JOC*Article*

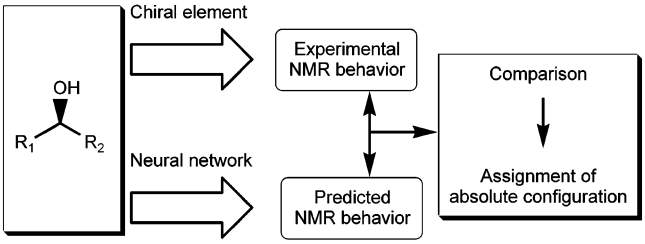
Automatic Assignment of Absolute Configuration from 1D NMR Data

Qing-You Zhang, Gonçalo Carrera, Mário J. S. Gomes, and João Aires-de-Sousa*

Departamento de Química, CQFB and REQUIMTE, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Monte de Caparica, Portugal

jas@fct.unl.pt

Received November 5, 2004



Opposite enantiomers exhibit different NMR properties in the presence of an external common chiral element, and a chiral molecule exhibits different NMR properties in the presence of external enantiomeric chiral elements. Automatic prediction of such differences, and comparison with experimental values, leads to the assignment of the absolute configuration. Here two cases are reported, one using a dataset of 80 chiral secondary alcohols esterified with (*R*)-MTPA and the corresponding ^{1}H NMR chemical shifts and the other with 94 ^{13}C NMR chemical shifts of chiral secondary alcohols in two enantiomeric chiral solvents. For the first application, counterpropagation neural networks were trained to predict the sign of the difference between chemical shifts of opposite stereoisomers. The neural networks were trained to process the chirality code of the alcohol as the input, and to give the NMR property as the output. In the second application, similar neural networks were employed, but the property to predict was the difference of chemical shifts in the two enantiomeric solvents. For independent test sets of 20 objects, 100% correct predictions were obtained in both applications concerning the sign of the chemical shifts differences. Additionally, with the second dataset, the difference of chemical shifts in the two enantiomeric solvents was quantitatively predicted, yielding r^2 0.936 for the test set between the predicted and experimental values.

1. Introduction

Determination of absolute configuration of new chiral compounds is usually the last step of structure elucidation. While the molecular formula, the connection table, and the relative stereochemistry of organic compounds can be normally deduced from mass spectrometry, IR, and NMR spectroscopy of pure samples, the assignment of absolute configuration from two enantiomeric possibilities is more problematic. Following the work of Bijvoet,[1] crystallographic methods can unequivocally assign absolute configuration, but they require at least a crystalline sample. Common alternatives include (a) chemical transformations leading to a compound of known configuration, (b) attachment of a chiral block of known configuration for further determination of relative stereochemistry, (c) interpretation of chiroptical properties,[2,3] and (d) interpretation of NMR behavior in the presence of chiral solvating agents.[4,5]

In the absence of an external chiral element, opposite enantiomers exhibit the same properties in NMR spectroscopy. But the situation changes with a chiral solvating agent or after covalent bonding to a standard chiral

(1) Bijvoet, J. M.; Peerdeman, A. F.; Vanbommel, A. J. Determination of the absolute configuration of optically active compounds by means of X-rays. *Nature* **1951**, *168*, 271−272.

(2) Berova, N.; Nakanishi, K.; Woody, R. W. *Circular Dichroism: Principles and Applications*, 2nd ed.; John Wiley & Sons: New York, 2000.

(3) Polavarapu, P. L. Optical rotation: recent advances in determining the absolute configuration. *Chirality* **2002**, *14*, 768−781.

(4) Seco, J. M.; Quinoa, E.; Riguera, R. The assignment of absolute configuration by NMR. *Chem. Rev.* **2004**, *104*, 17−117.

(5) Wenzel, T. J.; Wilcox, J. D. Chiral reagents for the determination of enantiomeric excess and absolute configuration using NMR spectroscopy. *Chirality* **2003**, *15*, 256−270.

10.1021/jo048029z CCC: $30.25
Published on Web 02/22/2005

block (transformation of enantiomers into diastereoisomers).[4,5] In those cases, opposite stereoisomers can be discriminated by NMR spectroscopy—their NMR properties become different. It should be possible to build models that correlate those properties with the absolute configuration. If reliable models are established, they can strongly support the assignment of absolute configuration, particularly if predictions can be simultaneously obtained on the basis of different techniques/models.[6]

Covalent bonding of enantiomers to homochiral Mosher acid is a common technique to discriminate enantiomers by NMR spectroscopy and to measure the enantiomeric excess. Kelly compiled an excellent data set of Mosher esters of secondary alkanols, including the ^{1}H NMR chemical shifts of the methoxide group of the Mosher esters.[7] Some rules were proposed to assign the absolute configuration of alkanols from the comparison between the chemical shifts of the Mosher esters of opposite enantiomers.[7] A certain configuration of the chiral carbon atom is associated with the enantiomer yielding the (*R*)-Mosher ester with higher chemical shift. However, this rule requires some considerations about conformation, subjective decisions on inverting or not CIP rules, and it has been tested more for explanation than for prediction. It is therefore hardly suitable for automatic implementation. In the first part of this paper, we present an automatically developed model that predicts if a given enantiomer, after esterification with (*R*)-Mosher acid, has a higher or lower chemical shift (of the Mosher acid methoxide group) than its enantiomer. Such a model can be used to assign the absolute configuration from the NMR chemical shifts of the two enantiomers. The approach is based on counterpropagation neural networks (CPG NNs), which are trained with chirality codes[8] of the enantiomers. The chirality codes are calculated from the molecular structure and represent the molecular chirality.[9,10] Such a numerical fixed-length representation of the molecular chirality is required for input to a NN. The NN learns the relationship between the chirality codes of the chiral structures and the corresponding NMR properties, from a set of training examples. It is then able to make predictions for an independent test set.

A second investigation is presented on the relationship between the chiral molecular structure of secondary alkanols and their ^{13}C NMR properties in chiral deuterated solvents. This study uses a database produced by Kishi and co-workers[11−13] with the ^{13}C NMR chemical shifts of atoms adjacent to the chiral hydroxymethine center, taken in (*R*,*R*)-BMBA-*p*-Me and (*S*,*S*)-BMBA-*p*-Me chiral solvents. Here we report the estimation of the difference between the chemical shifts of a given carbon atom in the two enantiomeric solvents by counterpropagation neural networks. The neural networks receive as input a representation of the atom (an atomic chirality code) and gives as output the difference of its chemical shift in the (*R*,*R*)-BMBA-*p*-Me and (*S*,*S*)-BMBA-*p*-Me solvent. For the practical assignment of absolute configuration to a pair of enantiomers, the difference of the experimental chemical shifts in the enantiomeric solvents would be compared with the predicted shifts for both enantiomers.

2. Methodology

2.1. Data Sets. Two data sets were investigated. The first data set consists of a series of 40 secondary alkanols (**1**−**40**) and their enantiomers, which are the alkanol moieties of α-methoxy-α-(trifluoromethyl)phenylacetic acid (MTPA or Mosher acid) esters retrieved from the literature.[7] If the ^{1}H NMR chemical shift of the methoxy group of (*R*)-MTPA ester of an alkanol was higher than that of the alkanol enantiomer, the chemical shift difference for this enantiomer is considered to be positive, while the chemical shift difference for the opposite enantiomer is considered negative. Thus, the absolute configuration of the alkanol is associated with a chemical shift difference. Figure 1 shows the enantiomers of the data set exhibiting a positive difference of chemical shifts. Figure 2 illustrates the procedure for the determination of the chemical shift difference. The 40 enantiomeric pairs were divided into a training set and a test set. The test set consists of 20 molecules (**3**, **8**, **14**, **19**, **23**, **25**, **26**, **31**, **35**, **40**, and their enantiomers) that were chosen to cover a variety of skeletons and were not used for training—the training set was composed of the remaining 60 compounds.

The second data set is based on 24 chiral alcohols (**41**−**64**) for which ^{13}C NMR spectra were taken in chiral bidentate solvents (*R*, *R*)- and (*S*, *S*)-BMBA-*p*-Me. The structures and chemical shift differences $\Delta\delta_{RR-SS} = \delta_{(R,R)} - \delta_{(S,S)}$ for carbon atoms adjacent to the hydroxymethine unit were retrieved from the literature[11−13] and are shown in Figure 3. The chemical shift differences $\Delta\delta_{RR-SS}$ have opposite signs for corresponding carbon atoms in opposite enantiomers. The data set includes non-zero $\Delta\delta_{RR-SS}$ values for 47 carbon atoms in 24 chiral alcohols. Therefore, 94 atoms (47 carbon atoms in each of the two series of enantiomers) were investigated. They were partitioned into a test set of 20 atoms (from the compounds **42**, **48**, **54**, **58**, **62**, and their enantiomers) and a training set consisting of the remaining 74 atoms.

It must be emphasized that in both studies the two enantiomers of each enantiomeric pair were always included together either in the training set or in the test set.

2.2. Conformation-Independent Chirality Code (CICC). We have introduced a conformation-independent chirality code (CICC) that quantitatively describes the stereochemical situation at chiral centers.[8] Here only a brief explanation of the descriptors is given. First, a value of e_{ijkl} was defined through eq 1 that considers atoms i, j, k, and l, each of them belonging to a different ligand of a chiral center.

$$e_{ijkl} = \frac{a_i a_j}{r_{ij}} + \frac{a_i a_k}{r_{ik}} + \frac{a_i a_l}{r_{il}} + \frac{a_j a_k}{r_{jk}} + \frac{a_j a_l}{r_{jl}} + \frac{a_k a_l}{r_{kl}} \qquad (1)$$

(6) For a review on chromatography in the context of determination of absolute configuration, see: Roussel, C.; Rio, A. D.; Pierrot-Sanders, J.; Piras, P.; Vanthuyne, N. *J. Chromatogr. A* **2004**, *1037*, 311−328.

(7) Kelly, D. R. A new method for the determination of the absolute stereochemistry of aromatic and heteroaromatic alkanols using Mosher's esters. *Tetrahedron: Asymmetry* **1999**, *10*, 2927−2934.

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

(9) Aires-de-Sousa, J.; Gasteiger, J. Prediction of enantiomeric selectivity in chromatography. Application of conformation-dependent and conformation-independent descriptors of molecular chirality. *J. Mol. Graph. Model.* **2002**, *20*, 373−388.

(10) Aires-de-Sousa, J.; Gasteiger, J.; Gutman, I.; Vidović, D. Chirality codes and molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 831−836.

(11) Kobayashi, Y.; Hayashi, N.; Kishi, Y. Toward the creation of NMR databases in chiral solvents: bidentate chiral NMR solvents for assignment of the absolute configuration of acyclic secondary alcohols. *Org. Lett.* **2002**, *4*, 411−414.

(12) Kobayashi, Y.; Hayashi, N.; Kishi, Y. Application of chiral bidentate NMR solvents for assignment of the absolute configuration of alcohols: scope and limitation. *Tetrahedron Lett.* **2003**, *44*, 7489−7491.

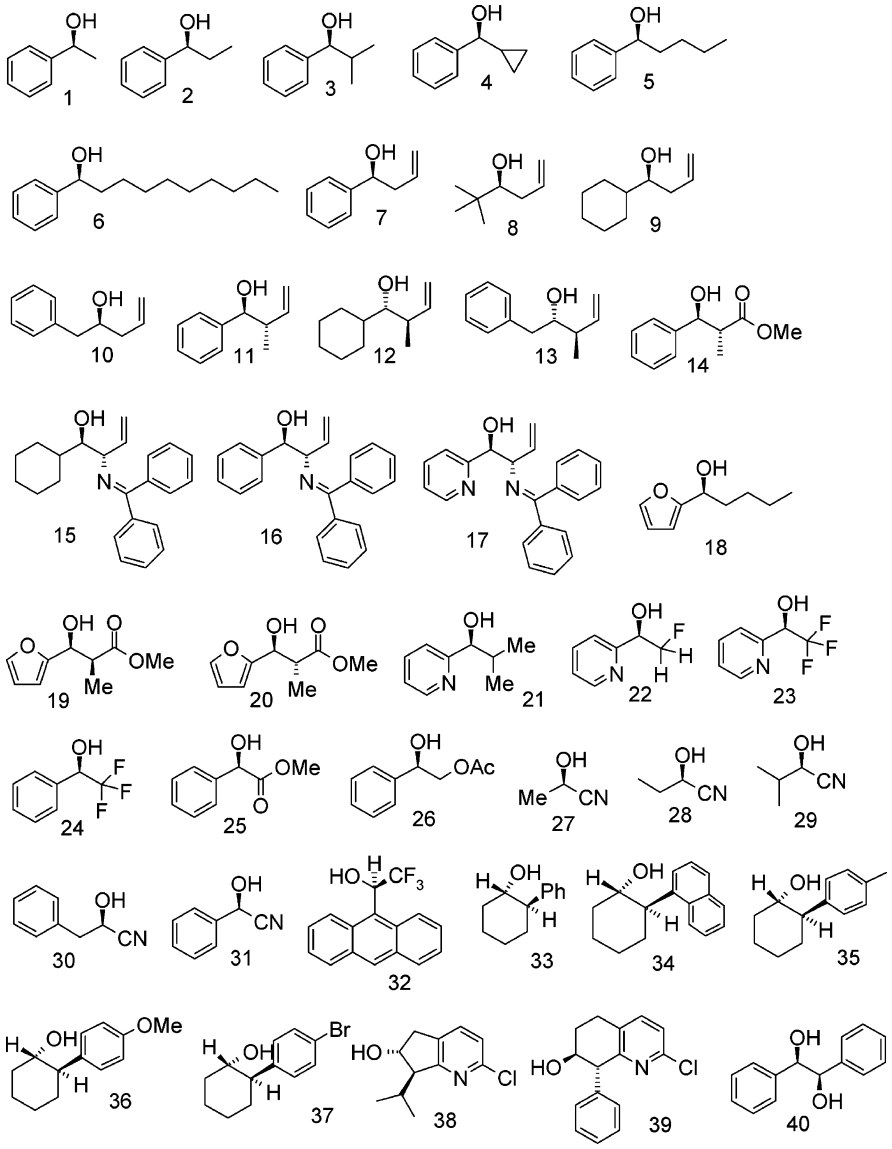


FIGURE 1. Data set of secondary chiral alkanols giving rise to (*R*)-MTPA ester derivatives with higher ^{1}H NMR chemical shifts (MeO group) than their enantiomers.

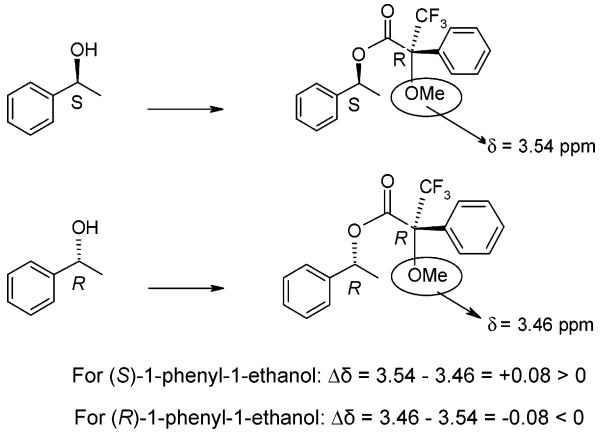


FIGURE 2. Definition of chemical shift difference (Δδ) after esterification with (*R*)-MTPA.

a_i is a property of atom i, such as atomic charge, and r_{ij} is a distance between atoms i and j. 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 the minimum number of bond counts.

Furthermore, a chirality sign, 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 sign, 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 three-dimensional 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}.

(13) Kobayashi, Y.; Czechtizky, W.; Kishi, Y. Complete stereochemisty of tetrafibricin. *Org. Lett.* **2003**, *5*, 93−96.

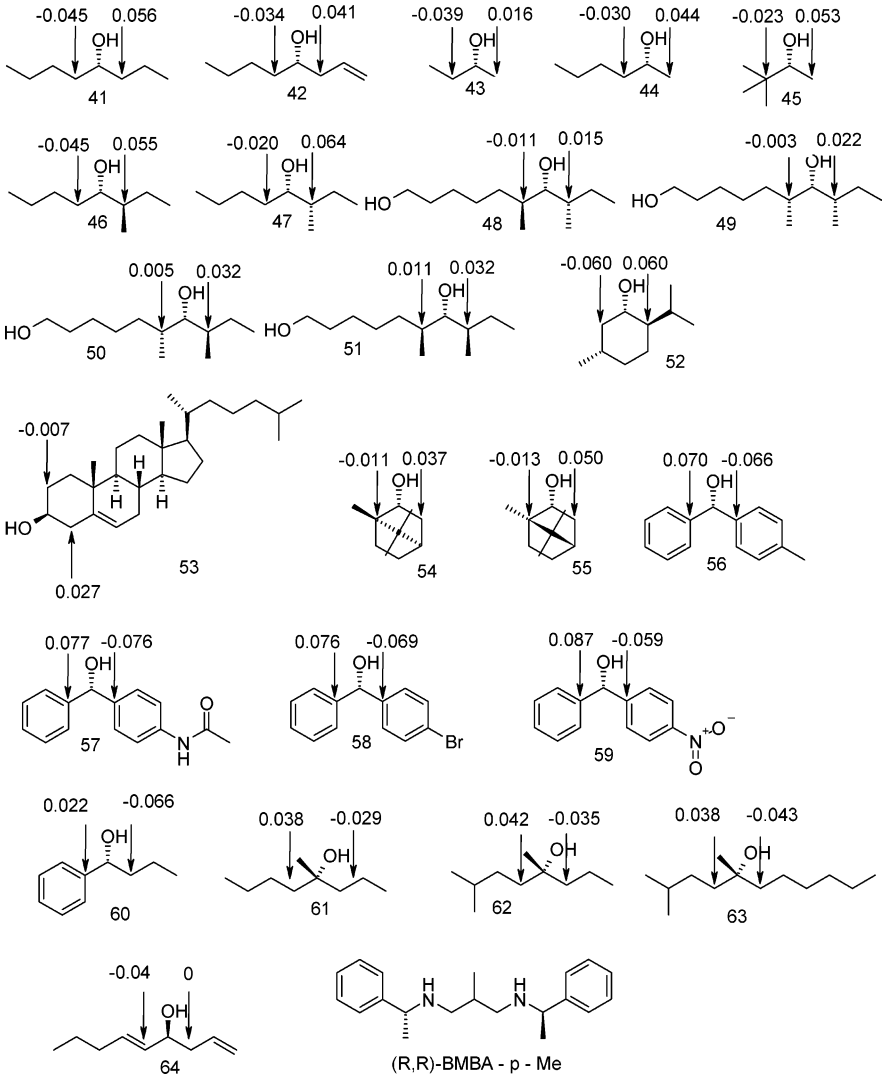


FIGURE 3. Data set of chiral alkanols and the differences of ^{13}C NMR chemical shifts of carbon atoms adjacent to the hydroxymethine unit in enantiomeric solvents (*R*,*R*)- and (*S*,*S*)-BMBA-*p*-Me.

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*, $f_{CICC}(u)$, using eq 2, where n_A, n_B, n_C, and n_D are the number of atoms belonging to ligands A, B, C, and D, respectively:

$$f_{CICC}(u) = \sum_{i}^{n_A}\sum_{j}^{n_B}\sum_{k}^{n_C}\sum_{l}^{n_D} s_{ijkl} \exp[-b(u - e_{ijkl})^2] \qquad (2)$$

$f_{CICC}(u)$ is calculated at a number of discrete points 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 points of $f_{CICC}(u)$ 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 $f_{CICC}(u)$ vs *u*.[8]

2.3. Conformation-Dependent Chirality Code (CDCC). CDCC is a more general conformation-dependent description of molecular chirality.[9,10] 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 1, with r_{ij} again being the sum of bond lengths between atoms on the path with minimum number of bond counts. *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 (the rules are described in ref 9). If a set of four atoms is an achiral set, i.e., it is superposable on its mirror image, then it is not further considered. The property, a_i, of an atom should have values that allow one to distinguish between nonequivalent atoms. For that purpose we have selected partial atomic charges,[14,15] or polarizabilities,[16,17] calculated by PETRA[18] because this software rapidly assigns

(14) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

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 expected to be of much higher influence on the properties 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 NMR chemical shift.

c is defined for each combination of atoms i, j, k, and l by eq 3, 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 3, 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.

$$c_{ijkl} = \frac{x_j y_k z_l}{x_j y_k + x_j|z_l| + y_k|z_l|} \quad (3)$$

The two values, e_{ijkl} and c_{ijkl}, calculated for all combinations of four atoms, are then combined to generate a *conformation-dependent chirality code*, f_{CDCC}, using eq 4, where n is the number of atoms in each molecule and c introduces the conformation dependence:

$$f_{\mathrm{CDCC}}(u) = \sum_{i}^{n}\sum_{j}^{n-1}\sum_{k}^{n-2}\sum_{l}^{n-3} c_{ijkl} \exp[-b(u - e_{ijkl})^2] \quad (4)$$

$f_{\mathrm{CDCC}}(u)$ 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. As for the CICC code, 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.

2.4. Atomic Conformation-Dependent Chirality Code (aCDCC). The chirality code CDCC described in section 2.3 is a molecular code; i.e., the code is the representation of the whole molecule. Introduction of an atomic chirality code is necessary to account for local chirality around an atom, which is relevant for modeling atomic properties such as the NMR chemical shift. An *atomic* code means that every atom rather than the whole molecule has its own chirality code. Two atomic codes derived from CDCC are introduced.[19] For the calculation of the atomic chirality code for an atom a, the same eq 4 is used, but now only combinations of four atoms including atom a are considered, i.e., $i = a$ or $j = a$ or $k = a$ or $l = a$.

Thus, the sum of atomic chirality codes aCDCC for all the atoms in a molecule equals four times the molecular chirality code CDCC.

To further emphasize the special role of atom a in eq 4, a second version of the atomic chirality code was put forward—aCDCC_2. In this case, the highest priority among the four atoms of any combination is assigned to atom a, for the calculation of c_{ijkl}. This can be represented by changing eq 4 into eq 5:

$$f_{\mathrm{aCDCC\_2}}(u) = \sum_{j}^{n-1}\sum_{k}^{n-2}\sum_{l}^{n-3} c_{ijkl} \exp[-b(u - e_{ijkl})^2] \; (i = a) \quad (5)$$

An example of an atomic chirality code aCDCC_2 for the enantiomers of 4-octanol (**41**) is shown in Figure 4.

A 3D molecular structure as well as atomic properties are required for calculating the chirality code. The Cartesian coordinates of the atoms were calculated from the connection tables of the molecules by the 3D structure generator CORINA.[20−23] The physicochemical atomic properties (partial atomic charge and effective polarizability) were calculated using fast empirical methods implemented in the program package PETRA 3.11,[18] charges by the PEOE method,[14,15] and effective polarizabilities by a published procedure.[16,17] Chirality codes were calculated with 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 Windows platform.

2.5. Counterpropagation Neural Network. To model the relationship between the chirality codes and the corresponding chemical shift differences counterpropagation neural networks (CPG NN)[24] were used. CPG networks were chosen because they are useful for the modeling of complex and nonlinear relationships and they yield maps that can be visually interpreted.

The input data for a CPG network are stored in a two-dimensional 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 chirality codes. In Figure 5 the upper block represents this part of the CPG network, which is basically a Kohonen[25] network, or self-organizing map (SOM). The output data (in this case the chemical shift differences) are stored in a second layer that acts as a look-up table.

Before the training of a CPG network starts, random weights are generated. During the training, 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 codes). The weights of the winning neuron are then adjusted to make them even more similar to the presented data, and the weight of the corresponding output neuron is adjusted to become closer to the experimental chemical shift difference. The neurons in the neighborhood of the winning neuron are also corrected, the extent of adjustment depending on the topological distance to the central neuron. The network is trained iteratively, i.e., all the objects of the training set

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

(16) Gasteiger, J.; Hutchings, M. G. Empirical models of substituent polarisability and their application to stabilisation effects in positively charged species. *Tetrahedron Lett.* **1983**, *24*, 2537−2540.

(17) Gasteiger, J.; Hutchings, M. G. Quantification of effective polarisability. Applications to studies of X-ray photoelectron spectroscopy and alkylamine protonation. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559−564.

(18) http://www2.chemie.uni-erlangen.de/software/petra/

(19) For a quantitative measure of atomic chirality see: Moreau, G. Atomic chirality, a quantitative measure of the chirality of the environment of an atom. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 929−938.

(20) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(21) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1992**, *3*, 537−547.

(22) Sadowski, J.; Rudolph, C.; Gasteiger, J. The generation of 3D-models of host−guest complexes. *Anal. Chim. Acta* **1992**, *265*, 233−241.

(23) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(24) For detailed description of neural networks, see: (a) Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527; *Angew. Chem.* **1993**, *105*, 510−536. (b) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.

(25) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer: Berlin, 1989.

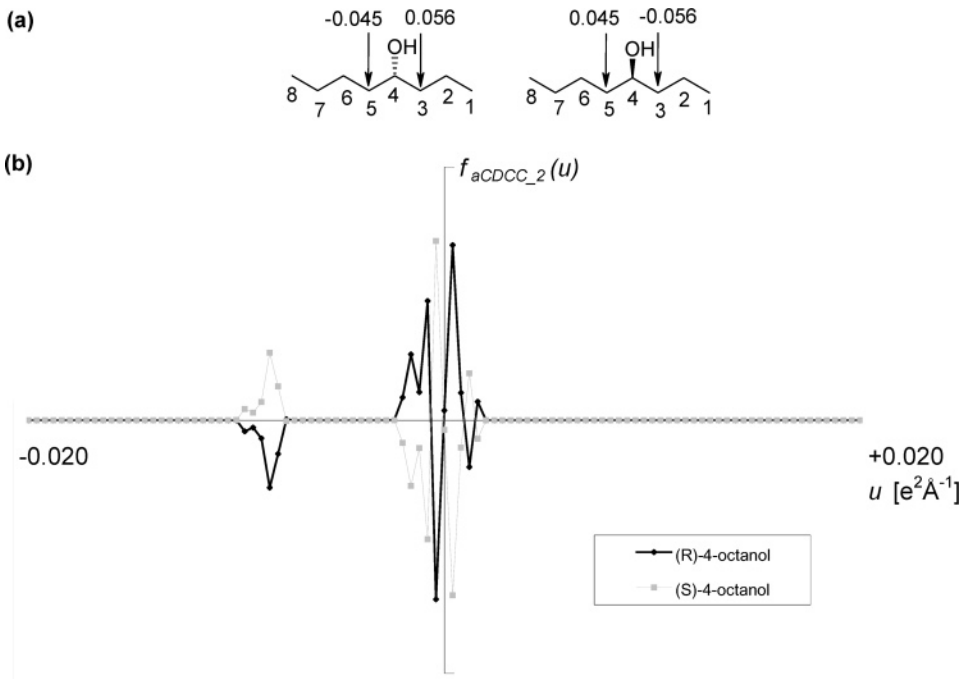


FIGURE 4. (a) Structure of (*R*)-4-octanol and (*S*)-4-octanol with chemical shift differences in NMR chiral solvents (*R*,*R*)-BMBA-*p*-Me and (*S*,*S*)-BMBA-*p*-Me. (b) Atomic chirality codes aCDCC_2 of carbon 5 of (*R*)-4-octanol and (*S*)-4-octanol.

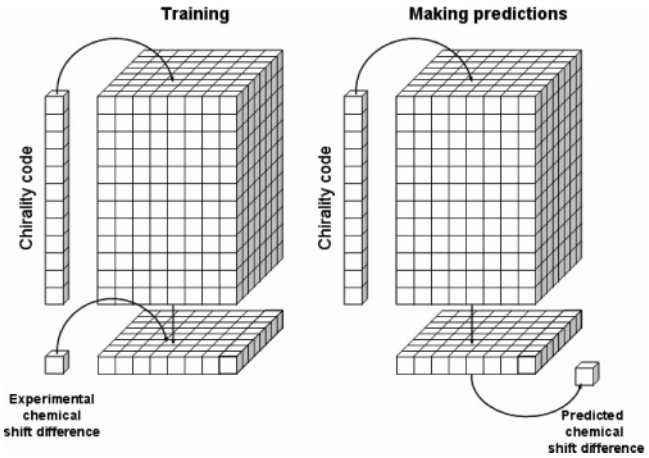


FIGURE 5. Representation of a counterpropagation neural network (CPG NN). Every small box of the network block represents a weight. The CPG NN is trained by iterative presentation of objects (chirality code and the corresponding observed NMR chemical shifts difference). After the training, the NN is able to predict the chemical shift difference on input of a chirality code.

are presented several times, and the weights are corrected, until the network stabilizes. Note that chemical shift differences are not used in determining the winning neuron.

After the training, the CPG NN is able to predict the chemical shift difference on input of an object represented by its chirality code. The winning neuron is chosen and the corresponding weight in the output layer is used for prediction (Figure 5).

2.6. Selection of Variables Using Genetic Algorithms. Chirality codes are spectrum-like representations with a number of descriptors that depends on the range and resolution. Some of the descriptors may be not relevant for our purposes, and can even introduce noise. Models with few descriptors are usually preferred for increased robustness. Here we used genetic algorithms for the selection of variables produced by the chirality codes. Genetic algorithms simulate the evolution of a population, where each individual of the population represents a subset of descriptors and its fitness is assessed by the ability to generated accurate models. Although thorough descriptions of the technique can be easily obtained from several sources,[26] an overview of the method is here included.

Each individual in the population represents a subset of descriptors and is defined by its chromosome. A chromosome has as many genes as there are possible descriptors (code length of chirality code)—each gene corresponds to one descriptor. One gene takes a value of 1 if the corresponding descriptor is included in the subset, and it takes a value of 0 if the descriptor does not belong to the subset represented by the individual (Figure 6a).

At the beginning of the evolution, the chromosomes are randomly generated. In order that the number of genes with value 1 is kept relatively low (small subsets of descriptors) the probability of generating 1 for a gene was set (randomly for each new chromosome) between 0 and 0.4.

A population of individuals is allowed to evolve over a number of generations. In each generation, half of the population die, and the other half survive (the fittest individuals). Each of the surviving individuals mates with another (randomly chosen) surviving individual, and two new offspring are generated. The chromosomes of the offspring result from crossover of their parents' chromosomes, followed by random mutation (Figure 6b,c). The population of the next generation consists of the new offspring and their parents.

Crossover occurs at a randomly chosen single point. Mutation is allowed to occur at every gene of the new offspring with a random probability. The probability of mutation 0→1 is set for each individual (randomly) between 0 and 5%. The probability of mutation 1→0 is set for each individual (randomly) between 4 and 6 times higher than the probability of mutation 0→1.

The evaluation (scoring) of each chromosome is made by a CPG neural network that uses the subset of descriptors

(26) Homeyer, A. V. Evolutionary Algorithms and their Applications in Chemistry. In *Handbook of Chemoinformatics,* Gasteiger, J., Engel, J., Ed.; Wiley-VCH: New York, 2003; Vol. 3, pp 1239−1280.

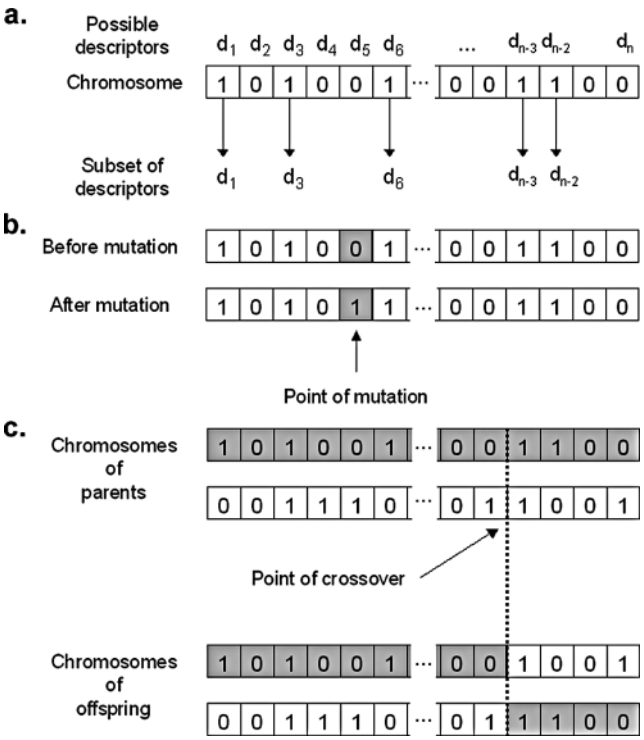


FIGURE 6. (a) Chromosome defines a subset of descriptors from the entire chirality code (descriptors d_1–d_n). (b) Mutation of one gene. (c) Crossover between chromosomes of parents, and the resulting chromosomes of the offspring.

encoded in the chromosome for predicting chemical shift differences. This CPG NN works in a similar way to those described above. A selection of the chirality code values (encoded by the chromosome) is fed to the NN obtaining a chemical shift difference as the output. The NN is trained with the training set, and the score of one chromosome (fitness function) is the root-mean-square of errors for the predictions obtained for the training set. Chromosomes giving lower errors are considered to be fitter than those giving higher errors and are selected for mating.

3. Results and Discussion

3.1. Prediction of Absolute Configuration of Secondary Alcohols from NMR Data of Their Mosher Esters. The 80 alcohols (40 enantiomeric pairs) were encoded by chirality codes (CDCC or CICC). These codes can be generated with different options such as resolution, range of u, smoothing parameter, atomic properties, and types of atoms considered. By changing these variables, 199 different codes were generated and screened. The different codes were calculated with a code length of 51 values by using (a) the partial atomic charge or effective polarizability as atomic property (b) all types of atoms or not considering hydrogen atoms; (c) values of u in the interval $[-r, +r]$ with r varying from 0.030 to 0.200 e^2 Å^{-1} for partial atomic charges or in the interval $[0, +r]$ with r varying between 50 and 500 Å for effective polarizabilities; (d) combinations of four atoms with maximum interatomic path distances of 5, 8 or any number of bonds; (e) the smoothing parameter b was set to (code length/range of u)^2. The 51-dimensional vectors were normalized by their vector sum.

The experimental output for a given enantiomer was set to +1 or −1 if its chemical shift difference is positive or negative, respectively. As described in the Methodology section, the chemical shift difference is calculated from the chemical shifts of the MeO group in the (*R*)-MTPA ester derivatives. A CPG NN size of 8 × 8 was trained with the chirality codes and chemical shift differences of 60 alcohols (training set). The trained network was then used to make predictions for the test set. If the output value was positive, the alcohol was predicted to give rise to a chemical shift higher than its enantiomer and vice versa.

TABLE 1. Prediction of NMR Behavior of 60 Mosher Acid Derivatives by CPG NNs of Size 8 × 8 Trained on the Basis of 199 Different Chirality Codes

atomic property	hydrogen considered	conformation-independent	no. of expts[a]	correct predictions—best codes[b]
partial atomic charge	no	no	24	54 (4)
				53 (1)
				52 (1)
	no	yes	24	52 (4)
				50 (2)
				49 (1)
	yes	no	24	52 (6)
				51 (2)
				50 (3)
	yes	yes	24	58 (1)
				57 (1)
				55 (2)
effective polarizability	no	no	23	52 (1)
				50 (1)
				49 (1)
	no	yes	26	54 (1)
				53 (1)
				51 (3)
	yes	no	27	51 (2)
				50 (2)
				49 (1)
	yes	yes	27	56 (1)
				55 (2)
				53 (3)

[a] Different codes were tested with different ranges and different limits for the maximum interatomic distance allowed within a set of four atoms; [b] Number of correct predictions for the training set of 60 compounds. The number of codes yielding the given result is displayed in parentheses.

The quality of clustering in a CPG NN can be assessed to a large extent 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 training set* without considering the results for the test set. To reduce the impact of fluctuations derived from the random values of the weights at the outset of the training, and the random order by which examples are presented during the training, five CPG networks were trained independently, and the average value of the five outputs was used as the output. Table 1 shows the best results obtained for the training set and the number of codes that yielded these results. The results are given separately for the codes using partial atomic charges or effective polarizabilities, for codes adopting CICC or CDCC algorithm, and for codes considering all the types of atoms or neglecting hydrogen atoms. The best three results for the training set were displayed in the last column of Table 1.

The chirality codes encoded by CICC with all types of atoms and partial atomic charges could be found to be the best code. It allowed for correct prediction of 58 out

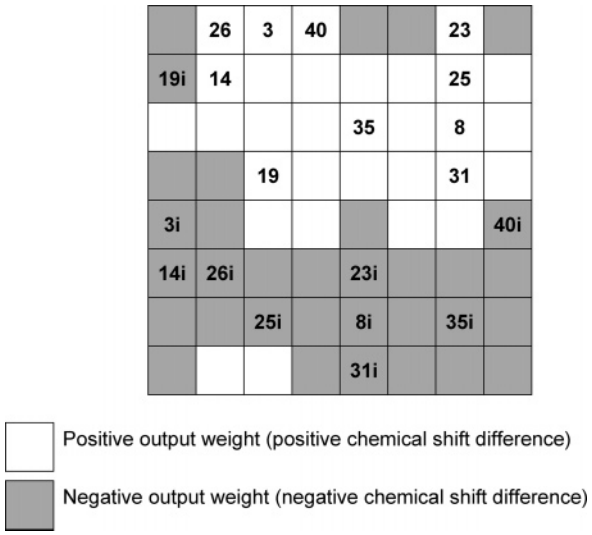


FIGURE 7. Representation of the output weights of a 8 × 8 CPG NN after training with 60 chiral alcohols encoded by CICC descriptors. After the training, the 20 molecules of the test set (**3**, **8**, **14**, **19**, **23**, **25**, **26**, **31**, **35**, **40**, and their enantiomers) were also mapped for classification—their labels include an "i" if the experimental chemical shift difference is negative.

of 60 cases (the five individual networks gave 57, 57, 56, 56, and 57 correct predictions, respectively). When an ensemble of five networks, trained with such a code, was applied to the test set, all the 20 objects in the test set were correctly predicted (individual networks giving 18, 20, 18, 18, and 20 correct predictions). The chirality code was generated with partial atomic charges as the atomic property, $f_{CICC}(u)$ was sampled at 51 evenly distributed values of u between −0.030 and +0.030 e^2 Å^{-1}, combinations of four atoms with maximum interatomic path distances larger than eight were neglected, and all type of atoms were considered. The results clearly show that the approach followed for prediction of the sign of NMR chemical shift difference was successful.

Figure 7 shows the surface of a CPG NN after being trained with the best chirality code, and colored according to the values of the weights in the output layer. The network consists of a toroidal surface (grid) of neurons. The map shows a characteristic region (in white) corresponding to the enantiomers whose chemical shift differences were positive, and a clearly distinct region for the opposite enantiomers (colored with gray). The objects of the test set were mapped into the same map, and were labeled with their reference numbers (and with an "i" when the experimental chemical shift difference is negative). It can be seen that all the test objects were correctly classified (an object is classified as giving a positive chemical shift difference if it activates a neuron with a positive output weight and vice versa).

Further validation of the method was pursued with a different test set consisting of molecules **18**, **19**, **20**, **38**, **40**, and their enantiomers. These compounds exhibit some substructural features that cannot be found in the remaining 35 structures (the new training set). The networks were trained using the best chirality code previously found. Correct predictions were obtained for 9 out of the 10 cases. The results show that the chirality codes/NN strategy was successful in the development of automatic predictions of chiral NMR behavior from the molecular structure, and revealed that the chirality codes contain relevant information for that purpose. In the application to compounds structurally different from those of the training set, the model has shown good robustness.

TABLE 2. Prediction of NMR Behavior in Chiral Solvents of Alkanols 41−64 and Their Enantiomers by CPG NNs of Size 9 × 9 Using the Best Chirality Codes as Input

code no.	aCDCC[a]	code no.	aCDCC_2[a]
24	68 (64, 66, 65, 68, 66)	24	68 (69, 68, 68, 66, 68)
36	65 (64, 57, 62, 67, 59)	28	68 (66, 67, 62, 60, 63)
35	65 (56, 64, 60, 59, 60)	1	68 (60, 58, 60, 60, 63)

[a] Number of correct predictions for the training set (74 atoms). The number of correct predictions obtained by the ensemble of five CPG NNs is displayed together with the five individual results (within parentheses).

3.2. Prediction of Absolute Configuration of Chiral Alcohols from NMR Data in Chiral Solvents (*R*,*R*)- and (*S*,*S*)-BMBA-*p*-Me. Using the database of NMR data in chiral solvents, the carbon atoms adjacent to chiral centers were encoded by the atomic versions of CDCC − aCDCC or aCDCC_2. A series of 90 different codes were generated with a code length of 101 by using (a) the partial atomic charge or effective polarizability as the atomic property (b) all types of atoms or not considering hydrogen atoms; (c) values of u in the interval $[-r, +r]$ with r varying from 0.030 to 0.200 e^2 Å^{-1} for partial atomic charges or in the interval $[0, +r]$ with r varying between 120 and 1000 Å for effective polarizabilities; (d) combinations of four atoms with maximum interatomic path distances of 4, 6, or 8 bonds; (e) the smoothing parameter b was set to (code length/range of u)^2.

The set of all codes comprise 94 objects (atoms) characterized by 101 variables. Constant variables were deleted and each variable was normalized.

If the sign of $\Delta\delta_{RR-SS}$ (see Methodology) was positive, a value of +1 was given to represent chemical shift difference of this compound, and if it is negative, a value of −1 was given. The chemical shift difference and the chirality codes were submitted to CPG NNs of size 9 × 9. For each chirality code, five independent networks were trained with the training set and incorporated into an ensemble of networks as in the first application. When a trained network was applied to make predictions, a positive value of the output was interpreted as a prediction of a positive chemical shift difference for the atom and vice versa.

The 90 chirality codes calculated with different options were evaluated by the quality of their predictions for the training set at the end of the training *without considering the results for the test set*. The best three codes are shown in Table 2. In assessing the quality of a code, not only the average of the five independent predictions was considered, but also the stability of the number of correct predictions over the five independent networks. It can be seen that the number of correct predictions with aCDCC_2 code 24 (68 correct predictions for the training set) is the same as for codes 28 and 1. However, code 24 is more stable than the other two and is therefore preferred.

TABLE 3. CPG NN Prediction of NMR Behavior in Chiral Solvents for the Training and Test Sets Using as Input Either Code 24 or a Selection of Values from Code 24

	correct prediction for the training set (74 atoms)	correct prediction for the test set (20 atoms)
aCDCC	68 (91.9%)	17 (85%)
aCDCC (subset)[a]	72 (97.3%)	16 (80%)
aCDCC_2	68 (91.9%)	20 (100%)
aCDCC_2 (subset)[a]	70 (94.6%)	20 (100%)

[a] A subset of code 24 was selected by genetic algorithms.

Code 24 was generated with (a) all types of atoms, i.e. including hydrogen atoms; (b) partial atomic charge as the atomic property; (c) value of u in the interval $[-r, +r]$ with $r = 0.200$ e^2 Å; (d) combination of four atoms with maximum interatomic path distance of 4 bonds.

To obtain more compact, robust, and accurate models, CPG neural networks were trained with subsets of code 24 descriptors, instead of all generated descriptors. The selection of the subsets was obtained by genetic algorithms as described in the Methodology section. CPG NNs trained with the selected descriptors were compared with the networks trained with all the descriptors. The same network size, training set, and test set were used—the results are displayed in Table 3.

Selection of descriptors has not significantly changed the accuracy of predictions both for aCDCC and aCDCC_2, but yielded much smaller models—8 descriptors instead of 31 for aCDCC and 11 instead of 31 for aCDCC_2.

Figure 8 shows the map of a CPG NN trained with selected descriptors from aCDCC_2 code 24 and colored according to the output weights. It reveals the trend for the two classes of enantiomers to excite neurons in two distinct regions (white and gray). The objects of the test set were also mapped, predictions were obtained based on the sign of the output value, and all the atoms were correctly classified. It should be noted that the NN recognized similarities between the enantiomers of the same class (the chemical shift difference of an atom) even though diverse structures were present.

Even with the best code (aCDCC_2 code 24), and selected descriptors, there are four atoms in the training set that are wrongly predicted (α carbon atoms on the left side of the hydroxymethine unit of **50** and **51** in Figure 3, and the corresponding carbon atoms in their enantiomers). These get wrong predictions with all the five independent networks. Significantly, they are also exceptions to the empirical rule proposed by Kishi and co-workers.[11] Two of them, however, were correctly predicted using the aCDCC code 24.

Conformation-dependent CDCC codes are based on single 3D molecular models −rigid conformation. For flexible molecules, more than one conformation is certainly relevant for NMR properties. This is probably the reason the best found code is a rather localized code—it is calculated only with combinations of four atoms with interatomic distances up to four bonds. Within such local environments, the universe of conformation diversity is much reduced. A possible additional reason is obviously the fact that chemical shifts basically depend on short-range interactions.

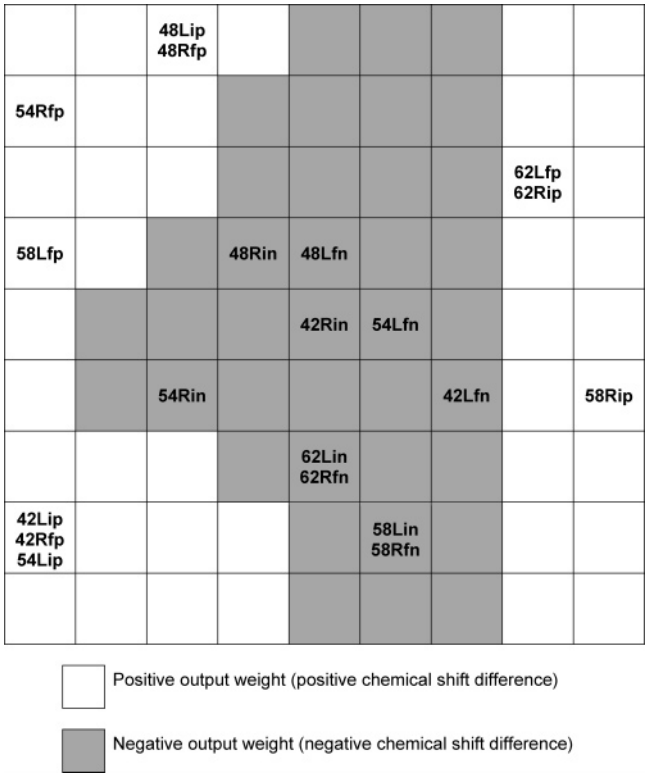


FIGURE 8. Representation of the output weights of a 9 × 9 CPG NN after training with 74 atoms encoded by aCDCC_2 descriptors. After the training, the 20 atoms of the test set were also mapped for classification—their labels include an "n" if the experimental chemical shift difference is negative or a "p" if it is positive; an "L" indicates an atom on the left side of the molecule and "R" represents an atom on the right side of the molecule as represented in Figure 3; "f" is for compounds shown in Figure 3 and "i" for their enantiomers.

The model incorporating the five networks can be used to assign the absolute configuration from the NMR data of unknown samples. For example, to assign the absolute configuration to a sample of compound **42** with unknown configuration, the following steps should be followed. NMR spectra are taken in chiral solvents (*R*,*R*)- and (*S*,*S*)-BMBA-*p*-Me. The differences between the chemical shifts in the (*R*,*R*) and the (*S*,*S*) solvents are calculated for the two carbon atoms adjacent to the hydroxymethine chiral center. The differences are compared with the predictions obtained by the model for the two possible enantiomers of **42**. The enantiomer with the predicted signs of the chemical shift differences matching the experimental data is identified. The same absolute configuration is assigned to the unknown sample.

An additional experiment was performed by changing the composition of the training and test sets. A new test set was defined with structures **53**, **54**, **55**, and their enantiomers, while the training set consisted of the remaining compounds. The structures in the test set were chosen as they have quite unique structural features. CPG NNs were trained on the basis of code nr. 24 and selection of variables was again performed by genetic algorithms. The 12 chemical shift differences of the test set were correctly predicted.

3.3. Quantitative Prediction of Chemical Shift Difference of Chiral Alcohols in Bidentate Chiral

TABLE 4. CPG NN Quantitative Prediction of NMR Behavior in Chiral Solvents for the Training Set (Results Obtained with the Best Codes)

code no.	aCDCC (r^2)[a]	code no.	aCDCC_2 (r^2)[a]
36	0.803 (0.581, 0.463, 0.667, 0.705, 0.733)	24	0.774 (0.681, 0.761, 0.752, 0.764, 0.737)
24	0.738 (0.661, 0.746, 0.703, 0.745, 0.674)	28	0.763 (0.711, 0.637, 0.694, 0.697, 0.691)
35	0.728 (0.512, 0.400, 0.686, 0.639, 0.648)	23	0.750 (0.707, 0.672, 0.670, 0.718, 0.704)

[a] r^2 value for the training set (76 atoms). The correlations obtained by the ensemble of five CPG NNs are displayed together with the five individual results (within parentheses).

TABLE 5. CPG NN Quantitative Prediction of NMR Behavior in Chiral Solvents for the Training and Test Sets Using as Input Either Code 24 or a Selection of Code 24 Values

	r^2 (training set)	r^2 (test set)
aCDCC	0.738	0.754
aCDCC (subset)[a]	0.828	0.743
aCDCC_2	0.774	0.865
aCDCC_2 (subset)[a]	0.839	0.936

[a] A subset of code 24 was selected by genetic algorithms.

Solvents (*R*, *R*)- and (*S*, *S*)-BMBA-*p*-Me. Encouraged by the good predictions of the sign of chemical shift differences, we tried to further predict their quantitative values. The same series of 90 chirality codes was screened as in section 3.2. The only difference is that the sign of chemical shift difference was replaced by its actual value shown in Figure 3, and now two more cases were included in the training set with a chemical shift difference of zero. The same test set was used.

The best three codes for aCDCC and aCDCC_2 are shown in Table 4. Judging by the r^2 value, and also by the stability of the predictions in the five networks, code 24 was again chosen as the best, both for aCDCC and aCDCC_2. The networks trained with these codes were tested with the independent test set (Table 5).

In a similar way to section 3.2, selection of variables was performed with genetic algorithms. The results for the networks trained with the obtained subsets are also shown in Table 5. Chirality codes aCDCC_2 were again better than aCDCC in the independent test, and selection of descriptors further improved the predictions while building more compact models—10 variables were selected in each case.

Good correlations were observed between the experimental and calculated chemical shift differences—r^2 = 0.839 for the training set and r^2 = 0.936 for the test set, Figure 9. These are particularly remarkable considering the narrow range of chemical shift differences (ca. 0.09 ppm in ^{13}C NMR spectra). It can also be seen that the sign of the chemical shift difference was correctly predicted for all the cases in the test set, and for all but six examples in the training set. The six wrong cases fell near the axes, and they include the four cases wrongly predicted in the qualitative approach (see section 3.2).

Quantitative prediction of the chemical shift difference can assist in the assignment of *relative* configuration. For example structures **46** and **47** are diastereoisomers. They have the same configuration of the carbon atom bonded to the hydroxyl group, and the two atoms adjacent to this chiral center have the same *sign* in **46** and **47**. However, the chemical shift differences are quantitatively different

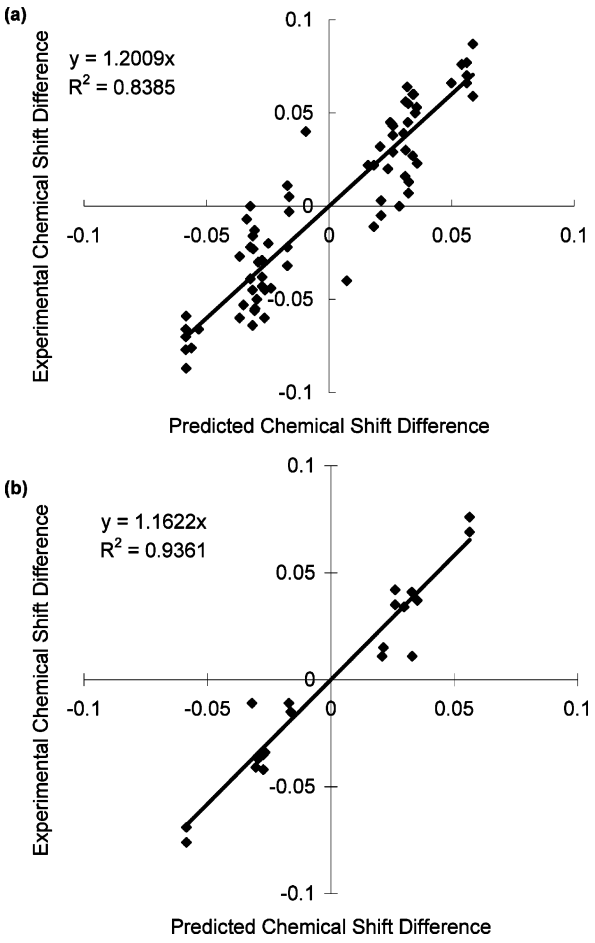


FIGURE 9. Quantitative prediction of chemical shifts difference (ppm) in two enantiomeric solvents for (a) training set and (b) test set.

in the two diastereoisomers due to the different configuration of the second chiral center. Reliable quantitative predictions would assign the relative configuration of the second chiral center. The relatively small size of the data set however precludes further experimentation on that direction.

4. Conclusion

An automatic system was developed on the basis of chirality codes, which integrates available 1D NMR data, builds models, and makes accurate predictions about chiral NMR properties without human definition of explicit rules. Selection of variables using genetic algorithms was useful for refinement of prediction ability and for reducing the size of the models.

Atomic chirality codes were shown to describe the chirality of an atom's environment in a way that can be correlated with its NMR properties in chiral solvent BMBA-*p*-Me.

In the two applications—modeling of chemical shift after covalent bonding to (*R*)-MTPA and modeling of chemical shift in chiral solvents—correct predictions were achieved for independent test sets. In the second application, quantitative predictions could be obtained with $r^2 = 0.936$.

This work is a contribution to the assignment of absolute configuration from NMR data, particularly for its implementation in automatic systems.

Acknowledgment. We are indebted to Prof. Johann Gasteiger for access to software developed by his research group. Z.Q.Y. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a postdoctoral grant under the POCTI program (SFRH/BPD/14476/2003).

JO048029Z