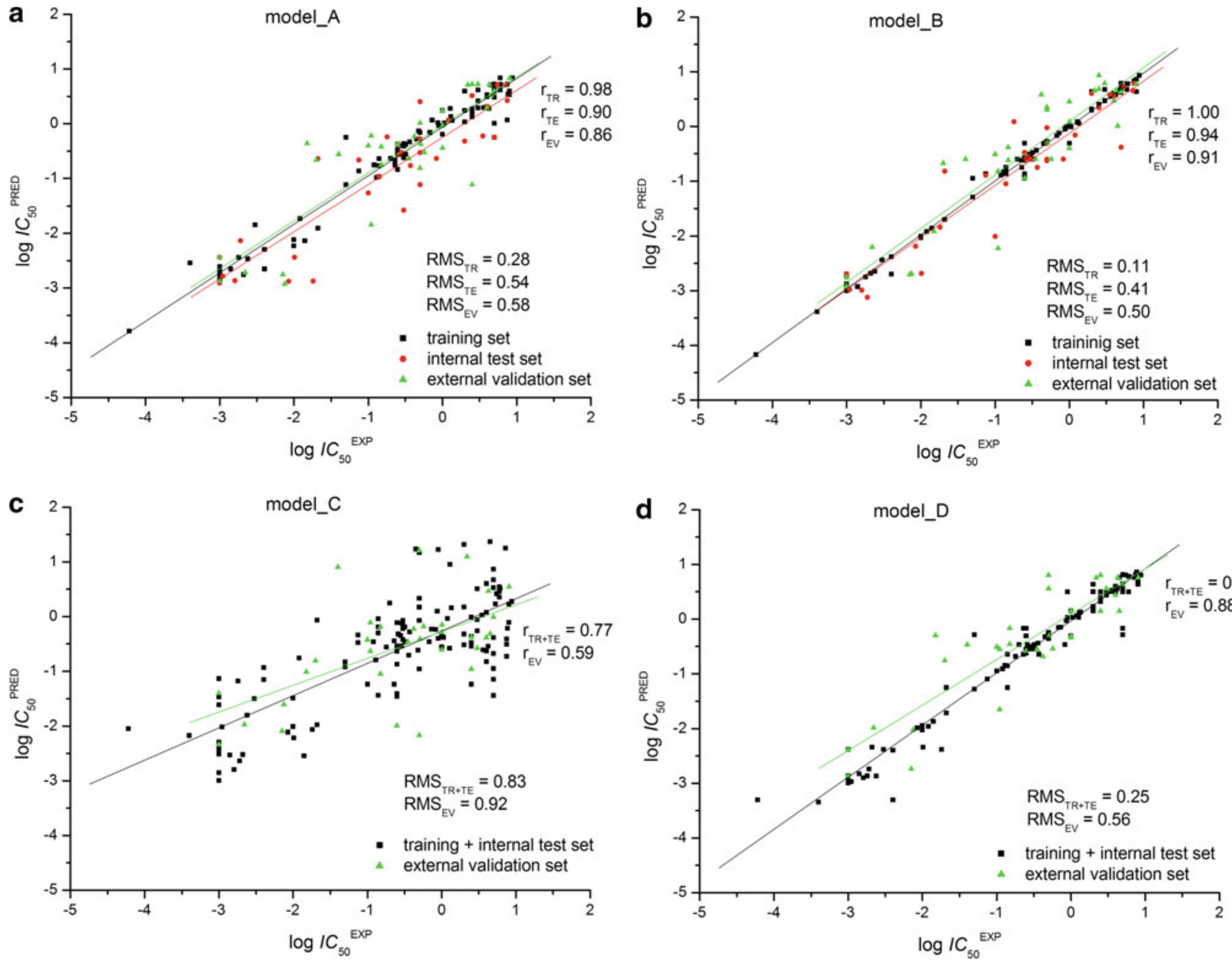

Fig. 3 Experimental versus predicted antiprion activity values (log IC_{50}) obtained by predictive models; **a** CP-ANN (model_**A**), **b** CP-ANN-GA (model_**B**), **c** MLR (model_**C**) and **d** CP-ANN-GA (model_**D**)

We have achieved a considerable reduction of variables from 87 to 9 MDs and reasonably low RMS errors for the training and the internal test set (model_**B**; $RMS_{TR} = 0.11$ and $RMS_{TE} = 0.41$). The parameters of the optimal predictive CP–ANN–GA model (Fig. 3b, model_**B**) are the following: network dimension 14 × 14, 400 learning epochs, maximal and minimal learning rates of 0.5 and 0.01, respectively. In the supplementary information (Table S2) the obtained RMS values for different CP–ANN and GA parameters are given.

The regression plots of the experimental versus predicted log IC_{50} obtained from the predictive model_**A** and model_**B** are shown in Fig. 3. Experimental and predicted IC_{50} values for all 158 active compounds from the model_**B** are given in Table 3.

All predictive models are validated with the external validation set (EV). Experimental versus predicted antiprion activity values for the external validation set obtained from the model_**A** and model_**B** are presented in Fig. 3. A concordance correlation coefficient (CCC) for the evaluation of the predictive ability of the models is calculated by Eq. 3 [56,57]. Any deviation of the results (predictions) from the regression line results in CCC value smaller than 1.

$$\hat{\rho}_c = \frac{2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{x}-\bar{y})^2} \qquad (3)$$

x corresponds to the experimental values, y corresponds to the predicted values, n is the number of compounds.

Obtained CCC values for the model_**A** are 0.92, 0.89, 0.86 for the training, internal test, and external validation set, respectively. Higher CCC values are obtained for the model_**B**; 0.95, 0.94, and 0.90 for the training, internal test, and external validation set, respectively. Furthermore, the external predictive power of models was verified with additional coefficients Q^2_{EV} and r^2_m [57–59]. The following values were obtained for $Q^2_{EV} = 0.81$ and 0.85, r^2_m (average) = 0.70 and 0.73, and $\Delta r^2_m = 0.07$ and 0.07 for the model_**A** and model **B**, respectively. Therefore, the implementation of the

Table 3 List of 158 active compounds with their experimental and predicted IC_{50} values and the literature source

No.	ID	Group	IC_{50}^{EXP} [μM]	IC_{50}^{PRED} [μM] model_**B**	IC_{50}^{PRED} [μM] MLR model	Reference
1TE	1	III	4.00	3.65	1.20	[33]
2^{TR}	2	III	4.00	3.65	4.01	[21,33]
3^{TR}	10	III	2.50	2.44	0.43	[26]
4^{TR}	12	III	7.50	6.88	0.19	[26]
5TE	13	III	0.50	0.24	0.11	[26]
6^{TR}	14	III	2.50	2.52	0.97	[26]
7^{TR}	15	III	5.00	4.85	3.37	[26]
8^{TR}	17	III	6.00	4.76	3.53	[27]
9^{TR}	19	III	3.00	4.76	3.23	[27]
10TE	20	III	3.50	3.83	1.06	[27]
11^{TR}	21	III	0.45	0.49	17.16	[27]
12TE	22	III	0.50	0.49	14.72	[27]
13^{TR}	23	III	0.90	0.95	16.18	[27]
14^{EV}	24	III	0.04	0.25	8.10	[27]
15^{TR}	25	III	0.01	0.01	0.18	[27]
16^{TR}	27	III	6.00	4.76	2.26	[27]
17^{TR}	29	III	0.003	0.004	0.03	[27]
18^{EV}	30	III	0.11	0.01	0.24	[27]
19^{EV}	33	III	0.008	0.002	0.02	[27]
20^{TR}	34	III	0.004	0.002	0.12	[27]
21TE	35	III	0.50	0.95	1.49	[27]
22^{TR}	36	III	8.00	7.23	0.79	[27]
23^{TR}	49	VI	0.30	0.31	0.84	[16,21,27,28,31–33,44]
24^{TR}	55	VI	5.00	4.33	7.42	[33]
25^{EV}	56	VI	4.00	2.93	0.41	[33]
26^{TR}	59	VI	3.00	2.93	0.55	[33]
27^{TR}	66	VI	2.00	4.33	1.50	[21,33]
28^{TR}	68	VI	8.00	4.33	1.68	[21]
29^{EV}	69	VI	5.00	4.33	0.98	[21]
30TE	85	VI	0.02	0.15	0.86	[35]
31^{TR}	86	VI	0.14	0.15	0.50	[35]
32^{EV}	87	VI	0.15	0.42	0.65	[35]
33^{EV}	88	VI	0.11	0.15	0.78	[35]
34TE	89	VI	0.25	0.34	0.78	[35]
35^{TR}	90	VI	0.32	0.34	0.35	[35]
36^{EV}	91	VI	0.51	0.50	0.39	[35]
37^{EV}	92	VI	1.01	1.22	0.71	[35]
38^{TR}	93	VI	0.48	0.50	0.24	[35]
39TE	94	VI	0.18	1.22	0.28	[35]
40 T^{R}	95	VI	4.24	3.83	0.29	[35]
41^{TR}	96	VI	0.90	0.98	0.40	[35]
42^{TR}	97	VI	1.28	1.22	0.26	[35]
43^{TR}	98	VI	0.29	0.31	0.76	[35]
44^{EV}	99	VI	0.10	0.31	0.38	[35]
45^{TR}	100	VI	1.06	0.98	0.42	[35]
46^{EV}	101	VI	0.42	3.83	0.60	[35]
47^{TR}	102	VI	0.36	0.37	0.20	[35]
48^{TR}	103	VI	0.13	0.15	0.16	[35]
49^{TR}	107	VI	1.00	0.49	0.28	[26]
50^{TR}	109	VI	0.25	0.49	0.23	[26]
51^{TR}	110	VI	2.50	2.46	0.06	[26]
52^{EV}	111	VI	2.50	8.65	0.11	[26]
53^{EV}	149	VI	0.32	0.40	0.24	[25]
54^{TR}	156	VI	0.10	0.25	0.24	[31,32]
55^{EV}	157	VI	0.02	0.21	0.16	[31]
56^{TR}	158	VI	0.23	0.18	0.65	[31]
57^{EV}	159	VI	0.15	0.24	0.09	[31]
58^{EV}	160	VI	0.57	0.25	0.66	[31]
59TE	161	VI	0.26	0.25	0.48	[31]
60TE	162	VI	0.37	0.18	0.53	[31]
61^{TR}	163	VI	0.14	0.18	0.91	[31]
62^{TR}	164	VI	0.30	0.30	0.38	[31]
63TE	165	VI	0.83	0.25	0.59	[31]
64^{TR}	166	VI	0.20	0.25	1.76	[31]
65^{EV}	167	VI	0.35	0.25	0.35	[31]
66TE	168	VI	0.08	0.13	0.33	[31]
67^{TR}	169	VI	0.12	0.13	0.35	[31]
68^{TR}	170	VI	0.23	0.24	0.06	[31]
69^{TR}	171	VI	0.29	0.28	0.29	[31]
70TE	172	VI	0.28	0.28	0.45	[31]
71^{TR}	173	VI	0.25	0.24	0.29	[31]
72^{TR}	174	VI	5.00	4.34	4.69	[32]
73^{TR}	175	VI	5.00	4.85	0.42	[31–33]
74TE	179	VI	2.50	2.20	0.34	[32]
75^{TR}	180	VI	2.00	2.00	0.31	[32]
76^{TR}	181	VI	5.00	6.13	0.21	[32]
77^{TR}	182	VI	7.50	6.13	0.31	[32]
78TE	183	VI	7.50	6.13	0.39	[32]
79^{EV}	185	VI	2.50	4.85	0.42	[32]
80^{EV}	186	VI	1.00	2.86	0.52	[32]
81^{TR}	187	VI	1.00	1.02	0.44	[32]
82^{EV}	188	VI	3.00	6.13	0.27	[32]
83^{EV}	189	VI	4.50	1.02	0.49	[32]
84^{TR}	191	IV	5.30	5.36	1.69	[16]

Table 3 continued

No.	ID	Group	IC_{50}^{EXP} [μM]	IC_{50}^{PRED} [μM] model_**B**	IC_{50}^{PRED} [μM] MLR model	Reference
85^{TE}	192	IV	5.50	5.36	2.62	[16]
86^{TR}	193	IV	6.10	6.04	3.25	[16]
87^{EV}	194	IV	4.30	4.56	2.93	[16]
88^{EV}	195	IV	2.20	4.56	12.43	[16]
89^{TR}	196	IV	4.50	4.56	23.34	[16]
90^{TE}	197	IV	7.20	4.56	17.81	[16]
91^{TR}	202	IV	8.80	8.65	1.89	[16]
92^{EV}	205	IV	8.10	5.97	3.51	[16]
93^{TR}	206	IV	6.00	6.04	2.86	[16]
94^{TR}	225	II	0.87	0.89	1.26	[13]
95^{TR}	226	II	3.60	3.54	0.25	[13]
96^{TR}	227	II	7.70	6.21	0.61	[13]
97^{TR}	230	II	0.68	0.68	0.35	[25]
98^{TR}	233	II	0.73	0.76	0.59	[25]
99^{TE}	234	II	1.20	0.69	0.92	[25]
100^{TR}	265	VII	0.0004	0.0002	0.007	[36]
101^{TE}	266	VII	0.0102	0.002	0.006	[36]
102^{TE}	267	VII	0.0016	0.001	0.002	[36]
103^{TR}	268	VII	0.0010	0.001	0.034	[36]
104^{TR}	269	VII	0.0010	0.001	0.004	[36]
105^{TR}	270	VII	0.0010	0.001	0.004	[36]
106^{TR}	271	VII	0.0010	0.001	0.001	[36]
107^{EV}	272	VII	0.0071	0.002	0.008	[36]
108^{TR}	273	VII	0.0024	0.002	0.016	[36]
109^{EV}	274	VII	0.0022	0.006	0.011	[36]
110^{TE}	275	VII	0.0085	0.006	0.008	[36]
111^{TR}	276	VII	0.0018	0.002	0.066	[36]
112^{TE}	277	VII	0.0010	0.002	0.073	[36]
113^{TR}	278	VII	0.0021	0.002	0.003	[36]
114^{EV}	279	VII	0.0010	0.002	0.040	[36]
115^{TE}	280	VII	0.0181	0.014	0.009	[36]
116^{TR}	281	VII	0.0014	0.001	0.003	[36]
117^{TR}	282	VII	0.0010	0.001	0.005	[36]
118^{EV}	283	VII	0.0010	0.001	0.005	[36]
119^{TR}	284	VII	0.0210	0.020	0.011	[36]
120^{TE}	285	VII	0.0011	0.001	0.010	[36]
121^{TR}	286	VII	0.0010	0.001	0.005	[36]
122^{TE}	287	VII	0.0019	0.001	0.002	[36]
123^{TR}	288	I	0.0140	0.014	0.003	[15,23,28]
124^{TR}	295	I	0.05	0.11	0.12	[24]
125^{TR}	296	I	0.25	0.11	0.04	[24]
126^{TE}	302	I	5.00	0.42	0.07	[24]
127^{EV}	305	I	0.25	0.11	0.01	[24]
128^{TR}	311	I	5.00	4.95	0.11	[24]
129^{TR}	316	I	0.25	0.27	0.04	[24]
130^{EV}	336	I	0.50	2.26	0.007	[24]
131^{TR}	337	I	0.25	0.14	0.14	[24]
132^{TR}	338	I	0.08	0.14	0.46	[24]
133^{TR}	340	I	2.50	2.54	0.29	[24]
134^{TE}	371	I	0.14	0.09	0.04	[22]
135^{TR}	372	I	0.05	0.05	0.15	[22]
136^{TR}	373	I	0.14	0.14	0.65	[22]
137^{TR}	376	IX	1.35	1.14	1.45	[7]
138^{TR}	379	IX	1.30	1.20	8.99	[20]
139^{TR}	464	IX	5.00	4.17	0.04	[28]
140^{TR}	467	IX	0.30	0.25	0.65	[28]
141^{TE}	468	IX	0.30	0.25	0.75	[28]
142^{TR}	471	IX	3.00	2.96	0.88	[28]
143^{TR}	476	IX	4.00	4.00	0.15	[28]
144^{TE}	477	IX	2.00	4.00	0.48	[28]
145^{TR}	478	IX	4.00	4.00	0.24	[28]
146^{TE}	479	IX	0.50	0.25	0.86	[28]
147^{EV}	480	IX	0.015	0.012	0.098	[28]
148^{TR}	481	IX	0.004	0.004	0.071	[38]
149^{TR}	483	IX	0.0001	0.000	0.009	[39]
150^{TE}	484	IX	0.10	0.01	0.06	[39]
151^{TR}	485	IX	0.01	0.01	0.01	[39]
152^{TR}	486	IX	0.001	0.001	0.024	[39]
153^{TR}	487	IX	0.001	0.002	0.003	[39]
154^{TE}	488	IX	0.001	0.002	0.001	[39]
155^{TR}	489	IX	0.01	0.01	0.03	[39]
156^{EV}	503	IX	0.50	2.02	16.13	[32]
157^{TR}	506	IX	0.50	0.54	2.16	[20]
158^{TR}	507	IX	2.00	2.07	20.98	[20]

TR training set; *TE* internal test set; *EV* external validation set

GA affords a more accurate and robust model for antiprion prediction.

For the purpose of comparison of nonlinear ANN method with a linear one, a multiple linear regression (MLR) method was applied. The MLR model was developed using the same data set as described previously for the CP–ANN models and the same set of input vectors (158 compounds with 87 MDs and the corresponding log IC_{50} values), taking into account that for the MLR modeling the internal division into the training and test set is not needed. Thus, we have used the combined TR and TE set for the linear model construction with QSARINS software [51], while EV set was used for the model validation. The internal validation of the models was performed with leave-one-out cross-validation of the training data (TR & TE). The best MLR model (model_**C**) was obtained with 7 variables and the following six parameters reflect model predictive performance: $Q^2_{\mathrm{LOO}} = 0.53$, $Q^2_{\mathrm{EV}} = 0.49$, $\mathrm{CCC_{TR}} = 0.74$, $\mathrm{CCC_{EV}} = 0.58$, r^2_{m} (average) = 0.28 and $\Delta r^2_{\mathrm{m}} = 0.04$. In Fig. 3c, experimental versus predicted log IC_{50} for the MLR model (model_**C**) are presented, and the predictions of model_**C** are given in Table 3.

We have repeated the training of the CP–ANN–GA models with a merged set of the training and the internal test set compounds (TR & TE), using the same network parameters as selected for the previously developed optimal model (model_**B**). The resulting new model (model_**D**; Fig. 3d) was tested for its predictive ability with the EV set.

As can be seen from Fig. 3, the nonlinear modeling method affords a significantly better model performance than the linear one, which implies a complex biological nature of antiprion activity.

Interpretation of selected variables

Reduction of the descriptors space by selecting the influential variables not only increases the predictive ability and robustness of models, but also provides valuable information about structural details with important influence on the biological activity. In our study, the best predictive model was obtained with nine variables (MDs) including all five types of structural descriptors (see Table 4). Furthermore, to gain quantitative information about the most influential variables, we have determined the percentage of occurrences of each descriptor in top 50 CP–ANN–GA models ranked by the following fitness criteria: $\mathrm{RMS_{TE}} \leq 0.45$, $\mathrm{RMS_{VA}} \leq 0.7$, and number of selected MDs ≤ 15. The top ten descriptors are listed in Table 4 (consensus MD set). On the basis of these 10 MDs the predictive CP–ANN model was developed, but its performance ability did not exceed the best CP–ANN–GA model (model_**B**).

The XY shadow/XY rectangle is the most frequently selected descriptor, presented in 80 % of the top 50 models, followed by maximal bond order of a C atom and total dipole of the molecule, 68 and 38 % of occurrence, respectively. The rest of descriptors have much lower percentage of occurrences as majority of the structural descriptors from the initial set are actually grouped into different clusters, each of them describing certain feature, therefore all descriptors of the same cluster reflect the same structural feature. Because of that, the overlapping of solely certain descriptors among models is low, but if we compare the structural features, we find higher consensus. The same trend was noticed also within the 85 MD set for the classification model and the 87 MD set for the predictive model, where we found in both sets the 24 same MDs, but the rest of descriptors are analogic, describing similar features, the only exceptions are the 9 MDs relating to nitrogen atom features.

In particular, for the investigated set of compounds, four features influence significantly the prediction of antiprion activity (in this study represented with value of IC_{50}). Firstly, among the constitutional descriptors, which depend on the atomic constitution of the chemical structure, types of bonds, the presence and type of rings in molecule, only the descriptor regarding to the number of nitrogen atoms was selected. Therefore, the variables associated with the N-atoms (number, bond order, and electron–electron repulsion), and maximal exchange energy for a C–N bond suggest that the presence of nitrogen atoms in the structure positively influences the activity. This is further supported with the comparison of active and inactive compounds; the nitrogen atom is not present in the majority of inactive compounds. Secondly, selected descriptors describing the complementary information content index and XY shadow/XY rectangle reflect the influence of the shape and the size of the molecule for the prion-binding affinity, indicating a possibility to interact with more than one binding site. Thirdly, descriptors accounting to terms of charged solvent-accessible atomic surface area and dipole of the molecule are indicative for the role of charge distribution. They especially highlight the influence of the polarity of molecules on the activity. Fourthly, the importance of the capability of forming hydrogen and covalent bonds, and electrostatic interactions are indicated by several descriptors (Kier & Hall valence connectivity index, hydrogen-bonding acceptor ability of the molecule, coulombic interaction, exchange energy and bond order among C, H, and N atoms). Hence, considering fundamental characteristics of the protein–protein interface hot spots as relatively planar and hydrophobic regions [12], the above selected descriptors account for the features that affect protein–protein interactions. As can be seen from Fig. 4, there are two sites in the prion structure that are highly influenced by the conformational changes of native and pathogenic mutants [10,11].

Compounds of Perrier et al. [15] ID 387–398 and May et al. [16] ID 191–214 were designed according to docking studies of Perrier proposed binding site [15] and are probably dealing with stabilization of β_2–α_2 loop region. Compounds of Hosokawa-Muto et al. [37] ID 376, 399–452 were designed according to docking studies of Kuwata et al. [7] proposed binding site and are probably dealing with stabilization of α_2–α_3 interhelical region. Compounds studied by Bongarazone et al. [25] ID 229–240 and Tran et al. [13] ID 215–228 are considerably longer bivalent structures and probably bind to both binding sites and stabilize α_2–α_3 interhelical and β_2–α_2 loop regions. Our predictive model considers all types of compounds mentioned, and the selected descriptors indicate the interactions of small molecules with one or several binding hot spots, thus preventing protein–protein interactions.

Table 4 List of the most influential variables (MDs) of the best CP–ANN–GA model (model_**B**) and a consensus MD set (from the analysis of the top 50 CP–ANN–GA model)

Type of MD	MD set of the best CP–ANN–GA model	Consensus MD set
Constitutional	Number of N atoms	
Topological	Kier & Hall index	
	Average complementary information content	
Geometrical		XY shadow/XY rectangle
Electrostatic	Surface weighted charged partial positive-charged surface area	Hydrogen-bonding acceptor ability of the molecule
Quantum -chemical	Surface weighted charged partial negative-charged surface area	Total dipole of the molecule
	Average bond order of a N atom	Total point-charge complementary of the molecular dipole
	Maximal bond order of a C atom	Total charge weighted partial positively charged surface area
	Minimal electron–electron repulsion energy for a N atom	Maximal PI–PI bond order
	Maximal coulombic interaction for a C–H bond	Maximal bond order of a C atom
		Minimal exchange energy for a C–C bond
		Maximal exchange energy for a C–N bond
		Minimal coulombic interaction for a C–H bond

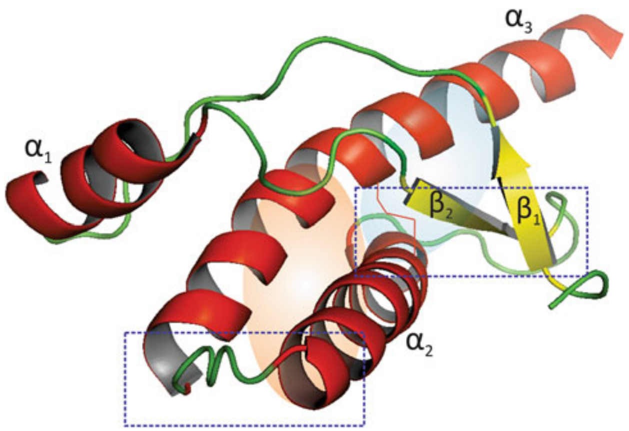

Fig. 4 Structure of native human prion (PDB ID: 1QM1) with representation of the conformation and the binding hot spots. Secondary structure elements: helices $\alpha_1-\alpha_3$, (residues 144–154, 173–194, and 200–228), β-sheets (residues 128–131 and 161–164). Disulfide-bond (C179–C214) connects $\alpha_2-\alpha_3$. *Dashed line* represents conformation hot spots according to the comparison with human pathogenic mutants V210I and Q212P (PDB ID: 2LEJ, 2KUN); $\alpha_2-\alpha_3$ interhelical and $\beta_2-\alpha_2$ loop regions. Proposed binding sites of ligands are in blue (Perrier et al. [15]) and orange (Kuwata et al. [7])

Applicability domain of predictive models

The applicability domain (AD) assessment is used to evaluate the reliability of prediction of objects [60]. In our models, reliable predictions of property can be achieved for compounds, which match with the physico-chemical and structural space restrictions of the models. The applicability domain of the CP–ANN predictive models was assessed with the Euclidean distance (ED)-based approach. In this approach the so-called minimum ED space (MEDS) defined by a number of compounds in the training and the internal test set together with standardized residuals describe a domain of model applicability [61]. For the MLR predictive model the leverage-based AD estimation was performed [53] giving us a plot of the leverage values as a function of standardized residuals (Williams plot).

The boundaries of the AD of CP–ANN prediction models (model_**B** and model_**D**) are defined by the threshold value ED_{crit}, which is defined by the maximal ED value of the training or the internal test set compound. ED value is calculated for each new compound tested by the CP–ANN models; all compounds with ED values higher than ED_{crit} are considered as structural outliers. For the MLR model (model_**C**) the threshold is defined by the hat value ($h*$), and all compounds with $h > h*$ are determined as being out of the model domain. In order to also asses the response outliers, the boundaries regarding standardized residuals are set between $\pm 3\sigma$ in both methods. The structural and response outliers are presented in Fig. 5 for two predictive models (model_**B**, model_**C**).

According to Fig. 5, both methods give comparable results as the majority of compounds are within the models AD. The leverage approach for MLR model is a bit less restrictive

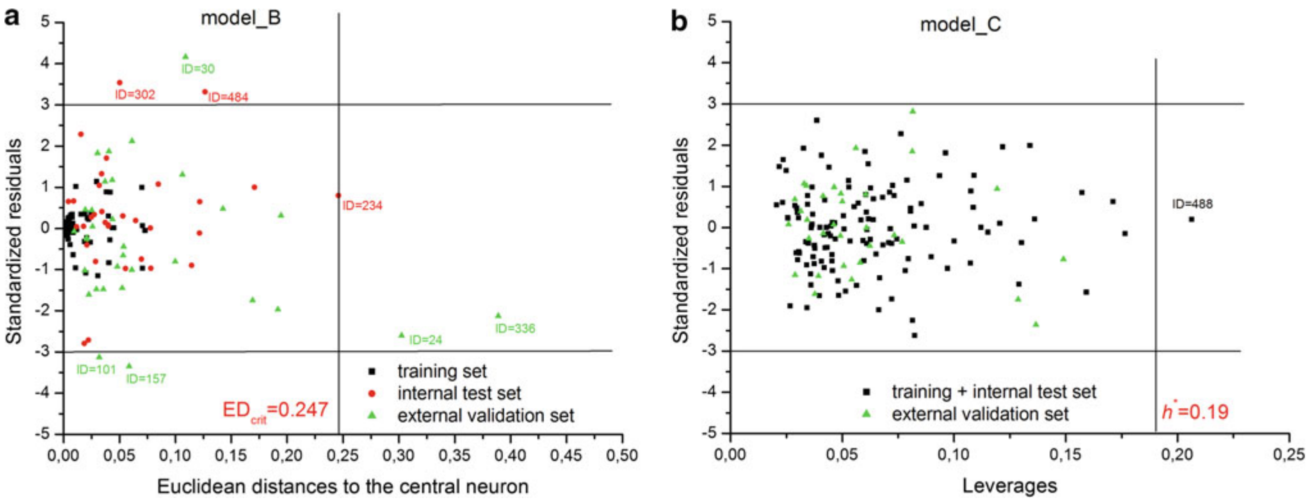

Fig. 5 Applicability assessment of predictive models; **a** CP–ANN–GA (model_**B**), and **b** MLR (model_**C**)

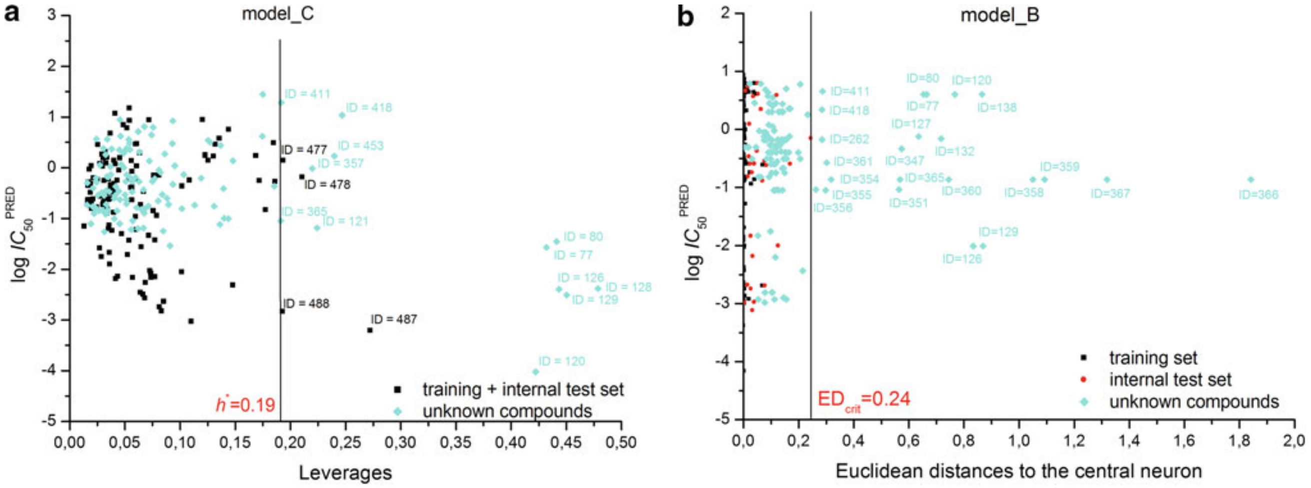

Fig. 6 Insubria and MEDS graphs of predictive models: **a** MLR (model_**C**) and **b** CP–ANN–GA (model_**B**)

compared to the ED approach for CP–ANN models regarding structural outliers. Only one compound (ID = 488) in MLR model (model_**C**) is out of AD, while in the case of CP–ANN models (model_**B**), 2 compounds (ID = 24, 336) are out of the model domain regarding structural outliers and 5 compounds (ID = 30, 111, 157, 302, and 484) regarding response outliers. The compounds with ID = 24 and 336 are considered as structural outliers in the CP–ANN predictive model_**B**. Indeed, their structures differ considerably from other compounds in the data set (see Table S1 in supplementary information).

Compounds with missing IC_{50} values for antiprion activity

The validated predictive models were used for the prediction of IC_{50} values of 120 compounds, which were experimentally determined as active ones, but their exact IC_{50} values were not yet estimated. The main objective is whether the model is applicable for compounds with unknown properties, which were not used for model development. For linear models, well known Insubria graphs [53] are used for the determination of model applicability for compounds with unknown properties. Graphical depiction of leverages as a function of predicted values is presented in Fig. 6a. For nonlinear CP–ANN predictive models we propose a new graphical approach on the basis of minimum ED space (MEDS, [61]), which gives an assessment of the applicability of CP–ANN predictive models regarding chemical structure only. In this way, the constructed MEDS graph is comparable to the Insubria graph. ED of all compounds (training, internal test, and the set of unknown compounds) are plotted as a function of predicted log IC_{50} values obtained from the CP–ANN–GA model (Fig. 6b). The boundary that determines the model applicability is defined by the maximal value of ED of the training and the internal test set. In the case of

MLR model the boundary is defined by the hat value (h^*). Therefore, the compounds falling out of the AD could not be considered.

As shown in Fig. 6a, Insubria graph for MLR model (model_C) shows that most of the compounds with unknown properties (92 %) are placed inside the model domain, defined with the value of $h* = 0.19$, and ten compounds are out of the model domain. In the case of the CP–ANN–GA model (model_B, Fig. 6b) 81 % of compounds are placed inside the boundaries, set with the value of EDcrit = 0.24. As demonstrated in Fig. 6a, b, 10 and 23 compounds from the model_C and model_B, respectively, are placed outside their restricted boundaries, and therefore, their predictions are considered unreliable. In both models, six compounds (ID = 77, 80, 120, 126, 129, 418) determined as outliers were the same. Structurally, these compounds have different chemical structures from those with which the model was build. For example, the Congo red derivatives (derivative group I) from a study by Rudyk et al. [23] (ID = 343–361, 365–367) were not included in the predictive model development. Therefore, these are compounds with the double nitrogen bond in linker sequence, which probably possess specific features than other compounds. Furthermore, it shows that compounds having two NO_2 functional group and two S-atoms (ID = 120, 126, 127, 138) or two CF_3 functional groups (ID = 132) in the same structure also indicate special features, as are placed outside. Obviously, the selected structural descriptors do not depict similarities noticed by visual comparison of these structures.

The results from the 120 compounds with unknown IC_{50} obtained by the best predictive CP–ANN–GA model (model_**B**) are listed in Table S3 and their 2D structures are shown in Table S1 (both in supplementary information). A rough range of activity for 120 compounds was estimated from the available literature data. The IC_{50} values predicted by the model_B coincide perfectly with the rough estimation of activity (see Table S3), 94 % of compounds are correctly classified.

Conclusions

The structure–activity models for the categorization of active and inactive small organic molecules and for the prediction of their inhibition potency for prion aggregations are constructed and available for further in silico modeling. The models are following the OECD principles, they are properly validated and show a good predictive ability. The best-performing models are nonlinear, based on the counter-propagation artificial neural networks, using molecular descriptors to input chemical structures of compounds. The comparison with linear predictive models was also carried out. The computational models developed could be effectively applied in the screening processes of further drug design of prion disease therapeutics. For each novel compound of interest, not only the prediction of inhibition constants could be given, but also the reliability of prediction would be evaluated, as all models have their applicability domain defined. In the process of model developing, the genetic algorithm was used, which highly increased the model performance by reducing the hundreds of descriptors of chemical structure to the set of only few most influential ones. Furthermore, the top ten selected variables were collected from the best 50 CP–ANN–GA models and were studied in more detail, as the influential variables are indicating chemical structure features affecting the antiprion activity. On the basis of the analysis of the set of investigated compounds, we can conclude that the features regarding molecular shape, charge distribution, capability of forming bonds, and the presence of nitrogen atoms are significantly influencing the antiprion activity of small organic compounds. These selected most influential features of small molecules are in general also important for maintaining protein–protein interactions; hence, the potential of the investigated compounds to reduce prion aggregation by competing with the prion–prion interactions is considered. Taking into account the very interesting finding in the joint experimental and computational study of Hosokawa-Muto et al. [37], which had shown that experimentally and computationally determined binding affinities of tested compounds to prion protein not always correlate with their antiprion activity, it seems that using only docking approaches in initial drug design screening is insufficient. Therefore, we propose to merge our developed models with molecular docking studies to result in an efficient strategy for future antiprion drug discovery.

Acknowledgments The financial support from the Grant of Transregional Network for Innovation and Technology Transfer to Improve Health Care (TRANS2CARE) and Slovenian research program P1-017 is gratefully acknowledged. The authors are grateful to Dr. Amrita Roy Choudhury for language corrections.

References

1. Prusiner SB (1998) Prions. Proc Natl Acad Sci USA 95: 13363–13383
2. van Rheede T, Smolenaars MMW, Madsen O, de Jong WW (2003) Molecular evolution of the mammalian prion protein. Mol Biol Evol 20:111–121. doi:10.1098/rspb.2005.3259
3. Taylor DR, Hooper NM (2006) The prion protein and lipid rafts. Mol Membr Biol 23:89–99. doi:10.1080/09687860500449994
4. Basakov IV, Legname G, Baldwin MA, Prusiner SB, Cohen FE (2002) Pathway complexity of prion protein assembly into amyloid. J Biol Chem 227(24):21140–21148. doi:10.1074/jbc.M111402200
5. Govaerts C, Wille H, Prusiner SB, Cohen FE (2004) Evidence for assembly of prions with left-handed β-helices into trimers. Proc Natl Acad Sci USA 101:8342–8347. doi:10.1073/pnas.0402254101
6. Caughey B, Caughey WS, Kocisko DA, Lee KS, Silveira JR, Morrey JD (2006) Prions and transmissible spongiform encephalopa-

thy (TSE) chemotherapeutics: A common mechanism for anti-TSE compounds? Acc Chem Res 39:646–653. doi:10.1021/ar050068p

7. Kuwata K, Nishida N, Matsumoto T, Kamatari YO, Hosokawa-Muto J, Kodama K, Nakamura HK, Kimura K, Kawasaki M, Takakura Y, Shirabe S, Takata J, Kataoka Y, Katamine S (2007) Hot spots in prion protein for pathogenic conversion. Proc Natl Acad Sci USA 104(29):11921–11926. doi:10.1073/pnas.0702671104
8. Tribouillard D, Gug F, Galons H, Bach S, Saupe SJ, Blondel M (2007) Antiprion drugs as chemical tools to uncover mechanisms of prion propagation. Prion 1/1:48–52 PMC2633708
9. Pamplona R, Naudi A, Gavin R, Pastrana MA, Sajnani G, Ilieva EV, del Rio JA, Portero-Otin M, Ferrer I, Requena JR (2008) Incrised oxidation, glycoxidation, and lipoxidation of brain proteins in prion disease. Free Radical Biol Med 45:1159–1166. doi:10.1016/j.freeradbiomed.2008.07.009
10. Ilc G, Giachin G, Jaremko M, Jaremko L, Benetti F, Plavec J, Zhukov I, Legname G (2010) NMR structure of the human prion protein with the pathological Q212P mutation reveals unique structural features. PLoS One 5:e11715. doi:10.1371/journal.pone.0011715
11. Biljan I, Ilc G, Giachin G, Raspadori A, Zhukov I, Plavec J, Legname G (2011) Toward the molecular basis of inherited prion diseases: NMR structure of the human prion protein with V210I mutation. J Mol Biol 412:660–673. doi:10.1016/j.jmb.2011.07.067
12. Xu Y, Shi J, Yamamoto N, Moss JA, Vogt PK, Janda KD (2006) A credit-card library approach for disrupting protein–protein interactions. Bioorg Med Chem 14:2660–2673. doi:10.1016/j.bmc.2005.11.052
13. Tran HNA, Bongarzone S, Carloni P, Legname G, Bolognesi ML (2010) Synthesis and evaluation of library of 2.5-bisdiamino-benzoquinone derivatives as probes to modulate protein–protein interactions in prions. Bioorg Med Chem Lett 20:1866–1868. doi:10.1016/j.bmcl.2010.01.149
14. Kranjc A, Bongarzone S, Rossetti G, Biarnés X, Cavalli A, Bolognesi ML, Roberti M, Legname G, Carloni P (2009) Docking ligands on protein surfaces: The case study of prion protein. J Chem Theory Comput 5:2565–2573. doi:10.1021/ct900257t
15. Perrier V, Wallace AC, Kaneko K, Safar J, Prusiner SB, Cohen FE (2000) Mimicking dominant negative inhibition of prion replication through structure-based drug design. Proc Natl Acad Sci USA 97:6073–6078. doi:10.1073/pnas.97.11.6073
16. May BCH, Zorn JA, Witkop J, Sherrill J, Wallace A, Legname G, Prusiner SB, Cohen FE (2007) Structure–activity relationship study of prion inhibition by 2-aminopyridine-3.5-dicarbonitrile-based compounds: parallel synthesis, bioactivity and in vitro pharmacokinetics. J Med Chem 50:65–73. doi:10.1021/jm061045z
17. Nicoll AJ, Trevitt CR, Risse E, Quarterman E, Ibarra AA, Wright C, Jackson GS, Sessions RB, Farrow M, Waltho JP, Clarke AR, Collinge J (2010) Pharmacological chaperone for the structured domain of human prion protein. Proc Natl Acad Sci USA 107:17610–17615. doi:10.1073/pnas.1009062107
18. Appleby BS, Lyketsos CG (2011) Rapidly progressive dementias and the treatment of human prion diseases. Expert Opin Pharmacother 12:1–12. doi:10.1517/14656566.2010.514903
19. Zerr I (2009) Therapeutic trials in human transmissible spongiform encephalo-pathies: recent advances and problems to address. Infect Disord Drug Target 9:92–99. doi:10.2174/1871526510909010092
20. Doh-ura K, Iwaki T, Caughey B (2000) Lysosomotropic agents and cysteine protease inhibitors inhibit scrapie-associated prion protein accumulation. J Virol 74:4894–4897. doi:10.1128/JVI.74.10.4894-4897
21. Kocisko DA, Baron GS, Rubenstein R, Chen J, Kuizon S, Caughey B (2003) New inhibitors of scrapie-associated prion protein formation in a library of 2.000 drugs and natural products. J Virol 77:10288–120294. doi:10.1128/JVI.77.19.10288-10294.2003
22. Demaimay R, Harper J, Gordon H, Weaver D, Chesebro B, Caughey B (1998) Structural aspects of Congo red as an inhibitor of protease-resistant prion protein formation. J Neurochem 71:2534–2541. doi:10.1046/j.1471-4159.1998.71062534.x
23. Rudyk H, Vasiljevic S, Hennion RM, Birkett CR, Hope J, Gilbert IH (2000) SIeeining Congo redand its analogues for their ability to prevent the formation of PrP-res in sIapie-infected cells. J Gen Virol 81:1155–1164
24. Sellarajah S, Lekishvili T, Bowring C, Thompsett AR, Rudyk H, Birkett CR, Brown DR, Gilbert IH (2004) Synthesis of analogues of Congo red and evaluation of their anti-prion activity. J Med Chem 47:5515–5534. doi:10.1021/jm049922t
25. Bongarzone S, Ai Tran HN (2010) Parallel synthesis, evaluation and preliminary structure–activity relationship of 2.5-diamino-1.4-benzoquinones as a novel class of bivalent anti-prion compound. J Med Chem 53:8197–8201. doi:10.1021/jm100882t
26. Cope H, Mutter R, Heal W, Pascoe C, Brown P, Pratt S, Chen B (2006) Synthesis and SAR study of acridine, 2-methylquinoline and 2-phenylquinazoline analogues as anti-prion agents. Eur J Med Chem 41:1124–1143. doi:10.1016/j.ejmech.2006.05.002
27. Kubo MI, Doh-ura K, Ishikawa K, Kawatake S, Sasaki K, Kira J, Ohta S, Iwaki T (2004) Quinoline derivatives are therapeutic candidates for transmissible spongiform encephalopathies. J Virol 78:1281–1288. doi:10.1128/JVI.78.3.1281-1288.2004
28. Doh-ura K, Tamura K, Karube Y, Naito M, Tsuruo T, Kataoka Y (2007) Chelating compound, chrysoidine, is more effective in both antiprion activity and brain endothelial permeability than quinacrine. Cell Mol Neurobiol 27:303–315. doi:10.1007/s10571-006-9122-0
29. Bolognesi ML, Ai Tran HN (2010) Discovery of class of diketopiperazines as antiprion compounds. Chem Med Chem 5:1324–1334. doi:10.1002/cmdc.201000133
30. Csuk R, Barthel A, Raschke C, Kluge R, Ströhl D, Trieschmann L, Böhm G (2009) Synthesis of monomeric and dimeric aIidine compounds as potential therapeutics in alzheimer and prion diseases. Arch Pharm Chem Life Sci 342:699–709. doi:10.1002/ardp.200900065
31. Dollinger S, Löber S, Klingenstein R, Korth C, Gmeiner P (2006) A chimeric ligand approach leading to potent antiprion active acridine derivatives: design, synthesis and biological investigations. J Med Chem 49:6591–6595. doi:10.1021/jm060773j
32. Klingenstein R, Löber S, Kujala P, Godsave S, Leliveld SR, Gmeiner P, Peters PJ, Korth C (2006) Tricyclic antidepressants, quinacrine and a novel, synthetic chimera thereof clear prions by destabilizing detergent-resistant membrane compartments. J Neurochem 98:748–759. doi:10.1111/j.1471-4159.2006.03889.x
33. Korth C, May BC, Cohen FE, Prusiner SB (2001) Acridine and phenothiazine derivatives as pharmacotherapeutics for prion disease. PNAS 98:9836–9841. doi:10.1073/pnas.161274798
34. May BC, Fafarman AT, Hong SB, Rogers M, Deady LW, Prusiner SB et al (2003) Protein inhibition of scrapie prion replication in cultured cells by bis-acridines. Proc Natl Acad Sci USA 100:3416–3421. doi:10.1073/pnas.2627988100
35. Thi HTN, Lee CY, Teruya K, Ong WY, Doh-ura K, Go ML (2008) Antiprion activity of functionalized 9-aminoacridines related to quinacrine. J Bioorg Med Chem 16:6737–6746. doi:10.1016/j.bmc.2008.05.060
36. Ishikawa K, Kudo Y, Nishida N, Suemoto T, Sawada T, Iwaki T, Doh-ura K (2006) Sterylbenzoazole derivatives for imaging of prion plaques and treatment of transmissible spongiform encephalopathies. J Neurochem 99:198–205. doi:10.1111/j.1471-4159.2006.04035.x
37. Hosokawa-Muto J, Kamatari YO, Nakamura HK, Kuwata K (2009) Variety of antiprion compounds discovered through an in silico sIeen based on cellular-form prion protein structure: correlation

between antiprion activity and binding affinity. Antimicrob Agents Chemother 53:765–771. doi:10.1128/AAC.01112-08

38. Ishikawa K, Doh-ura K, Kudo Y, Nishida N, Murakami-Kubo I, Ando Y, Sawada T, Iwaki T (2004) Amyloid imaging probes are useful for detection of prion plaques and treatment of transmissible spongiform encephalopathies. J Gen Virol 85:1785–1790. doi:10.1099/vir.0.19754-0
39. Kawasaki Y, Kawagoe K, Chen CJ, Teruya K, Sakasegawa Y, Doh-ura K (2007) Orally administered amyloidophilic compound is effective in prolonging the incubation periods of animals cerebrally infected with prion diseases in a prion strain-dependent manner. J Virol 81:12889–12898. doi:10.1128/JVI.01563-07
40. Light DW, Warburton R (2011) Demythologizing the high costs of, pharmaceutical research. BioSoceties 6:1–17. doi:10.1057/biosoc.2010.40
41. OECD (2007) Guideance document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models. ENV/JM/MONO(2007) 2, www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en. Accessed 16 Aug 2013
42. Todeschini R, Consonni V, Gramatica P (2009) Chemometrics in QSAR. In: Brown S, Tauler R, Walczak R (eds) Comprehensive chemometrics, vol 4. Elsevier, Oxford, pp 129–172
43. Tareq M, Khan H (2012) Recent trends on QSAR in the pharmaceutical perceptions. University of Illinois, Chicago, USA. doi:10.2174/97816080537971120101
44. Ryou C, Legname G, Peretz D, Craig JC, Baldwin MA, Prusiner SB (2003) Differential inhibition of prion propagation by enantiomers of Quinacrine. Lab Invest 83:837–843. doi:10.1097/01.LAB.0000074919.08232.A2
45. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 107:3902–3909. doi:10.1021/ja00299a024
46. Katritzky AR, Lobanov VS, Karelson M (1994) Codessa 2.0, Comprehensive descriptors for structural and statistical analysis. University of Florida, USA
47. Massart DL, Vandeginste BGM, Budgens LM, Dejong S, Lewi PJ, Smeyers-verbeke J (1997) Handbook of chemometrics and qualimetrics: Part A. Elsevire Science, Amsterdam
48. Zupan J, Novič M, Ruisánchez I (1997) Kohonen and counterpropagation artificial neural networks in analytical chemistry: tutorial. Chemometr Intell Lab Syst 38:1–23
49. Novič M, Zupan J (1995) Investigation of infrared spectra-structure correlation using Kohonen and counter-propagation neural network. J Chem Inf Comput Sci 35:454–466. doi:10.1021/ci00025a013
50. Leardi R (2001) Genetic algorithms in chemometrics and chemistry: a review. J. Chemom 15:559–569
51. Chirico N, Papa E, Kovarich S, Cassani S, Gramatica P (2012) QSARINS, software for QSAR MLR model development and validation. QSAR Res Unit in Environ Chem and Ecotox, DiSTA, University of Insubria, Varese, Italy. http://www.qsar.it.[CCC] Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268. doi:10.2307/2532051
52. Gramatica P, Pilutti P, Papa E (2004) Validated QSAR prediction of OH tropospheric degradability: splitting into training-test set and consensus modeling. J Chem Inf Comp Sci 44:1794–1802. doi:10.1021/ci049923u
53. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701. doi:10.1002/qsar.200610151
54. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 6:412–424. doi:10.1093/bioinformatics/16.5.412
55. Mlinšek G, Novič M, Hodoscek M, Šolmajer T (2001) Prediction of enzyme binding:Human thrombin inhibition study on quantum chemical and artificial intelligence methods based on X-ray structures. J Chem Inf Comput Sci 41:1286–1294. doi:10.1021/ci000162e
56. Chirico N, Gramatica P (2011) Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51:2320–2335. doi:10.1021/ci200211n
57. Chirico N, Gramatica P (2012) Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J Chem Inf Model 52:2044–2058. doi:10.1021/ci300084j
58. Roy K, Mitra I (2012) On the use of the metric rm^2 as an effective tool for validation of QSAR models in computational drug design and predictive toxicology. Med Chem 12:419–504. doi:10.2174/138955712800493861
59. Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q2 parameter for QSAR validation. J Chem Inf Model 49:1669–1678. doi:10.1021/ci900115y
60. Eriksson L, Jaworska J, Worth AP, Cronin MTD, 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
61. Minovski N, Župerl Š, Drgan V, Novič M (2013) Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum euclidean distance space analysis: a case study. Anal Chim Acta 759:28–42. doi:10.1016/j.aca.2012.11.002