RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2016, **6**, 18452

A QSPR study on the liquid crystallinity of five-ring bent-core molecules using decision trees, MARS and artificial neural networks†

Jelena Antanasijević,*[a] Davor Antanasijević,[b] Viktor Pocajt,[a] Nemanja Trišović[a] and Katalin Fodor-Csorba[c]

Received 7th October 2015
Accepted 5th February 2016

DOI: 10.1039/c5ra20775d

www.rsc.org/advances

Accelerating progress in the discovery of new bent-core liquid crystal (LC) materials with enhanced features relies on the understanding of structure–property relationships that underline the formation of LC phases. The aim of this study was to develop a model for the prediction of LC behaviour of five-ring bent-core systems using a QSPR approach that combines dimension reduction techniques (*e.g.* genetic algorithms *etc.*) for the selection of molecular descriptors and decision trees, multivariate adaptive regression splines (MARS) and artificial neural networks (ANN) as classification methods. A total of 27 models based on separate pools of calculated molecular descriptors (2D; 2D and 3D) and published experimental outcomes were evaluated. Overall, the results suggest that the acquired ANN LC classifiers are usable for the prediction of LC behaviour. The best of these models showed high accuracy and precision (91% and 97%). Since the best classifier is able to successfully capture trends in a homologous series, it can be used not only to screen new bent-core structures for potential LCs, but also for the estimation of influence of structural modifications on LC phase formation, as well as for the evaluation of LC phase stability.

Introduction

The outstanding feature of bent-core liquid crystals (LCs) is the spontaneous formation of polar order even without molecular chirality.[1] It originates from the bent shape of the aromatic core which restricts the rotation around the long axis and causes the molecules to be tightly packed in the bent direction.[2] This leads to a macroscopic polarization of smectic layers providing ferroelectric and antiferroelectric properties with potential applications for electro-optical switches, as optical phase modulators, nonlinear optical materials, *etc.*[3–5]

Extensive efforts have been made to determine a relationship between the mesomorphic properties of bent-core liquid crystals and their molecular structure.[6] Although some general understandings about the matter have been established,[7–10] designing of the LC molecular structure with favourable properties is still a great challenge for chemists, concerning that those molecules need to exhibit LC behaviour at lower temperatures. Also, it should be noted that the mesophase behaviour of the bent-core compounds is more sensitive to structural modifications than that of calamitic ones concerning the number of the rings, type and orientation of the connecting groups, substituents on the central and outer rings as well as the length of the terminal chains.[7] In addition, the synthesis of bent-core LCs is often very complex, expensive and time consuming, and therefore the use of statistical classification techniques may be helpful in order to reduce the ratio of synthesized bent-core molecules that does not exhibit LC properties.

Although there are various studies related to the prediction of a particular LC property,[11–15] only a limited number of quantitative structure–property relationship (QSPR) models for the prediction of liquid crystallinity can be found in the literature.[16–19] In those papers, the LC behaviour of ferrocene derivatives, copolyethers, polyazomethines and calamitic compounds was predicted using different statistical methods, mainly artificial neural networks (ANNs).

Because of the complex relationship between the bent-core structure and its LC behaviour, the use of nonlinear classifiers is required to achieve an accurate prediction. In this study,

[a]*University of Belgrade, Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia. E-mail: jantanasijevic@tmf.bg.ac.rs*

[b]*University of Belgrade, Innovation Center of the Faculty of Technology and Metallurgy, Karnegijeva 4, 11120 Belgrade, Serbia*

[c]*Wigner Research Centre for Physics, Institute for Solid State Physics and Optics of the Hungarian Academy of Sciences, P.O. Box 49, H-1525 Budapest, Hungary*

† Electronic supplementary information (ESI) available: Molecular structure and liquid crystal behaviour of the modelled five-ring bent-core compounds. Basic terms of molecular descriptors. The LC equations of created 2D and 2&3D MARS models. List of descriptors used in the MARS models with short description. Short description of descriptors selected using feature selection and genetic algorithms. The SKNN and CPNN optimization results. Prediction of LC behaviour of the external test set compounds for all 27 created models. The 2D-FSL-SKNN output map. The 2&3D-GA-PNN output with probabilities. See DOI: 10.1039/c5ra20775d

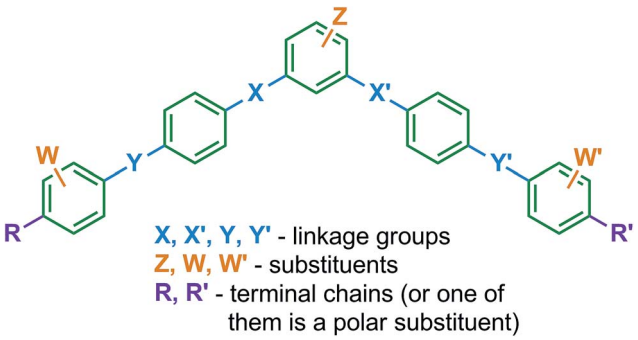

Fig. 1 Schematic representation of the modelled five-ring bent-core system.

decision tree (DT), multivariate adaptive regression splines (MARS) and three different ANN architectures, namely supervised Kohonen (SKNN), counter propagation (CPNN) and probabilistic neural network (PNN), were applied for the prediction of LC behaviour of five-ring system (Fig. 1). In addition, feature selection (FSL) as well as genetic algorithms (GA) and principal component analysis (PCA) as dimension reduction methods were employed for the selection of descriptors.

Methods

Dataset

In this study, bent-core compounds and their LC behaviour (see Table S1, ESI†) used in the development of QSPR models were taken from literature (references S1–S18 presented in ESI†). The dataset consisted of 294 five-ring aromatic compounds with linkage groups of different type and orientation, terminal chains of different type and length, and variety of substituents on the central and outer rings. There were 243 LC compounds and 51 compounds for which LC behaviour was not observed (NLC). For the purpose of developing the model, the dataset was randomly divided into the training set and external prediction set, in the ratio 85 : 15. The training set, which consisted of 206 LCs and 44 NLCs, was used to adjust the parameters of the models. The prediction set, which consisted of 37 LCs and 7 NLCs, was used to test the developed models and to evaluate their generalization ability. Thus the proportion of LC and NLC compounds in the two subsets was almost identical as in the original dataset. Additionally, for the purposes of training PNN and optimizing the architectures of SKNN and CPNN, the training dataset was divided into learning set and internal validation set, in the ratio 4 : 1.

Structure generation and optimization

The molecular structures of the LC compounds used in this study were sketched using ChemDraw,[20] and each structure was stored in the individual (.mol) file. The structures were initially optimized using MMFF94 optimization routine (ChemAxon, Marvin[21]) and the final geometries of the minimum energy conformation were obtained.

Descriptor generation

A series of 2D and 3D descriptors was generated using PaDEL-Descriptor software,[22] including a variety of constitutional, topological, geometric, electrostatic, steric, quantum-chemical and hybrid descriptors. A detailed description and examples of these descriptors can be found in the literature.[23] Any descriptor whose values were identical for all compounds was eliminated in order to reduce the number of descriptors that contained irrelevant information. The reduced pool of 501 descriptors (360 2D and 141 3D) was further used for the development of the model.

Dimension reduction

In this study, QSPR models were created separately with 2D descriptors and with 2D and 3D descriptors together (2&3D), concerning that even after careful handling of the possible conformations, 2D descriptor based models can outperform the

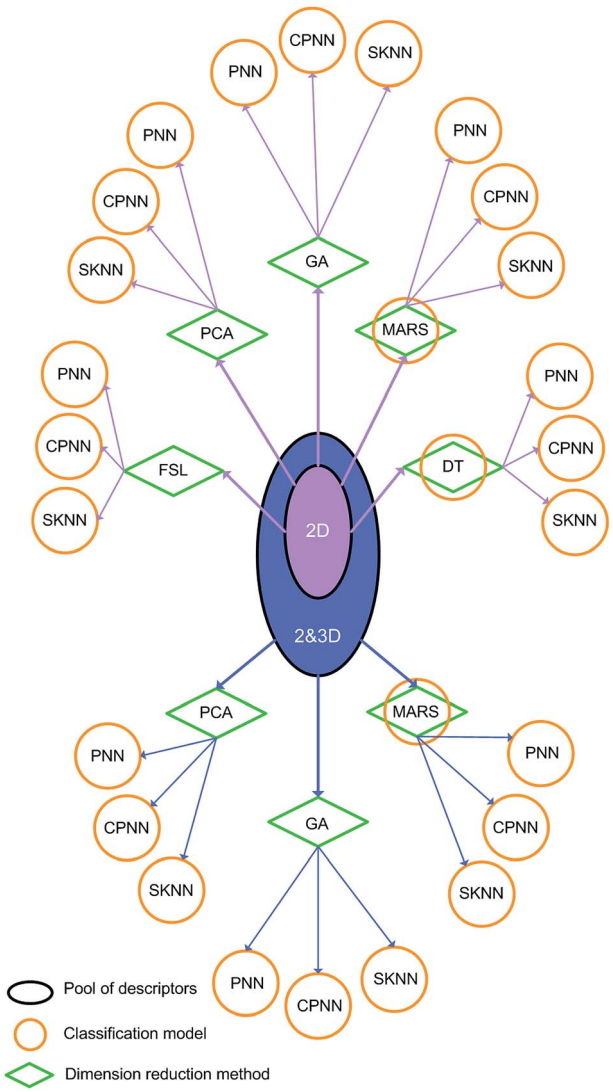

Fig. 2 A schematic summary of the applied techniques and developed classification models.

3D ones,[24–28] and for generation of the latter larger computational resources are needed.

The selection of proper molecular descriptors is a difficult and target-dependent task that can be handled using dimension reduction methods.[29] Therefore, after the initial descriptor removal a further dimension reduction was carried out using feature selection, genetic algorithms and principal component analysis. It should be noted that this dimension reduction step is necessary only in the case of ANN models. Decision trees and MARS are capable to select the most important descriptors during the training of the model, thus the subsets of descriptors selected by DT and MARS were also used for the development of ANN model. The Statistica[30] feature selection routine was utilized to select 20 most significant descriptors, based on the computed Chi-square statistic and p value (significance) for each descriptor. The PCA was performed also in Statistica[30] by extracting the principal components (PCs) with the eigenvalues higher than 1. A genetic algorithm descriptor selection was performed using Neuroshell 2 Genetic Adaptive module[31] by applying the input smoothing factors (ISFs) (see section Artificial neural networks) as a sensitivity tool. After this ISF sensitivity analysis, the minimum number of incorrect classifications (MNIC) from the PNN training was used as the measure of subset quality. Descriptor subsets with fewer descriptors, but equivalent MNIC values, were favoured in the process of reduction.

The considered pools of descriptors, applied dimension reduction and classification techniques and all 27 developed models of the present study have been schematically presented in Fig. 2.

Decision tree

Decision tree is a conclusion scheme which partitions feature space into a set of hyperrectangles and models the output as a constant in each partition.[32] The main advantages of DT are: invariance to monotone transformations (normalization of the data is not required), irrelevant features do not severely detriment performance, relatively high robustness to outliers and interpretability (easily quantification of the importance of each feature).[33]

Several DT algorithms have been used in practice: Chi-Squared Automatic Interaction Detector (CHAID), Quick, Unbiased, Efficient Statistical Trees (QUEST), and Classification and Regression Trees (CART). Among them, CART, which is a non-parametric binary recursive tree structure developed by Breiman *et al.*,[34] was adopted for this study. CART is an efficient tree induction method for large data sets, and it has been used as a classification[35] and a feature selection method.[36] The DT was built by splitting the root node into two child nodes which were then split repeatedly until the terminal nodes were reached. Each split was evaluated using Gini measure as an impurity function.[37] In order to avoid overfitting, the obtained DT has been pruned at the end of the categorization process. The pruning procedure develops a sequence of smaller trees, based on the cost-complexity parameter, and determines the DT with higher accuracy.[36]

In this study, the CART implementation in Statistica (C&RT module) was used for DT model development. A 5-fold cross-validation is performed in order to obtain a stable tree with the smallest overall misclassification rate.

Multivariate adaptive regression splines (MARS)

MARS, developed by Friedman,[38] is a multivariate nonparametric classification/regression technique well suited for high-dimensional problems (*i.e.*, a large number of inputs).[32] MARS combines the strengths of decision trees and spline fitting by replacing the step functions normally associated with DTs with piecewise linear basis functions.[39] The MARS algorithm builds models from two side truncated functions (basis functions) of the inputs (x) separated by the "knots" (t):

$$(x-t)_+ = \begin{cases} (x-t) & x > t \\ 0 & x \le t \end{cases} \quad (1)$$

$$(t-x)_+ = \begin{cases} (t-x) & x < t \\ 0 & x \ge t \end{cases} \quad (2)$$

A knot marks the border of the data region where the behaviour of the function significantly changes and marks the edge of a pair of basis functions, thus building contiguous plane surfaces by summing up basis functions (B_m) with suitable coefficient (a_m):[40]

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \quad (3)$$

where $\hat{y}$ is the predicted output, a_0 the coefficient of the constant basis function and M the number of basis functions. Eqn (3) describes the MARS model with the order of interactions (K) equal to 1. For the order of interactions $K \ge 2$, the B_m denotes the product of basis functions ($b_{m,k}$):

$$B_m(x) = \prod_{k=1}^{K} b_{m,k}(x) \quad (4)$$

The procedure for finding the best MARS model includes the forward selection and backward elimination procedures. In the forward stepwise addition procedure, the pairs of basis functions were added until the performance of MARS model was improved. Such model is often a very complex and overfitting. During the backward elimination, the model is pruned by removing the redundant basis functions using the generalized cross-validation (GCV).[41] The GCV is the mean squared residual error divided by a penalty dependent on the model complexity.[42] Further details on MARS can be found elsewhere.[32,38]

For classification purposes, MARS can be implemented in two manners: (1) the pairwise classification, with output coded as 0 or 1, is handled as a regression, and (2) the classification of more than two classes need to be performed using a hybrid of MARS called POLYMARS.[43]

In this study, the first technique is adopted and the MARS models were produced in Statistica using MARSpline routine. The developed MARS models had a maximum of 40 basis

functions, allowed backward pruning, and a GCV penalty of 2. In order to determine the optimal order of interaction of the spline basis functions, the models with the order of interaction restricted to 2, 3 and 4 were compared. A lower-order model, with similar accuracy as a higher-order one, was favoured, as suggested by Zhang and Singer.[44]

Artificial neural networks

ANNs, which simulate functioning of the human brain, are frequently applied for regression[45,46] and classification purposes.[47] An ANN is consisted of artificial neurons organized in layers with intra- or inter-layer connections, resulting in feed-forward (standard) or feed-back networks. Each neuron is characterized by the numeric weights, which are adjusted (trained) using either supervised, if target (output) values are needed, or unsupervised algorithm.[48] Only a brief description of the ANN architectures used is presented here, since all the details can be found in the quoted papers and other literature.

In this work, the SKNN and CPNN models were created using the Kohonen and CP-ANN MATLAB toolbox 3.6 (ref. 49) released by Milano Chemometrics and QSAR research Group, while the PNN models were created using NeuroShell 2 software.[31]

SKNN is based on a self-organizing map (SOM) learning algorithm developed by Kohonen.[50,51] The SOM is a single layered network and this (Kohonen) layer is often visualized as a square or hexagonal toroidal space, which is consisted of a grid of N^2 neurons, where N is the number of neurons for each side. Each neuron contains as many weights as the number of inputs. The weights of each neuron are updated on the basis of the input vectors, for a certain number of times (epochs). Both the N and epochs must be defined by the user.[52]

SKNN consists of the input and output map, which are 'glued' together forming a combined input–output map which is updated according to the SOM training procedure. After training, the input and output maps are decoupled. The topological formation of the combined input–output map is performed in a supervised way, since the input and output values are used explicitly during the SKNN training. The prediction of unknown class of a new sample is performed by locating the winning neuron in the input map, which is followed by locating of the class of the corresponding neuron in the output map. The maximum value of this neuron's weight vector determines the actual class membership.[53]

CPNN can be also considered as an extension of SOMs, but it combines characteristics from both supervised and unsupervised learning. The theoretical concept of the CPNN was founded by Hecht-Nielsen.[54] CPNN consists of two layers: an input layer (called Kohonen layer), which performs the mapping of the input data, and an output layer (called Grossberg layer) that serves as a "pointing device"[55] and whose neurons have as many weights as the number of classes that need to be learned. In contrast to the learning in the Kohonen layer, the correct response is needed for the correction of the weights in the Grossberg layer, thus the learning is performed in the supervised manner.[56] At the end of the CPNN training, each neuron of the Kohonen layer can be assigned to a class on the basis of the output weights and all the samples placed in that neuron are automatically assigned to the corresponding class.[57] The class of a new sample is estimated following the same procedure as in the case of SKNN.

The overfitting of both SKNN and CPNN is prevented by the optimization of architecture in terms of the number of neurons in output layer and the number of epochs of using genetic algorithms as it is described by Ballabio *et al.*[52] For this purpose the following GA fitness function is used:

$$F = \mathrm{acc_v}(1 - |\mathrm{acc_v} - \mathrm{acc_t}|) \quad (5)$$

where $\mathrm{acc_t}$ and $\mathrm{acc_v}$ are accuracies calculated on the training and internal validation set, respectably. After the optimal architectural parameters were obtained, the best SKNN and CPNN models were selected using the cross-validation method of 5 folds.

PNN, invented by Specht,[58] is a one-pass feed-forward supervised learning neural network consisting of four layers: input, pattern, summation and decision layer. PNN approximates Bayes classifier where the class conditional probabilities are estimated by using the Parzen's window approach.[59] In a binary classification problem, PNN predicts the class of samples using the Bayes decision rule:

$$h_k c_k f_k(x) > h_m c_m f_m(x) \quad (6)$$

where class k and m have the prior probabilities of h_k and h_m, costs of misclassification of c_k and c_m, and probability density function (PDF) of $f_k(x)$ and $f_m(x)$, respectively. In the PNN algorithm, the PDF of each class is estimated from the available training samples using Gaussian kernel,[60] the fundamental equation of PNN being the following:[61]

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} y_i \exp(-D(x, x_i))}{\sum_{i=1}^{n} \exp(-D(x, x_i))} \quad (7)$$

where y_i is the class vector, and $D(x, x_i)$ is Euclidean distance between an observation x and each of the other observations x_i in the training set belonging to the class k (eqn (8)).

$$D(x, x_i) = \sum_{j=1}^{p} \left[(x_j - x_{ij})/\sigma_j\right]^2 \quad (8)$$

In eqn (8) p is the number of inputs, while the σ is so-called smoothing factor, which is only adjustable parameter that needs to be optimized during the PNN training. The σ represents the width of the calculated Gaussian curve for each PDF. One of the major issues associated with the PNN is the selection of optimal smoothing factor.[62] In this study, genetic algorithms were used for searching the optimal σ. When GA is used, beside the overall σ, the so-called individual smoothing factors (ISFs), for each input, are also calculated. The ISF quantifies the importance of a given input to the model, thus ISFs were used as a sensitivity tool.

Table 1 Confusion table

	Predicted class	
Actual class	LC	NLC
LC	*a*	*b*
NLC	*c*	*d*

During the PNN training the learning dataset is used to set the network weights, while the validation data was utilized for the determination of optimal smoothing factor and corresponding ISFs.

Classifiers performance metrics

The performance of created classifiers was analysed only on the basis of classification results obtained for the prediction set. The used performance metrics are defined as follows:

$$\text{Accuracy acc} = \frac{a+d}{a+b+c+d} \quad (9)$$

$$\text{Precision Pr} = \frac{a}{a+c} \quad (10)$$

$$\text{Recall } r = \frac{a}{a+b} \quad (11)$$

$$\text{Geometric mean } G_{\text{mean}} = \sqrt{rd/(c+d)} \quad (12)$$

where a is true positive, b is false negative, c is false positive, and d is true negative predictions (Table 1).

Accuracy gives the percentage of LCs and NLCs correctly classified, while the precision gives the percentage of correctly classified LCs among all compounds which are classified as LCs. The numerical value of recall represents the probability of identifying compounds that exhibit the LC phases. In addition to the acc the G_{mean} is used, since the acc can be misleading in cases where the classes are unequally represented in the training set. G_{mean} has two distinctive properties of being independent of the distribution of examples between classes and being nonlinear. The second property means that the "cost" of misclassifying each positive example increases the more often positive examples are misclassified.[63]

Results

Prediction of LCs using decision tree

Decision tree methodology was applied separately to the pool of 2D and 2&3D descriptors, and in both cases the same DT model containing only 2D descriptors has been obtained. The DT model has five terminal nodes distributed over three levels (Fig. 3). This DT model is based on 4 molecular descriptors, which short description is also presented in Fig. 3. The terminology used for explanation of molecular descriptors is presented in ESI (page S16†).

The root node was split using the JGI9 descriptor, which is related to the charge transfer between the pairs of atoms and therefore it indicates the charge transfer over the molecule. The 30 compounds with higher JGI9 values were finally split by the SM1_Dzp descriptor to two terminal nodes: the first with the ratio of LCs of 72% and the second containing only NLC compounds. The SM1_Dzp descriptor was calculated as

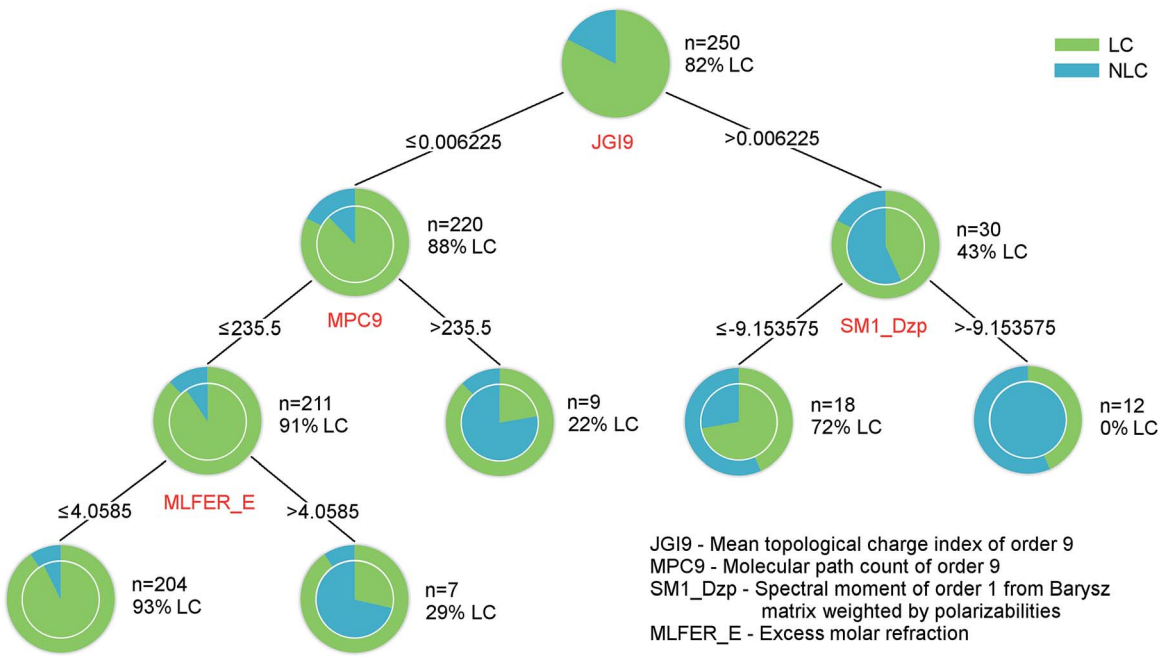

Fig. 3 The decision tree model obtained for the classification of LCs. The outer ring of each node presents the "purity" of the parent node. The number of compounds in a node (*n*) and ratio of LCs is displayed near the corresponding node. The descriptors and their splitting values are presented between two levels.

Table 2 Performance metrics values for the DT model and corresponding confusion table

Des. type	acc (%)	Pr (%)	r (%)	G_{mean} (%)	Actual class	Predicted class: LC	Predicted class: NLC
2D	80	87	89	50	LC	33	4
					NLC	5	2

a molecular spectral moment of order 1 from Barysz matrix weighted by its polarizability. Barysz matrix is a symmetric weighted distance matrix accounting contemporarily for the presence of heteroatoms and multiple bonds in a molecule. The spectral moment of order 1 from Barysz matrix is equal to the sum of the matrix eigenvalues.[23]

The remaining 220 compounds were additionally split using the MPC9 and after that using the MLFER_E descriptor, which resulted in one LC and two NLC terminal nodes. The root node containing 82% of LCs was "purified" to the ratio of LCs of 93%. The MPC9 is a molecular path count of order 9 topological descriptor that counts the total number of paths of length m (in this case 9) in the molecule. The length m of the path is the number of edges along the path and it is called path order.[23] The MLFER_E, which quantifies the excess molar refraction, is one of the molecular linear free energy relation (MLFER) descriptors. The excess molar refraction represents polarizability contributions from n- and π-electrons and can be calculated from the refractive index and characteristic molecular volume.[64]

The pairwise correlation coefficient among these descriptors has an average value of 0.25, a minimum value of 0.05 (between JGI9 and MPC9) and a maximum of 0.47 (between SM1_Dzp and MLFER_E).

The classification results obtained for the prediction set are presented in Table 2. The compounds in the prediction dataset were classified correctly with the accuracy of 80% and precision of 87%, while the G_{mean} had a lower value of only 50%. Low G_{mean} value indicates a significant misclassification of NLC compounds by the DT model, which is obvious from the confusion table (Table 2).

Prediction of LCs using MARS

As it is mentioned above, the MARS algorithm with the order of interaction restricted to 2, 3 and 4 was applied to the separate pool of descriptors (2D and 2&3D) and the obtained models were compared. The two MARS models with the smallest overall misclassification rate were selected, one for the each pool of descriptors. The 2D-MARS model (eqn (S1) in ESI†) was generated with 15 2D descriptors, 27 basis functions and with the order of interaction of 4, while the 2&3D-MARS model (eqn (S2) in ESI†) was generated with 13 2D and 5 3D descriptors, 26 basis functions and with the order of interaction of 2.

A list of basis functions for each of the two MARS models is shown in Table 3, while the corresponding coefficients are presented in ESI (eqn (S1) and (S2)†). The values of performance metrics and corresponding confusion tables for both MARS

Table 3 Basic functions of the 2D- and 2&3D-MARS model

Basic function	2D-MARS model	2&3D-MARS model
B_1	max(0; 6.10×10^{-1} − AVP-0)	max(0; AVP-0 − 6.10×10^{-1})
B_2	max(0; 2.09×10^{2} − MPC10)	max(0; 6.10×10^{-1} − AVP-0)
B_3	max(0; MPC10 − 2.09×10^{2})	max(0; MPC10 − 2.09×10^{2})
B_4	max(0; WPOL − 9.50×10^{1})	max(0; WPOL − 9.50×10^{1})
B_5	max(0; MDEC-12 − 4.39)	max(0; E3s − 3.20×10^{-1})
B_6	max(0; 9.50×10^{1} − WPOL)	max(0; 3.20×10^{-1} − E3s)
B_7	max(0; 4.39 − MDEC-12)	max(0; 3.75×10^{-1} − RPCS)
B_8	max(0; BCUTp-1h − 9.46)	max(0; ETA_EtaP_F − 1.10)
B_9	max(0; 9.46 − BCUTp-1h)	max(0; 1.10 − ETA_EtaP_F)
B_{10}	max(0; SCH-7 − 6.60×10^{-1})	max(0; 3.84 − IC2)
B_{11}	max(0; 6.60×10^{-1} − SCH-7)	max(0; IC2 − 3.84)
B_{12}	max(0; 2.11 − VP-6)	max(0; MLFER_BO − 1.63)
B_{13}	max(0; MIC5 − 4.29×10^{1})	max(0; geomDiameter − 4.49×10^{1})
B_{14}	max(0; 4.29×10^{1} − MIC5)	max(0; MLFER_L − 2.53×10^{1})
B_{15}	max(0; VE1_Dt − 3.58×10^{-2})	max(0; WNSA-2 + 2.05×10^{3})
B_{16}	max(0; MLFER_S − 3.66)	max(0; -2.05×10^{3} − WNSA-2)
B_{17}	max(0; 3.66 − MLFER_S)	max(0; VP-4 − 5.90)
B_{18}	max(0; SpAbs_Dzp − 1.75×10^{3})	max(0; 2.53×10^{1} − MLFER_L)
B_{19}	max(0; 1.75×10^{3} − SpAbs_Dzp)	max(0; 5.90 − VP-4)
B_{20}	max(0; VE3_Dzs + 4.72×10^{1})	max(0; VE3_Dzv + 8.84)
B_{21}	max(0; AVP-0 − 6.10×10^{-1})	max(0; −8.84 − VE3_Dzv)
B_{22}	max(0; VE3_Dzv + 8.84)	max(0; P2m − 3.23×10^{-1})
B_{23}	max(0; −8.84 − VE3_Dzv)	max(0; VE3_Dt + 1.62×10^{1})
B_{24}	max(0; VE2_Dt − 2.81×10^{-4})	max(0; 1.47×10^{2} − ETA_Eta_R)
B_{25}	max(0; 2.81×10^{-4} − VE2_Dt)	max(0; 8.75×10^{-2} − VC-5)
B_{26}	max(0; 7.32 − MDEC-12)	max(0; -3.57×10^{1} − VE3_Dze)
B_{27}	max(0; VPC-4 − 1.52)	N/A

Table 4 Performance metrics values for the MARS models and corresponding confusion tables

Des. type	acc (%)	Pr (%)	r (%)	G_{mean} (%)	Actual class	Predicted class	
						LC	NLC
2D	80	89	86	61	LC	32	5
					NLC	4	3
2&3D	84	88	95	52	LC	35	2
					NLC	5	2

models are shown in Table 4. The 2&3D-MARS model has higher accuracy and excellent recall of 95%, thus it was able to predict almost all LCs from the prediction set. Both MARS models have low G_{mean} values, similar as the DT model, because of the substantial misclassification of NLC compounds.

A list of descriptors used in the MARS models with short description is presented in ESI (Table S2†). The pairwise correlation coefficient among descriptors used in the 2D-MARS model has an average value of 0.36, while those coefficients average value between the descriptors of 2&3D-MARS model is 0.34.

Among selected 2D descriptors, the MPC is described in previous section (Prediction of LCs using decision tree). Three new MLFER descriptors are used in the MARS models: MLFER_S quantifies dipolarity/polarizability, MLFER_L is a solute gas-hexadecane partition coefficient and MLFER_BO represents overall solute hydrogen bond basicity. Descriptors labelled as VP-4, VP-6, AVP-0, SCH-7, VPC-4 and VC-5 are topological descriptors that give information regarding the connectivity of various atoms in the molecule and they are referred as connectivity indices calculated using Chi operator. Those descriptors are able to take into account the presence of heteroatoms in a molecule, as well as double and triple bonds, molecular size, degree of branching and flexibility.

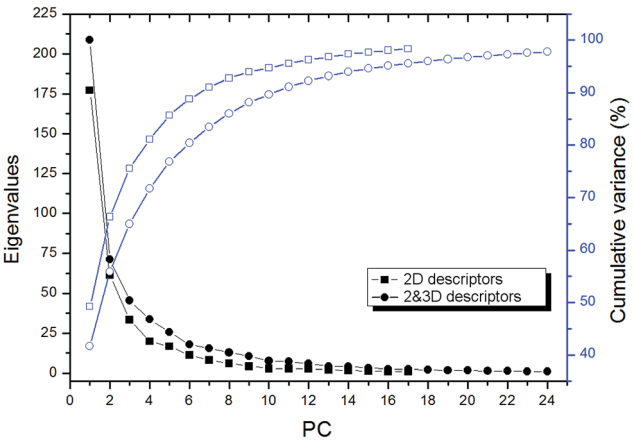

Fig. 4 PCA components with eigenvalues and variance.

Next group of descriptors are those calculated from Barysz matrix, namely the logarithmic coefficient sum of the last eigenvector weighted by van der Waals volumes (VE3_Dzv), by I-state (VE3_Dzs) or by Sanderson electronegativities (VE3_Dze) and graph energy weighted by polarizabilities (SpAbs_Dzp). Descriptors determined from the detour matrix (also known as a matrix of maximal topological distances), *i.e.* coefficient sum of the last eigenvector (VE1_Dt) and its average (VE2_Dt) and logarithmic (VE3_Dt) values, were also used in the models.

The Wiener polarity number[65] (WPOL) is equal to the number of bonds around which free rotations can take place. Moreover, it is related to the flexibility and steric properties of molecules.

Information content (IC) descriptors are based on the calculation of equivalence classes from the molecular graph.

Table 5 List of descriptors selected using feature selection and genetic algorithms

Descriptor group	2D-FSL	2D-GA	2&3D-GA
Barysz matrix	SM1_Dzi; SM1_Dzs	VR2_Dzs	SM1_Dzi; SM1_Dzp; EE_Dzi; EE_Dzm; EE_Dzv
BCUT	BCUTp-1h	BCUTp-1h	BCUTc-1h
Carbon types	C1SP3	C3SP2	
Chi chain	SCH-6; SCH-7; VCH-6; VCH-7		
Chi cluster	VC-5		VC-3
Chi path cluster	VPC-4; VPC-5	VPC-6	
Path count	piPC5; TpiPC	piPC7; TpiPC	MPC8
Extended topochemical atom	ETA_dAlpha_B; ETA_dPsi_A	ETA_dAlpha_B; ETA_BetaP	ETA_Beta; ETA_Beta_ns_d
Molecular distance edge	MDEC-13		
Molecular linear free energy relation	MLFER_E		
Topological distance matrix		VE3_D	
Topological charge	GGI6; GGI9; JGI9		
Information content		SIC3	
Constitutional descriptor		Mare	
WHIM[a]			Dp; L1s

[a] 3D descriptors.

Table 6 List of parameters used for GAs

Parameter	SKNN and CPNN optimization	PNN smoothing factor determ.
Fitness function	Eqn (5)	MNIC
Population size	10	200
Mutation prob.	0.05	NeuroShell
Crossover prob.	0.50	2 default value
Stop criterion	25 evaluations	20 generations[a]
Number of runs	10	1

[a] With no improvement of 1%.

Among them, the IC indices of neighbourhood symmetry take into account also neighbour degree and edge multiplicity. The Modified Information Content index (MIC) is the IC index weighted by the corresponding atomic masses of all atoms in the molecule. The MDEC-12 descriptor counts the molecular distance edge between all primary and secondary carbons.

BCUTs (Burden – CAS – University of Texas eigenvalues) are extensions of the Burden descriptors, which are based on a combination of atomic numbers for each atom and a description of nominal bond-types for adjacent and nonadjacent atoms. The BCUT descriptors expand the number and types of atomic features that can be considered and also provide a greater variety of proximity measures and weighting schemes. The result is a new whole-molecule descriptor that has proved useful in measuring molecular diversity and related tasks.[66]

The last two selected 2D descriptors (ETA_Eta_R and ETA_EtaP_F) belong to the group of extended topochemical atom (ETA) indices. ETA_Eta_R is a composite index that consider both bonded and non-bonded interactions and describes overall topological environment of a molecule relative to the molecular size. ETA_EtaP_F is a functionality index, which accounts the presence of heteroatoms and multiple bonds.[67]

As mentioned above, five descriptors derived from 3D molecular geometry are chosen by the 2&3D-MARS model. One of them is geometric diameter (geomDiameter), defined as the maximum geometric eccentricity in a molecule, and it represents the longest geometric distance between two atoms in the molecule. The other two (WNSA-2 and RPCS) are charged partial surface area descriptors that combine shape and electronic information to characterize molecules and, therefore, encode features responsible for polar interactions between molecules. The WNSA-2 is related to the negative charge surface area, while the RPCS is related to the positive one. Finally, the descriptors labelled as E3s and P2m are WHIM (Weighted Holistic Invariant Molecular) descriptors that give a relevant molecular 3D information regarding the molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames.

More details on the descriptors, which are briefly presented in this and next section, are available in literature.[23]

Table 7 Optimal architectural and training parameters of SKNN and CPNN models

Descriptor			Optimal SKNN		Optimal CPNN	
Type	Select.	N_D[a]	Out. layer	Epochs	Out. layer	Epochs
2D	DT	4	12 × 12	350	8 × 8	200
	MARS	15	12 × 12	350	10 × 10	400
	GA	11	12 × 12	200	12 × 12	300
	FSL	20	12 × 12	300	12 × 12	350
	PCA	360 (17)[b]	12 × 12	350	12 × 12	250
2&3D	MARS	18	12 × 12	200	12 × 12	250
	GA	12	10 × 10	500	12 × 12	350
	PCA	501 (24)[b]	12 × 12	400	12 × 12	350

[a] N_D – number of descriptors. [b] Number of PCs.

Dimension reduction for ANN development

In addition to the selection of descriptors, which was performed during the DT and MARS model development, set of descriptors were obtained by FSL and GA. Also, descriptors were transformed into PCs using the PCA.

A list of descriptors selected by FSL and GA is presented in Table 5, while the short description is provided in ESI (Table S3†). 17 PCs from the pool of 2D descriptors and 24 PCs from the pool of 2&3D descriptors, both with cumulative variance of 98%, were extracted using the PCA. The eigenvalue of each PC along with corresponding variance is presented in Fig. 4.

The same set of 20 2D descriptors was obtained for both considered pools of descriptors by the FSL. The application of GA yielded a set of 11 2D descriptors selected from the corresponding 2D pool, and a set of 10 2D and 2 3D descriptors chosen from the 2&3D pool of descriptors. Pairwise correlations among descriptors selected by FSL have an average value of 0.44, while the average value of pairwise correlation coefficients

Table 8 Performance metric values for the SKNN models and corresponding confusion tables

Model	acc (%)	Pr (%)	r (%)	G_{mean} (%)	Actual class	Predicted class: LC	Predicted class: NLC
2D-DT	80	87	89	50	LC	33	4
					NLC	5	2
2D-MARS	80	87	89	50	LC	33	4
					NLC	5	2
2D-GA	86	90	95	64	LC	35	2
					NLC	4	3
2D-FSL	91	92	97	75	LC	36	1
					NLC	3	4
2D-PCA	68	81	81	0	LC	30	7
					NLC	7	0
2&3D-MARS	84	88	95	52	LC	35	2
					NLC	5	2
2&3D-GA	84	89	92	63	LC	34	3
					NLC	4	3
2&3D-PCA	77	89	84	60	LC	31	6
					NLC	4	3

Table 9 Performance metric values for the CPNN models and corresponding confusion tables

Model	acc (%)	Pr (%)	r (%)	G_{mean} (%)	Actual class	Predicted class: LC	Predicted class: NLC
2D-DT	82	89	89	62	LC	33	4
					NLC	4	3
2D-MARS	82	87	92	51	LC	34	3
					NLC	5	2
2D-GA	82	89	89	62	LC	33	4
					NLC	4	3
2D-FSL	86	92	92	72	LC	34	3
					NLC	3	4
2D-PCA	75	84	86	35	LC	32	5
					NLC	6	1
2&3D-MARS	86	90	95	64	LC	35	2
					NLC	4	3
2&3D-GA	91	95	95	82	LC	35	2
					NLC	2	5
2&3D-PCA	80	85	92	36	LC	34	3
					NLC	6	1

among the descriptors selected by GA from the pool of 2D and 2&3D descriptors was 0.26 and 0.28, respectively.

A relatively high average value of pairwise correlation coefficients between descriptors selected by FSL is consistent with the fact that FSL measures the significance of single descriptors, one by one, and, in contrast to the GA, FSL does not select the "best" combination of descriptors.

Only new descriptor types that haven't been previously mentioned will be described in this section. Carbon-type descriptors calculate the carbon connectivity in terms of hybridization: C1SP3 represents the number of singly bound carbon bound to one other carbon, while C3SP2 represents the number of doubly bound carbon bound to three other carbons. The descriptors labelled as piPC5, piPC7 and TpiPC are conventional bond order ID number descriptors, and they belong to the path count descriptor group. The ID number is a molecular weighted path sum which accounts for multiple bonds in the molecule. One of the selected descriptors is a mean atomic Allred–Rochow electronegativity (Mare), scaled on the carbon atom, and it is a constitutional descriptor. Different Estrada indices calculated from Barysz matrix were selected by GAs. The Estrada indices encode information on complexity of molecular graphs and are also used to describe characteristic physicochemical features of complex systems. They are based on the exponential function and consider both positive and negative eigenvalues at the same time, without compensation effects.

At this point it can be summarized that the selected descriptors encode information about molecular geometry, polarity, flexibility, intermolecular interactions and distribution of the electronic charge. Each of these features alters molecular packing and results in the formation and properties of bent-core LC phases. Considering that molecular packing is determined by a sensitive balance between many competing factors, a variety of descriptors is required for a satisfactory prediction of LC behaviour.

Prediction of LCs using ANNs

Prior to the creation of SKNN and CPNN models, GAs were used to select the most suitable numbers of neurons in the output layer and training epochs. Other parameters of ANN architecture, such as the boundary condition and the neuron shape, were fixed, thus the toroidal boundary condition and hexagonal neuron shape were selected, whereas 20% of training samples were randomly extracted and used as internal validation set in each GA run. Other GAs settings are summarized in Table 6. The results of SKNN and CPNN optimization obtained for different descriptor sets are shown in Table 7. An example of resulting plot of GA optimization (so-called "bubble plot"), which is obtained during the optimization of 2D-FSL-SKNN model, is shown in ESI (Fig. S1†).

The PNN architecture parameters, *i.e.* the number of neurons in each layer, are solely dependent on the features of training dataset. More precisely, in this case the number of neurons in the input layer corresponds to the number of descriptors, while the pattern layer has as many neurons as the number of compounds in the learning set. The number of neurons in the summation layer is equal to the number of classes, while the decision layer has only one neuron in the case of binary classification. Since the same learning set was used, all PNN models had 200 patterns, 2 summation and 1 decision neuron, while the number of input neurons was varied from 4 to 24, in order to match the number of descriptors used. The GAs were employed for the determination of optimal value of smoothing factor during the PNN training and their parameters are presented in Table 6. The values of performance metrics and corresponding confusion tables for ANN models are shown in Tables 8–10. It can be noticed that the performance of majority of ANN models was good, with 2D-FSL-SKNN, 2&3D-GA-CPNN and 2&3D-GA-PNN performing better than others, achieving accuracy higher than 90%.

Table 10 Performance metric values for the PNN models and corresponding confusion tables

Model	acc (%)	Pr (%)	r (%)	G_{mean} (%)	Actual class	Predicted class: LC	Predicted class: NLC
2D-DT	75	91	78	67	LC	29	8
					NLC	3	4
2D-MARS	86	97	86	86	LC	32	5
					NLC	1	6
2D-GA	84	94	86	79	LC	32	5
					NLC	2	5
2D-FSL	80	97	78	82	LC	29	8
					NLC	1	6
2D-PCA	70	85	78	47	LC	29	8
					NLC	5	2
2&3D-MARS	66	87	70	55	LC	26	11
					NLC	4	3
2&3D-GA	91	97	92	89	LC	34	3
					NLC	1	6
2&3D-PCA	77	86	86	50	LC	32	5
					NLC	5	2

A detailed evaluation of ANN classifiers is presented in the next section.

Discussion

Classifier comparison

A comparison of accuracy obtained on the prediction set of all tested classifiers is presented in Fig. 5. In total, there were 17 2D models (DT, MARS and 15 ANNs based on different set of descriptors) and 10 2&3D models (MARS and 9 ANNs also based on different descriptor sets) applied for the prediction of LC behaviour of five-ring bent-core systems.

It can be seen from Fig. 5 that ANN models based on DT or MARS descriptor selection, in most cases (7/9), have the same or better performance in comparison with the corresponding DT or MARS models. However, the best ANN classifiers were created using the descriptor sets obtained by FSL and GA, which are dedicated selection techniques.

The models based on PCs were outperformed by all other models, whilst the 2D descriptors based models proved to be less accurate than the corresponding models created with both 2D and 3D descriptors.

The average performance of classifiers created with descriptors selected using different techniques is presented in Fig. 6. The division of classifiers, in respect to the dimension reduction technique used, which emerges from Fig. 5 is more obvious in Fig. 6. Considering overall performance, the obtained models can be divided into three groups: (1) PCA based, (2) DT/MARS based and (3) FSL/GA based. A difference of about 5% in terms of acc and Pr, between the groups can be observed, and, among the best performing ones, the FSL based ANN models have better precision, while the GA ones are slightly more accurate.

Predictive power of the best classifiers

The prediction results for each pool of descriptors and for the best DT, MARS and ANN models are given in Fig. 7 (the results for other models are presented in Fig. S2–S5 in ESI†).

Since two 2&3D ANN models have the same accuracy, GA-PNN was regarded as better, owing to its higher G_{mean} value (89%).

Although the 2D-DT model demonstrates inferior predictive power in comparison with the models presented in Fig. 7, it can be considered as very convenient for the practical use. Namely, the simplicity of DT approach allows the execution of the model even in a spreadsheet environment (Microsoft Excel and similar) by applying the if-then rules obtained from DT. Thus, during the molecular design, new structures can be easily screened for a potential LC behaviour and the influence of structural units varied in homologous series can be quickly evaluated. The fact that DT model uses only 2D descriptors, further favour its usage.

In the case of best 2D and 2&3D models (2D-FSL-SKNN and 2&3D-GA-PNN), the prediction results exhibit four misclassifications. The 2D-FSL-SKNN has given three false positives (that is, a NLC is classified as the LC) and one false negative (*i.e.* a LC is classified as the NLC), while the 2&3D-GA-PNN has predicted

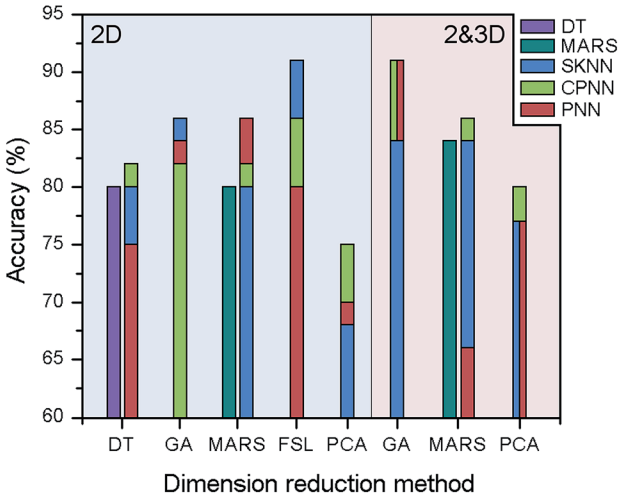

Fig. 5 Accuracy of models obtained using different classification and dimension reduction techniques. The dimension reduction techniques are arranged by ascending number of descriptors.

one false positive and three false negatives. Neither of the false predictions is common to the both ANN models.

The output map for the 2D-FSL-SKNN is presented in ESI (Fig. S6†). As can be seen, the compounds **74** and **267** (P10 and P37 in Fig. S6†) are classified with the probability of 50%.

This means that the model simply does not have enough information from the available training set and selected descriptors to make a confident prediction of the LC property of those molecules. The compound **92** is most likely misclassified as LC, by 2D-FSL-SKNN model, because it is more similar to the

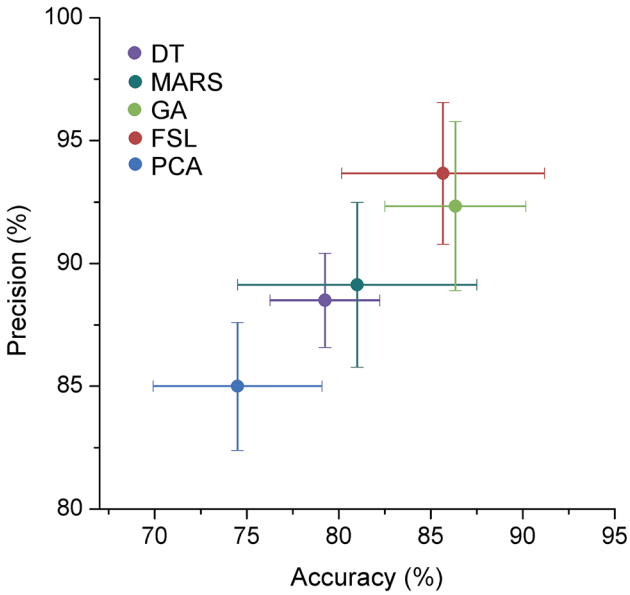

Fig. 6 The average performance of classifiers created with descriptors selected by different techniques. The lines represent standard deviation. The number of DT and MARS based models is 4 and 8, respectively, while the number of GA, FSL and PCA models is the same, 6 of each.

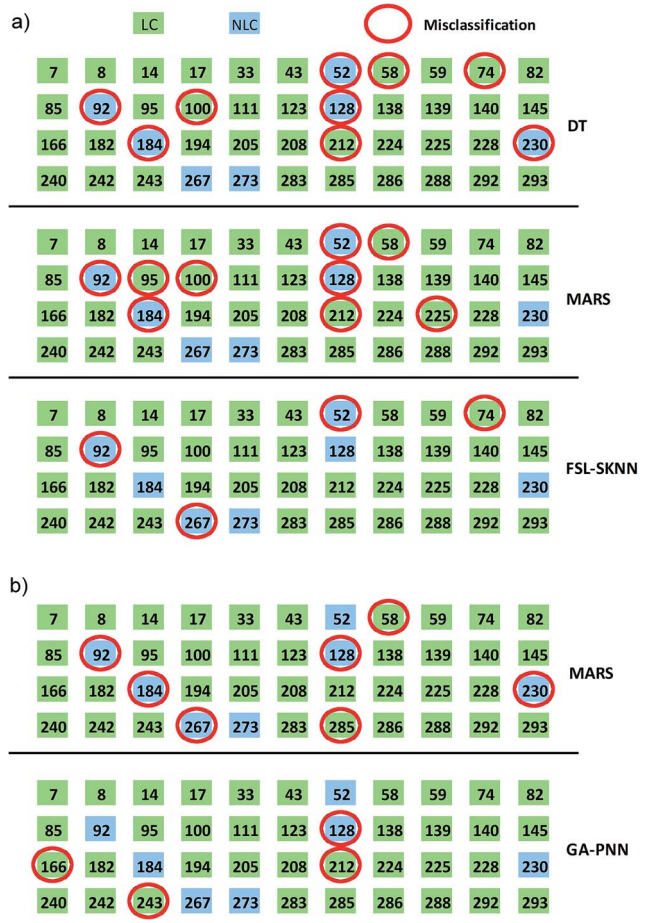

Fig. 7 Predicted LC behaviour of the test compounds by the best DT, MARS and ANN models based on: (a) 2D and (b) 2&3D descriptor sets.

LC compound **93** in the term of values of selected descriptors, than with other compounds from the same homologous series, which are NLC compounds.

The NLC compound **128**, which is misclassified as LC by the 2&3D-GA-PNN model with the probability of 81% (according to the PNN output, Table S4 in ESI†), can be consider as an outlier since all other compounds from the same homologous series are actually LC compounds. This is also supported by the fact that all models, except the 2D-FSL-SKNN, have misclassified this compound. The other three LC compounds, namely **166**, **212** and **243**, were classified as such with the probability of 47%, 19% and 6%, respectively.

Further decrease of misclassification rate could be achieved by using a more balanced dataset with an increased number of compounds, especially the NLC ones. Unfortunately, there is a deficit of reported NLC structures available in literature since published papers contain series of compounds most of which being liquid crystals. Since LCs are often synthesised in series of 5 to 10 compounds, from a practical point of view it is necessary only to make a confident identification of LC series among the candidate series, while the classifiers are actually trained for a more complex task, *i.e.* to predict individual "losers" in the whole "winning" series.

Influence of the terminal chain length

The ability of classifier to capture the trends in homologous series of compounds is another measure of its quality. The influence of the terminal chain length on the 2&3D-GA-PNN probabilities (p_{LC}) (Table S4 in ESI†) is accessed by analysing the obtained p_{LC} for compounds that belong to the same homologous series (Fig. 8).

In order to highlight the p_{LC} trend, the p_{LC} for compounds other than those from the prediction set were inter- and extrapolated, *i.e.* for the NLC compounds p_{LC} is set to 0%, while for the LC compounds p_{LC} is estimated to be >50% according to the observed trend. Fig. 8 shows the p_{LC} trends, which are observed as the chain length is increased by addition of carbons to the chain end, for 4 series of compounds that contain two or more molecules from the prediction set.

In the case of series of compounds **284–294**, the GA-PNN model has captured well-known effect of LC phase stabilization by the increase of terminal chains length. The increase of terminal chains length results in increased lateral attractive forces which stabilize the LC phases.[68] Therefore, the bent-core compounds with longer terminal chains have higher p_{LC} values, because they are more like to be LCs than shorter-chained ones.

For the series of compounds **5–15** and **133–143** the observed p_{LC} trend shows no influence of the terminal chain length. This is probably related with the fact that in those homologous series the length of only one terminal chain is varied, while the second one had fixed length (a long dodecyloxy chain). Apparently the presence of dodecyloxy chain maintains the stability of LC phase, and therefore all homologous exhibit LC behaviour with the same or similar values of p_{LC}.

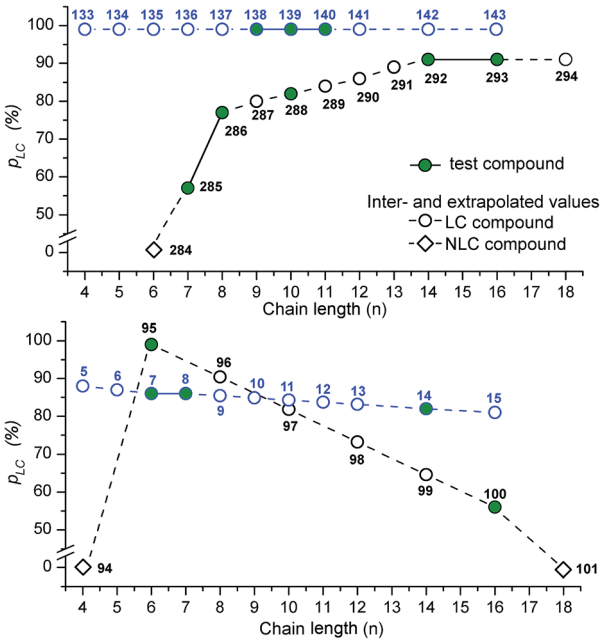

Fig. 8 The p_{LC} trend in a homologous series as the number of carbons (n) in the terminal chain is increased.

Compounds from the homologous series **94–101** form dark conglomerate (DC) mesophases. For this particular homologous series, it was determined that the homologous with medium alkyl chain length are the most stable and that upon further chain elongation the DC phases become instable. For example, the crystallization of compound **100** takes place immediately after the formation of DC, while the longest homologous (**101**) doesn't exhibit DC mesophase.[69] This behaviour is well captured by the 2&3D-GA-PNN model: the p_{LC} decreases from the medium homologous to the longest one.

Conclusions

This work demonstrates that the complex phenomena of LC phase formation by five-ring bent-core molecules can be effectively modelled using a decision trees, MARS and artificial neural networks together with dimension reduction techniques. Using molecular descriptors from the pools of 2D and 2&3D chosen by feature selection and genetic algorithms or by classification techniques itself (DT and MARS), a total of 27 models are created and evaluated. For each pool of descriptors, several models with the accuracy of prediction of unknown compounds from the prediction set greater than 90% were obtained. Also, the dedicated descriptor selection approaches (FSL and GAs) proved their advantage by outperforming the models based on other selection techniques applied.

Overall, the results suggest that the each tested ANN architecture (SKNN, CPNN and PNN) is usable for the prediction of LC behaviour. Especially, the 2D-FSL-SKNN and 2&3D-GA-PNN models demonstrated to be practical and effective tools for the LCs prediction, demonstrating a high accuracy of 91% and precision of 92% and 97%, respectively.

Finally, the analysis of ability of the best classifier (2&3D-GA-PNN) to capture the trends in homologous series showed that this model is capable to predict the stability of potential LC compound, as the function of PNN output probability. Therefore chemists can use the proposed PNN approach: (1) to screen the new bent-core structures in their quest for new LCs, (2) to estimate the stability of LC mesophase, and (3) to quantify the influence of structural modifications on the LC phase formation and its stability. Although the created models do not provide an understanding of the LC phase formation mechanism itself, they represent a rational and practical approach for the prediction of liquid crystallinity with high accuracy.

Further research is planned in expanding the proposed approach for the prediction of particular type of mesophase, as well as in developing regression models for the prediction of transition temperatures of LC phases of various five-ring bent-core systems.

Acknowledgements

The authors are grateful to the Ministry of Education, Science and Technological Development of the Republic of Serbia, Project No. 172007 and 172013 for financial support.

Notes and references

1 T. Niori, T. Sekine, J. Watanabe, T. Furukawa and H. Takezoe, *J. Mater. Chem.*, 1996, **6**, 1231–1233.
2 D. R. Link, G. Natale, R. Shao, J. E. Maclennan, N. A. Clark, E. Körblova and D. M. Walba, *Science*, 1997, **278**, 1924–1927.
3 A. Eremin and A. Jákli, *Soft Matter*, 2013, **9**, 615–637.
4 H. Takezoe and Y. Takanishi, *Jpn. J. Appl. Phys.*, 2006, **45**, 597–625.
5 R. Amaranatha Reddy, U. Baumeister, J. L. Chao, H. Kresse and C. Tschierske, *Soft Matter*, 2010, **6**, 3883–3897.
6 R. A. Reddy and C. Tschierske, *J. Mater. Chem.*, 2006, **16**, 907–961.
7 S. A. R. Krishnan, W. Weissflog, G. Pelzl, S. Diele, H. Kresse, Z. Vakhovskaya and R. Friedemann, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1170–1177.
8 G. Pelzl, S. Diele and W. Weissflog, *Adv. Mater.*, 1999, **11**, 707–724.
9 C. V. Yelamaggad, M. Mathews, S. A. Nagamani, D. S. S. Rao, S. K. Prasad, S. Findeisen and W. Weissflog, *J. Mater. Chem.*, 2007, **17**, 284–298.
10 N. Gimeno, J. Barbera, J. L. Serrano, M. B. Ros, M. R. De La Fuente, I. Alonso and C. L. Folcia, *Chem. Mater.*, 2009, **21**, 4620–4630.
11 S. R. Johnson and P. C. Jurs, *Chem. Mater.*, 1999, **11**, 1007–1023.
12 H. Kranz, V. Vill and B. Meyer, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 1173–1177.
13 R. Schroder, H. Kranz, V. Vill and B. Meyer, *J. Chem. Soc., Perkin Trans. 2*, 1996, 1685–1689.
14 R. Berardi, L. Muccioli and C. Zannoni, *ChemPhysChem*, 2004, **5**, 104–111.
15 G. Tiberio, L. Muccioli, R. Berardi and C. Zannoni, *ChemPhysChem*, 2009, **10**, 125–136.
16 F. Leon, S. Curteanu, C. Lisa and N. Hurduc, *Mol. Cryst. Liq. Cryst.*, 2007, **469**, 1–22.
17 C. Lisa, S. Curteanu, V. Bulacovschi and D. Apreutesei, *Rev. Roum. Chim.*, 2008, **53**, 283–290.
18 S. Curteanu, C. Racles and V. Cozan, *J. Optoelectron. Adv. Mater.*, 2008, **10**, 3382–3391.
19 F. Leon, C. Lisa and S. Curteanu, *Mol. Cryst. Liq. Cryst.*, 2010, **518**, 129–148.
20 *ChemDraw 12.0*, CambridgeSoft, USA, 2009, http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/ChemDrawProfessional/.
21 *Marvin 14.11.3*, ChemAxon Ltd., Budapest, Hungary, 2014, http://www.chemaxon.com/download/marvin-suite/.
22 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.
23 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH Verlag, Weinheim, Germany, 2009.
24 A. Peragovics, Z. Simon, I. Brandhuber, B. Jelinek, P. Hári, C. Hetényi, P. Czobor and A. Málnási-Csizmadia, *J. Chem. Inf. Model.*, 2012, **52**, 1733–1744.
25 K. Roy and P. P. Roy, *Chem. Biol. Drug Des.*, 2008, **72**, 370–382.

26 R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1–9.
27 J. H. Nettles, J. L. Jenkins, A. Bender, Z. Deng, J. W. Davies and M. Glick, *J. Med. Chem.*, 2006, **49**, 6802–6810.
28 V. Venkatraman, V. I. Pérez-Nueno, L. Mavridis and D. W. Ritchie, *J. Chem. Inf. Model.*, 2010, **50**, 2079–2093.
29 L. Bernazzani, C. Duce, A. Micheli, V. Mollica, A. Sperduti, A. Starita, M. R. Tine, V. Risorgimento, I. Pisa, L. B. Pontecorvo, V. Belzoni and I. Padova, *J. Chem. Inf. Model.*, 2006, **46**, 2030–2042.
30 *Statistica 10*, Statsoft, Inc., USA, 2010, http://www.statsoft.com/.
31 *Neuroshell 2*, Ward Systems Group, Inc., MD, USA, 2007, http://www.wardsystems.com/neuroshell2.asp.
32 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2009.
33 C. M. Simon, R. Mercado, S. K. Schnell, B. Smit and M. Haranczyk, *Chem. Mater.*, 2015, **27**, 4459–4475.
34 L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and regression trees*, Wadsworth, Inc., Monterey, CA, 1984.
35 N. S. H. N. Moorthy, S. A. Martins, S. F. Sousa, M. J. Ramos and P. A. Fernandes, *RSC Adv.*, 2014, **4**, 61624–61630.
36 W. A. Young II, G. R. Weckman, V. Hari, H. S. Whiting II and A. P. Snow, *Neural Computing and Applications*, 2012, **21**, 1477–1489.
37 B. Debska and B. Guzowska-Świder, *Anal. Chim. Acta*, 2011, **705**, 261–271.
38 J. H. Friedman, *Ann. Stat.*, 1991, **19**, 1–141.
39 J. R. Leathwick, J. Elith and T. Hastie, *Ecol. Modell.*, 2006, **199**, 188–196.
40 G. Amatulli, A. Camia and J. San-Miguel-Ayanz, *Sci. Total Environ.*, 2013, **450–451**, 209–222.
41 J. B. Ghasemi and E. Zolfonoun, *Spectrochim. Acta, Part A*, 2013, **115**, 357–363.
42 P. J. G. Nieto, J. C. Á. Antón, J. A. V. Vilán and E. García-Gonzalo, *Environ. Sci. Pollut. Res.*, 2015, **22**, 6642–6659.
43 C. J. Stone, M. H. Hansen, C. Kooperberg and Y. K. Truong, *Ann. Stat.*, 1997, **25**, 1371–1470.
44 H. Zhang and B. H. Singer, *Recursive Partitioning and Applications*, Springer, New York, 2010.
45 L. Pogliani and J. V. de Julian-Ortiz, *RSC Adv.*, 2013, **3**, 14710–14721.
46 L. Pogliani and J. V. de Julian-Ortiz, *RSC Adv.*, 2014, **4**, 44733–44740.
47 D. Bianchi, R. Calogero and B. Tirozzi, *Math. Comput. Model.*, 2007, **45**, 34–60.
48 A. M. Peres, P. Baptista, R. Malheiro, L. G. Dias, A. Bento and J. A. Pereira, *Chemom. Intell. Lab. Syst.*, 2011, **105**, 65–73.
49 *Kohonen and CP-ANN toolbox*, Milano Chemometrics and QSAR Research Group, 2015, http://michem.disat.unimib.it/chm/download/softwares/help_cpann/index.htm#sub_1.
50 T. Kohonen, *Biol. Cybern.*, 1982, **43**, 59–69.
51 T. Kohonen, *Biol. Cybern.*, 1982, **44**, 135–140.
52 D. Ballabio, M. Vasighi, V. Consonni and M. Kompany-Zareh, *Chemom. Intell. Lab. Syst.*, 2011, **105**, 56–64.
53 W. Melssen, R. Wehrens and L. Buydens, *Chemom. Intell. Lab. Syst.*, 2006, **83**, 99–113.
54 R. Hecht-Nielsen, *Appl. Opt.*, 1987, **26**, 4979–4984.
55 I. Kuzmanovski, M. Novič and M. Trpkovska, *Anal. Chim. Acta*, 2009, **642**, 142–147.
56 V. S. Šelih, M. Šala and V. Drgan, *Food Chem.*, 2014, **153**, 414–423.
57 D. Ballabio, V. Consonni and R. Todeschini, *Chemom. Intell. Lab. Syst.*, 2009, **98**, 115–122.
58 D. F. Specht, *Neural Network.*, 1990, **3**, 109–118.
59 E. Parzen, *Ann. Math. Stat.*, 1962, **33**, 1065–1073.
60 A. Gelman, J. Carlin, H. Stern and D. Rubin, *Bayesian Data Analysis*, CRC Press, New York, 2003.
61 P. D. Mosier and P. C. Jurs, *J. Chem. Inf. Model.*, 2002, **42**, 1460–1470.
62 M. Zhong, D. Coggeshall, E. Ghaneie, T. Pope, M. Rivera, M. Georgiopoulos, G. C. Anagnostopoulos, M. Mollaghasemi and S. Richie, *Neural Comput.*, 2007, **19**, 2840–2864.
63 M. Kubat, R. C. Holte and S. Matwin, *Mach. Learn.*, 1998, **30**, 195–215.
64 C. F. Poole and S. K. Poole, in *Solid-Phase Extraction: Principles, Techniques, and Applications*, ed. N. J. K. Simpson, CRC Press, New York, 2000.
65 H. Wiener, *J. Am. Chem. Soc.*, 1947, **69**, 17–20.
66 D. T. Stanton, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 11–20.
67 K. Roy and R. N. Das, *J. Hazard. Mater.*, 2013, **254–255**, 166–178.
68 D. Coates, Thermotropic Liquid Crystals, in *Organic Molecular Solids: Properties and Applications*, ed. W. Jones, CRC Press, Boca Raton, USA, 1997, p. 34.
69 M. Alaasar, M. Prehm, M. Brautzsch and C. Tschierske, *Soft Matter*, 2014, **10**, 7285–7296.