Journal of Molecular Graphics and Modelling 20 (2002) 373–388

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Prediction of enantiomeric selectivity in chromatography Application of conformation-dependent and conformation-independent descriptors of molecular chirality

João Aires-de-Sousa[a], Johann Gasteiger[b,*]

[a] Secção de Química Orgânica Aplicada, Departamento de Química, CQFB and SINTOR-UNINOVA, Campus Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Monte de Caparica, Portugal

[b] Computer-Chemie-Centrum, Institute for Organic Chemistry, University of Erlangen-Nürnberg, Nägelsbachstraße 25, D-91052 Erlangen, Germany

Received 6 August 2001; received in revised form 2 November 2001; accepted 6 November 2001

Abstract

In order to process molecular chirality by computational methods and to obtain predictions for properties that are influenced by chirality, a fixed-length conformation-dependent chirality code is introduced. The code consists of a set of molecular descriptors representing the chirality of a 3D molecular structure. It includes information about molecular geometry and atomic properties, and can distinguish between enantiomers, even if chirality does not result from chiral centers.

The new molecular transform was applied to two datasets of chiral compounds, each of them containing pairs of enantiomers that had been separated by chiral chromatography. The elution order within each pair of isomers was predicted by means of Kohonen neural networks (NN) using the chirality codes as input. A previously described conformation-independent chirality code was also applied and the results were compared.

In both applications clustering of the two classes of enantiomers (first eluted and last eluted enantiomers) could be successfully achieved by NN and accurate predictions could be obtained for independent test sets. The chirality code described here has a potential for a broad range of applications from stereoselective reactions to analytical chemistry and to the study of biological activity of chiral compounds.

Keywords: Chirality; Enantioselectivity; Molecular descriptors; Chiral chromatography; HPLC; Chirality code; Conformation; Neural networks; Structure–property relationships

1. Introduction

Chirality is a fundamental aspect of molecular structure, sometimes having profound influence on the properties of compounds. Enantiomers quite often exhibit different chemical and physical properties as well as different biological activity.

Molecular representations incorporating information about chirality are crucial in molecular diversity studies and QSAR/QSPR that are influenced by chiral properties. The 3D atomic coordinates implicitly and accurately represent chirality, but are dependent on the position and orientation of the molecule in Cartesian space. Neural networks (NN) in particular need a fixed number of values as input, independently of the molecular size and number of atoms. Representations of the 3D structure by specification of interatomic distance, on the other hand, cannot differentiate between enantiomers. Binary systems such as R/S or D/L are obviously more efficient for labeling than for representing chirality in a general and chemically meaningful way.

Several quantitative measures of chirality have been developed in the past and were extensively reviewed [1–3]. Mislow et al. [1] distinguished between two classes of measures. In the first “the degree of chirality expresses the extent to which a chiral object differs from an achiral reference object”. In the second “it expresses the extent to which two enantiomorphs differ from one another”. These methods yield a single real value, usually an absolute quantity that is the same for both enantiomorphs.

Recently, Benigni et al. [4] proposed a chirality measure for molecules in a data set. This measure is based on the comparison of the 3D structure for a molecule with all the others in a data set, in terms of electrostatic potential and shape indices. Moreau [5] described a quantitative measure

* Corresponding author. Tel.: +49-9131-85-26570; fax: +49-9131-85-26566.
E-mail address: gasteiger@chemie.uni-erlangen.de (J. Gasteiger).

1093-3263/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S1093-3263(01)00136-X

of the chirality of the environment of *each atom*. Applications of quantitative measures of chirality to the prediction of experimental observables have been quite limited.

A different idea was to incorporate R/S labels into conventional topological indices [6]. Derived chirality descriptors were correlated with biological activity by Julián-Ortiz et al. [7] and more recently by Golbraikh et al. [8].

Mason et al. [9] developed molecular fingerprints coding information about the 3D geometry of four-point pharmacophores, including chirality. The fingerprints were used to define a measure of pharmacophoric similarity/diversity between molecules. Improved ability to recognize compounds with similar biological activity, and to identify ligands to target active sites, was achieved using four-point pharmacophores by comparison to similar three-point methods.

Recently we proposed [10] a chirality code that represents the chirality generated by chiral carbon atoms and is independent of conformation. Instead of *measuring* chirality by a single value, this code is a molecular transform that *represents* chirality using a spectrum-like, fixed-length code, and includes information about the geometry of chiral centers, properties of the atoms in their neighborhoods, bond lengths, and distinguishes between enantiomers. Additionally, we demonstrated that such a code can be successfully applied to the prediction of the enantiomeric selectivity in chemical reactions using artificial NN. This code has the advantage of describing chirality without being influenced by the conformation. However, it is restricted to applications in which the chirality arises from a chiral carbon (or at least a chiral atom). For example it cannot be applied to axially chiral compounds in which a locked conformation generates chirality.

In this paper a second chirality code is described that characterizes the chirality of a 3D structure considered as a rigid set of points (atoms) with properties (atomic properties) and connected by bonds. The code includes information about the molecular geometry (including 3D interatomic distances), connectivity, atomic properties, and can distinguish between enantiomers. It depends on the conformation and has the form of radial distribution functions as used in X-ray structure determination.

It will be shown that the conformation-dependent chirality code (CDCC) can be correlated by means of Kohonen NN with the elution order of enantiomers in chiral chromatographic separations.

Chromatography is one of the best methods available to separate pairs of enantiomers. It is extensively used both in academia and industry for analytical and preparative purposes. Prediction of the elution order for a given pair of enantiomers on a specific chromatographic system can be a useful complementary method to determine the absolute configuration of new compounds.

Modeling of the docking processes responsible for enantio-differentiation in chiral chromatography has been performed by various computational methods [11], namely using molecular mechanics and semi-empirical studies [12–14]. However, their application to highly complex structures or to large datasets of compounds is precluded by the large computation time required. The molecular structure of the stationary phase must also be fully characterized which is rarely the case.

Quantitative structure-enantioselective retention relationships were established with success for datasets of compounds with closely related configurations at a chiral carbon atom. In these cases the (*R*)- and (*S*)-series (or the first eluting and the second eluting series) were studied independently by regression analysis or NN on the basis of molecular descriptors computed at the semi-empirical quantum mechanical level [15,16].

We followed a different strategy that uses NN to learn from known examples represented by the chirality descriptors (CDCC) mentioned above. Two series of cases were studied. In each case, the elution order for a series of enantiomeric pairs had been determined [17,18] on a chiral HPLC column under the same experimental conditions. Kohonen NN were used to correlate the CDCC with the elution order. When possible, the previously reported [10] conformation-independent chirality code (CICC) was also used and compared with the CDCC. The first application involves a chiral stationary phase of high complexity that is untreatable by conventional molecular modeling. The second example was selected to illustrate the ability of the CDCC to represent the chirality caused by locked conformations instead of chiral atoms.

2. Methodology

2.1. Conformation-independent chirality code (CICC)

The outset of our approach was a mathematical transformation of the 3D structure of a molecule into an atomic radial distribution function (Eq. (1)) [19]

$$g(r) = \sum_{i}^{n}\sum_{j}^{n-1} a_i a_j \exp[-b(r - r_{ij})^2] \tag{1}$$

In this equation, a_i and a_j are properties of atoms i and j such as atomic number, r_{ij} the distance between the atoms i and j, b a smoothing parameter, and n is the number of atoms in each molecule. The value of r is a running variable for the function $g(r)$.

Such representations of the 3D structure of a molecule are used in X-ray structure powder diffraction experiments. We have shown that a molecular encoding scheme can be developed on the basis of Eq. (1) and have utilized such a representation of the 3D structure of a molecule for the simulation of infrared spectra [19].

Eq. (1) and the encoding of a molecule derived from it can, however, not distinguish between enantiomers. In order to take account of the chirality, we have recently introduced a CICC [10] that quantitatively describes the stereochemical situation at chiral centers. First, a value of e_{ijkl} was defined through Eq. (2) that considers atoms *i*, *j*, *k*, and *l*, each of them belonging to a different ligand of a chiral center.

$$e_{ijkl} = \frac{a_i a_j}{r_{ij}} + \frac{a_i a_k}{r_{ik}} + \frac{a_i a_l}{r_{il}} + \frac{a_j a_k}{r_{jk}} + \frac{a_j a_l}{r_{jl}} + \frac{a_k a_l}{r_{kl}} \qquad (2)$$

where a_i is a property of atom *i*, such as atomic charge, and r_{ij} is a distance between atoms *i* and *j*. In order to consider the 3D structure but make the chirality code independent of a specific conformer, r_{ij} was taken as the sum of the bond lengths between atoms *i* and *j* on the path with minimum number of bond counts.

Furthermore, a chirality signal, s_{ijkl}, was defined that can attain values of +1 or −1. For the computation of s_{ijkl}, atoms *i*, *j*, *k*, and *l* are ranked according to decreasing atomic property a_i (when the property of two atoms is the same, the properties of the atoms directly bonded to the chiral center, A, B, C, or D, and belonging to the same two ligands, are used for ranking). Then the 3D coordinates of A are used for atom *i*, those of B for *j*, those of C for *k*, and those of D for *l*. The first three atoms (in the order established by ranking) define a plane. If they are ordered clockwise and the fourth atom is behind the plane, the chirality signal, s_{ijkl}, obtains a value of +1. If the geometric arrangement is opposite, s_{ijkl} obtains a value of −1.

The value of e_{ijkl} embodies the conformation-independent 3D arrangement of the atoms of the ligands of a chirality center in distance space and thus cannot distinguish between enantiomers. This distinction is introduced by the descriptor s_{ijkl}.

The two values, *e* and *s*, calculated for all the combinations of four atoms *i*, *j*, *k*, and *l* (each of the four atoms sampled from a different ligand of a chiral center) are then combined to generate a conformation-independent chirality code $f_{\mathrm{CICC}}(u)$, using Eq. (3), where n_{A}, n_{B}, n_{C}, and n_{D} are the number of atoms belonging to ligands A, B, C, and D, respectively:

$$f_{\mathrm{CICC}}(u) = \sum_{i}^{n_{\mathrm{A}}}\sum_{j}^{n_{\mathrm{B}}}\sum_{k}^{n_{\mathrm{C}}}\sum_{l}^{n_{\mathrm{D}}} s_{ijkl} \exp[-b(u - e_{ijkl})^2] \qquad (3)$$

$f_{\mathrm{CICC}}(u)$ is calculated at a number of discrete values with defined intervals to obtain the same number of descriptors, irrespective of the size of the molecule. The actual range of *u* used in an application is chosen according to the range of atomic properties related to the range of observed interatomic distances for the given molecules.

The number of discrete values of $f_{\mathrm{CICC}}(u)$ determines the resolution of the chirality code; *b* is a smoothing factor; in practice *b* controls the width of the peaks obtained by a graphical representation of $f_{\mathrm{CICC}}(u)$ versus *u* [10].

2.2. Conformation-dependent chirality code (CDCC)

We now introduce a more general and conformation-dependent description of molecular chirality, which is formally comparable to the CICC. One main difference is that chiral carbon atoms are now not explicitly considered, and combinations of *any* four atoms are now used, independently of the existence or not of chiral centers, and of their belonging or not to ligands of chiral centers. Every combination of four atoms (A, B, C, and D) is characterized by two parameters, *e* and *c*. As for the CICC, *e* is a parameter that depends on atomic properties and on distances, and is calculated by Eq. (2), with r_{ij} again being the sum of bond lengths between atoms on the path with minimum number of bond counts. The parameter *c* is now a geometric parameter (dependent on conformation) that takes real values, and it takes opposite values for the correspondent set of four atoms in opposite enantiomers.

For the computation of *c*, atoms A, B, C, and D are ranked according to decreasing atomic property (and renamed according to ranking in the order *i*, *j*, *k*, and *l*). When the atomic property of two atoms is the same, they are ranked according to a set of rules based on geometric arguments and atomic properties (Fig. 1). If a set of four atoms is an achiral set (as defined by the rules of Fig. 1, which means there is an element of symmetry making the mirror image of the set superimposable on it) then it is not further considered. The property, a_i, of an atom should have values that allow one to distinguish between non-equivalent atoms. For that purpose we have selected partial atomic charges [20,21], or polarizabilities [22,23], calculated by PETRA [24], because this software rapidly assigns highly selective values to the atoms of large molecules and sizable datasets. Furthermore, we decided to rank atoms using atomic physicochemical properties since these are of much higher influence on the physical, chemical, or biological behaviour of molecules than other conventionally used values (such as atomic numbers in the CIP rules). For the applications described here, partial atomic charges and effective polarizabilities were chosen because of their expected influence on the chromatographic behaviour of molecules. An example of ranking two atoms in a combination of four atoms is described in Fig. 2.

The parameter *c* is defined for each combination of atoms *i*, *j*, *k*, and *l* by Eq. (4), where x_j, y_j, and z_j are the coordinates of atom *j* in the Cartesian system defined in such a way that atom *i* is at position (0,0,0), atom *j* lies on the positive side of the *x*-axis, and atom *k* lies on the *xy* plane and has a positive *y* coordinate. On the right-hand side of Eq. (4), the numerator represents the volume of a rectangular prism with edges x_j, $|y_k|$, and $|z_l|$, while the denominator represents the surface of the same solid. If x_j, y_k, or z_l have a very small absolute value, the set of four atoms is only slightly deviating from an achiral situation. That is reflected in *c*, which would then take a small absolute value; *c* is conformation-dependent because it is a function of 3D atomic coordinates.

$$c_{ijkl} = \frac{x_j y_k z_l}{x_j y_k + x_j |z_l| + y_k |z_l|} \qquad (4)$$

The two values, e_{ijkl} and c_{ijkl}, calculated for all combinations of four atoms, are then combined to generate a

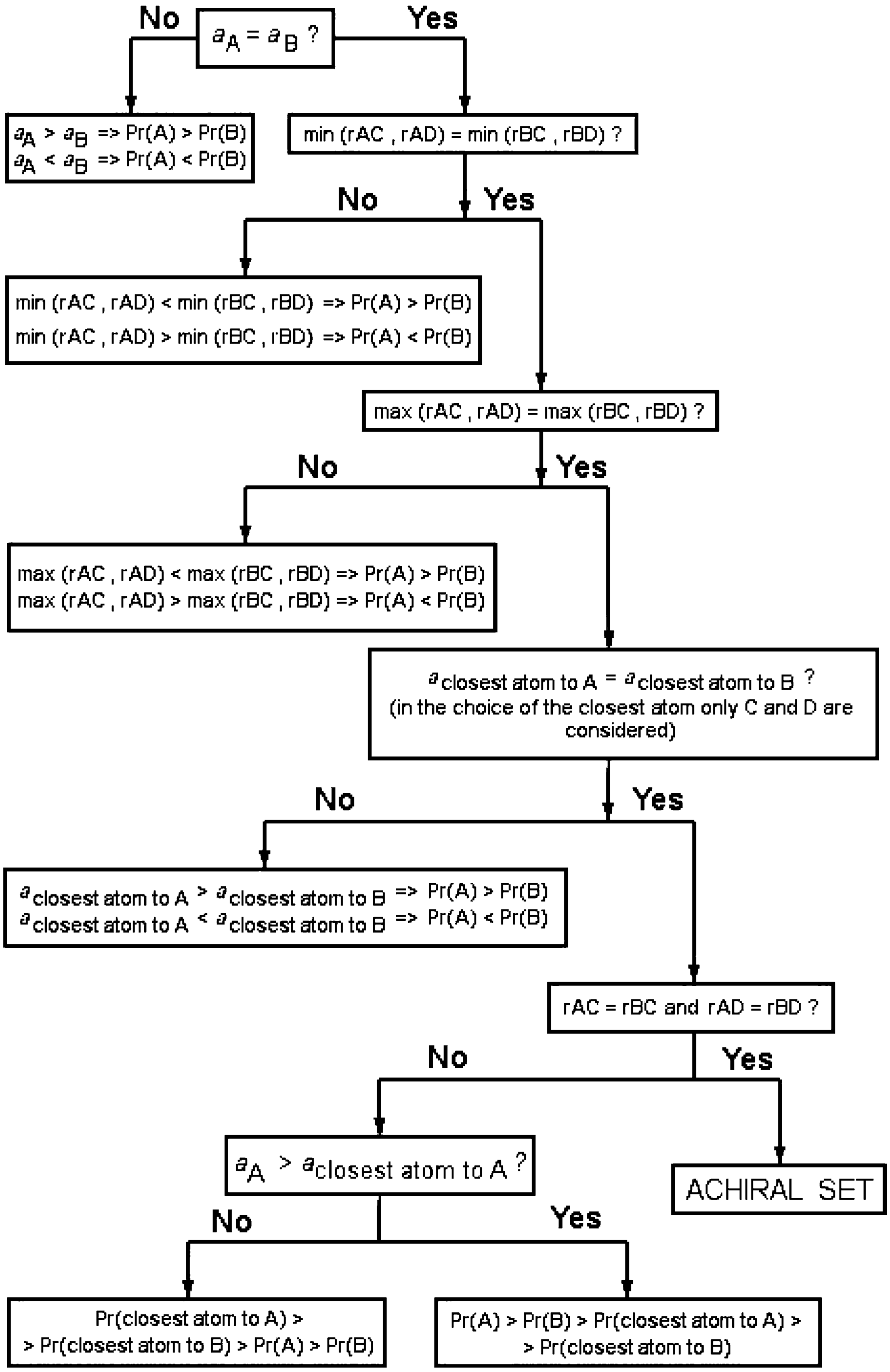
$a_A = a_B$?
No
Yes
$a_A > a_B$ => Pr(A) > Pr(B)
$a_A < a_B$ => Pr(A) < Pr(B)
min (rAC , rAD) = min (rBC , rBD) ?
No
Yes
min (rAC , rAD) < min (rBC , rBD) => Pr(A) > Pr(B)
min (rAC , rAD) > min (rBC , rBD) => Pr(A) < Pr(B)
max (rAC , rAD) = max (rBC , rBD) ?
No
Yes
max (rAC , rAD) < max (rBC , rBD) => Pr(A) > Pr(B)
max (rAC , rAD) > max (rBC , rBD) => Pr(A) < Pr(B)
$a_{\text{closest atom to A}} = a_{\text{closest atom to B}}$?
(in the choice of the closest atom only C and D are considered)
No
Yes
$a_{\text{closest atom to A}} > a_{\text{closest atom to B}}$ => Pr(A) > Pr(B)
$a_{\text{closest atom to A}} < a_{\text{closest atom to B}}$ => Pr(A) < Pr(B)
rAC = rBC and rAD = rBD ?
No
Yes
$a_A > a_{\text{closest atom to A}}$?
ACHIRAL SET
No
Yes
Pr(closest atom to A) > > Pr(closest atom to B) > Pr(A) > Pr(B)
Pr(A) > Pr(B) > Pr(closest atom to A) > > Pr(closest atom to B)

Fig. 1. Procedure for establishing the relative priority of two atoms (A and B) in a combination of four atoms (A, B, C, and D). Here, Pr(A) is the priority of atom A; a_A the atomic property of atom A; r_{AB} the 3D distance between atoms A and B; min(w, z) the minimum of values w and z; and max(w, z) is the maximum of values w and z.

conformation-dependent chirality code f_{CDCC}, using Eq. (5), where n is the number of atoms in each molecule, and c introduces the conformation dependence:

$$f_{\mathrm{CDCC}}(u) = \sum_{i}^{n}\sum_{j}^{n-1}\sum_{k}^{n-2}\sum_{l}^{n-3} c_{ijkl}\exp[-b(u-e_{ijkl})^2] \qquad (5)$$

$f_{\mathrm{CDCC}}(u)$ is calculated at a number of discrete values of u, with defined intervals to obtain the same number of descriptors, irrespective of the size of the molecule. The actual range of u used in an application is chosen according to the range of atomic properties related to the range of observed interatomic distances for the given molecules.

The number of discrete values of $f_{\mathrm{CDCC}}(u)$ determines the resolution of the chirality code; b is a smoothing factor; in

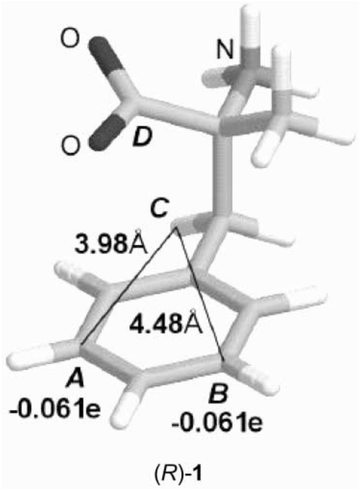

Fig. 2. Assignment of relative priorities to atoms A and B in the combination of atoms A, B, C, and D belonging to molecule (*R*)-**1**. A and B have the same atomic charge (−0.061e) so the next rule should be applied; $\min(r_{AC}, r_{AD}) = r_{AC} = 3.98$ Å; $\min(r_{BC}, r_{BD}) = r_{BC} = 4.48$ Å; $\min(r_{AC}, r_{AD}) < \min(r_{BC}, r_{BD})$, so the priority of A is higher than the priority of B.

practice *b* controls the width of the peaks obtained by a graphical representation of $f_{CDCC}(u)$ against *u*.

The f_{CDCC} is a function of *u*. For all the molecules in a data set, f_{CDCC} is calculated at the same values of *u*, and therefore the chirality code has the same length for all the molecules. At every value of *u*, f_{CDCC} is a summation of several terms (one for each set of atoms *i*, *j*, *k*, and *l*). Each term is the product of *c* and an exponential function; *c* can take positive or negative values and is independent of *u*. But the exponential function is always positive and takes different values depending on *u*. Thus, each of the summation terms (one for each set of atoms *i*, *j*, *k*, and *l*) has a constant sign over *u*, but different magnitude (because of the exponential), which causes the global sum to be positive at some values of *u* and negative at others.

An example of the chirality code for the enantiomers of α-methylphenylalanine (**1**) in two different conformations is shown in Fig. 3.

In some chiral compounds, chirality arises from a locked conformation (well known examples are binaphthalene derivatives such as BINAP). Even most of the achiral compounds at a given temperature, have some chiral conformations. They behave as achiral because the conformations are fast interconverting at the given temperature, making the average number of opposite enantiomeric conformations the same (e.g. butane). Because the CDCC uses a single (frozen) conformation, it generates non-zero values for molecules represented in a chiral conformation, even if the conformation is unlocked and the molecule behaves as achiral. For example, if butane is represented in a chiral conformation, the CDCC will not be zero. The CDCC was designed to account for any type of molecular chirality given a 3D molecular structure.

2.3. Kohonen neural network

To model the relationship between the chirality code, $f_{CDCC}(u)$, of a given chiral compound and the enantiomeric preference of retention of this compound by a chiral chromatographic column, a Kohonen NN (self-organizing map) [25–27] was used. The problem was defined in terms of a classification task: the goal was to classify chiral molecules as 'first eluting enantiomer' or 'last eluting enantiomer' in an HPLC system. NN were chosen because they are especially useful for the modeling of complex and non-linear relationships. Kohonen self-organizing maps are particularly fast to train, and learn by unsupervised training, revealing similarities between objects (chirality codes). This is an advantage for assessing the chemical significance of the new chirality code.

The input data for a Kohonen network are stored in a grid of neurons, each containing as many elements (weights) as there are input variables. In the investigations described in this paper the input variables are discrete values of the chirality codes (Fig. 4).

During the training of a Kohonen network, each individual object (chirality code) is mapped into that neuron of the Kohonen layer (central neuron or winning neuron) that contains the most similar weights compared to the input data (chirality code). It is said that the central neuron was *excited* by the object. The weights are then adjusted to make them even more similar to the presented data. The extent of adjustment depends on the topological distance to the central neuron—the closer a neuron is to the central neuron the larger is the adjustment of its weights. The objects of the training set are iteratively fed to the network until the training is stopped (the end of training is defined, e.g. by a fixed number of cycles or by a measure of stability).

After training, all the objects of the training set are mapped, and, in this work, a neuron is assigned to a class if exclusively molecules of that class excite it. A trained Kohonen NN will reveal similarities in the objects of a data set in the sense that similar objects (similar chirality codes) are mapped into the same or closely adjacent neurons. When a new molecule is presented to the trained network, it can be classified according to the most frequent class of the winning and adjacent neurons. As mentioned before, in the two applications here described, the molecules were classified as 'first eluting enantiomer' or 'last eluting enantiomer'.

2.4. Computational details

The Cartesian coordinates of the atoms in a molecule were calculated from the connection tables of the molecules by the 3D structure generator CORINA [28–31]. The physicochemical atomic properties (partial atomic charge and effective polarizability) were calculated using fast empirical methods implemented in the program package PETRA 3.0 [24], charges by the PEOE method [20,21] and effective polarizabilities by a previously published procedure [22,23].

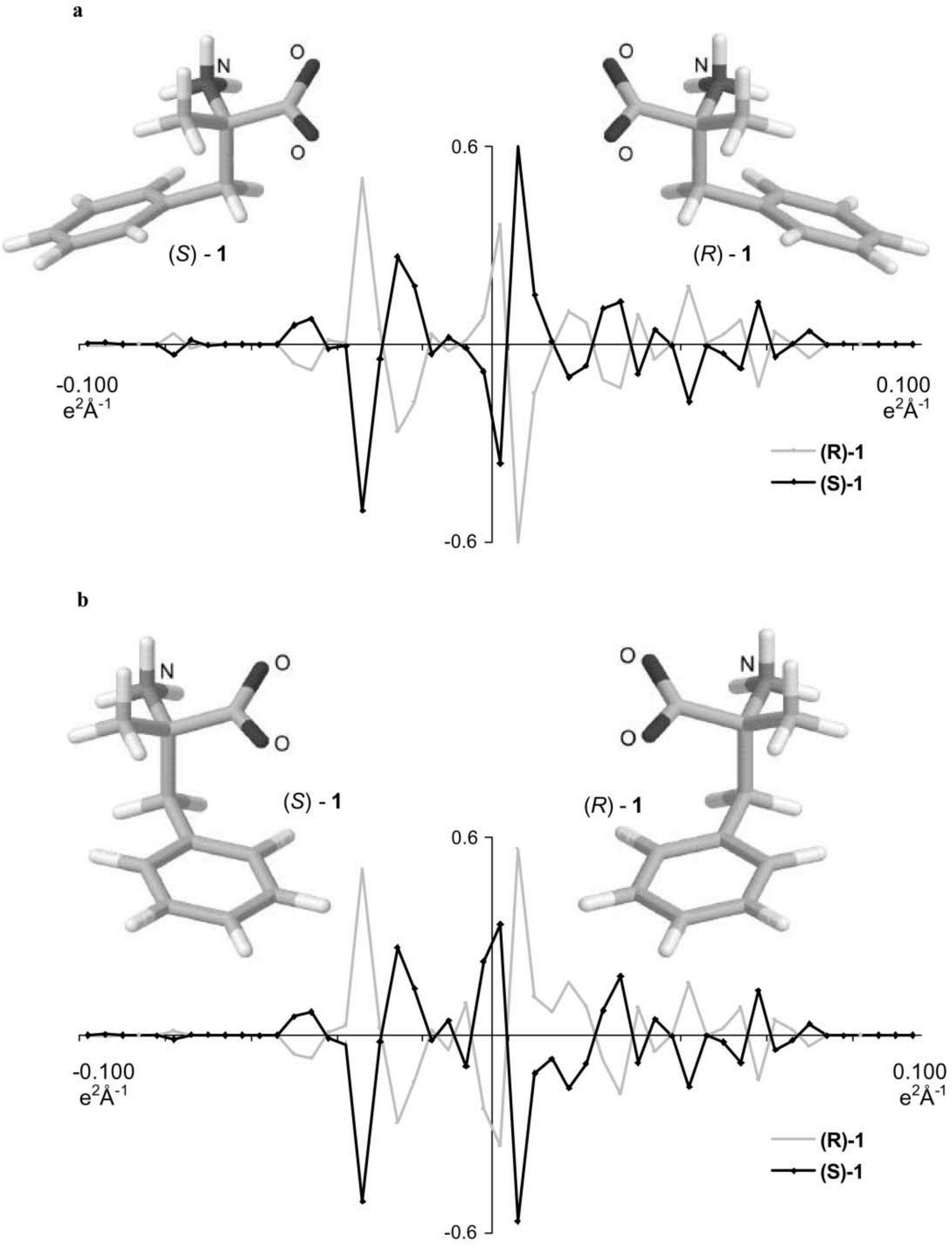

Fig. 3. Three-dimensional structure and representation of $f_{\mathrm{CDCC}}(u)$ vs. u for (*R*)-**1** and (*S*)-**1** at two different conformations (a and b) sampled at 50 evenly separated values between -0.100 and $+0.100\,\mathrm{e}^2\,\text{Å}^{-1}$. Partial atomic charge was used as the atomic property.

The calculation of chirality codes has been performed by a computer program especially developed for this task. The program was written using the C programming language and, for the experiments described here, was compiled for the SGI IRIX 6.5 platform.

The Kohonen networks used in this investigation were simulated with the simulator KMAP 3.0 [32]. The networks consist of a grid of neurons in a toroidal shape. Each data set used in this investigation was divided into a training and a test set. The number of data for the individual applications is mentioned in the text. Training of the Kohonen network was performed by using a linear decreasing triangular scaling function used with the maximum possible initial learning span and an initial learning rate of 0.7. The weights were initialized with normally distributed pseudo random numbers that were calculated using the mean and standard deviation

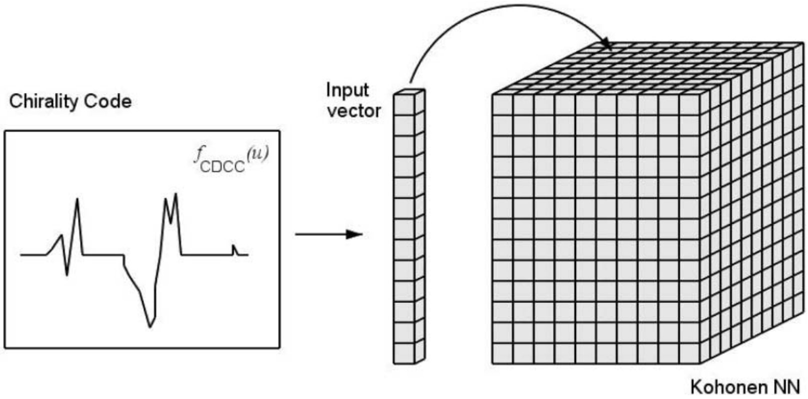

Fig. 4. Illustration of the Kohonen NN method. A chirality code is transformed into a one-dimensional vector with a number of elements equal to the number of weights in a neuron. When a chirality codes is presented to the network, the neuron with the most similar weights to the chirality code is excited (the winning or central neuron).

of the input data set as parameters. For the selection of the central neuron, the minimum Euclidean distance between the input vector and the neuron weights were used. Training was performed until the learning span was reduced to zero and the learning rate was less than 0.1, or a maximum of 4000 objects was presented to the network.

3. Results and discussion

Two cases of chiral chromatographic systems were selected for application of the chirality codes. The first application involves a chiral stationary phase of high complexity that is only partially characterized. Conventional molecular modeling of the possible interactions between the stationary phase and the analytes in order to decide which enantiomer of a pair has more affinity to the stationary phase, and is eluted last, is impossible. The second example includes compounds in which chirality arises from restricted rotation about a single bond and was therefore selected to illustrate the ability of the CDCC to represent such type of chirality.

3.1. *HPLC on a teicoplanin chiral stationary phase*

Twenty-eight chiral compounds represented in Fig. 5 were separated from their enantiomers by HPLC on a teicoplanin chiral stationary phase [17]. The macrocyclic glycopeptide teicoplanin has twenty chiral centers surrounding four pockets or cavities, three sugar moieties, and a molecular weight of ∼1885. It is bonded to silica gel through multiple covalent linkages and it can interact with analytes by inclusion complexation, π–π interactions, and hydrogen bonding [33]. This is a very complex stationary phase and modeling of the possible interactions with the analytes is impracticable. In such a situation, learning from known examples seems more appropriate, and the new chirality codes looked quite appealing for representing such data.

The reduced conformational freedom of many compounds (**11**–**28**), the structural similarities between the others (**2**–**10**), and the fact that the same program generated all 3D models allowed the use of the conformation-dependent chirality code. Therefore the 28 analytes (Fig. 5) and their enantiomers were encoded by the CDCC and submitted to Kohonen NN. They were divided into a test set of six compounds (**10**, **16**, **21** and their enantiomers) that were chosen to cover a variety of skeletons and were not used for the training. That left a training set containing the remaining 50 compounds.

The CDCC can be adjusted by choosing different parameters (such as resolution, range of u, smoothing parameter), atomic properties, and types of atoms considered. By changing these variables, 6615 different codes were generated and automatically screened. The different codes were calculated using (a) all types of atoms or not considering hydrogen atoms; (b) partial atomic charges or effective polarizabilities as atomic properties; (c) code lengths between 20 and 75; (d) values of u in the interval ($-s$, $+s$) with s varying between 0.020 and 0.190 e^2 Å^{-1} (when charges were used) or in the interval (0, $+s$) with s varying between 30 and 290 Å^5 (when polarizabilities were used); (e) combinations of four atoms with maximum interatomic path distances of 5, 6, 7, 8, 9, 10, or 11 bonds. The smoothing parameter b was set to (code length/range of u)^2.

In the automatic screening, each code was quantitatively evaluated using a Kohonen NN. Each of the 50 compounds from the training set was encoded using the CDCC, and the data set was used to train a 10 × 10 Kohonen NN. Once the network was trained, the training set and the test set were fed to the network and the number of correct classifications (as first or last eluting enantiomer) obtained for both sets were used as quality measures for the code.

The quality of clustering in a Kohonen NN can be assessed by the number of correct predictions (classifications) for the training set. We therefore chose, as the best codes, those with higher number of correct predictions for the

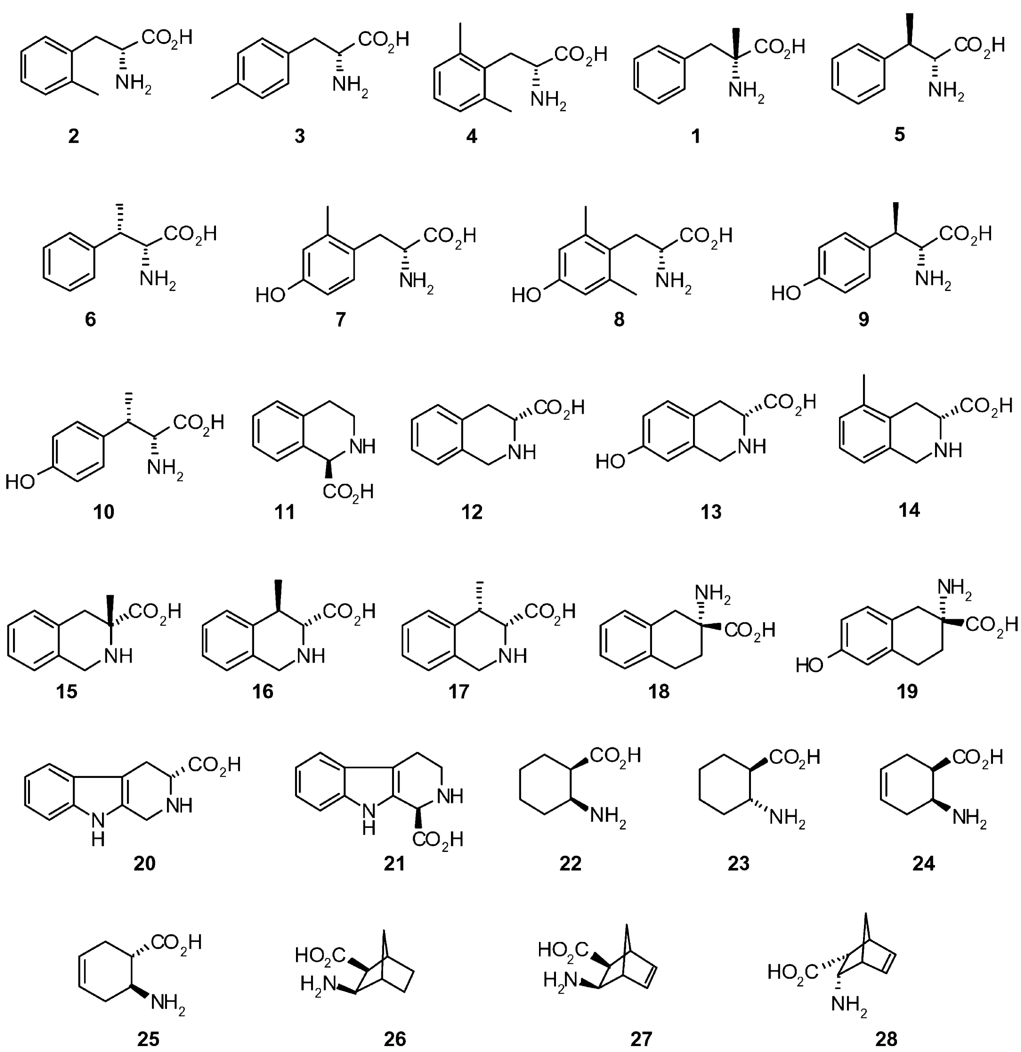

Fig. 5. Last eluted enantiomers in a chromatographic separation on chiral HPLC with teicoplanin stationary phase.

Table 1
Results of the screening of CDCC applied to the classification of molecules **1**–**28** and their enantiomers by Kohonen NN of size 10 × 10

	Atomic property	H atoms considered	Number of experiments[a]	Best codes		
				Correct predictions for the training set (50 cpds)[b]	Predictions for the test set[c]	
					Correct (%)	Wrong (%)
(1)	Partial atomic charge	Yes	1512	49 (1) 48 (1) 47 (8)	95	2
(2)	Partial atomic charge	No	1512	49 (1) 48 (25)	46	38
(3)	Effective polarizability	Yes	1862	49 (3) 47 (9)	49	25
(4)	Effective polarizability	No	1729	46 (1) 45 (5) 44 (3)	54	26

[a] The number of experiments is not the same in the four series because some combinations of atomic property, range, atoms considered, and resolution yielded a null chirality code for some compounds and were not used.

[b] The number of codes yielding the given result is displayed in parentheses.

[c] Global percentage of right and wrong predictions obtained for the test set using the best codes referred to in the preceding column. The remaining cases were undecided.

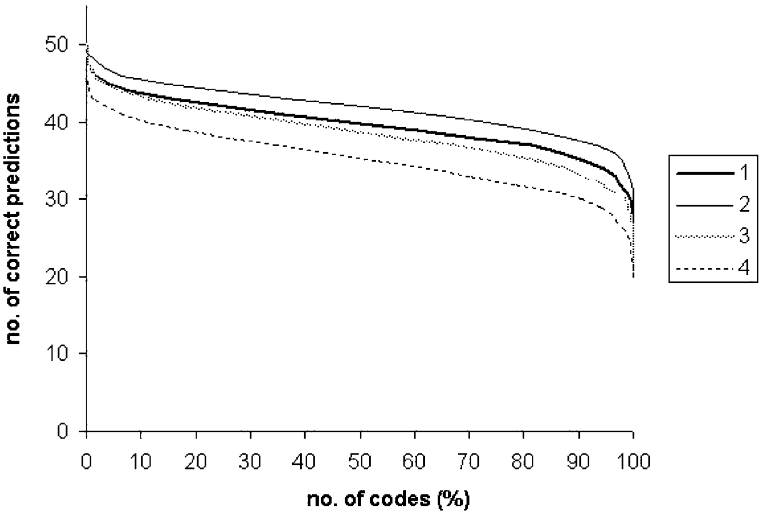

Fig. 6. Graphical representation of the distribution of correct predictions (for the training set) over the four series of experiments described in Table 1. The values on the abscissa are the percentage of codes yielding the number of correct predictions (or more) given by the ordinate.

training set without considering the results for the test set at this phase. Table 1 shows the best results obtained and the number of codes that afforded these results (fifth column). The results are given separately for the codes using partial atomic charges or polarizabilities, and for the codes considering all types of atoms or neglecting hydrogen atoms. Some combinations of atomic property, atoms considered, and resolution, yielded, for some compounds, a chirality code with all the non-zero values falling outside of the chosen range for *u*. These experiments were not used, which explains why the number of experiments is not the same in the various series of Table 1. For every set of best results, the average of correct and wrong predictions for the test set were calculated and displayed in the last two columns of Table 1. A representation of the distribution of correct predictions for each of the four series of experiments is displayed in Fig. 6.

The results clearly show that the approach followed for the prediction of elution order was successful. Chirality codes could be found that allowed a Kohonen NN to separate well the first eluted enantiomers from the last eluted enantiomers, with a maximum number of 49 correct predictions (out of 50 objects in the training set). The distributions represented in Fig. 6 indicate that, for each of the four series of experiments, there are ca. 5% of the codes that perform considerably

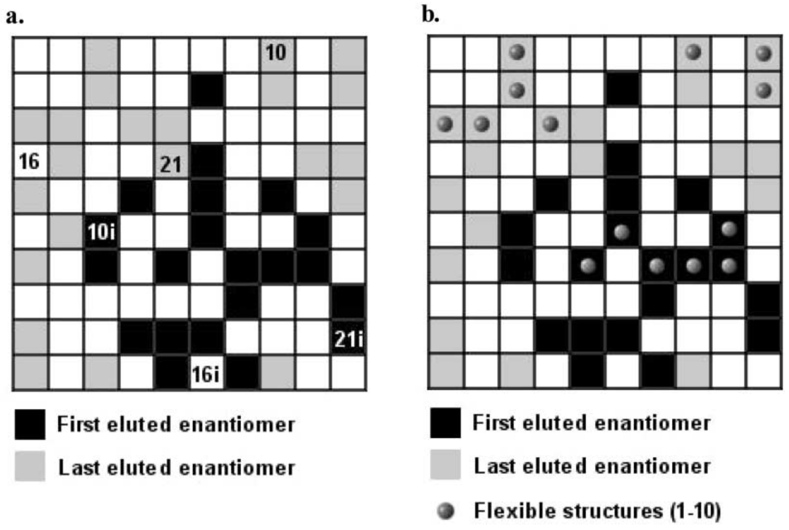

Fig. 7. Mapping of chiral amino acids encoded by CDCC into a 10 × 10 Kohonen NN with toroidal surface. After training, the 50 molecules from the training set were mapped and the neurons were colored accordingly (some neurons were empty and some were excited by more than one molecule). (a) Test structures **10**, **16** and **21** and the inverted enantiomers (labeled with ‘i’) were presented to the trained network in order to be classified. All test examples were correctly classified as ‘first eluted’ or ‘last eluted’ enantiomer in the given HPLC system. (b) The neurons excited by the flexible structures (**1**–**10**) are highlighted with solid spheres.

Table 2
Results of the screening of CDCC applied to the classification of molecules **11**–**28** and their enantiomers by Kohonen NN of size 8 × 8

	Atomic property	H atoms considered	Number of experiments[a]	Best codes		
				Correct predictions for the training set (32 cpds)[b]	Predictions for the test set[c]	
					Correct (%)	Wrong (%)
(1)	Partial atomic charge	Yes	1512	31 (1) 30 (7)	72	3
(2)	Partial atomic charge	No	1512	32 (3) 31 (28)	79	5
(3)	Effective polarizability	Yes	1862	32 (1) 31 (15)	28	52
(4)	Effective polarizability	No	1729	31 (5) 30 (8)	38	12

[a] The number of experiments is not the same in the four series because some combinations of atomic property, range, atoms considered, and resolution yielded a null chirality code for some compounds and were not used.

[b] The number of codes yielding the given result is displayed in parentheses.

[c] Global percentage of right and wrong predictions obtained for the test set using the best codes referred to in the preceding column. The remaining cases were undecided.

better than the others, ca. 5% of the codes that perform much worse than the others, and for the majority of the codes the performance only varies within a small range. Comparing the results obtained using partial atomic charges (series 1 and 2) with those obtained using polarizabilities (series 3 and 4) it can be said that globally partial atomic charges gave better results than polarizabilities, even if the scale and number of codes tested was different in the two situations. That is apparent in Table 1 and becomes clearer in Fig. 6. Table 1 suggests that better clustering was achieved for the training set when hydrogen atoms were neglected (series 2 versus 1), and Fig. 6 reveals the same trend for the global picture. However, the best codes of series 2 exhibited a much-reduced predictive ability for the test set (Table 1, two right columns) comparing to series 1.

The best results (Table 1, entry 1) were obtained with partial atomic charge as atomic property and including hydrogen atoms in the calculations. One code gave 49 correct predictions for the training set, one code gave 48 correct predictions and eight codes gave 47 correct predictions. These 10 best codes yielded an average of 95% of correct predictions for the test set.

An example of a trained Kohonen NN surface, colored according to the mapping of the training set, is shown in Fig. 7. The objects of the test set were also mapped into it and are identified by their reference numbers. It can be seen that all the test objects were correctly classified (an object is classified according to the most frequent class of the winning and adjacent neurons) and the map resulting from the training shows that the first eluted enantiomers were placed

Table 3
Results of the screening of CICC applied to the classification of molecules **1**–**28** and their enantiomers by Kohonen NN of size 10 × 10

	Atomic property	H atoms considered	Number of experiments[a]	Best codes		
				Correct predictions for the training set (50 cpds)[b]	Predictions for the test set[c]	
					Correct (%)	Wrong (%)
(1)	Partial atomic charge	Yes	1463	50 (5) 49 (27)	84	4
(2)	Partial atomic charge	No	1407	50 (2) 49 (13)	94	1
(3)	Effective polarizability	Yes	1984	48 (1) 47 (6)	76	12
(4)	Effective polarizability	No	1596	46 (3) 45 (10)	96	0

[a] The number of experiments is not the same in the four series because some combinations of atomic property, range, atoms considered, and resolution yielded a null chirality code for some compounds and were not used.

[b] The number of codes yielding the given result is displayed in parentheses.

[c] Global percentage of right and wrong predictions obtained for the test set using the best codes referred to in the preceding column. The remaining cases were undecided.

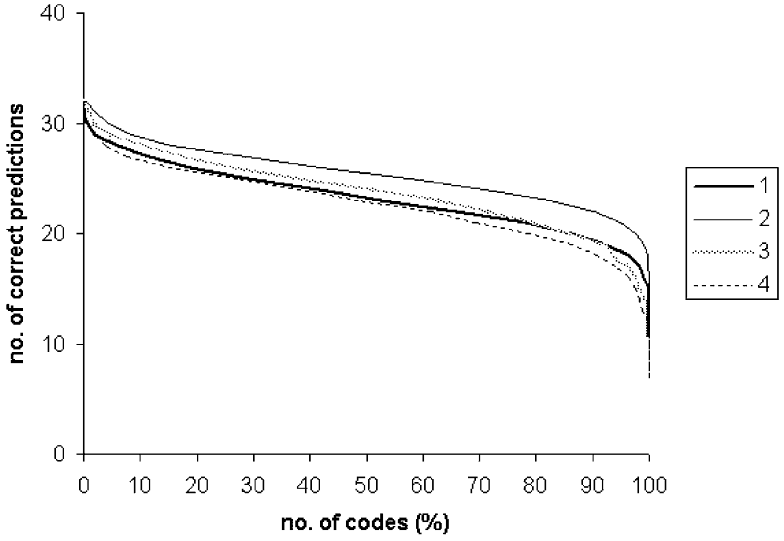

Fig. 8. Graphical representation of the distribution of correct predictions (for the training set) over the four series of experiments described in Table 2. The values on the abscissa are the percentage of codes yielding the number of correct predictions (or more) given by the ordinate.

in a characteristic region (in black) clearly different from that of the second eluted enantiomers (colored with gray). In this case, partial atomic charges were used as atomic property for the computation of e, $f_{\mathrm{CDCC}}(u)$ was sampled at 41 evenly distributed values of u between -0.060 and $+0.060\,\mathrm{e}^2\,\text{Å}^{-1}$. Combinations of four atoms in Eq. (5) with interatomic distances larger than eight bonds were neglected. The resulting 41-dimensional vectors were normalized by their vector sum.

Some of the molecules (**1**–**10**) in the data set are expected to exhibit high conformational flexibility, which might reduce the relevance of the information encoded by the CDCC (as only one conformation is used to generate the code). The importance of this fact was evaluated by: (a) repeating the experiments with only the more rigid structures (**11**–**28**); and (b) applying the CICC [10] described in Section 2.1 instead of CDCC. The results are summarized in Tables 2 and 3, and the distributions of correct predictions are represented in Figs. 8 and 9.

Use of the CICC was reasonable because the chirality in all the molecules of the data set arises from chiral carbon atoms. Globally better predictions and clustering were obtained with the CICC, and the results were less sensitive to the atomic property or to the inclusion of hydrogen atoms (Table 3). However, the best results obtained were similar to those yielded with the CDCC (Table 1). Now, series 1 and 2,

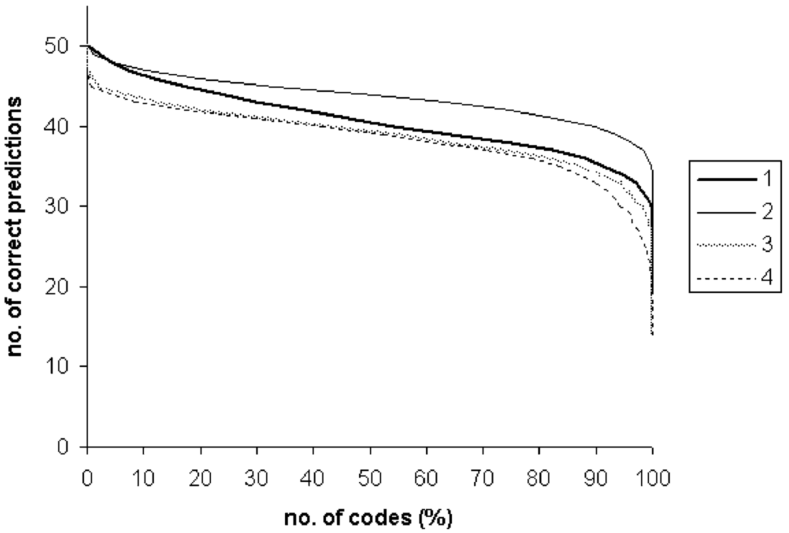

Fig. 9. Graphical representation of the distribution of correct predictions (for the training set) over the four series of experiments described in Table 3. The values on the abscissa are the percentage of codes yielding the number of correct predictions (or more) given by the ordinate.

gave very good clustering of the training set (Table 3 and Fig. 9) and the best codes of each series also revealed high predictive ability for the test set (Table 3, two right columns). Using only the more rigid structures (Table 2) good clustering was observed, although predictions for the test set were somewhat less accurate than with CICC. These facts suggest that, in this application, the conformational flexibility of some molecules does not prevent the CDCC from giving high quality results, probably because the flexible molecules (**1–10** and their enantiomers) are structurally similar and form a sub-cluster within each class of enantiomers. The results displayed in Fig. 7b support this interpretation.

The ability of the chiral stationary phase to separate a pair of enantiomers is due to the different interactions between each of the isomers and the stationary phase. The complexity of the teicoplanin structure generates a situation that is hard

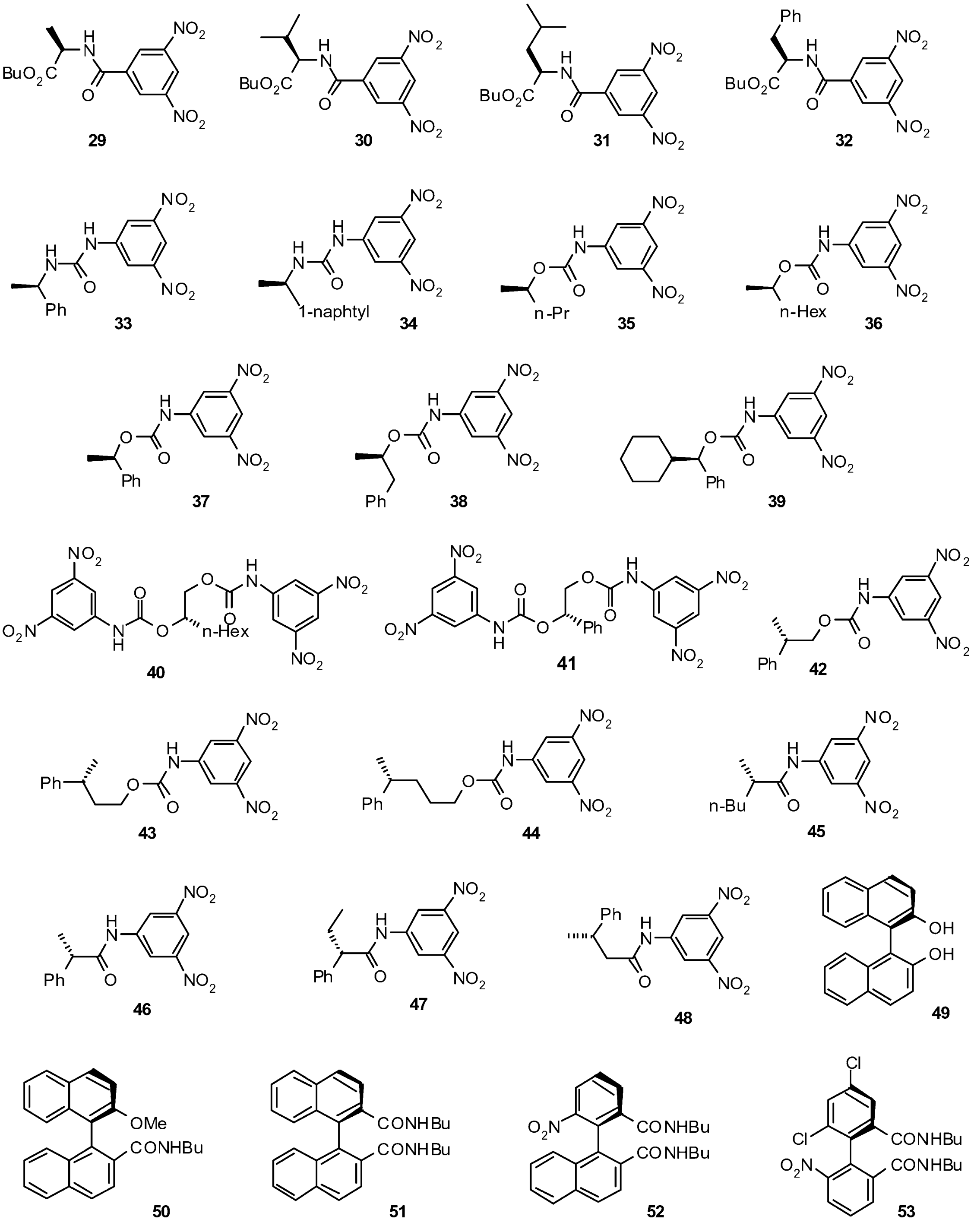

Fig. 10. First eluted enantiomers in a chromatographic separation on chiral HPLC with a bianthracene-based stationary phase.

Table 4
Results of the screening of CDCC applied to the classification of molecules **29–53** and their enantiomers by Kohonen NN of size 10 × 10

	Atomic property	H atoms considered	Number of experiments[a]	Best codes		
				Correct predictions for the training set (44 cpds)[b]	Predictions for the test set[c]	
					Correct (%)	Wrong (%)
(1)	Partial atomic charge	Yes	720	44 (15)	81	4
(2)	Partial atomic charge	No	720	44 (1) 43 (3) 42 (7)	52	26
(3)	Effective polarizability	Yes	510	43 (2) 42 (4)	57	23
(4)	Effective polarizability	No	510	41 (1) 40 (1) 39 (2)	42	33

[a] The number of experiments is not the same in the four series because some combinations of atomic property, range, atoms considered, and resolution yielded a null chirality code for some compounds and were not used.

[b] The number of codes yielding the given result is displayed in parentheses.

[c] Global percentage of right and wrong predictions obtained for the test set using the best codes referred to in the preceding column. The remaining cases were undecided.

to model and to rationalize because of the very large structure involved and the many types of possible interactions. The chirality codes/NN strategy was quite successful in this case and it revealed that both the conformation-dependent and the CICC contain relevant information in the context of enantiomeric differentiation by chromatography.

3.2. HPLC on an axially dissymmetric chiral stationary phase

In a second application we investigated the 25 chiral compounds represented in Fig. 10 that were separated from their enantiomers by HPLC on a bianthracene-based chiral stationary phase [18]. Chirality in structures **29–48** results from a chirality center (a chiral carbon atom), but a different situation occurs with compounds **49–53** that are chiral although they have no chiral carbon atom. In the latter case, chirality is the result of a chirality axis and it depends on conformation. CDCC must therefore be used in order that the chirality of the compounds can be represented and the elution order of enantiomeric pairs can be predicted.

A test set was defined with structures **30**, **46**, **50**, and their enantiomers. These were chosen to represent molecules exhibiting different kinds of chirality and possessing different functional groups. The training set consisted of the remaining 44 molecules. All the compounds were classified as ‘first eluting enantiomer’ or ‘last eluting enantiomer’.

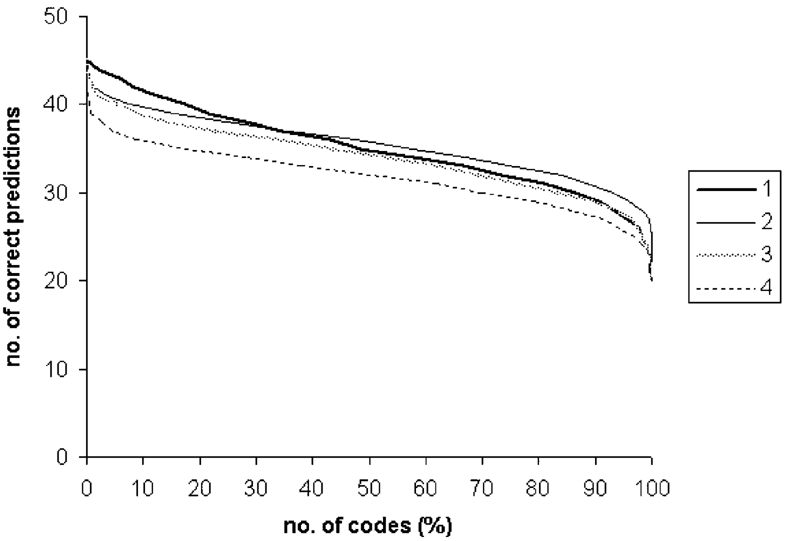

Fig. 11. Graphical representation of the distribution of correct predictions (for the training set) over the four series of experiments described in Table 4. The values on the abscissa are the percentage of codes yielding the number of correct predictions (or more) given by the ordinate.

Several codes were generated, used to train Kohonen NN, and evaluated in a similar fashion to the first application. They were calculated with (a) all types of atoms or not considering hydrogen atoms; (b) partial atomic charges or effective polarizabilities as atomic properties; (c) code lengths between 20 and 70; (d) values of u in the interval $(-s, +s)$ with s varying between 0.010 and $0.240\,e^2\,\text{Å}^{-1}$ (when charges were used) or in the interval $(0, +s)$ with s varying between 40 and $200\,\text{Å}^5$ (when polarizabilities were used); (e) combinations of four atoms with maximum interatomic path distances of 6–9 bonds, or no limit. The results are summarized in Table 4 and Fig. 11.

The best results were achieved including hydrogen atoms in the generation of the codes, and using partial atomic charges as the atomic property (Table 4, entry 1). Fifteen codes could be found that gave correct predictions for all the objects of the training set. For these 15 best codes, the average of correct predictions for the test set was 81%.

An example of a trained Kohonen NN surface, colored according to the mapping of the training set, is shown in Fig. 12. It can be seen that the two classes of enantiomers excited neurons in two distinct regions (black and gray). The objects of the test set (identified by their reference numbers) were also mapped and all of them were correctly classified (an object is classified according to the most frequent class of the winning and adjacent neurons). It should be noted that the NN recognized similarities between the enantiomers of the same class (order of elution) even though diverse structures are present, as well as different types of chirality. In this case, partial atomic charges were used as atomic property for the computation of e, $f_{\text{CDCC}}(u)$ was sampled at 41 evenly distributed values of u between -0.160 and $+0.160\,e^2\,\text{Å}^{-1}$. The resulting 41-dimensional vectors were normalized by their vector sum.

The data set can also be employed to compare the CDCC with the CICC if structures **49–53** (axially chiral) are excluded. A comparable number of experiments with Kohonen NN of 9 × 9 neurons were then performed using compounds **29–48** (Table 5 and Fig. 13). The best codes of series 1 (calculated using partial atomic charges and including hydrogen atoms) gave again the best predictions for the test set, but the experiments showed that the CICC did not generally improve the results. This observation should be connected with the structural similarities between the molecules, as was observed in the first application. It shows that, in these circumstances, the CDCC can even be applied to conformationally flexible molecules.

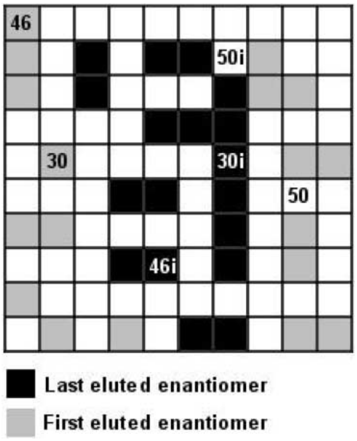

Fig. 12. Mapping of the chiral compounds from the second dataset (Fig. 6) encoded by CDCC into a 10 × 10 Kohonen NN with toroidal surface. After training, the 44 molecules from the training set were mapped and the neurons were colored accordingly (some neurons were empty and some were excited by more than one molecule). Test structures **30**, **46** and **50** and the inverted enantiomers (labeled with 'i') were presented to the trained network in order to be classified. All test examples were correctly classified as 'first eluted' or 'last eluted' enantiomer on the given bianthracene-based chiral HPLC column.

Table 5
Results of the screening of CICC applied to the classification of molecules **29–48** and their enantiomers by Kohonen NN of size 9 × 9

	Atomic property	H atoms considered	Number of experiments[a]	Best codes		
				Correct predictions for the training set (36 cpds)[b]	Predictions for the test set[c]	
					Correct (%)	Wrong (%)
(1)	Partial atomic charge	Yes	576	35 (1) 34 (5)	79	4
(2)	Partial atomic charge	No	576	36 (7) 35 (19)	48	24
(3)	Effective polarizability	Yes	648	34 (4) 33 (11)	53	22
(4)	Effective polarizability	No	545	34 (2) 33 (5)	50	39

[a] The number of experiments is not the same in the four series because some combinations of atomic property, range, atoms considered, and resolution yielded a null chirality code for some compounds and were not used.

[b] The number of codes yielding the given result is displayed in parentheses.

[c] Global percentage of right and wrong predictions obtained for the test set using the best codes referred to in the preceding column. The remaining cases were undecided.

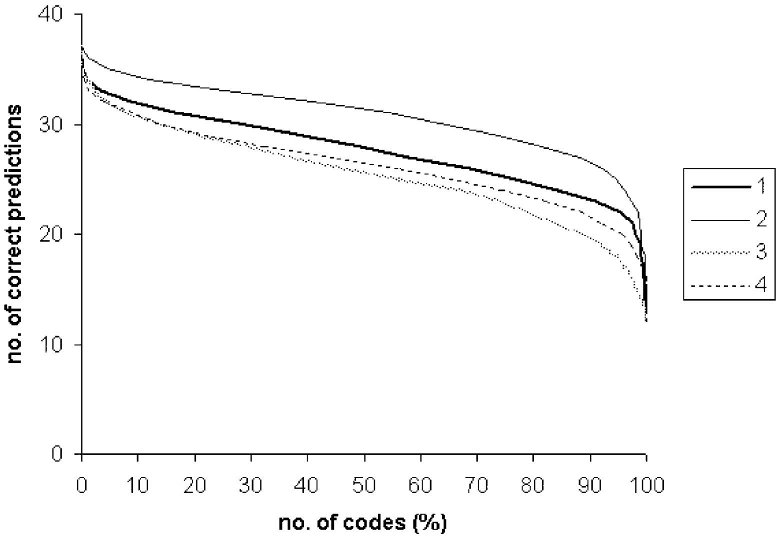

Fig. 13. Graphical representation of the distribution of correct predictions (for the training set) over the four series of experiments described in Table 5. The values on the abscissa are the percentage of codes yielding the number of correct predictions (or more) given by the ordinate.

4. Conclusion

A representation of molecular chirality by a fixed-length code was introduced, which was able to account for both central chirality and axial chirality. Its chemical significance was demonstrated by two applications in chiral chromatography.

Comparing the CDCC with the CICC for the studied applications (excluding molecules with axial chirality), it was observed that the CDCC gave results of at least the same accuracy as those obtained by the CICC, but not decisively better. The optimum type of code, set of parameters and atomic properties cannot be known a priori for a given application, as is generally the case when applying a set of molecular descriptors. Even though, for the applications described here, the use of partial atomic charge as the atomic property and inclusion of hydrogen atoms yielded good results in every case, and was usually the best choice. The set of functional parameters should also be adapted for a specific data set, the meaningful domain of $f_{\mathrm{CDCC}}(u)$ depending on the size of the molecules and range of atomic properties. The majority of the codes allowed the Kohonen NN to produce reasonable clustering of the training sets (Figs. 6, 8, 9, 11 and 13) but, to obtain more perfect self-organization, the codes had to be optimized for the specific tasks.

As a guideline, it can be said that it is adequate to represent the chirality of a molecule with CDCC when: (a) the chirality does not result from chiral carbon (or other tetrahedral) atoms, such as in atropoisomers; (b) the conformational flexibility of the compounds is low or the compounds are in a locked conformation, as in strained ring systems or in the crystalline state; (c) the code is used to compare compounds with related structural features for which similar low energy conformations can be proposed, as in a series of α-amino acids.

The CICC can be used when the chirality is due to chiral carbon atoms. It is particularly useful when it is not possible to work with a single conformation.

The method promises to have a wide range of applications, from enantioselective reactions to analytical chemistry and to the study of biological activity of chiral compounds.

Acknowledgements

J.A.S. acknowledges Fundação para a Ciência e a Tecnologia (Lisboa, Portugal) for a post-doctoral fellowship. J.G. thanks the Fonds der Chemischen Industrie for financial support.

References

[1] A.B. Buda, T. Heyde, K. Mislow, On quantifying chirality, Angew. Chem. Int. Ed. Engl. 31 (1992) 989–1007; Angew. Chem. 104 (1992) 1012–1031.

[2] D. Avnir, H.Z. Hel-Or, P.G. Mezey, Symmetry and chirality: continuous measures, in: P.V.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer III, P.R. Schreiner (Eds.), The Encyclopedia of Computational Chemistry, Vol. 4, Wiley, Chichester, 1998, pp. 2890–2901.

[3] H. Zabrodsky, D. Avnir, Continuous symmetry measures. 4. Chirality, J. Am. Chem. Soc. 117 (1995) 462–473.

[4] R. Benigni, M. Cotta-Ramusino, G. Gallo, F. Giorgi, A. Giuliani, M.R. Vari, Deriving a quantitative chirality measure from molecular similarity indices, J. Med. Chem. 43 (2000) 3699–3703.

[5] G. Moreau, Atomic chirality, a quantitative measure of the chirality of the environment of an atom, J. Chem. Inf. Comput. Sci. 37 (1997) 929–938.

[6] H.P. Schultz, E.B. Schultz, T.P. Schultz, Topological organic chemistry. 9. Graph theory and molecular topological indices of stereoisomeric organic compounds, J. Chem. Inf. Comput. Sci. 35 (1995) 864–870.

[7] J.V. Julián-Ortiz, C.G. Alapont, I. Ríos-Santamarina, R. García-Doménech, J. Gálvez, Prediction of properties of chiral compounds by molecular topology, J. Mol. Graph. Model. 16 (1998) 14–18.

[8] A. Golbraikh, D. Bonchev, A. Tropsha, Novel chirality descriptors derived from molecular topology, J. Chem. Inf. Comput. Sci. 41 (2001) 147–158.

[9] J.S. Mason, I. Morize, P.R. Menard, D.L. Cheney, C. Hulme, R.F. Labaudiniere, New four-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures, J. Med. Chem. 42 (1999) 3251–3264.

[10] J. Aires-de-Sousa, J. Gasteiger, A new description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions, J. Chem. Inf. Comput. Sci. 41 (2001) 369–375.

[11] K.B. Lipkowitz, Atomistic modeling of enantioselective binding, Acc. Chem. Res. 33 (2000) 555–562.

[12] K.B. Lipkowitz, D.A. Demeter, R. Zegarra, R. Larter, T. Darden, A protocol for determining enantioselective binding of chiral analytes on chiral chromatographic surfaces, J. Am. Chem. Soc. 110 (1988) 3446–3452.

[13] R. Däppen, H.R. Karfunkel, F.J.J. Leusen, Computational chemistry applied to the design of chiral stationary phases for enantiomeric separation, J. Comp. Chem. 11 (1990) 181–193.

[14] J. Aerts, An improved molecular modeling method for the prediction of enantioselectivity, J. Comp. Chem. 16 (1995) 914–922.

[15] T.D. Booth, K. Azzaoui, I.W. Wainer, Prediction of chiral chromatographic separations using combined multivariate regression and neural networks, Anal. Chem. 69 (1997) 3879–3883.

[16] R. Kaliszan, T.A.G. Noctor, I.W. Wainer, Quantitative structure-enantioselective retention relationships for the chromatography of 1,4-benzodiazepines on a human serum albumin based HPLC chiral stationary phase: an approach to the computational prediction of retention and enantioselectivity, Chromatographia 33 (1992) 546–550.

[17] A. Péter, G. Török, D.W. Armstrong, High-performance liquid chromatographic separation of enantiomers of unusual amino acids on a teicoplanin chiral stationary phase, J. Chromatogr. A 793 (1998) 283–296.

[18] S. Oi, H. Ono, H. Tanaka, M. Shijo, S. Miyano, Axially dissymmetric bianthracene-based chiral stationary phase for the high-performance liquid chromatographic separation of enantiomers, J. Chromatogr. A 679 (1994) 35–46.

[19] M.C. Hemmer, V. Steinhauer, J. Gasteiger, The prediction of the 3D structure of organic molecules from their infrared spectra, Vibrat. Spectrosc. 19 (1999) 151–164.

[20] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, Tetrahedron 36 (1980) 3219–3228.

[21] J. Gasteiger, H. Saller, Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept, Angew. Chem. Int. Ed. Engl. 24 (1985) 687–689; Angew. Chem. 97 (1985) 699–701.

[22] J. Gasteiger, M.G. Hutchings, Empirical models of substituent polarisability and their application to stabilisation effects in positively charged species, Tetrahedron Lett. 24 (1983) 2537–2540.

[23] J. Gasteiger, M.G. Hutchings, Quantification of effective polarisability. Applications to studies of X-ray photoelectron spectroscopy and alkylamine protonation, J. Chem. Soc. Perkin 2 (1984) 559–564.

[24] http://www2.chemie.uni-erlangen.de/software/petra/.

[25] T. Kohonen, Self-Organization and Associative Memory, Springer, Berlin, 1988.

[26] J. Gasteiger, J. Zupan, Neural networks in chemistry, Angew. Chem. Int. Ed. Engl. 32 (1993) 503–527; Angew. Chem. 105 (1993) 510–536.

[27] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, 2nd Edition, Wiley, Weinheim, 1999.

[28] J. Sadowski, J. Gasteiger, From atoms and bonds to three-dimensional atomic coordinates: automatic model builders, Chem. Rev. 93 (1993) 2567–2581.

[29] J. Gasteiger, C. Rudolph, J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules, Tetrahedron Comput. Methods 3 (1992) 537–547.

[30] J. Sadowski, C. Rudolph, J. Gasteiger, The generation of 3D-models of host–guest complexes, Anal. Chim. Acta 265 (1992) 233–241.

[31] J. Sadowski, J. Gasteiger, G. Klebe, Comparison of automatic three-dimensional model builders using 639 X-ray structures, J. Chem. Inf. Comput. Sci. 34 (1994) 1000–1008.

[32] A. Teckentrup, Ph.D. thesis, University of Erlangen-Nürnberg, Erlangen, Germany, 2000.

[33] Chirobiotic Handbook, 3rd Edition, Advanced Separation Technologies Inc. (ASTEC), Whippany, NJ, 1999.