Journal of Chromatography A 1672 (2022) 463048

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Artificial neural networks to model the enantioresolution of structurally unrelated neutral and basic compounds with cellulose tris(3,5-dimethylphenylcarbamate) chiral stationary phase and aqueous-acetonitrile mobile phases

Mireia Pérez-Baeza[a], Yolanda Martín-Biosca[a,**], Laura Escuder-Gilabert[a], María José Medina-Hernández[a], Salvador Sagrado[a,b,*]

[a] Departamento de Química Analítica, Universitat de València E- 46100 Burjassot, Valencia, Spain
[b] Instituto Interuniversitario de Investigación de Reconocimiento Molecular y Desarrollo Tecnológico (IDM), Universitat Politècnica de València, Universitatde València, Valencia, Spain

ARTICLE INFO

Article history:
Received 28 February 2022
Revised 5 April 2022
Accepted 8 April 2022
Available online 11 April 2022

Keywords:
Enantioresolution
Cellulose chiral stationary phase
Heterogeneous dataset
Artificial neural networks
Feature selection
Relative variable importance

ABSTRACT

Artificial neural networks (ANN; feed-forward mode) are used to quantitatively estimate the enantioresolution (*Rs*) in cellulose tris(3,5-dimethylphenylcarbamate) of chiral molecules from their structural information. To the best of our knowledge, for the first time, a dataset of structurally unrelated compounds is modelled using ANN, attempting to approach a model of general applicability. After setting a strategy compatible with the data complexity and their relatively limited size (56 molecules), by prefixing initial ANN inner weights and the validation and cross-validation subsets, the ANN optimisation based on a novel quality indicator calculated from 9 ANN outputs allows selecting a proper (predictive) ANN architecture (a single hidden layer of 7 neurons) and performing a forward-stepwise feature selection process (8 variables are selected). Such relatively simple ANN offers reasonable good general performance in predicting *Rs* (e.g. validation plot statistics: mean squared error = 0.047 and R = 0.98 and 0.92, for all or just the validation molecules, respectively). Finally, a study of the relative importance of the selected variables, combining the estimation from two approaches, suggests that the surface tension (positive overall contribution to *Rs*) and the –NHR groups (negative overall contribution to *Rs*) are found to be the main variables explaining the enantioresolution in the current conditions.

1. Introduction

Chiral stationary phases (CSPs) based on amylose and cellulose polysaccharide derivatives are by far the preferred choice for the enantioseparation of chiral compounds in high-performance liquid chromatography (HPLC). Besides the wide availability of commercially CSPs, their broad chiral recognition capacity enables the separation of the enantiomers of a high variety of compounds [1]. When using polysaccharide-based CSPs in HPLC, separations can be carried out in normal phase (NPLC), in reversed phase (RPLC) as well as in polar organic mode [1]. For the analytical scale, RPLC has some advantages in the analysis of pharmaceuticals and aqueous biological samples and coupling with mass spectrometry detection [2].

There are numerous studies on the fundamentals of enantioseparation mechanisms involving polysaccharide-based CSPs [3–13]. Spectroscopic methods and molecular modelling approaches have contributed to elucidating the chiral recognition mechanism [3–13]. Despite the insight provided by these studies, finding the right CSP/mobile phase (MP) combination that allows the separation of a given pair of enantiomers remains a very challenging task. Different strategies have been proposed for this purpose. The most common consists of screening a series of CSP/MP combinations, a trial-and-error approach that often requires a considerable experimental and economical effort [14–20]. Alternatively, quantitative structure-enantioselective retention relationships (QSERR), which

* Corresponding author.
** Corresponding author.
E-mail addresses: yolanda.martin@uv.es (Y. Martín-Biosca), sagrado@uv.es (S. Sagrado).

https://doi.org/10.1016/j.chroma.2022.463048

correlate relevant separation parameters with molecular descriptors derived from molecular structure, have emerged as a useful strategy to select a suitable CSP/MP combination in chiral RPLC optimisation processes [21,22].

For polysaccharide-based CSPs, QSERR have been built using chemometric tools, like multiple linear regression (MLR) [23–26], partial least squares regression (PLS) [25,27] or random forest approach [28,29]. These models, most of them constructed using chromatographic data of structurally related compounds, allow the prediction of retention times [25], logarithm of the retention factor (log k) [22,26,27], separation factor α [28,29] and log α [26–28].

In previous papers, our research group modelled the enantioresolution (as a categorical variable) of structurally unrelated compounds in some commercial cellulose- and amylose-based stationary phases using data in RPLC mode [30,31]. 56 structural variables, including chiral topological parameters, molecular descriptors and molecular topological parameters were used. Discriminant partial least squares for one categorical response variable (DPLS1) was used for feature selection and modelling purposes. The models obtained for the different CSPs presented adequate predictive and descriptive capacity. The descriptive ability (importance – contribution– of structural variables to explain enantioresolution) was explored in two stages: (i) feature selection, FS, (eliminating unnecessary variables) and (ii) calculation of the normalised regression coefficients of variables.

Artificial neural networks (ANN) methodology is a flexible machine learning (artificial intelligence) approach capable of modelling complex/non-linear relationships between input and response variable(s) [32,33]. ANN were designed to learn from a training dataset (until the predictive ability of a different validation dataset improves) by using a neural architecture that consists of interconnected neurons arranged in several layers. In the so-called feed-forward modality, the architecture comprises an input layer (independent variables), one or more hidden layers, and an output layer (response variable(s)). The ANN behaviour depends on the weights of the connections. During ANN training, these weights are optimised using an iterative algorithm to obtain a final model that provides the minimal prediction error on the response variable(s) [34].

However, due to the inner complexity of ANN, it is difficult to extract information on how models work to predict the output from input variables (in ANN, the difficulty in establishing the relationship between input-output variables is commonly referred to as a black box). Different approaches have been developed to approximate the relative importance of the input variables to predict the response, although all of them have limitations [34]. For instance, the approach from Olden et al. comprises the sum of the product of the inner weights connecting the input-response variables through the hidden layer [35] but is sensitive to the neural architecture and the initial starting inner weights [34]. The Lek et al. approach changes the value of one input variable while maintaining the other variables at a constant value (at different quantiles) [36], but it could provide different data compared to those used for training [34]. However, we have found no evidence on which approach could provide a correct estimation of the importance of variables, due to the lack of a way to verify the estimation quality up to now.

After an exhaustive search, only three ANN applications to predict enantioselective parameters in chiral liquid chromatography have been found [37–39]. All these QSERR studies have been carried out for a limited number of structurally related compounds using macrocyclic antibiotics [37], cellulose [38] and brush-type [39] CSPs. The homogeneity (simplicity) of the data sets used probably allows performing standard conditions for ANN (that normally applies to big datasets) as well as a comparison with a conventional regression model.

The main objectives of this work can be summarised as designing a global strategy, based entirely on ANN, able to select a (preferably short) set of structural variables capable of satisfactorily estimating enantioresolution in cellulose tris(3,5-dimethylphenylcarbamate) chiral stationary phase and aqueous-acetonitrile mobile phases, working with a relatively limited dataset of 56 compounds from 14 different chemical families. An additional aim is to approximate the relative importance of the selected variables to quantitatively predict *Rs*. To the best of our knowledge, ANN have not been used to directly predict the enantioresolution values (*Rs*) of structurally unrelated chemical compounds nor to perform a selection of the more relevant structural parameters describing *Rs* in such a complex situation due to the heterogeneity of the dataset.

2. Experimental

2.1. ANN nomenclature

Fig. 1 depicts the scheme of ANN used in this work. For all compounds, the known structural variables (**X**) are set as the input layer and the predicted *Rs* values (**y**) are the output layer (to be compared with the known experimental *Rs* data (**t**) during the training process). In the middle, we arranged up to 2 hidden layers and tried up to 30 neurons in each layer; so, a total of 930 ANN were evaluated. Each ANN architecture was identified by *Nv*-[*k*-*kk*]−1, where Nv is the number of structural variables and k and kk are the number of neurons in the first and second hidden layer, respectively. In the case of only one hidden layer, *kk*=0, the ANN were identified by *Nv*-[*k*]−1. Note that *k* and *kk* should be optimised, that *Nv* could be reduced after a feature selection process, and finally, from the ANN information (e.g. W_i, **y**) the relative importance of the *Nv* input variables could be approximated.

2.2. Structural and chromatographic data

To study the structure-enantioresolution relationships for each compound under study, 56 structural variables (**X**) were used. These variables include chiral carbon-related parameters, molecular descriptors and topological parameters predicted by software (ChemSpider Database) [40], and calculated hydrophobicity parameters as the octanol-water partition coefficient (log *P*) and the apparent log *P* at pH = 8 (log *D*) [30]. These structural variables were described in detail in previous papers [30,41,42]. Briefly, variables $\mathbf{x_1}$ to $\mathbf{x_7}$ correspond to chiral carbon (*C**) related parameters (e.g. number of aromatic heterocycles groups bonded to the chiral carbon (*C***hA*) as $\mathbf{x_3}$); variables $\mathbf{x_8}$ to $\mathbf{x_{12}}$ correspond to predicted molecular descriptors from ACD/Labs and ChemAxon calculations [40] (e.g. surface tension (*ST*, dyn•cm^{-1}) as $\mathbf{x_{12}}$); variables $\mathbf{x_{13}}$ to $\mathbf{x_{29}}$ correspond to molecular topological parameters predicted by ChemAxon (e.g. Balaban index (*Bi*) and number of sp^3 hybridized carbons divided by the total carbon count (fsp^3) as $\mathbf{x_{28}}$ and $\mathbf{x_{29}}$, respectively), variables $\mathbf{x_{30}}$ to $\mathbf{x_{54}}$ are obtained as the count of atoms/groups present in the compound (e.g. number of -NHR and -NR$_2$ amino groups in the molecule (*NHR, NR*$_2$), number of R-O-R ether groups in the molecule (*ROR*) and number of - NHCOR amide groups bonded to an aromatic ring in the molecule (*ANHCOR*) as $\mathbf{x_{36}}$, $\mathbf{x_{37}}$, $\mathbf{x_{38}}$ and $\mathbf{x_{46}}$, respectively), while variables $\mathbf{x_{55-56}}$ are the hydrophobicity parameters.

Experimental *Rs* values of 56 structurally unrelated chemical compounds under study were obtained as described in [31] using the chiral column Lux Cellulose-1 (cellulose tris(3,5-dimethylphenylcarbamate) and mobile phases consisting of binary mixtures of ammonium bicarbonate buffer (5 mM, pH 8) and acetonitrile (ACN) in varying proportions (10–98% in volume of ACN).

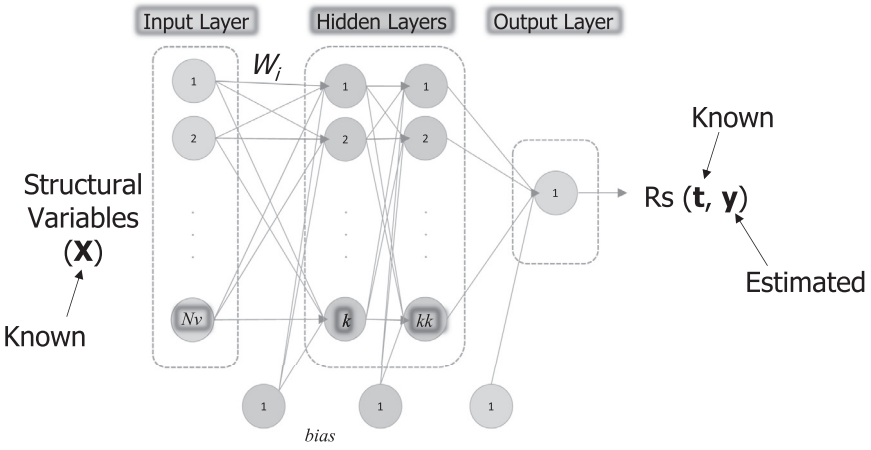

Fig. 1. Architecture of the artificial neural networks (ANN) in this work. The known structural variables (**X**) are set in the *Nv* neurons of the input layer. The resolution (*Rs*) estimated values (**y**) will be obtained in a single neuron (output layer). In the middle, the so-called hidden layers must be arranged, defining the complete architecture. ANN architectures assayed (930) include up to 2 hidden layers and up to 30 neurons (*k, kk*) in each layer. Sigmoid (hidden layers) and linear (output layer) activation functions are used. The weights (W_i; connecting all the neurons, including those side neurons introducing bias) are adjusted during the learning process, minimising the differences between **y** and the known experimental *Rs* values (**t**). Once an ANN has been trained is able to estimate **y**-values for novel compounds.

For each compound, amongst the various *Rs* values obtained using the different mobile phases tested, the maximum *Rs* value was chosen to be used as the response variable (**t**). Other separations conditions were: flow-rate, 1 mL•min^{-1} and temperature, 25 °C. These *Rs* values are shown in Table 1. Both **X** and **t** were autoscaled as pretreatment.

2.3. Software and calculations

The global approach developed in this paper includes three automatic strategies based on ANN outputs: (i) architecture optimisation, (ii) feature selection (FS) and (iii) approximation of the relative importance of variables. The first two strategies involve the comparison of the results from different ANN. To ensure a reliable ANN comparison, all aspects that could lead to different outputs due to the limited number of compounds, combined with the presumable complexity of the **t**-**X** relationship, were prefixed in ANN calculations. So, all the inner weights (W_i in Fig. 1) were set to one, and the same subsets of compounds (training, validation and cross-validation) were used in all ANN. The validation subset for overfitting control includes 7 compounds covering a representative range of *Rs* values (see Table 1). The rest of the 49 compounds corresponds to the training block (for learning). Additionally, 9 sets of 5 different training compounds were used in a cross-validation process implemented to improve robustness. Training compounds in these cross-validation subsets were punctually converted into test compounds, i.e., they did not participate in the training nor the validation task and were just predicted by the ANN as external compounds.

To compare the performance of any ANN tested, a quality indicator *Q* (Eq. (1)) was introduced. *Q* depends on 9 different statistics related to the differences between the estimated (**y**) and experimental (**t**) *Rs* values, in terms of mean squared errors (*MSE*) and regression statistics of the validation plot (correlation coefficient, *R*; slope b_1, and intercept, b_0), for all, training (subscript *train*) and validation (subscript *v*) compounds. *MSE* statistics were 'penalised' (identified using *p* before the statistic), adding to the mean the standard deviation, both estimated from 10 ANN outputs (one with 49 training compounds plus 9 cross-validation runs).

$$Q = pMSE + pMSE_v + W1|pMSE_v - pMSE_{train}| + W2((1-R) + |1-b_1| + |b_0|) + W3((1-R_v) + |1-b_{1v}| + |b_{0v}|) \quad (1)$$

The term $|pMSE_v - pMSE_{train}|$ characterises the (penalised) overfitting degree. Note that penalised values aim to take decisions in ANN comparisons including both the estimation of external compounds (test) as well as the uncertainty arising from the different cross-validated outputs. Note also that the regression terms in Eq. (1) were set in a format where lower values mean better quality (as *MSE* does intrinsically); thus, lower *Q* values imply higher ANN quality. The *Q*-weights (*W1, W2* and *W3*) in Eq. (1) were set at values 1, 10 and 5, respectively, to provide an almost similar contribution of all the terms in Eq. 1.

Algorithms were written in MATLAB® R2019a (Mathworks®) [43]. MATLAB has a Toolbox containing algorithms and tools for creating and training artificial neural networks, called Deep Learning Toolbox. The programs used in this work include routines provided by this Toolbox: feedforwardnet.m, which generates a feedforward (regression-focused) ANN from a defined [*k-kk*] hidden layers architecture; train.m, that trains a shallow ANN from given **X**-**t** data and architecture and net.m, which predicts response values (**y**) from a trained ANN. To carry out the different strategies, novel algorithms were also developed, such as fANN.m, a MATLAB function that builds and trains an ANN from **X**-**t** data with a defined [*k-kk*] hidden layers architecture by prefixing validation and test blocks and initial W_i and, optionally, eliminates variables or objects (the last allows to perform cross-validation studies); StuANN.m, which numerically and graphically compares *Q* values (and their parameters) of different ANN architectures up to *k* and *kk* neurons in the hidden layers.

The ANN final architecture was chosen from the 930 studied, in two steps. First, selecting those ANN showing $R > 0.9$ (filter-1), and from them, those whose *Q* values were below the 10th percentile (filter-2). Finally, a visual inspection of the validation plots (**y** vs. **t**) and the statistics of Eq. (1) allowed to select the definitive architecture, preferably one with a lower *k* x *kk* value (see Fig. 1), assuming that simpler ANN, with a lower number of connecting coefficients (W_i) should be more robust.

The optimal ANN architecture was chosen to conduct the subsequent FS process. The function sequentialsfs.m (part of the Machine Learning/Artificial Intelligence suite of MATLAB®) was employed. This function uses a sequential strategy (i.e. adds or removes one variable in each run). In this case, a forward mode (adding variables) was used. As before, the variables selected in

Table 1
Compounds of different families, with their maximum enantioresolution values (*Rs*) obtained in cellulose tris(3,5-dimethylphenylcarbamate) CSP using aqueous-acetonitrile mobile phases. Compounds are identified by their name and numbered order (*N*). Validation set compounds (indicated by +) and test compounds (in 9 cross-validation subsets indicated by its number in parenthesis) are shown.

N	Family[a]	Compound	*Rs*[b]	Validation	Test
1	AAD	Disopyramide	0.0		(1)
2	AAD	Mexiletine	0.0		(2)
3	AAD	Propafenone	0.0		(3)
4	ACO	Warfarin	0.0	+	
5	AD	Nomifensine	1.8		(4)
6	AD	Citalopram	0.0		(4)
7	AD	Fluoxetine	0.0		(5)
8	AD	Viloxazine	0.5		(3)
9	AD	Trimipramine	0.7		(7)
10	AD	Bupropion	0.0		(6)
11	AD	Mianserin	1.8	+	
12	AF	Benalaxyl	1.2		(2)
13	AF	Imazalil	0.6		(5)
14	AF	Penconazole	0.9		(1)
15	AF	Hexaconazole	2.5		(7)
16	AF	Myclobutanil	3.0		(8)
17	AF	Metalaxyl	5.5		
18	AH	Doxylamine	0.0		(7)
19	AH	Brompheniramine	0.0		(8)
20	AH	Chlorpheniramine	0.0		
21	AH	Orphenadrine	2.3		(6)
22	AH	Carbinoxamine	0.0		(9)
23	AH	Hydroxyzine	1.8		(5)
24	AH	Terfenadine	0.5		(4)
25	AH	Cetirizine	0.0		(1)
26	AH	Fexofenadine	0.0		(2)
27	LA	Mepivacaine	0.0		(3)
28	LA	Propanocaine	0.0		(4)
29	LA	Prilocaine	0.4	+	
30	LA	Bupivacaine	0.0		(5)
31	ANP	Aminoglutethimide	2.7	+	
32	ANP	Bicalutamide	0.0		(6)
33	BB	Pindolol	1.6		(3)
34	BB	Propranolol	0.0		(7)
35	BB	Metoprolol	0.0		(8)
36	BB	Acebutolol	0.0		(9)
37	BB	Atenolol	0.0		
38	BB+BD	Salbutamol	0.0		
39	BB	Timolol	0.0		
40	BD	Bambuterol	0.0		(9)
41	BD	Isoprenaline	0.5	+	
42	BD	Orciprenaline	0.0		(2)
43	BD	Clenbuterol	0.3		(1)
44	BD	Terbutaline	0.0		(1)
45	CaB	Verapamil	0.7		(8)
46	CaB	Felodipine	0.0		(2)
47	CaB	Cilnidipine	0.7	+	
48	ACD	Procyclidine	0.0		(3)
49	APD	Trimeprazine	0.0		(4)
50	APD	Ethopropazine	0.0	+	
51	APD	Promethazine	0.0		(6)
52	APD	Thioridazine	0.0		(7)
53	PPI	Pantoprazol	0.0		(8)
54	ANAL	Methadone	0.0		(9)
55	PPI	Rabeprazol	0.7		(9)
56	PPI	Lansoprazol	0.6		(6)

[a] Drug/Pesticide families: AAD (antiarrhythmic drugs), ACO (anticoagulants), AD (antidepressants), AF (antifungals), AH (antihistamines), LA (local anaesthetics), ANP (antineoplastics), BB (β-blockers), BD (bronchodilators), CaB (calcium channel blockers), ACD (anticholinergic drugs), APD (antipsychotic drugs), PPI (proton pump inhibitors), and ANAL (analgesics).

[b] Maximum *Rs* value obtained using the mobile phases tested consisting of binary mixtures of ammonium bicarbonate buffer (5 mM, pH 8) and acetonitrile (ACN) in varying proportions (10–98% in volume of ACN).

each run were those that provided the lowest *Q*. Since FS is based on ANN outputs, the whole process can be called ANN-FS process.

To approximate the relative importance of the selected variables two approaches adapted from [35,36] were used. They are based on the use of connecting weights (ANNcw) and **y**-estimates to calculate the variable effect (ANNef). Since there is no literature evidence on which approach could provide correct (and the best) estimations, we used the mean and standard deviation of the outputs from these two approaches. Both approaches include as a first step the estimation of a vector of regression-like coefficients (**b;** one coefficient per variable as occurs e.g., in PLS). In the ANNcw case, **b** was calculated by multiplying the W_i-inner matrices (hidden-output layers and input-hidden layers) calculated by the ANN algorithm. In the ANNef case, **b** was estimated, creating two 'artificial extreme molecules' per **X** variable, using the maximum and minimum values of this variable and the mean value for the rest of the variables. The half difference between **y**-estimations associated with the 'extreme molecules' was an estimation of **b**. Finally, the **b**-vector was normalised with respect to its maximum value ($b_{\max}$) as:

$$\mathbf{I} = \mathbf{b}/b_{max} \tag{2}$$

We calculated the **I** vectors using Eq. (2) for both approaches. Then, the mean **I** and the standard deviation were used as an approximation of the relative importance of the variables (including the sign) on the *Rs* estimation.

3. Results and discussion

3.1. Previous results and data complexity

The use of a PLS model for predicting *Rs* values did not provide satisfactory results. Only by dividing the compounds into two categories (resolved and non-resolved) and using a discriminant approach (DPLS1 model) adequate predicted categories were found (results not shown). In our opinion, the data complexity is too high to be solved by this model (despite that, in general, PLS can deal with moderate non-linear problems). This was the main reason to try ANN to reach a quantitative level of prediction.

ANN is intrinsically less robust than PLS (since much more weights must be estimated). Moreover, ANN outputs can vary depending on the complexity of the ANN architecture, but also on the initial W_i values and validation and test subsets, mainly when the number of objects (here compounds) is limited. Thus, fixing the initial W_i values (Fig. 1) and pre-fixing the validation (see Table 1), as well as the cross-validation subsets (see Table 1), was necessary for consistently comparing the ANN results (i.e., architecture comparison and FS-ANN process). An extra element of complexity is the fact that the *Rs* values present a very asymmetric distribution (e.g. 62% of the *Rs* values are 0). Consequently, a random selection of subsets could be inadequate. As can be seen in Table 1, the validation set contains *Rs* values ($\boldsymbol{t}$ = 0, 0, 0.4, 0.5, 0.7, 1.8 and 2.7), while the entire data set ranges between 0 and 4.86.

3.2. ANN architecture optimisation

ANN architecture optimisation was carried out according to the procedure described in the experimental section based on the application of two consecutive filters. Fig. 2 shows the quality values (*Q*; Eq. (1)) for the 930 architectures assayed. As it can be seen, the ANN architecture has a notable influence on the quality of the ANN outputs. In the plot, dots show the ANN architectures that pass filter-1 ($R > 0.9$), while circles indicate which of these also pass filter-2 (*Q* below 10th percentile). This last group exhibit similar *Q*-values, close to the target line ($Q = 0$). The simplest architecture

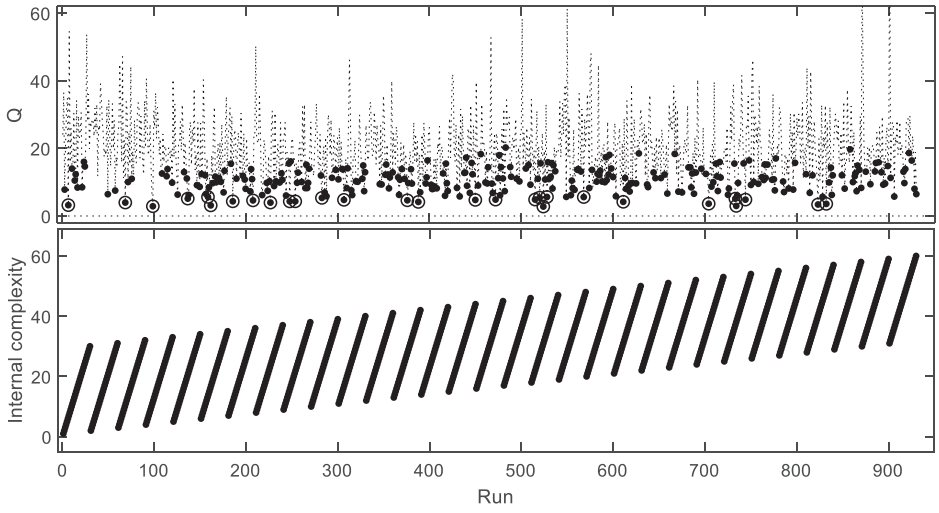

Fig. 2. Quality (Q; upper part) values (Eq. (1)) through the ANN architecture study, where the runs correspond to different ANN architectures obtained by increasing the number of neurons in the first hidden layer (k, from 1 to 30; See Fig. 1) for a given number of neurons in the second hidden layer (kk, increasing from 0 (no layer) to 30, See Fig. 1). Thus, runs 1–30 correspond to ANN with a hidden layer with k from 1 to 30 neurons; runs 31–60 values correspond to ANN with two hidden layers with $kk = 1$ with k from 1 to 30 neurons, and so on. Internal ANN complexity (number of hidden neurons, $k + kk$, units; lower part) of each run. Symbols: dotted line = Q values for the 930 architectures assayed; dot = ANN architectures that pass filter-1 ($R > 0.9$). Circle = ANN architectures that pass both filter-1 and filter-2 (Q below 10th percentile). $Q = 0$ (target line) is also included as a horizontal dotted line.

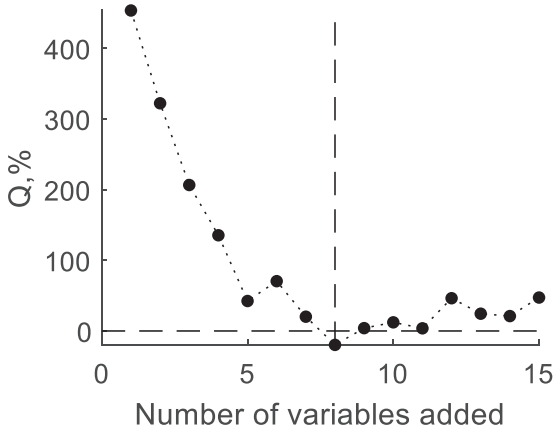

Fig. 3. Quality (Q) values (Eq. (1)), in percentage respect to the Q value obtained with the initial 56 variables (horizontal dotted line; $Q = 0\%$), through the ANN-based feature selection (FS-ANN) process, where input variables are sequentially added. The vertical dashed line indicates the number of variables added that provides a $Q < 0\%$.

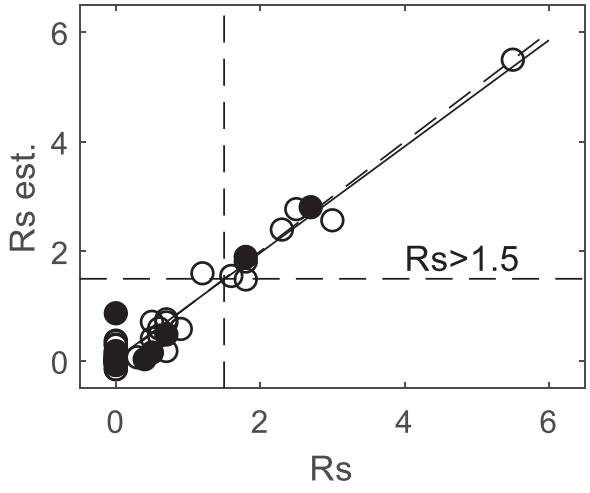

Fig. 4. *Rs* validation plot of the ANN: 8-[7]−1. Selected variables are: x_3, *C*hA*; x_{12}, *ST*; x_{28}, *Bi*; x_{29}, fsp^3; x_{36}, *NHR*; x_{37}, *NR2*; x_{38}, *ROR* and, x_{46}, *ANHCOR* (See Section 2.2), including all the compounds (empty circles). Validation compounds are highlighted using full circles. Vertical and horizontal dashed lines indicate $Rs = 1.5$. The diagonal dashed line reflects the target line while the solid line is the regression line.

fitting the two filters corresponds to run 7 ($Q = 3.2$). This architecture consists of 7 neurons in a single hidden layer (56-[7]−1) and was selected and used in the next step (FS-ANN process). The minimum Q value (2.8) was found at run 524, which corresponds to an ANN with two-hidden layers (56-[14 17]−1). However, as could be expected, this more complex ANN showed a higher penalised overfitting (0.49; i.e. extremely good prediction for training but worse prediction of the validation set) than the selected one (penalised overfitting 0.13). Thus, this ANN was not considered for the next step (feature selection).

3.3. Feature selection

The FS-ANN process, applied to the previously chosen ANN, aimed to reduce the current Nv value (56 variables) to (i) improve the current performance and (ii) be a first stage in defining the importance of variables by eliminating the unnecessary ones. Fig. 3 shows the evolution of the FS-ANN process, as variables are added (forward mode, 1 variable added in each run) by means of the Q-values (Eq. (1)), calculated in percentage respect the Q value obtained using all the variables (previous section). After adding only 8 variables we observed a Q value lower than the initial one obtained with 56 variables (a satisfactory result). Since this fact fitted the main aim of this work, i.e., to select a reduced set of variables able to estimate the enantioresolution, we stopped the search for a higher number of variables at this point. Thus, the final architecture becomes 8-[7]−1. The selected variables are: x_3, *C*hA*; x_{12}, *ST*; x_{28}, *Bi*; x_{29}, fsp^3; x_{36}, *NHR*; x_{37}, *NR2*; x_{38}, *ROR* and, x_{46}, *ANHCOR* (see Section 2.2).

3.4. Performance of the ANN (predictive ability)

Fig. 4 shows the validation plot (**y** vs. **t**; rescaled values representing *Rs*) obtained with ANN 8-[7]−1 for all compounds (the validation set is highlighted by full circles). As can be observed, a good agreement between experimental and predicted values was

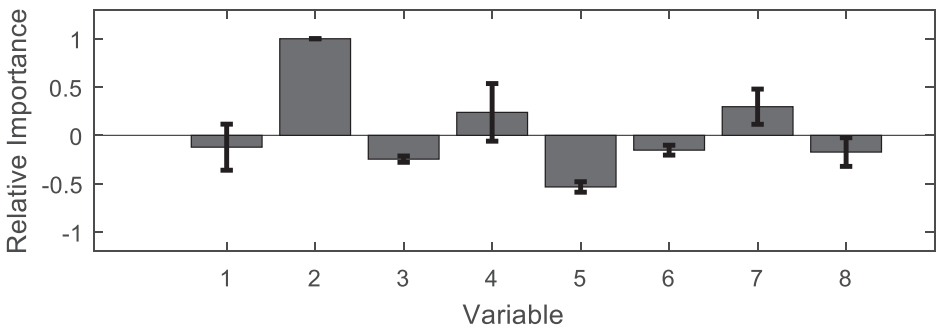

Fig. 5. Relative importance of selected variables (**I**; Eq. (2)) to estimate *Rs* of the ANN: 8-[7]−1. Selected variables in the order: 1-x_3, *C*hA*; 2-x_{12}, *ST*; 3-x_{28}, *Bi*; 4-x_{29}, fsp^3; 5-x_{36}, *NHR*; 6-x_{37}, *NR2*; 7-x_{38}, *ROR* and 8-x_{46}, *ANHCOR*. Data include the mean (bar) and standard deviation (error bar) of two strategies (ANNcw, based on inner connecting weights and ANNef, based on the variable effect).

obtained, according to the low *Q* value found for this relatively simple architecture. Only for one compound, with experimental $Rs = 0$, a low *Rs* estimate is observed, still sufficiently below the $Rs = 1.5$ critical line (indicating full enantioresolution). A good agreement is also visible between the regression line and the target (diagonal) line of the validation plot. These observations fit well with the first aim fixed for the present work since a relatively simple ANN (architecture, few structural variables) becomes capable of satisfactorily estimating enantioresolution in the current conditions.

On the other hand, to check the final ANN performance, it is advisable to use not only a validation subset (mandatory) but also a test (external) subset. When using a big dataset (the conventional situation in ANN models), this option normally does not result in a loss of representativeness of the remaining training data to build the ANN (due to the large number of cases). However, if the amount of data available is not high (as in many situations in the scientific context), separating an extra test subset (probably with particular structural information) would likely affect the representativeness in the learning process, providing a less general ANN.

3.5. Importance of variables

The importance of the variables (also known as sensitivity analysis) for describing the *Rs* data is an important issue in all the QSERR models. Fig. 5 shows an approximation of the relative importance (**I**; Eq. (2)) of the 8 selected variables in predicting the enantioresolution. **I** is the mean of the results of two adapted/simplified approaches, ANNcw and ANNef [35,36]. **I** absolute magnitude and sign approximate the importance of the variable and the positive or negative overall contribution to *Rs*, respectively. It must be pointed out that **I** values do not act as real regression coefficients. On the other hand, all the selected variables are necessary to obtain the *Rs*-estimations (Fig. 4), however, not all should be equally important (Fig. 5).

The standard deviation is an indication of the uncertainty arising from the two different approaches. The uncertainty for some variables was low, which could confer more reliability of their **I**-estimations. Fig. 5 results suggest that the contribution of variables $\mathbf{x_3}$ (*C*hA*) and $\mathbf{x_{29}}$ (fsp^3) is unclear since the error bar is higher than their magnitude. On the other hand, $\mathbf{x_{12}}$ (*ST*) is the most important, with a global positive effect on the enantioresolution. The second important variable is $\mathbf{x_{36}}$ (*NHR*), which should globally inversely affect *Rs*. Other less important variables are: $\mathbf{x_{38}}$ (*ROR*), positive overall contribution to *Rs*, and $\mathbf{x_{28}}$ (*Bi*), $\mathbf{x_{37}}$ (*NR2*) and $\mathbf{x_{46}}$ (*ANHCOR*), negative overall contribution to *Rs*.

For a given compound, the balance between positive and negative contributions determines enantioresolution in that CSP. The results suggest that the presence of dipoles (variable *ST*) and ether groups (variable *ROR*) in the molecule (e.g. compounds $N = 13$, 17, 21, 22, 23, 25, 33, 35, 47, 55; Table 1) favours enantioresolution, which indicates the importance of van der Waals forces and hydrogen-bond interactions with the phenylcarbamate moiety in the CSP on enantiomer separation.

On the other hand, high steric factors (high values of Babalan index), together with the presence of secondary ($N = 3$, 7, 10, 29, 33–44), tertiary amines ($N = 1$, 6, 9, 18–22, 28, 45, 49–51, 54) and amide groups ($N = 27$, 29, 30, 32, 36) in the molecule decrease enantioresolution. As can be observed in Table 1, this second group of compounds, except $N = 21$ and 33, show null or partial resolution. However, both compounds present ether groups in their molecules and are fully enantioresolved.

3.6. Additional remarks

Since this work constitutes the first attempt to obtain ANN models to estimate *Rs* of a heterogeneous group of chiral compounds, we focused on adapting conventional ANN strategies to the current complex problem. Considering the introductory context of this work, we opted for simple fit-for-purpose strategies instead of more sophisticated ones. For instance, in a problem with a large number of input variables, an independent criterion to perform a previous feature selection is a logical strategy. However, in our case, in addition to the limited number of variables, we were interested in performing the feature selection based on the ANN outputs themselves (as an alternative to other approaches described [44]); thus, we preferred to initiate the FS-ANN process with a suitable ANN (architecture optimized; note that many architectures provide inadequate outputs; probably connected with the complexity of the current data; Fig. 2). Then, in FS, variables were added progressively until an improved ANN with a reduced number of variables was obtained (8 input variables; Fig. 3), which also facilitates establishing the relative importance of variables. On the other hand, after the ANN-FS process, to verify the consistency of the previous optimized architecture ($k = 7$), an additional (focused) architecture optimization was performed, now using the 8 input variables selected (probing architectures with *k* from 1 to 7 combined with *kk* from 0 to 3). The results were comparable to those obtained previously with 56 variables, and again, a single hidden layer of 7 neurons provided the best *Q* value, which supports the validity of our strategy.

In the same context, there are more sophisticated strategies proposed for the data division on subsets, e.g. those based on a clustering approach [44], that uses both descriptors and the response variable to divide compounds into subsets. A k-means neighbour clustering approach was used to separate 7 subgroups of compounds using the autoscaled **y** and **X** variables, followed by a

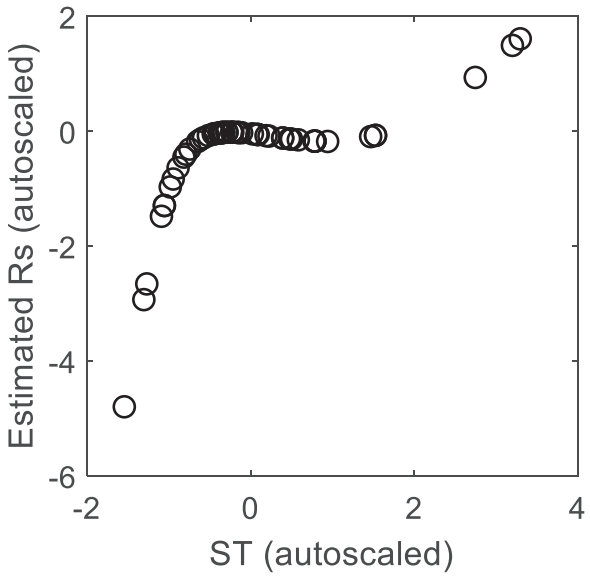

Fig. 6. Estimated autoscaled *Rs* response curve for variable $\mathbf{x}_{12}$ (Surface tension; *ST*). The autoscaled *ST* values are used to derive the curve, keeping the rest of the autoscaled variables in Fig. 5 in their mean value.

principal component analysis (PCA). The main (PC1-PC2) PCA plots were evaluated. The PCA score plot showed rough segregation between the clusters. On the other hand, the *Rs* variable was located close to the origin of coordinates in the PCA loading plot (negligible contribution), suggesting that it is not well represented by the clusters. Thus, taking into account the extreme asymmetry of the *Rs* data, we have prioritized the selection of the subsets based only on the response variable, for simplicity and representativeness of the current available *Rs*-space.

One of the approaches used in Section 3.5 (ANNef approach) for evaluating the impact of a single descriptor on the response variable uses the extreme values of the descriptor understudy. A more sophisticated strategy, involving the whole range under examination, provides a response curve [44]; broadening the ANNef approach. In this work, five of the selected variables in Fig. 5 are binary (so a response curve can not be obtained) and only variables 2–4 are continuous. For the most important variable, *ST*, a response curve was obtained (Fig. 6). As it can be observed, the autoscaled *Rs*-*ST* data corresponding to the current ANN model show a complex-non linear relationship, which indicates (in the extremes of the *ST* data) the overall positive contribution of *ST* to *Rs* already shown in Fig. 5.

Conclusion

For the first time (to the best of our knowledge), ANN have been used to quantitatively estimate the enantioresolution (*Rs*) of a heterogeneous dataset of chiral molecules from their structural information. Furthermore, the relative importance of a reduced number of predicting variables was established. The strategy has been designed for consistently comparing different ANN (architecture, feature selection) for a relatively short dataset (here 56 compounds), which implies prefixing some parameters (ANN initial inner weights, validation and cross-validation subsets) to strengthen the decision-making process.

The proposed strategy has been tested in the prediction of enantioresolution values of the 56 chiral compounds from 14 different chemical families chromatographed in cellulose tris(3,5-dimethylphenylcarbamate) and aqueous-acetonitrile mobile phases, from selected structural information. The entire approach only uses outputs provided by ANN. The key of the ANN comparison strategy is a novel 9-parameters quality indicator (*Q*) which relies on penalised (by the cross-validated uncertainty) *MSE* statistics (including an overfitting level) as well as on regression statistics.

After selecting the optimum architecture (first ANN comparison stage), it was used to reduce the **X** dimension (sequential FS-ANN approach; second ANN comparison stage). The final ANN is ready to: (i) estimate further enantioresolution data for novel non-chromatographed compounds and (ii) establish the relative importance of the selected variables. For the current dataset used, a simple architecture of 8 structural variables input layer and 7 neurons in a single hidden layer previous to the response variable layer (8-[7]–1), offers good general performance (e.g. $MSE = 0.047$ or $R = 0.98$ and 0.92, for all or just the validation compounds, respectively) in predicting *Rs*. The analysis of the relative importance of the selected variables in this final ANN suggests that the surface tension (positive overall contribution to *Rs*) and the groups –NHR in the molecule (negative overall contribution to *Rs*) are the most relevant for enantioresolution prediction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Mireia Pérez-Baeza: Investigation, Validation, Visualization, Formal analysis, Writing – original draft. **Yolanda Martín-Biosca:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – original draft, Writing – review & editing. **Laura Escuder-Gilabert:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – original draft, Writing – review & editing. **María José Medina-Hernández:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Salvador Sagrado:** Conceptualization, Methodology, Supervision, Formal analysis, Writing – original draft, Writing – review & editing, Software, Data curation.

Acknowledgments

Mireia Pérez Baeza is grateful to the Generalitat Valenciana and the European Social Fund for the financial support (ACIF/2019/158 research contract).

References

[1] B. Chankvetadze, Polysaccharide-Based Chiral Stationary Phases for Enantioseparations by High-Performance Liquid Chromatography: an Overview, in: G.K.E. Scriba (Ed.), Chiral Separations, Humana, New York, NY, 2019 Methods in Molecular Biology, vol 1985.
[2] A. Tarafder, L. Miller, Chiral chromatography method screening strategies: past, present and future, J. Chromatogr. A 1638 (2021) 461878.
[3] G.K.E. Scriba, Chiral recognition in separation sciences. Part I: polysaccharide and cyclodextrin selectors, Trends Anal. Chem. 120 (2019) 115639.
[4] B. Chankvetadze, Recent developments on polysaccharide-based chiral stationary phases for liquid-phase separation of enantiomers, J. Chromatogr. A 1269 (2012) 26–51.
[5] B. Cerra, A. Macchiarulo, A. Carotti, E. Camaioni, I. Varfaj, R. Sardella, A. Gioiello, Enantioselective HPLC Analysis to Assist the Chemical Exploration of Chiral Imidazolines, Molecules 25 (2020) 640–651.
[6] M. Lämmerhofer, Chiral recognition by enantioselective liquid chromatography: mechanisms and modern chiral stationary phases, J. Chromatogr. A 1217 (2010) 814–856.
[7] G.K.E. Scriba, Chiral recognition mechanisms in analytical separation sciences, Chromatographia 75 (2012) 815–838.
[8] J. Shen, Y. Okamoto, Efficient separation of enantiomers using stereoregular chiral polymers, Chem. Rev. 116 (2016) 1094–1138.
[9] Y. Okamoto, T. Ikai, Chiral HPLC for efficient resolution of enantiomers, Chem. Soc. Rev 37 (2008) 2593–2608.

[10] T. Ikai, Y. Okamoto, Structure control of polysaccharide derivatives for efficient separation of enantiomers by chromatography, Chem. Rev 109 (2009) 6077–6101.
[11] J. Shen, T. Ikai, Y. Okamoto, Synthesis and application of immobilized polysaccharide-based chiral stationary phases for enantioseparation by high-performance liquid chromatography, J. Chromatogr. A 1363 (2014) 51–61.
[12] R.B. Kasat, S.Y. Wee, J.X. Loh, E.I. Franses, N.-H.L. Wang, Effect of the solute molecular structure on its enantioresolution on cellulose tris(3,5-dimethylphenylcarbamate), J. Chromatogr. B 875 (2008) 81–92.
[13] A.Di Michele I.Varfaj, F. Ianni, M. Saletti, M. Anzini, C. Barola, B. Chankvetadze, R. Sardella, A. Carotti, Enantioseparation of novel anti-inflammatory chiral sulfoxides with two cellulose dichlorophenylcarbamate-based chiral stationary phases and polar-organic mobile phase(s), J. Chromatogr. Open 1 (2021) 100022.
[14] C. Perrin, N. Matthijs, D. Mangelings, C. Granier-Loyaux, M. Maftouh, D. Massart, Y.V. Heyden, Screening approach for chiral separation of pharmaceuticals part II. Reversed-phase liquid chromatography, J. Chromatogr. A 966 (2002) 119–134.
[15] L. Zhou, C. Welch, C. Lee, X. Gong, V. Antonucci, Z. Ge, Development of LC chiral methods for neutral pharmaceutical related compounds using reversed phase and normal phase liquid chromatography with different types of polysaccharide stationary phases, J. Pharm. Biomed. Anal. 49 (2009) 964–969.
[16] M.E. Andersson, D. Aslan, A. Clarke, J. Roeraade, G. Hagman, Evaluation of generic chiral liquid chromatography screens for pharmaceutical analysis, J. Chromatogr. A 1005 (2003) 83–101.
[17] J. Lin, C. Tsang, R. Lieu, K. Zhang, Method screening strategies of stereoisomers of compounds with multiple chiral centers and a single chiral center, J. Chromatogr. A 1624 (2020) 461244.
[18] V.S. Sharp, N. Hicks, J. Stafford, A multimodal liquid and supercritical fluid chromatography chiral separation screening and column maintenance strategy designed to support molecules in pharmaceutical development (Part 1), LCGC Europe 26 (2013) 608–618.
[19] C.L. Barhate, L.A. Joyce, A.A. Makarov, K. Zawatzky, F. Bernardoni, W.A. Schafer, D.W. Armstrong, C.J. Welch, E.L. Regalado, Ultrafast chiral separations for high throughput enantiopurity analysis, Chem. Commun. 53 (2017) 509.
[20] A. Tarafder, L. Miller, Chiral chromatography method screening strategies: past, present and future, J. Chromatogr. A 1638 (2021) 461878.
[21] P. De Gauquier, K. Vanommeslaeghe, Y. Vander Heyden, D. Mangelings, Modelling approaches for chiral chromatography on polysaccharide-based and macrocyclic antibiotic chiral selectors: a review, Anal. Chim. Acta 1198 (2022) 338861.
[22] N.M. Maier, P. Franco, W. Lindner, Separation of enantiomers: needs, challenges, perspectives, J. Chromatogr. A 906 (2001) 3–33.
[23] S. Khater, M.A. Lozac'h, I. Adam, E. Francotte, C. West, Comparison of liquid and supercritical fluid chromatography mobile phases for enantioselective separations on polysaccharide stationary phases, J. Chromatogr. A 1467 (2016) 463–472.
[24] J. Shen, T. Ikai, Y. Okamoto, Synthesis and application of immobilized polysaccharide-based chiral stationary phases for enantioseparation by high-performance liquid chromatography, J. Chromatogr. A 1363 (2014) 51–61.
[25] H. Barfeii, Z. Garkani-Nejad, A comparative QSRR study on enantioseparation of ethanol ester enantiomers in HPLC using multivariate image analysis, quantum mechanical and structural descriptors, J. Chin. Chem. Soc. 64 (2017) 176–187.
[26] L. Pisani, M. Rullo, M. Catto, M. de Candia, A. Carrieri, S. Cellamare, C.D. Altomare, Structure–property relationship study of the HPLC enantioselective retention of neuroprotective 7-[(1-alkylpiperidin-3-yl) methoxy] coumarin derivatives on an amylose-based chiral stationary phase, J. Sep. Sci. 41 (2018) 1376–1384.
[27] C. Luo, G. Hu, M. Huang, J. Zou, Y. Jiang, Prediction on separation factor of chiral arylhydantoin compounds and recognition mechanism between chiral stationary phase and the enantiomers, J. Mol. Graph. Model. 94 (2020) 107479.
[28] R. Sheridan, W. Schafer, P. Piras, K. Zawatzky, E.C. Sherer, C. Roussel, C.J. Welch, Toward structure-based predictive tools for the selection of chiral stationary phases for the chromatographic separation of enantiomers, J. Chromatogr. A 1467 (2016) 206–213.
[29] P. Piras, R. Sheridan, E.C. Sherer, W. Schafer, C.J. Welch, C. Roussel, Modeling and predicting chiral stationary phase enantioselectivity: an efficient random forest classifier using an optimally balanced training dataset and an aggregation strategy, J. Sep. Sci. 41 (2018) 1365–1375.
[30] Y. Martín-Biosca, L. Escuder-Gilabert, M.J. Medina-Hernández, S. Sagrado, Modelling the enantioresolution capability of cellulose tris (3, 5-dichlorophenylcarbamate) stationary phase in reversed phase conditions for neutral and basic chiral compounds, J. Chromatogr. A 1567 (2018) 111–118.
[31] M. Pérez-Baeza, L. Escuder-Gilabert, Y. Martín-Biosca, S. Sagrado, M.J. Medina-Hernández, Comparative modelling study on enantioresolution of structurally unrelated compounds with amylose-based chiral stationary phases in reversed phase liquid chromatography-mass spectrometry conditions, J. Chromatogr. A 1625 (2020) 461281.
[32] J. Zupan, J. Gaisteiger, Neural Networks in Chemistry and Drug Design, 2nd ed., Willey, Weinheim, 1999.
[33] J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, Pearson Education Limited, Harlow, 2005.
[34] J. Pizarroso, J. Portela, A. Muñoz, NeuralSens: sensitivity Analysis of Neural Networks. https://arxiv.org/abs/2002.11423 (accessed 28 February 2022).
[35] J.D. Olden, M.K. Joy, R.G. Death, An Accurate Comparison of Methods for Quantifying Variable Importance in Artificial Neural Networks Using Simulated Data, Ecol Modell 178 (2004) 389–397.
[36] S. Lek, Delacoste M, P. Baran, I. Dimopoulos, J. Lauga, S. Aulagnier, Application of Neural Networks to Modeling Nonlinear Relationships in Ecology, Ecol Modell 90 (1996) 39–52.
[37] K. Boronová, J. Lehotay, K. Hroboňová, D.W. Armstrong, Study of physicochemical interaction of aryloxyaminopropanol derivatives with teicoplanin and vancomycin phases in view of quantitative structure-property relationship studies, J. Chromatogr. A 1301 (2013) 38–47.
[38] M. Szaleniec, A. Dudzik, M. Pawul, B. Kozik, Quantitative structure enantioselective retention relationship for high-performance liquid chromatography chiral separation of 1-phenylethanol derivatives, J. Chromatogr. A 1216 (2009) 6224–6235.
[39] T. Suzuki, S. Timofei, B.E. Iuoras, G. Uray, P. Verdino, W.M.F. Fabian, Quantitative structure-enantioselective retention relationships for chromatographic separation of arylalkylcarbinols on Pirkle type chiral stationary phases, J. Chromatogr. A 922 (2001) 13–23.
[40] ChemSpider Database. Royal Society of Chemistry. http://www.chemspider.com/ (accessed 28 February 2022).
[41] L. Asensi-Bernardi, L. Escuder-Gilabert, Y. Martín-Biosca, M.J. Medina-Hernández, S. Sagrado, Modelling the chiral resolution ability of highly sulfated-β-cyclodextrin for basic compounds in electrokinetic chromatography, J. Chromatogr. A 1308 (2013) 152–160.
[42] L. Escuder-Gilabert, Y. Martín-Biosca, M.J. Medina-Hernández, S. Sagrado, Enantioresolution in electrokinetic chromatography-complete filling technique using sulfated gamma-cyclodextrin. Software-free topological anticipation, J. Chromatogr. A 1467 (2016) 391–399.
[43] MATLAB®R2019a (Mathworks®). https://matlab.mathworks.com/ (accessed 28 March 2022).
[44] M. Szaleniec, Prediction of enzyme activity with neural network models based on electronic and geometrical features of substrates, Pharmacol. Rep. 64 (2012) 761–781.