Chemometrics approach for the prediction of structure–activity relationship for membrane transporter bilitranslocase

R. Martinčič, K. Venko, Š. Župerl and M. Novič

Laboratory of Chemometrics, National Institute of Chemistry, Ljubljana, Slovenia

(Received 6 June 2014; in final form 28 July 2014)

Membrane transport proteins are essential for cellular uptake of numerous salts, nutrients and drugs. Bilitranslocase is a transporter, specific for water-soluble organic anions, and is the only known carrier of nucleotides and nucleotide-like compounds. Experimental data of bilitranslocase ligand specificity for 120 compounds were used to construct classification models using counter-propagation artificial neural networks (CP-ANNs) and support vector machines (SVMs). A subset of active compounds with experimentally determined transport rates was used to build predictive QSAR models for estimation of transport rates of unknown compounds. Several modelling methods and techniques were applied, i.e. CP-ANN, genetic algorithm, self-organizing mapping and multiple linear regression method. The best predictions were achieved using CP-ANN coupled with a genetic algorithm, with the external validation parameter $Q^2_{\rm{ext}}$ of 0.96. The applicability domains of the models were defined to determine the chemical space in which reliable predictions can be obtained. The models were applied for the estimation of bilitranslocase transport activity for two sets of pharmaceutically interesting compounds, antioxidants and antiprions. We found that the relative planarity and a high potential for hydrogen bond formation are the common structural features of anticipated substrates of bilitranslocase. These features may serve as guidelines in the design of new pharmaceuticals transported by bilitranslocase.

Keywords: QSAR models; bilitranslocase; transmembrane transport activity; antioxidants; antiprions

1. Introduction

Historically, the primary route of drug uptake into cells was considered to be passive diffusion through the lipid bilayer and this notion was widely accepted in the pharmaceutical industry for a long time. Recently a considerable amount of evidence and implications indicate the essential role of carrier-mediated transport for drug uptake into cells [1,2,4,5]. For most of pharmaceutical drugs in use nowadays these mechanisms were not specifically considered during drug design, but rather occurred serendipitously, by hitchhiking on membrane carriers normally used for the transport of nutrients and intermediary metabolites [4]. However, the importance of carrier-mediating properties of potential agents is now becoming seriously considered because of their high potential for improving the effectiveness, selectivity and safety of drugs, and dietary supplements [1,5]. One of such membrane carriers
is bilitranslocase, a transmembrane protein basically involved in bilirubin transfer from blood to liver cells.

Bilitranslocase (BTL) is a plasma membrane transporter (T.C. 2.A.65.1.1) specific for water-soluble organic anions (bile pigments, dietary flavonoids, nucleotides) [6–11]. It has been experimentally confirmed that BTL is localized on the vascular endothelium, absorptive (gastrointestinal) and excretory (hepatic and renal) epithelia [12,13]. BTL functions do not overlap with other organic anion transporters and probably the most unique feature of BTL is the ability to transport nucleotides, which no other known membrane transporter is able to do [11,14]. However, several transporters of nucleosides and nucleobases are reported [15,16], i.e. concentrative nucleoside transporters belonging to the SLC28 gene family, which enable a unidirectional high affinity transport of nucleosides and their analogues into the cells by coupling their transport to the inward directed sodium gradient [17], and equilibrative nucleoside transporters from the SLC29 gene family, which mediate a bidirectional, sodium-independent facilitated diffusion with a low affinity [18]. The capacity of BTL to transport anthocyanins [6,12], which are natural glycosylated molecules, suggests the possible development of water-soluble orally effective drugs. Thus, a comprehensive knowledge of BTL would offer new space for solutions in drug discovery, which could be of great interest for the pharmaceutical industry. Several studies have focused on advancing BTLs structure, but its three-dimensional (3D) structure remains enigmatic. The structures of the four transmembrane subunits of BTL have been successfully predicted by chemometrics modelling [20]. Furthermore, the stability of these four BTL transmembrane regions has been studied with the use of molecular dynamics simulations and the 3D structures of two transmembrane regions have been resolved by nuclear magnetic resonance (NMR) spectroscopy [22,23].

To gain some insights into the transport mechanism of structurally unresolved transmembrane proteins, the so-called quantitative structure–activity relationship (QSAR) methods can be applied [24–27], which are independent of the host protein structure. In the past decade the transport activity of BTL for many different ligands has been characterized with a biological assay [6,7,10–14], and the experimental data obtained represented the basis for the development of several QSAR models for prediction of BTL transport ability for specific sets of polyaromatic organic compounds. The models were built with the use of counter-propagation artificial neural networks (CP-ANN) for sets of anthocyanin and flavonol derivatives [10] and various endogenous compounds, xenobiotics, and purine and pyrimidine derivatives [11,28].

In this study we present novel QSAR models for the prediction of BTL transport activity that were developed using the combined data from our previous studies [10,11]. These models cover a wider chemical space than previous ones and can be used for the prediction of BTL transport activity of new drug candidates such as nucleotide-like anticancer drugs [29,30], antivirals [31], and phenol-based drugs [32,33] with potential antioxidant or anticancer activity. The modelling procedure is based on the nonlinear CP-ANN and support vector machines (SVM), and linear multiple linear regression (MLR) methods, following the principles set out by the Organisation for Economic Co-operation and Development (OECD) [34].

Furthermore, two sets of pharmaceutically interesting compounds are evaluated with the newly developed QSAR models, including several antioxidants (pulvinic acid and coumarine derivatives) [35,36] and small aromatic organic compounds with the potential of being therapeutic agents for prion diseases [37], with the aim of creating a list of the most promising active compounds for further experimental studies regarding bioavailability. Antioxidants are important for their reactive oxygen species scavenging potential, which is a high priority domain in medical research today, while antiprion compounds are widely investigated for their potential therapeutic effects in neurodegenerative diseases. In order to set guidelines for
further drug design studies the most promising compounds from the two studied datasets are highlighted and their structural properties having impact on the BTL transport ability are discussed.

2. Materials and methods

2.1 Dataset and structural descriptors

2.1.1 Dataset for model construction

The dataset contains 22 anthocyanins and 21 flavonols [10], and 77 compounds belonging to different chemical classes (nucleotides, nucleosides, nucleobases, various endogenous molecules, dyes and drugs) [11], together with the corresponding experimental data on BTL transport activity from previous studies. Among 120 compounds included in this study 70 compounds are inactive (I), while 50 compounds are active (A). Two-dimensional (2D) chemical structures for all 120 compounds can be found in Table S1 (in the supplementary material which is available via the multimedia link on the online article webpage), while Table 1 presents structure names and experimental values only for the active compounds. For the modelling purposes all $K_i$ (mM) values are expressed in logarithmic units as $pK_i$ (M).

2.1.2 Two pharmaceutically interesting datasets for model application

The data include two sets of pharmaceutically interesting compounds with undetermined BTL transport activity: (1) a set of 111 antioxidants [35,36] (ID = 201–311; see Table S2 in the supplementary material which is available via the multimedia link on the online article webpage), and (2) a set of 109 active small antiprion compounds [37] (ID = 401–509; see Table S3 in the supplementary material which is available via the multimedia link on the online article webpage).

2.2 Descriptors of chemical structures

The equilibrated 3D structures were obtained using the 2D chemical structures as an input to the AM1 semi-empirical method of the MOPAC software [40]. The Codessa software [41] was used to calculate the molecular descriptors, which are used for the mathematical representation of molecules and are generally grouped into five categories: constitutional, topological, geometrical, electrostatic and quantum-chemical. The Codessa output resulted in more than 380 molecular descriptors for each molecule.

2.3 Computational methods and software

For the modelling purposes, QSARINS software was applied for the development of multiple linear regression (MLR) models [42], our in-house FORTRAN based software was employed for the development of counter-propagation artificial neural network (CP-ANN) models [43–45], and the LIBSVM software package [46] was used for the construction of support vector machines (SVM) models. Kohonen neural networks were applied for the reduction of the number of variables and splitting of the data, while a genetic algorithm (GA) was employed for the selection of influential variables [47].
Table 1. List of 50 active compounds with their assigned ID numbers, and the experimental and predicted $pK_i$ values. Compounds selected in the training, test and validation set are labelled TR, TE and V, respectively.

| No. | ID | Name | Activity | $K_i \pm SE$ (mM) | $pK_i$ exp (M) | $pK_i$ pred (M) model M1 | AD (IN/OUT) model M1 |
|-----|----|------|----------|-------------------|---------------|------------------------|----------------------|
| 1   | 4 TR | Adenosine 3′-monophosphate | C | 0.95 ± 0.07 | 3.022 | 2.956 | IN |
| 2   | 5 TE | Adenosine 5′-monophosphate | C | 2.63 ± 0.19 | 2.580 | 2.956 | IN |
| 3   | 9 TE | Adenosine 5′-diphosphate | C | 1.42 ± 0.10 | 2.848 | 2.956 | IN |
| 4   | 11 V | Adenosine 5′-triphosphate | C | 0.385 ± 0.03 | 3.415 | 3.684 | OUT |
| 5   | 14 TR | Adenosine 5′-(α,β-methylene) diphosphate | C | 1.31 ± 0.14 | 2.883 | 2.956 | IN |
| 6   | 15 TE | Adenine 9-β-D-arabinofuranoside | NC | 3.76 ± 0.38 | 2.424 | 2.956 | IN |
| 7   | 16 TE | Adenosine 3′-phosphate 5′-phosphosulfate | C | 0.148 ± 0.01 | 3.830 | 3.684 | IN |
| 8   | 17 V | S-(5′-Adenosyl)-L-homocysteine | C | 0.44 ± 0.02 | 3.357 | 3.195 | IN |
| 9   | 18 TR | S-(5′-Adenosyl)-L-methionine chloride | C | 0.408 ± 0.03 | 3.389 | 3.422 | IN |
| 10  | 23 TR | Guanosine 5′-monophosphate | C | 13.92 ± 1.38 | 1.856 | 1.882 | IN |
| 11  | 26 TR | Guanosine 5′-diphosphate | C | 4.55 ± 0.34 | 2.342 | 2.351 | IN |
| 12  | 28 TR | Guanosine 5′-triphosphate | C | 7.66 ± 0.09 | 2.116 | 2.209 | IN |
| 13  | 31 TR | Uridine | C | 2.58 ± 0.24 | 2.588 | 2.580 | IN |
| 14  | 34 TR | Uridine 5′-monophosphate | C | 4.13 ± 0.38 | 2.384 | 2.580 | IN |
| 15  | 36 TR | Uridine 5′-diphosphate | C | 3.10 ± 0.37 | 2.509 | 2.524 | IN |
| 16  | 38 TR | Uridine 5′-triphosphate | C | 1.425 ± 0.08 | 2.846 | 2.853 | IN |
| 17  | 41 TR | Uridine 5′-diphosphogalactose | C | 2.47 ± 0.27 | 2.607 | 2.567 | IN |
| 18  | 47 V | Thymidine 5′-monophosphate | C | 3.71 ± 0.18 | 2.431 | 2.458 | IN |
| 19  | 49 TR | Thymidine 5′-diphosphate | C | 2.23 ± 0.15 | 2.652 | 2.628 | IN |
| 20  | 51 TR | Thymidine 5′-triphosphate | C | 1.45 ± 0.08 | 2.839 | 2.792 | IN |
| 21  | 68 TR | Uric acid | C | 1.50 ± 0.18 | 5.460 | 5.452 | IN |
| 22  | 75 TR | Cibacron Blue F3G-A | C | 0.00347 ± 0.00032 | 4.653 | 4.443 | IN |
| 23  | 78 V | Pelargonidin | C | 0.02221 ± 0.0017 | 4.756 | 4.794 | IN |
| 24  | 79 TE | Cyanidin | C | 0.01755 ± 0.00168 | 4.756 | 4.794 | IN |

(Continued)
Table 1. (Continued).

| No. | ID | Name                                  | Activity | $K_i \pm SE$ exp (mM) | $pK_i$ exp (M) | $pK_i$ pred (M) model M1 | AD (IN/OUT) model M1 |
|-----|----|---------------------------------------|----------|------------------------|----------------|--------------------------|---------------------|
| 25  | 80 TR | Delfinidin                            | C        | 0.00527 ± 0.00038      | 5.278          | 5.192                    | IN                  |
| 26  | 81 TR | Peonidin                              | C        | 0.00623 ± 0.00051      | 5.206          | 5.301                    | IN                  |
| 27  | 82 V  | Petunidin                              | C        | 0.00757 ± 0.00099      | 5.121          | 4.795                    | IN                  |
| 28  | 83 TE | Malvidin                              | C        | 0.0072 ± 0.0004        | 5.143          | 4.794                    | IN                  |
| 29  | 84 TE | Pelargonidin 3-O-β-d-glucopyranoside   | C        | 0.0028 ± 0.0002        | 5.553          | 5.144                    | IN                  |
| 30  | 85 TR | Cyanidin 3-O-β-d-glucopyranoside       | C        | 0.00578 ± 0.00039      | 5.238          | 5.197                    | IN                  |
| 31  | 86 TR | Delfinidin 3-O-β-d-glucopyranoside     | C        | 0.00857 ± 0.0002       | 5.067          | 5.087                    | IN                  |
| 32  | 87 TR | Peonidin 3-O-β-d-glucopyranoside       | C        | 0.00183 ± 0.0002       | 5.738          | 5.675                    | IN                  |
| 33  | 88 TR | Petunidin 3-O-β-d-glucopyranoside      | C        | 0.00403 ± 0.0002       | 5.395          | 5.380                    | IN                  |
| 34  | 89 TE | Malvidin 3-O-β-d-glucopyranoside       | C        | 0.00142 ± 0.00013      | 5.848          | 5.503                    | IN                  |
| 35  | 90 TE | Pelargonidin 3,5-di-O-β-d-glucopyranoside | C | 0.00642 ± 0.00029 | 5.192 | 5.247 | IN |
| 36  | 91 TR | Cyanidin 3,5-di-O-β-d-glucopyranoside  | C        | 0.00577 ± 0.00040      | 5.239          | 5.215                    | IN                  |
| 37  | 92 TE | Peonidin 3,5-di-O-β-d-glucopyranoside  | C        | 0.00681 ± 0.00077      | 5.167          | 5.215                    | IN                  |
| 38  | 93 TR | Malvidin 3,5-di-O-β-d-glucopyranoside | C        | 0.00636 ± 0.00045      | 5.197          | 5.197                    | IN                  |
| 39  | 94 TR | Cyanidin 3-O-α-L-arabinopyranoside     | C        | 0.00916 ± 0.00099      | 5.038          | 5.035                    | IN                  |
| 40  | 95 V  | Cyanidin 3-O-β-d-galactopyranoside     | C        | 0.03572 ± 0.00058      | 4.447          | 5.197                    | OUT                 |
| 41  | 96 TR | Malvidin 3-O-(6-O-acetoyl)-β-d-glucopyranoside | C | 0.05833 ± 0.00009 | 4.234 | 4.342 | IN |
| 42  | 100 TR | Galangin                               | NC       | 0.0606 ± 0.0010        | 4.218          | 4.164                    | IN                  |
| 43  | 101 TE | Kaempferol                             | NC C     | 0.0639 ± 0.00340.1316  | 4.1943.881     | −3.940                    | IN                  |
| 44  | 102 TR | Quercetin                              | NCC      | 0.0796 ± 0.00360.0211  | 4.0994.676     | −4.710                    | IN                  |
| 45  | 139 TE | Hydrochlorothiazide                    | C        | 0.056 ± 0.004          | 4.252          | 3.998                    | IN                  |
| 46  | 151 TE | Sulindac                               | C        | 0.062 ± 0.008          | 4.208          | 3.998                    | IN                  |
| 47  | 152 TR | Triiodo-l-thyronine (T3)               | NC       | 0.1743 ± 0.0002        | 3.758          | 3.998                    | IN                  |

(Continued)
3. Results and discussion

3.1 Initial reduction of the number of descriptors

Since the reduction in the number of initially calculated descriptors is crucial for computational modelling, we first removed all descriptors with low variance ($s^2 < 0.03$). The remaining pool of 175 molecular descriptors was normalized to zero mean and unit standard deviation, resulting in a $120 \times 175$ data matrix. Further reduction of the number of variables was achieved based on the similarity criterion by the mapping of their transposed dataset ($175 \times 120$ data matrix) into a $7 \times 7$ dimensional Kohonen neural network, where descriptors were mapped according to their similarity and consequently several of them were placed onto the same neurons. From each neuron only two descriptors, those with the minimal and maximal Euclidean distances (EDs) were selected into the final pool of 66 molecular descriptors used for the modelling.

3.2 Splitting of the data into the training, test and validation set

To avoid inconsistent results a careful splitting of the initial data is needed. This was performed following the recommended methodology for building QSAR models [48]; the training set (TR) was used for the learning of the network, the test set (TE) for defining the optimal model parameters (internal validation) and the validation set (V) for the external validation of the developed models. The division of compounds into the TR, TE, and V sets was done according to the optimal distribution of 120 compounds described with 66 molecular descriptors on the Kohonen top map, which was obtained by varying different parameters of the network until the maximal occupancy of neurons and the minimal average error at one object was achieved. The optimal distribution was attained using a network with following parameters: $13 \times 13$ neuron grid, 1000 learning epochs, 0.5 maximal learning rate, 0.01 minimal learning rate, non-toroidal boundary conditions, and with the use of a triangular correction function of the neighbourhood. The optimal Kohonen top map with the distribution of 120 compounds is presented in Figure 1.

The compounds are distributed on the Kohonen top map in accordance with their structural similarity; similar structures are positioned on the same or neighbouring neurons. Figure 1 depicts the grouping of compounds in three clear clusters: the first one contains nucleobases and their derivatives (ID = 1–65), the second one consists of structurally related

| No. | ID | Name         | Activity | $K_i \pm SE$ (mM) | $pK_i$ exp (M) | $pK_i$ pred (M) model M1 | $AD$ (IN/OUT) model M1 |
|-----|----|--------------|----------|------------------|----------------|-------------------------|------------------------|
| 48  | 154 | Sulfobromophthalaein | C        | 0.00532 ± 0.00063 | 5.274          | 5.465                   | OUT                    |
| 49  | 156 | Bilirubin    | C        | 0.00111 ± 0.00001 | 5.950          | 5.901                   | IN                     |
| 50  | 157 | Biliverdin   | C        | 0.00111 ± 0.00002 | 5.950          | 5.901                   | IN                     |

Note: aExperimental data obtained from references [6,7,10–14,38,39].
The division of 120 compounds into the TR (70), TE (31) and V (19) set was performed on the basis of their distribution on the Kohonen map (shown in Figure 1), taking careful consideration to cover entire chemical spaces of each of the three initial datasets. The ratio of the division was 58:26:16 for the TR, TE and V set, respectively. Data used for classification purposes enclosed both active and inactive compounds, represented equally in all three sets. For predictive models only the active compounds were considered and consequently the 70 inactive compounds were omitted. Therefore, only the remaining 50 active compounds were

|   | 120_c | 44_a | 6_a | 24_a | 3_a | 5_a | 22_a | 11_a | 13_a | 41_a | 42_a |
|---|-------|------|-----|------|-----|-----|------|------|------|------|------|
| 73_e | 54_a | 31_a | 20_a | 2_a | 56_a | 36_a | 51_a | 16_a | 28_a |
| 107_e |     |      |      |      | 47_a |      |      |      |      |
| 109_e | 111_e | 98_b | 86_b | 84_b | 85_b | 92_b | 94_b | 18_a | 144_d |
| 110_e | 111_c | 117_c |      | 132_a |      |      |      |      |
| 112_e |      |      |      |      |      |      |      |      |
| 91_b | 90_b | 88_b | 89_b | 87_b | 141_d | 128_d | 131_d | 136_a | 139_d |
| 92_b | 118_b | 96_b |      |      |      |      |      | 154_d | 152_d |
| 93_b |      |      |      |      |      |      |      |      | 121_d |
| 108_c | 130_d | 134_d |      |      |      |      |      |      |
| 116_c | 133_d | 149_d | 153_d | 100_c |      |      |
| 75_e | 77_e | 147_d | 140_d | 104_c | 78_b | 101_c | 13_d | 125_e |
| 76_e | 123_d | 156_a | 157_a | 124_d | 150_d | 70_c | 71_c | 72_c | 73_c | 79_b | 80_b | 102_c | 103_c |

Figure 1. Distribution of 120 compounds on the Kohonen top map used for the splitting of the data for classification purposes. Different chemical classes are marked with different letters; nucleobases, nucleotides and nucleosides with ‘a’; anthocyanins with ‘b’; flavonols with ‘c’; various drugs with ‘d’; and dyes and other compounds with ‘e’. The ID numbers of compounds in the TR set are written in a regular font, compounds in the TE set with a bold font, and compounds in the V set with a bold and italic font.

anthocyanins and flavonols (ID = 78–96 and 100–120, respectively), and the third one is comprised of various drugs (ID = 121–153).
distributed in the TR (30), TE (13) and V (7) set. The corresponding ratio of the data division was 60:26:14 for the TR, TE and V set, respectively.

3.3 Development of classification models

With the aim of developing a reliable and accurate classification model that could classify the compounds into active (A) and inactive (I) ones, two types of supervised learning algorithms were used: counter-propagation neural networks (CP-ANN) and support vector machines (SVM). Both algorithms are well known and frequently used methods giving good classification results [49].

The input data for both algorithms were 120 compounds, described with 66 calculated molecular descriptors and the corresponding BTL inhibition constants ($pK_i$). These were the basis for the separation of active from inactive compounds, the threshold for the separation being set at a $pK_i$ of 1.7 log units; this value had already been determined in the previous study of Župerl et al. [11].

All the obtained models were internally evaluated with the TE set and the V set was used for the external validation. The efficiencies of all the models were evaluated using classification coefficients based on the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions; accuracy (equation (1)), sensitivity (equation (2)), specificity (equation (3)) and Matthews correlation coefficient (equation (4)) [50].

\[
AC = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)
\]
\[
SE = \frac{TP}{TP + FN} \quad (2)
\]
\[
SP = \frac{TN}{TN + FP} \quad (3)
\]
\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)\times(TP + FN)\times(TN + FP)\times(TN + FN)}} \quad (4)
\]

To obtain the classification model with a minimal error for the internal test set (root mean square, RMSE), the network parameters were varied. The best CP-ANN classification model was constructed with 93 epochs of training, neuron grid size of $13 \times 13$, maximal and minimal learning rate of 0.5 and 0.01, respectively, non-toroidal boundary conditions, and a triangular correction function of the neighbourhood. The cumulative AC, SE, SP, and MCC of all objects obtained by the best CP-ANN classification model were 0.91, 0.84, 0.97, and 0.82 respectively.

The SVM classification was first performed using different kernel functions (linear, polynomial, Gaussian). In each case the algorithm parameters were determined using a grid search, where the optimal set of parameters was defined as the one that produced the lowest mean square error for the cross-validation of the TR set. The optimal SVM classification model was selected based on the accuracy of the V set. Best performance was achieved using a third-degree polynomial kernel function, parameter C set at one, and gamma at 0.03. The cumulative AC, SE, SP and MCC of all objects obtained by the top SVM classification model were 0.88, 0.86, 0.90, and 0.76, respectively. The classification results (AC, SE, SP, and MCC) for all three sets (TR, TE and V) obtained by the best CP-ANN and SVM classification models are given in Table 2.
The best CP-ANN and SVM classification models were applied to two sets of compounds with unknown transport activity: (1) a set of 111 antioxidants (ID = 201–311, Table S2 in the supplementary material which is available via the multimedia link on the online article webpage); and (2) a set of 109 antiprion compounds (ID = 401–509, Table S3 in the supplementary material which is available via the multimedia link on the online article webpage). The compounds were classified as active only if their predicted p\(K_i\) values were higher than 1.7.

The classification results (i.e. list of compounds predicted as active) for both datasets are given in Table 2. With both classification methods, 48 compounds were predicted as active (29 antioxidants and 19 antiprions). Although the SVM model gives a lower number of active compounds (20 compounds) than the CP-ANN model (41 compounds), the results of the two methods mostly overlap.

### Table 2. Performances of the best SVM and CP-ANN classification model and a list of compounds classified as active (labelled with ID values).

| No. | Antioxidants | Antiprions | Antioxidants | Antiprions |
|-----|--------------|------------|--------------|------------|
| 1   | 201          | 423        | 201          | 422        |
| 2   | 203          | 451        | 203          | 423        |
| 3   | 207          | 481        | 204          | 437        |
| 4   | 208          | 483        | 205          | 451        |
| 5   | 226          | 485        | 206          | 452        |
| 6   | 234          | 490        | 207          | 455        |
| 7   | 235          | 460        | 226          |           |
| 8   | 239          | 480        | 230          |           |
| 9   | 251          | 483        | 233          |           |
| 10  | 262          |            | 234          | 485        |
| 11  | 263          |            | 235          | 490        |
| 12  | 265          |            | 242          | 492        |
| 13  | 270          |            | 243          | 494        |
| 14  | 272          |            | 246          | 495        |
| 15  | 248          |            | 248          | 496        |
| 16  | 251          |            | 251          | 497        |
| 17  | 252          |            | 252          | 498        |
| 18  | 259          |            | 259          | 503        |
| 19  | 260          |            | 260          |           |
| 20  | 268          |            | 268          |           |
| 21  | 269          |            | 269          |           |
| 22  | 270          |            | 270          |           |
| 23  | 280          |            | 280          |           |

Note: Bold indicates compounds classified as active by both classification models.

### 3.4 Application of classification models on two pharmaceutically interesting datasets

The best CP-ANN and SVM classification models were applied to two sets of compounds with unknown transport activity: (1) a set of 111 antioxidants (ID = 201–311, Table S2 in the supplementary material which is available via the multimedia link on the online article webpage); and (2) a set of 109 antiprion compounds (ID = 401–509, Table S3 in the supplementary material which is available via the multimedia link on the online article webpage). The compounds were classified as active only if their predicted p\(K_i\) values were higher than 1.7. The classification results (i.e. list of compounds predicted as active) for both datasets are given in Table 2. With both classification methods, 48 compounds were predicted as active (29 antioxidants and 19 antiprions). Although the SVM model gives a lower number of active compounds (20 compounds) than the CP-ANN model (41 compounds), the results of the two methods mostly overlap.
3.5 Development of predictive models

The models for predicting BTL transport activity were built on the basis of 50 active compounds from the initial dataset (Table 1), which were divided into 30, 13 and 7 compounds for the TR, TE and V set, respectively. All compounds were described with 66 molecular descriptors and the corresponding $pK_i$ values, which were directly used as an input for the modelling algorithms.

First the MLR predictive model was built. Since the MLR modelling strategy in the QSARINS software incorporates leave-one-out cross-validation for the internal validation of the models, the division into the TR and TE set was not necessary. Therefore a merged TR and TE set (43 compounds) was used for the model development and the V set (seven compounds) for the external model validation. The best predictive ability of the MLR model was achieved using seven molecular descriptors (Table S4 in the supplementary material which is available via the multimedia link on the online article webpage), with reasonably high internal ($Q_{LOO}^2 = 0.92$) and external ($Q_V^2 = 0.77$) validation parameters. All $Q^2$ validation parameters in this study were calculated using the $Q^2_{F3}$ formula, proposed by Consonni et al. [51]. The best MLR model is presented in Figure 2(a) as a plot of experimental versus predicted $pK_i$ values.

Next the CP-ANN algorithm was applied. In order to obtain the predictive model with a minimal error of the internal test set (RMS$_{TE}$), different network parameters were varied. Within more than a thousand tested models, the optimal model (RMS$_{TE} = 0.75$) was obtained with the network dimension of $7 \times 7$, trained for 14 epochs, with maximal and minimal learning rates of 0.5 and 0.01, respectively. The low internal ($Q_{TE}^2 = 0.66$) and external ($Q_V^2 = 0.57$) validation parameters were ascribed to the high number of variables (66) used for the model construction. To further reduce the number of molecular descriptors used in the modelling procedure, we coupled the CP-ANN with a genetic algorithm (GA). More than 200 GA runs with different origin parameters (4–20 initial genes, 10–30 survivals, 0.5–1% of mutations) were performed with the goal of defining the set of most significant variables based on the lowest error of the TR and TE set, minimal RMS$_{TR+TE}$. During 600 generations of each
GA run a population of 100 chromosomes was evolving, with varying parameters of the network: number of neurons (5–25), number of learning epochs (50–1000) and learning rate (0.1–0.7). The best CP-ANN_GA model with high internal ($Q_{TE}^2 = 0.93$) and external ($Q_{l}^2 = 0.96$) validation parameters was obtained using 12 molecular descriptors, with a network dimension of 10 × 10, 100 learning epochs, 0.5 maximal and 0.01 minimal learning rates. Figure 2(b) depicts the regression plot of the experimental versus predicted pKi values obtained by the best predictive CP-ANN model coupled with GA. The predicted pKi values obtained from the best CP-ANN_GA model for all 50 active compounds used for the modelling are given in Table 1. According to Figure 2, the best nonlinear modelling method (CP-ANN_GA) resulted in a model with better performance compared with the best linear one (MLR).

To compare CP-ANN predictive modelling with some other non-linear modelling method, the support vector regression (SVR) algorithm was applied. When all 66 initially selected molecular descriptors were used for the modelling, the internal ($Q_{TE}^2$) and external ($Q_{l}^2$) validation parameters of the resulting SVR model reached values of 0.83 and 0.67, respectively, which is comparable with or even slightly better than in the case of the corresponding CP-ANN model. The SVR algorithm was also applied to the development of a model using only the 10 molecular descriptors from the best CP-ANN_GA model. The resulting internal ($Q_{TE}^2$) and external ($Q_{l}^2$) validation parameters were 0.46 and 0.43, respectively, which is quite poor and significantly worse than in the case of the best CP-ANN_GA model.

For further study the best CP-ANN_GA model was selected, not only for the reasons of a superior performance with the reduced set of descriptors, but also for the well-established assessment of the applicability domain, which is lacking in case of SVR modelling.

### 3.6 Applicability domains of the developed models

In QSAR modelling the applicability domain (AD) represents the chemical space (chemical information with regard to the properties and structures of compounds) that was used in the development of the model. It has to be defined for each model to determine the reliability of predictions for compounds not used in the model development [34]. The structural space that the model covers is defined with the compounds in the training and internal test set; consequently the model’s predictions are reliable only for compounds within this space. The distance-based approach for the applicability assessment is widely used and applicable for both linear and nonlinear modelling methods [48,52,53]. The compound is labelled as out of the domain when the distance between the compound and the centre of the training data set exceeds a defined threshold. For the AD evaluation of the MLR predictive model, a leverage based approach was used [48,52] where a so-called Williams plot is obtained (Figure 3(a)), which shows standardized residuals as a function of leverage values, while for the CP-ANN_GA predictive model, the Euclidean distance (ED) approach was used [53], where the compounds of the training and internal test set define the minimum ED space (MEDS) (Figure 3(b)).

The boundaries of the residuals of predicted properties are set at ±3σ for both methods and they determine the response outliers. The boundaries concerning the structural similarity of molecules differ for both methods. In case of the MLR predictive model, the leverage boundary was determined by the critical hat value $h^* = 0.56$ [48], while for the CP-ANN_GA predictive model, the threshold distance was determined as a maximal ED value from the TR and TE set ($ED_{crit} = 0.17$) [53]. The critical hat value $h^*$ and $ED_{crit}$ determine structural outliers of the MLR and CP-ANN_GA model, respectively.
According to Figure 3(a) (MLR model), two structural (ID = 75, 78) and one response (ID = 96) outliers are detected; two compounds (ID = 75, 96) are from the TR set and one (ID = 78) from the V set. In Figure 3(b) (CP-ANN_GA model), two compounds (ID = 154, 11) from the V set are considered as structural outliers. Compound with ID = 154 is sulphobromophthalein and is indeed structurally quite different to the compounds used for the model construction. Additionally one compound (ID = 95) from the V set is recognized as a response outlier.

3.7 Structural variables influencing the BTL transport activity

The influential variables can give guideline information on which chemical features correlate with inhibition constants, which can highlight structural details of compounds that have an important influence on the BTL transport activity. To interpret and understand this mechanism a detailed analysis of the selected descriptors was carried out. Furthermore, quantitative information about the most influential descriptors was collected on the basis of the top 121 CP-ANN_GA models selected by the following criteria: $\text{RMSTE} \leq 0.4$, $\text{RMSV} \leq 1.3$, and the number of selected descriptors $\leq 20$. The percentage of the occurrence of molecular descriptors included in 121 models was determined and the 11 most frequent variables are listed as a consensus descriptor set in Table S4 in the supplementary material which is available via the multimedia link on the online article webpage.

The most frequently selected molecular descriptor is quantum-chemical, describing minimal nuclear-electron attraction for a carbon atom, presented in 80% of 121 models. The second most common descriptor (72%) is topological, describing the average bonding information content and the third most common (60%) is the hydrogen acceptor dependent area-weighted surface charge of hydrogen bonding donor atoms (HDSA-2), which is directly connected to the hydrogen bonding between the molecules and encodes features responsible for polar interactions between molecules. A comparison of 12 molecular descriptors from the best CP-ANN_GA predictive model M1 and the 11 most frequently selected variables shows that most of them are overlapping and thus describing the same structural features.

In general the selected descriptors are in an agreement with the influential variables published in our previous studies [10,11,28], which highlighted descriptors related to the ability
of compounds to form hydrogen bonds (HDSA-2), the shape and compactness of compounds (Information content, Wiener index) and the ionic properties of compounds (FPSA-2, PPSA-2, PPSA-3). Therefore, with this new robust model, which covers a bigger structural domain, we can generalize and support the hypothesis that active molecules should be capable of forming hydrogen bonds and that BTL has the capacity to transport ionic species. Furthermore, a detailed screening of the shape of active compounds (competitive or non-competitive inhibitors) reveals a third important feature which is the planarity of molecules. This is mostly reflected in two selected descriptors: the number of aromatic bonds and Gravitation index. The importance of the planarity of molecules for the BTL transport had been already observed and experimentally confirmed in previous studies [10,11,28,39].

3.8 Application of the predictive models on two pharmaceutically interesting datasets

Two pharmaceutically interesting datasets (antioxidant and antiprion compounds) with undetermined BTL transport activity were evaluated with the proposed prediction models, MLR and CP-ANN_GA. Only the compounds predicted as active by the two above described classification models (SVM and CP-ANN, Table 2) were tested with the developed predictive models (MLR and CP-ANN_GA). The results of the classification (SVM and CP-ANN) and predictive models (CP-ANN_GA; model M2) for the set of antioxidants and antiprions are given in Tables S2 and S3 in the supplementary material which is available via the multimedia link on the online article webpage.

The main concern of the newly developed models is the reliability of predictions for the compounds of diverse chemical structures; therefore we first had to check if they fall into the AD. Because the biological properties of these compounds are not yet determined, the applicability assessment was possible only for the structural domain, so the modified AD plots had to be used; these were the so-called Insubria graphs in case of MLR [48,52], and MEDS graphs in case of CP-ANN_GA models [53]. The boundaries for the chemical space of the MLR and CP-ANN_GA models are again determined by the critical hat value ($h^*$) and $ED_{crit}$ (x-axis same as in Figure 3), respectively, while the y-axis shows predicted values instead of standardized residuals (Figure 4). Predictions for compounds that fall out of the model structural domain are considered unreliable and thus should not be included in further investigations.

The Insubria graph (Figure 4(a)) of the MLR model indicates that 20 out of 29 antioxidants and 10 out of 19 antiprions fall within the boundaries of the structural domain. In the MEDS graph of the CP-ANN_GA model (Figure 4(b)) all 29 antioxidants are positioned out of the AD and only two antiprions (ID = 451, 503) are placed inside the structural AD.

Even though the best CP-ANN_GA model has good predictive ability (i.e. low prediction error for the V set), it is not applicable for the set of antioxidant and antiprion compounds. This can be improved by choosing the model parameters that broaden the AD at the price of reducing the accuracy of predictions. Therefore, several other predictive CP-ANN_GA models were evaluated regarding the new criteria, and only the models with broader AD and still acceptable accuracy were considered. First, the models were ranked according to the decreasing $Q^2_I$; six best models are shown in Table 3. Then the compounds from two pharmaceutically interesting datasets were input to the models to check if they fall into the ADs (see Table 3).

Model M2 is the second best CP-ANN_GA model with slightly lower validation parameter then M1 ($Q^2_I = 0.91$), but with significantly broader AD for the antioxidants. The new MEDS graph for model M2 applied to the two pharmaceutically interesting datasets is shown
in Figure 5. As can be seen in Table 3 and in Figure 5, 17 out of 29 antioxidants are located in the model AD. Model M4 with lower validation parameter ($Q^2_v = 0.86$) has even more antioxidants (20) inside the AD. For antiprion compounds, however, only two compounds are placed in AD of models M1 and M4, and 3 compounds are placed in AD of model M2. Hence, for various untested compounds it is important that the appropriate model is selected as a consensus among its accuracy and applicability.

In Figure 5 the MEDS graph (AD of model M2) is presented along with a part of the Kohonen top map (in the upper right corner). The partial top map depicts nine neurons of the trained CP-ANN_GA model with all the compounds located on these neurons, including the antioxidant and antiprion compounds that hit these neurons in the prediction process. If we look at the top middle neuron where the TR compound with ID = 100 (Galangin; $pK_r = 4.22$) is placed, 12 antioxidant and two antiprion compounds can also be found there. Thus, the predicted $pK_r$ values for all the untested compounds on this neuron are defined by the experimental $pK_r$ value of Galangin. These compounds are placed on the dashed middle horizontal line of the MEDS graph in Figure 5. The dashed upper and bottom horizontal lines are associated with the predictions from the middle and low-left corner neurons, respectively.

Antioxidants with ID = 201, 203–206, 208, 226, 233–235, 246, 248, 251–252, 259–260, 263, 265, 270 and 280 (Table S2 in the supplementary material which is available via the

Table 3. Performances of the top six CP-ANN_GA models.

| Model no. | $Q^2_v$ | $r^2_{TR}$ | No. of descriptors | No. of compounds in the AD | Antioxidants | Antiprions |
|-----------|---------|------------|--------------------|---------------------------|-------------|-----------|
| M1        | 0.957   | 0.996      | 12                 | 0/29                      | 2/19        |
| M2        | 0.910   | 0.999      | 10                 | 17/29                     | 3/19        |
| M3        | 0.908   | 0.998      | 9                  | 5/29                      | 1/19        |
| M4        | 0.863   | 0.996      | 5                  | 20/29                     | 2/19        |
| M5        | 0.859   | 0.995      | 11                 | 0/29                      | 1/19        |
| M6        | 0.856   | 1.000      | 8                  | 5/29                      | 1/19        |
multimedia link on the online article webpage) are predicted to pass the membrane through BTL and lie within the AD of the prediction model M2. All these compounds have large systems of conjugated double bonds, which are the cause of the planarity of these molecules. The least planar compounds are 204 and 206, with relatively flexible sugar moieties attached; since it is known that the attachment of sugar groups can influence the transport of substances through various transporters [54], this could also be the case with BTL, with sugar moieties not being disruptors of planarity of these compounds but rather the facilitators of transport. However, it is interesting that none of the coumarine based antioxidants (ID = 295–311) were selected as active, despite their planarity and relative resemblance to compounds 203, 204, 206 and 208. All the active compounds also contain an abundance of oxygen moieties attached, mostly in the form of hydroxyl groups, but also lactones, ethers, esters and free carboxylic groups. Conjugated double bonds are probably crucial for the antioxidant activity through the stabilization of the free radical, whereas the abundance of oxygen atoms may suggests a high capacity for hydrogen bonding, which is thought to be important in the BTL transport. Besides this, compound 226 contains a cyano group, compound 252 a couple of chlorine atoms, compound 259 two nitro groups, compound 260 a bromine atom, and compound 263 a thiophene group; however, none of these seem to have a major influence on the transport activity, since no other compound with these substituents was classified as active. Compound 201 (norbadione) is a polyphenol, extracted from mushrooms; it was found to be an effective in vivo antioxidant, but unfortunately rather toxic [55]. Compounds 203 (quercetin), 204 ((+)-rutin trihydrate) [56], 205 (curcumin) [57], 206 (esculin) [58], 208 (scopoletin)
and 234 (gallic acid) are also known antioxidants. Since our results indicate these compounds can be transported through BTL, we suggest a further experimental evaluation of these findings because BTL transport could turn out to be an effective delivery system for such compounds, especially due to its ubiquity, which would in turn give better support to the clinical use of these antioxidants.

Compounds 451 (diketopiperazine derivative), 480 (Congo red derivative), and 503 (hematein) from the antiprion dataset (Table S3 in the supplementary material which is available via the multimedia link on the online article webpage) are also predicted as active by our classification models and are placed within the AD of the prediction model M2. Again the most noteworthy common features of these compounds are their planarity, and the abundance of oxygen and/or nitrogen atoms, indicating their high capacity for forming hydrogen bonds. Compounds 451 and 480 were synthesized for the purposes of the prion disease studies and were experimentally determined as active against prion accumulation [61,62]. Hematein (CAS 475-25-2) is a natural compound isolated from Caesalpinia sappan and is known to be used as an analgesic and anti-inflammatory agent in oriental medicine [63,64]. Furthermore, its potential for anti-inflammatory and anti-atherogenic activity was experimentally studied and shows promising results [65,66]. A study by Kocisko et al. [67] estimated that hematein acts also as an antiprion therapeutic agent. Due to new findings on the ability of BTL to transport these compounds, they could serve as leads in the future design of antiprion therapeutics for their potential of improving the bioavailability.

4. Conclusions

In this study QSAR models were developed in order to predict the transport activities of some pharmaceutically interesting compounds through the membrane transporter BTL; the ability of BTL to transport heterogeneous organic molecules as well as some drugs had been already determined in previous studies. First classification models for the prediction of BTL transport ability were built using CP-ANN and SVM algorithms on the basis of 120 compounds and the corresponding experimental data on the BTL ligand specificity. CP-ANN and SVM models performed with similar statistical parameters for cumulative accuracy (0.91 and 0.88, respectively), selectivity (0.84 and 0.86, respectively) and specificity (0.97 and 0.90, respectively). Construction of the predictive models was based only on active compounds, using CP-ANN and MLR methods. The best model performance was achieved with CP-ANN coupled with a GA, with a good external validation parameter ($Q_{\text{ext}}^2 = 0.96$) and a low number of variables used (12). The applicability domains of the predictive models were assessed using the Euclidean distance based approach for CP-ANN models and the leverage based approach for MLR models. The proposed models give reliable predictions of the transport ability by BTL for various small aromatic compounds.

The classification models were applied to the experimentally untested compounds from the antioxidant and antiprion datasets. We had estimated that some of them are likely to be transported by BTL. After challenging the developed predictive models with the compounds classified as active, we checked if they lay in the applicability domain. On the basis of a compromise between the accuracy of the model and the applicability domain area, it was concluded that for antioxidants the CP-ANN GA models were better than MLR models, while for most of the antiprions, CP-ANN GA models were not applicable due to severe structural differences. The MLR model with a lower number of descriptors proved to be more suitable for the prediction of antiprion transport ability because of its broader applicability domain.
The analysis of the resulting predictions for the compounds within the applicability domains of the CP-ANN_GA models shows that they have some common features, most notably the planarity of large moieties of these molecules and an abundance of functional groups capable of hydrogen bond formation. These findings, together with the newly developed models, could serve as guidelines in current design of different therapeutics, not only of antioxidants or antiprions but also nucleotide-like and phenol-based anticancer and antiviral drugs.

**Supplemental material**

Supplemental material for this article can be accessed here: http://dx/doi.org/10.1080/1062936X.2014.962082

**Acknowledgements**

The study was financially supported through the grants P1-0017 and 34547 from the Agency of Research of the Republic of Slovenia (ARRS) and an EU grant from the Transregional Network for Innovation and Technology Transfer to Improve Health Care (CB 101 – TRANS2CARE). Professor P. Gramatica is greatly acknowledged for providing the QSARINS software.

**References**

[1] D.B. Kell, P.D. Dobson, and S.G. Oliver, *Pharmaceutical drug transport: The issues and the implications that it is essentially carrier-mediated only*, Drug Discov. 16 (2011), pp. 704–714.

[2] P.D. Dobson, K. Lanthaler, S.G. Oliver, and D.B. Kell, *Implications of the dominant role of transporters in drug uptake by cells*, Curr. Top. Med. Chem. 9 (2009), pp. 163–181.

[3] D.B. Kell and P.D. Dobson, *The cellular uptake of pharmaceutical drugs is mainly carrier-mediated and is thus an issue not so much of biophysics but of systems biology*, Syst. Chem. (2008), pp. 149–168.

[4] P.D. Dobson and D.B. Kell, *Carrier-mediated cellular uptake of pharmaceutical drugs: An exception or the rule?*, Nat. Rev. Drug Discov. 7 (2008), pp. 205–220.

[5] N. Mizuno, T. Niwa, Y. Yotsumoto, and Y. Sugiyama, *Impact of drug transporter studies on drug discovery and development*, Pharm. Rev. 55 (2003), pp. 425–461.

[6] S. Passamonti, U. Vrhovsek, and F. Mattivi, *The interaction of anthocyanins with bilitranslocase*, Biochem. Biophys. Res. Commun. 296 (2002), pp. 631–636.

[7] S. Passamonti, M. Terdoslavich, N. Medic, A. Margon, A. Cocolo, F. Micalli, G. Decorti, and M. Franco, *Uptake of bilirubin into HepG2 cells assayed by thermal lens spectroscopy*, Function of bilitranslocase, FEBS J. 272 (2005), pp. 5522–5535.

[8] S. Passamonti, M. Terdoslavich, R. Franca, A. Vanzo, F. Tramer, E. Braidot, E. Petrussa, and A. Vianello, *Bioavailability of flavonoids: A review of their membrane transport and the function of bilitranslocase in animal and plant organisms*, Curr. Drug Metab. 10 (2009), pp. 369–394.

[9] M. Terdoslavich, *Molecular mechanisms of organic anion transport*, PhD disseration, University of Trieste, Italy, PhD disseration, 2005.

[10] A. Karawajczyk, V. Drgan, N. Medic, G. Oboh, S. Passamonti, and M. Novič, *Properties of flavonoids influencing the binding to bilitranslocase investigated by neural network modelling*, Biochem. Pharm. 73 (2007), pp. 308–320.

[11] Š. Župerl, S. Fornasaro, M. Novič, and S. Passamonti, *Experimental determination and prediction of bilitranslocase transport activity*, Anal. Chim. Acta. 705 (2011), pp. 322–333.

[12] S. Passamonti, U. Vrhovsek, A. Vanzo, and F. Mattivi, *The stomach as a site for anthocyanins absorption from food*, FEBS Lett. 544 (2003), pp. 210–213.
Expression of bilitranslocase in the vascular endothelium and its function as a flavonoid transporter, Cardiovasc. Res. 85 (2010), pp. 175–183.

Bilitranslocase is involved in the uptake of bromosulfophthalein in rat and human liver, Drug Metab. Lett. 6 (2012), pp. 165–173.

Transport of nucleoside analogs across the plasma membrane: A clue to understanding drug-induced cytotoxicity, Curr. Drug Metab. 10 (2009), pp. 347–58.

The human concentrative and equilibrative nucleoside transporter families, SLC28 and SLC29, Mol. Aspects. Med. 34 (2013), pp. 529–547.

The concentrative nucleoside transporter family, SLC28, Pflugers Arch. 447 (2004), pp. 728–734.

The equilibrative nucleoside transporter family, SLC29, Pflugers Arch. 447 (2004), pp. 735–743.

The transporter classification database, Nucleic Acids Res. 42 (2014), pp. 251–258; database available at http://www.tcdb.org.

Data-driven model for the prediction of protein transmembrane regions, SAR QSAR Environ. Res. 20 (2009), pp. 741–754.

Development of transmembrane region prediction algorithms and structural elucidation of bilitranslocase, PhD dissertation, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, 2013.

Structural analysis of a peptide fragment of transmembrane transporter protein bilitranslocase, PLoS One 7 (2012), p. e38967.

Structural elucidation of transmembrane transporter protein bilitranslocase: Conformational analysis of the second transmembrane region TM2 by molecular dynamics and NMR spectroscopy, Biochim. Biophys. Acta. 1828 (2013), pp. 2609–2619.

Ligand-based approaches to in silico pharmacology, Methods Mol. Biol. 672 (2011), pp. 489–502.

Structure elucidation of transmembrane proteins using public-available databases and experimental data on competitive inhibition, Acta Chim. Slov. 58 (2011), pp. 385–392.

‘Reverse’ carbocyclic fleximeres: Synthesis of a new class of adenosine deaminase inhibitor, Nucleosides, Nucleotides Nucleic Acids 32 (2013), pp. 137–154.

Synthesis and biological evaluation of a series of thieno-expanded tricyclic purine 2′-deoxy nucleoside analogues, J. Bioorg. Med. Chem. 20 (2012), pp. 3009–3015.
[32] M.M. Dias, N.F.L. Machado, and M.P.M. Marques, *Dietary chromones as antioxidant agents – the structural variable*, Food Funct. 2 (2011), pp. 595–602.

[33] N.F.L. Machado, R. Calheiros, S.M. Fiuza, F. Borges, A. Gaspar, J. Garrido, and M.P. Marques, *Phenolic esters with potential anticancer activity – the structural variable*, J. Mol. Model. 13 (2007), pp. 865–877.

[34] Organisation for Economic Co-operation and Development, *OECD guidance document on the validation of (quantitative) structure-activity relationship ([Q]SAR) models*, ENV/JM/MONO 2, OECD, Paris, 2007.

[35] D. Habrant, S. Poigny, M. Ségur-Derai, Y. Brunel, B. Heurtaux, T. Le Gall, A. Strehle, R. Saladin, S. Meunier, C. Mioskowski, and A. Wagner, *Evaluation of antioxidant properties of monoaromatic derivatives of pulvinic acids*, J. Med. Chem. 52 (2009), pp. 2454–2464.

[36] A. Le Roux, I. Kuzmanovski, D. Habrant, S. Meunier, P. Bischoff, B. Nadal, S.A.L. Thetiot-Laurent, T. Le Gall, A. Wagner, and M. Novič, *Design and synthesis of new antioxidants predicted by the model developed on a set of pulvinic acid derivatives*, J. Chem. Inf. Model. 51 (2011), pp. 3050–3059.

[37] K. Venko, Š. Župerl, and M. Novič, *Prediction of antiprion activity of therapeutic agents with structure-activity models*, Mol. Divers. 18 (2014), pp. 133–148.

[38] G.L. Sottocasa, S. Passamonti, L. Battiston, L. Pascolo, and C. Tiribelli, *Molecular aspects of organic anion uptake in liver*, J. Hepatol. 24 (1996), pp. 36–41.

[39] S. Passamonti and G.L. Sottocasa, *The quinoid structure is the molecular requirement for recognition of phthaleins by the organic anion carrier at the sinusoidal plasma membrane level in the liver*, Biochim. Biophys. Acta. 943 (1988), pp. 119–125.

[40] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, and J.J.P. Stewart, *Development and use of quantum mechanical molecular models. AM1: A new general purpose quantum mechanical molecular model*, J. Am. Chem. Soc. 107 (1985), pp. 3902–3909.

[41] A.R. Katritzky, V.S. Lobanov, and M. Karelson, *Codessa 2.0, Comprehensive descriptors for structural and statistical analysis*, University of Florida, USA, 1994; software available at http://www.semichem.com/codessa.

[42] P. Gramatica, N. Chirico, E. Papa, S. Cassani, and S. Kovarich, *QSARINS: A new software for the development, analysis, and validation of QSAR MLR models*, J. Comput. Chem. 34 (2013), pp. 2121–2132.

[43] J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim, Germany, 1999.

[44] J. Zupan, M. Novic, and I. Ruisanchez, *Kohonen and counterpropagation artificial neural networks in analytical chemistry*, Chemomert. Intell. Lab. 38 (1997), pp. 1–23.

[45] M. Novič and J. Zupan, *Investigation of infrared spectra-structure correlation using Kohonen and counter-propagation neural network*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 454–466.

[46] C.C. Chang and C.J. Lin, *LIBSVM: A library for support vector machines*, ACM TIST 2 (2011), pp. 1–27; software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[47] R. Leardi, *Nature-inspired methods in chemometrics: Genetic algorithms and artificial neural networks*, Elsevier, Amsterdam and Boston, 2003.

[48] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.

[49] M. Vasighi and M. Kompany-Zareh, *Classification ability of self organizing maps in comparison with other classification methods*, MATCH Commun. Math. Comput. Chem. 70 (2013), pp. 29–44.

[50] P. Baldi, S. Brunak, Y. Chauvin, C.A.F. Andersen, and H. Nielsen, *Assessing the accuracy of prediction algorithms for classification: An overview*, Bioinformatics 6 (2000), pp. 412–424.

[51] V. Consonni, D. Ballabio, and R. Todeschini, *Comments on the definition of the Q(2) parameter for QSAR validation*, J. Chem. Inf. Model. 49 (2009), pp. 1669–1678.
[52] A. Tropsha, P. Gramatica, and V.K. Gombar, The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSRR Models, QSAR Comb. Sci. 22 (2003), pp. 69–77.

[53] N. Minovski, Š. Župerl, V. Drgan, and M. Novič, 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 (2013), pp. 28–42.

[54] P.C.H. Hollman, M.N.C.P. Bijsman, Y. van Gameren, E.P.J. Cnossen, J.H.M. de Vries, and M.B. Katan, The sugar moiety is a major determinant of the absorption of dietary flavonoid glycosides in man, Free Radical Res. 31 (1999), pp. 569–573.

[55] A. Le Roux, E. Josset, S. Benzina, B. Nadal, M.D.E. Murr, B. Heurtaux, F. Taran, J.M. Denis, T. Le Gall, S. Meunier, and P. Bischoff, Evaluation of the radioprotective potential of the polyphenol norbadione A, Lett. Drug Des. Discov. 9 (2012), pp. 48–53.

[56] A. Saija, M. Scalese, M. Lanza, D. Marzullo, F. Bonina, and F. Castelli, Flavonoids as antioxidant agents – importance of their interaction with biomembrane, Free Radical Bio. Med. 19 (1995), pp. 481–486.

[57] V.P. Menon and A.R. Sudheer, Antioxidant and anti-inflammatory properties of curcumin, Adv. Exp. Med. Biol. 595 (2007), pp. 105–125.

[58] S. Biljali, V.A. Hadjimitova, M.N. Topashka-Ancheva, D.B. Momkova, T.T. Traykov, and M.H. Karaivanova, Antioxidant and antiradical properties of esculin, and its effect in a model of episrubin-induced bone marrow toxicity, Folia Med (Plovdiv). 54 (2012), pp. 42–49.

[59] C.Y. Shaw, C.H. Chen, C.C. Hsu, C.C. Chen, and Y.C. Tsai, Antioxidant properties of scopoletin isolated from Sinomonium acutum, Phytother. Res. 17 (2003), pp. 823–825.

[60] Y. Yilmaz and R.T. Toledo, Major flavonoids in grape seeds and skins: Antioxidant capacity of catechin, epicatechin, and gallic acid, J. Agr. Food Chem. 52 (2004), pp. 255–260.

[61] H. Rudyk, S. Vasiljevic, R.M. Hennon, C.R. Birkett, J. Hope, and I.H. Gilbert, Screening Congo Red and its analogues for their ability to prevent the formation of PrP-res in scrapie-infected cells, J. Gen. Virol. 81 (2000), pp. 1155–1164.

[62] M.L. Bolognesi, H.N.A. Tran, M. Staderini, A. Monaco, A. Lopez-Cobenas, S. Bongarzone, X. Biarnes, P. Lopez-Alvardo, N. Cabezas, M. Caramelli, P. Carloni, J.C. Menendez, and G. Legname, Discovery of a class of diketopiperazines as antiprion compounds, Chem. Med. Chem. 5 (2010), pp. 1324–1334.

[63] M.S. Hung, Z. Xu, Y.C. Lin, J.H. Mao, C.T. Yang, P.J. Chang, D.M. Jablons, and L. You, Identification of hematein as a novel inhibitor of protein kinase CK2 from a natural product library, BMC Cancer. 9 (2009), p. 135.

[64] J.J. Hong, T.S. Jeong, J.H. Choi, J.H. Park, K.Y. Lee, Y.J. Seo, S.R. Oh, and G.T. Oh, Hematein inhibits tumor necrotic factor-alpha-induced vascular cell adhesion molecule-1 and NF-kappaB-dependent gene expression in human vascular endothelial cells, Biochem. Biophys. Res. Commun. 281 (2001), pp. 1127–1133.

[65] J.H. Choi, T.S. Jeong, D.Y. Kim, and Y.M. Kim, H.J. Na, K.H. Nam, S.B. Lee, H.C. Kim, S.R. Oh, Y. Choi, S.H. Bok, and G.T. Oh, Hematein inhibits atherosclerosis by inhibition of reactive oxygen generation and NF-kappaB-dependent inflammatory mediators in hyperlipidemic mice, J. Cardiovasc. Pharmacol. 42 (2003), pp. 287–295.

[66] G.T. Oh, J.H. Choi, J.J. Hong, D.Y. Kim, S.B. Lee, J.R. Kim, C.H. Lee, B.H. Hyun, S.R. Oh, S.H. Bok, and T.S. Jeong, Dietary hematein ameliorates fatty streak lesions in the rabbit by the possible mechanism of reducing VCAM-1 and MCP-1 expression, Atherosclerosis 159 (2001), pp. 17–26.

[67] D.A. Kocisko, G.S. Baron, R. Rubenstein, J. Chen, S. Kuizon, and B. Caughey, New inhibitors of scrapie-associated prion protein formation in a library of 2.000 drugs and natural products, J. Virol. 77 (2003), pp. 10288–10294.