Physicochemical Stereodescriptors of Atomic Chiral Centers†

Qing-You Zhang and João Aires-de-Sousa*

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

Received June 13, 2006

Physicochemical atomic stereodescriptors (PAS) were implemented that represent the chirality of an atomic chiral center on the basis of empirical physicochemical properties of the ligands. The ligands are ranked according to a specific property, and the chiral center takes an *S*/*R*-like descriptor relative to that property. The procedure is performed for a series of properties, yielding a chirality profile. Application of the PAS descriptors to the prediction of enantioselectivity in chemical reactions, from the molecular structures, is illustrated here. The relationship between the molecular structures, represented by the PAS descriptors, and the enantioselectivity was learned by neural networks, decision trees, or random forests. In a first application, a data set was employed with chiral amino alcohols that enantioselectively catalyze the addition of diethylzinc to benzaldehyde. Prediction of the major enantiomer obtained in the reaction, from the molecular structure of the catalyst, was achieved with accuracy up to 90%. The second application investigated the enantiopreference of *Pseudomonas cepacia* lipase (PCL) toward primary alcohols. The learned models could make correct predictions about the preferred enantiomer, from the molecular structure of the substrate, in up to 93% of the cases. These included substrates with and without O-atoms bonded to the chiral center. The properties automatically selected to build the models can give indications on the relevant factors guiding the observed chemical behavior.

INTRODUCTION

Deriving chiral QSARs or QSPRs requires specific descriptors. Among the impressive number and diversity of molecular descriptors currently available for the establishment of structure−property relationships and assessment of molecular similarity, only a very few are capable of discriminating between enantiomers, i.e., encode chirality.[1] And yet chiral properties of chemical compounds are increasingly relevant in analytical chemistry, organic synthesis, materials science, drug development, toxicology, or environmental chemistry.

Several authors have incorporated *R/S* descriptors obtained by the Cahn−Ingold−Prelog (CIP)[2−4] rules into conventional topological descriptors[5−7] and applied such descriptors to chiral QSAR studies. Although the CIP descriptors are excellent for labeling and identifying the configuration of chiral centers, they have a fundamental weakness for developing structure−property relationships—they were not designed to bear any intrinsic chemical meaning. Observable chiral properties basically depend on the 3D arrangement of molecular fragments bearing certain physicochemical features. Accordingly, chiral structure−property relationships are in principle sounder if based on chirality descriptors incorporating physicochemical as well as geometrical features. More chemical significance should be obtained if the geometrical arrangement is defined in terms of physicochemical or topological[8] features. Since 1999 we have been involved in the development of chirality codes that represent the molecular chirality by a spectrumlike descriptor generated from 3D structures and empirical physicochemical properties.[9−11] More recently, other authors have also recognized the usefulness of Gasteiger charges in the definition of chirality sensitive descriptors for QSAR applications.[12] Spectrumlike chirality codes have been applied to the prediction of chiral properties such as chromatographic enantioselectivity in chiral HPLC,[10,13] enantioselectivity of chemical reactions,[9,14] and NMR chemical shifts in chiral environments.[15] Although successful in a number of applications, chirality codes are difficult to interpret. A method is usually preferred if it highlights which parameters are important for the studied property, and if these parameters somehow describe structural features in a language chemists can understand.

Here we present CIP-type descriptors that can be easily interpreted and incorporate a series of physicochemical features.[16] In the CIP rules, atomic number is the decisive property in the establishment of the ligands priorities. The main idea of this study is to replace atomic number by other properties of the ligands, more relevant to the prediction of observable molecular properties. Ranking ligands according to different properties result in different descriptors. Examples of such properties are size of ligands, charges, electronegativities, or polarizabilities. This approach yields, for a single chirality center, several '*R*/*S*-like' descriptors—each based on a different property of the ligands. Furthermore, the real-number properties of the four ligands themselves can be used as additional descriptors of the chirality center.

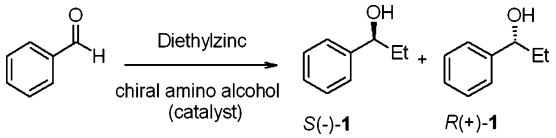
The resulting physicochemical atomic stereo (PAS) descriptors were tested in two chiral structure−property studies related to enantioselectivity in organic reactions and biocatalysis. Neural networks, decision trees, and random forests were investigated to predict enantioselectivity from the PAS descriptors.

† Dedicated to Professor Johann Gasteiger.

* Corresponding author phone: (+351) 21 2948300; fax: (+351) 21 2948550; e-mail: jas@fct.unl.pt.

10.1021/ci600235w CCC: $33.50
Published on Web 11/01/2006

Scheme 1. Enantioselective Catalytic Addition of Diethylzinc to Benzaldehyde



METHODOLOGY

Data Sets. Two data sets were investigated with a binary property assigned to a specific enantiomer of an enantiomeric pair—catalyst yielding the R/S product of a given reaction or preferred/not preferred enantiomer of a biocatalytic system.

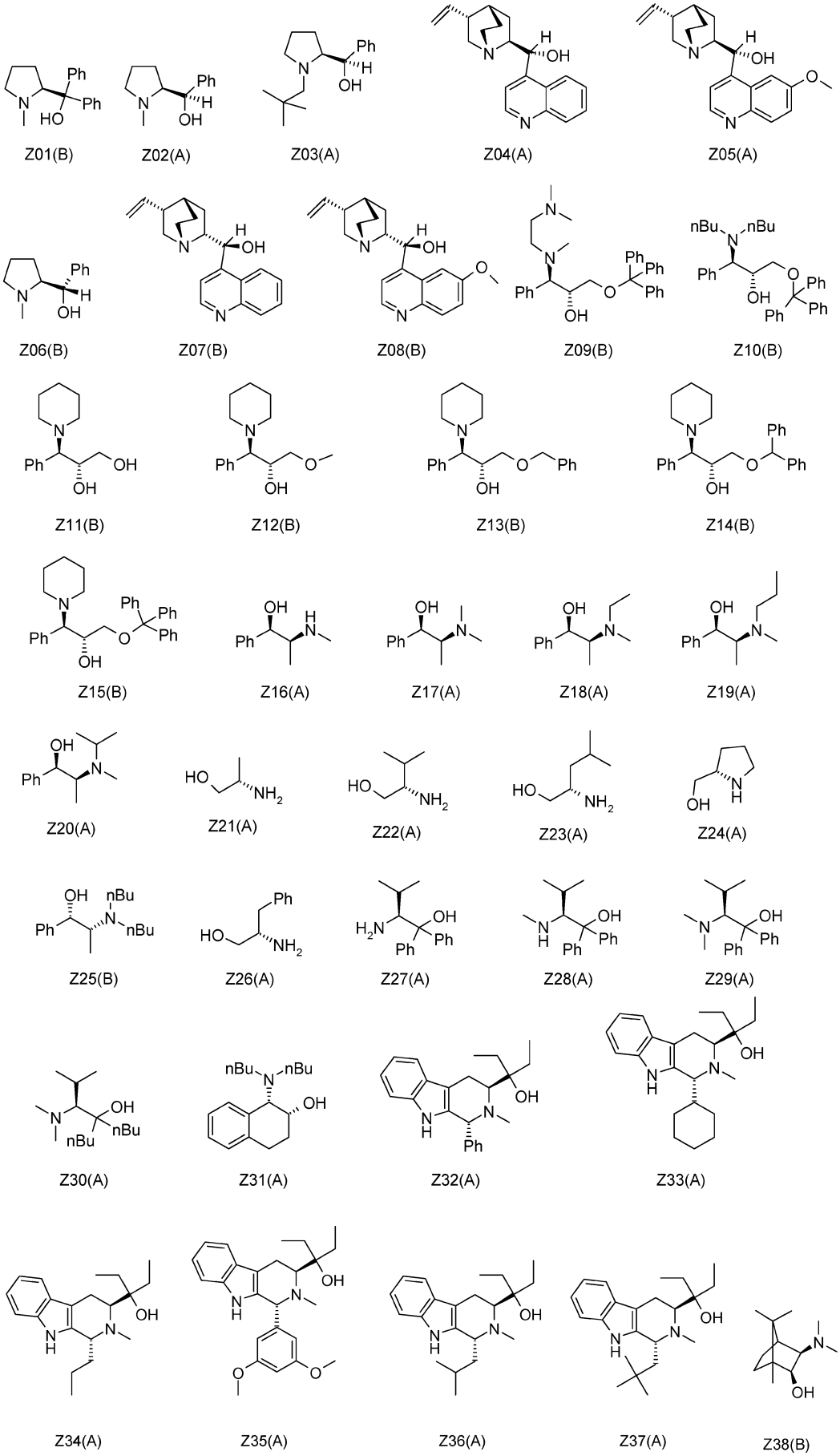
The first data set consists of a series of 48 enantiomeric pairs of chiral amino alcohols that enantioselectively catalyze the addition of diethylzinc to benzaldehyde and had been compiled for a previous study.[9] The reaction is shown in Scheme 1. Each catalyst yields preferentially either *S*(−)-**1** or *R*(+)-**1**. If one enantiomer produces *S*(−)-**1**, then the opposite enantiomer produces necessarily *R*(+)-**1**. The structures of the 48 chiral amino alcohols are shown in Figure 1. The data set was partitioned into a test set with 10 enantiomeric pairs of amino alcohols (**Z01, Z04, Z11, Z19, Z25, Z30, Z33, Z40, Z43, Z48**, and their enantiomers) and a training set with the remaining 38 pairs of enantiomers. The test set was chosen such that it covers the different types of compounds namely in terms of skeleton, type and number of chiral centers, and size.

The second data set is a series of 86 enantiomeric pairs of primary alcohols involved in racemic resolutions by transesterifications, or hydrolyses catalyzed by *Pseudomonas cepacia* lipase. Each enantiomer of a pair is identified as the preferred or not preferred enantiomer of the reactions.[17] Some of these alcohols have an oxygen atom bonded to the chiral center. The structures are shown in Figure 2. They were divided into a test set with 28 compounds (**P06**, **P08**, **P21**, **P23**, **P32**, **P39**, **P48**, **P51**, **P52**, **P59**, **P60**, **P76**, **P77**, **P78**, and their enantiomers) and a training set with the remaining compounds. Some of the structures are not chiral (e.g. **P53**) but result from the enantioselective hydrolysis of chiral esters. In those esters, only one of various primary alcohol groups was esterified (esterification of only one of the OH groups renders the compound chiral). Other achiral structures produce a chiral ester by selective esterification of only one of the primary alcohol groups. Because in the investigation of this data set the highest priority was always assigned to the CH_2OH ligand involved in the reaction (identified with an arrow in Figure 2), it was possible to define PAS descriptors even for the achiral structures.

Physicochemical Atomic Stereo (PAS) Descriptors. In CIP rules, the ligands are first ranked according to the atomic number of the atoms bonded to the chiral center. If two atoms have the same atomic number, a series of rules is applied to assign priorities to the ligands. If another atomic property is considered, instead of the atomic number, another descriptor can be obtained. For example, if the four ligands are ranked according to the σ-electronegativity of the atom bonded to the chiral center, new ranking results. With this ranking, an '*S*/*R*-like' descriptor can be derived for the chiral center. Figure 3 compares the assignment of the CIP descriptor and a 'CIP-like' descriptor (based on σ-electronegativity) to one enantiomer of lactic acid.

'CIP-like' descriptors were calculated—physicochemical atomic stereodescriptors (PAS)—on the basis of 21 different properties. Additionally, the properties of the four ligands were also investigated as descriptors of the chirality center. The following 21 properties of a ligand were used: 1. number of atoms (if a cycle is involved, an atom is assigned to a ligand depending on its topological distance to the atoms directly bonded to the chiral center); 2. number of atoms that are within the third sphere of bonds from the chiral center; 3. distance (in number of bonds) from the chiral center to the farthest atom in the ligand; 4. maximum distance (in number of bonds) between two atoms in the ligand; 5. partial atomic charge of the atom bonded to the chiral center; 6. maximum partial atomic charge within the third sphere of bonds from the chiral center; 7. minimum partial atomic charge within the third sphere of bonds from the chiral center; 8. sum of the σ and π residual electronegativity of the atom bonded to the chiral center; 9. maximum of the sum of the σ and π residual electronegativity within the third sphere of bonds from the chiral center; 10. minimum of the sum of the σ and π residual electronegativity within the third sphere of bonds from the chiral center; 11. effective atomic polarizability of the atom bonded to the chiral center; 12. maximum effective atomic polarizability within the third sphere of bonds from the chiral center; 13. minimum effective atomic polarizability within the third sphere of bonds from the chiral center; 14. maximum of the difference in electronegativity between two atoms of a bond within the third sphere of bonds from the chiral center; 15. maximum of the difference in partial atomic charge between two atoms of a bond within the third sphere of bonds from the chiral center; 16. maximum of the mean bond polarizability within the third sphere of bonds from the chiral center; 17. minimum of the mean bond polarizability within the third sphere of bonds from the chiral center; 18. maximum of the resonance stabilization of a positive charge after bond breaking within the third sphere of bonds from the chiral center; 19. maximum of the resonance stabilization after bond breaking within the third sphere of bonds from the chiral center; 20. maximum of the delocalization stabilization of a positive charge after bond breaking within the third sphere of bonds from the chiral center; and 21. maximum of the delocalization stabilization after bond breaking within the third sphere of bonds from the chiral center.

The PAS descriptors were encoded with three different mechanisms: (a) by a 21-positions code, each position corresponding to one property and assuming a value of +1 (*R*-like), −1 (*S*-like), or 0 (two or more ligands with the same property); (b) by a 42-positions code, each two positions corresponding to one property and assuming values (1,0) if *R*-like, (0,1) if *S*-like, or (0,0) if two or more ligands have the same property; and (c) by a 105-positions code, each five positions corresponding to one property, the first position being the same as in (a), and the last four the properties of the four ligands in decreasing order. If more than one chiral center belonging to the same molecule are to be encoded by PAS descriptors, the 21- or the 42-positions codes of the different chiral centers were summed. In the second application, in which a certain chiral center is expected to be particularly relevant for the independent variable (the chiral center bonded to the CH_2OH group), only the PAS descrip-


Z01(B)
Z02(A)
Z03(A)
Z04(A)
Z05(A)
Z06(B)
Z07(B)
Z08(B)
Z09(B)
Z10(B)
Z11(B)
Z12(B)
Z13(B)
Z14(B)
Z15(B)
Z16(A)
Z17(A)
Z18(A)
Z19(A)
Z20(A)
Z21(A)
Z22(A)
Z23(A)
Z24(A)
Z25(B)
Z26(A)
Z27(A)
Z28(A)
Z29(A)
Z30(A)
Z31(A)
Z32(A)
Z33(A)
Z34(A)
Z35(A)
Z36(A)
Z37(A)
Z38(B)

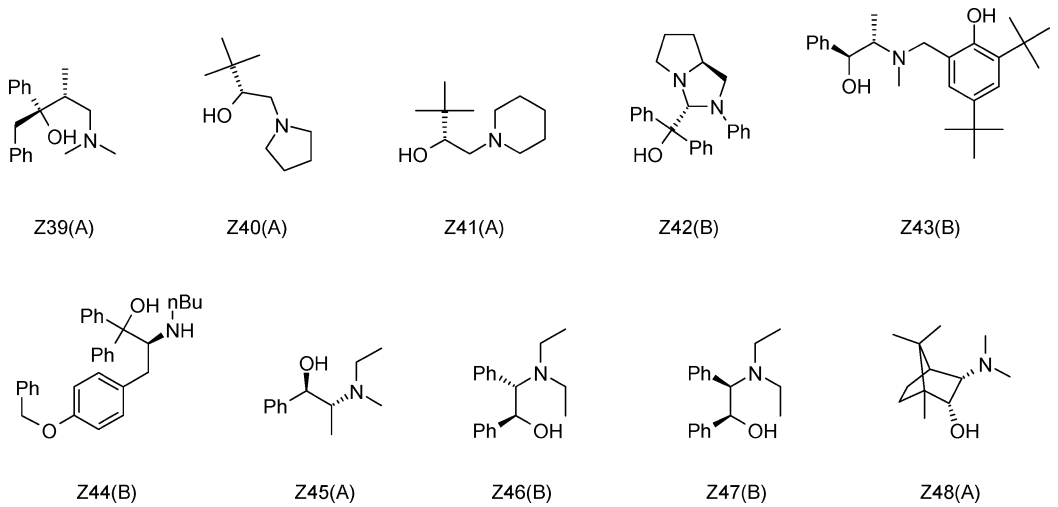


Figure 1. Amino alcohols described as catalysts in the addition of diethylzinc to benzaldehyde. Those yielding preferentially *R*(+)-**1** were labeled with 'A', and those yielding *S*(−)-**1** were labeled with 'B'.

tors of that chiral center were used even if the molecule had more than one chiral center.

The various physicochemical atomic properties were calculated using fast empirical methods[18,19] implemented in the program PETRA (Molecular Networks GmbH, Erlangen, Germany). The PAS codes were generated by in-house developed software written in the C programming language. The assignment of *S*/*R*-like configurations is based on Cartesian coordinates of the atoms, which were calculated from the connection tables of the molecules by the 3D structure generator CORINA.[20−23]

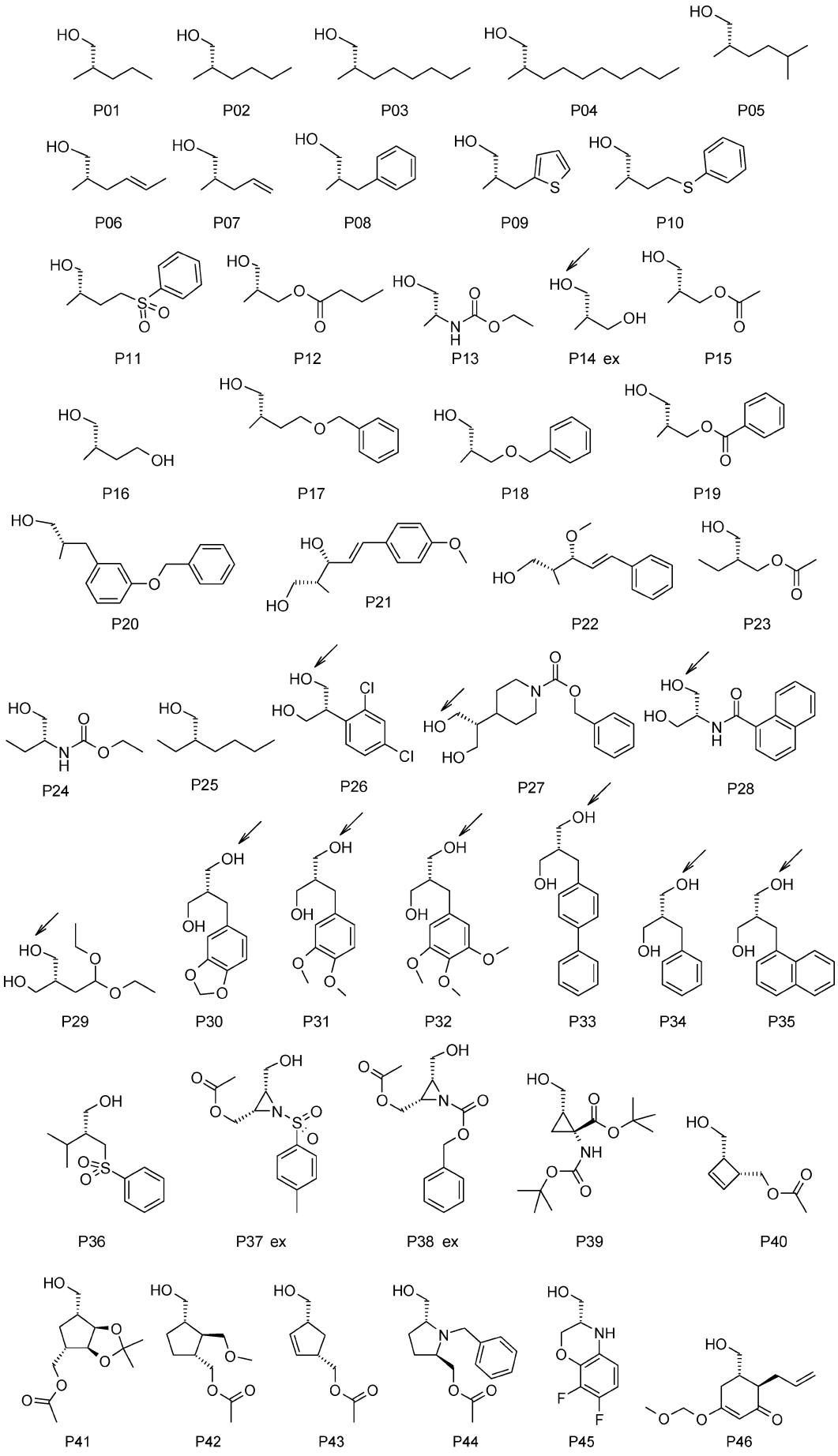
Counterpropagation Neural Networks. Counterpropagation neural networks (CPG NN)[24] were investigated to model the relationship between PAS codes and the chiral independent property. The input data for a CPG network are stored in a two-dimensional grid of neurons, each containing as many elements (weights) as there are input variables. In the investigations described in this paper the input variables are PAS descriptors. In Figure 4 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 (the chiral independent variable) 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 (PAS 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 (PAS 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 (numerical representation of the) class of the presented molecule. 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 are presented several times, and the weights are corrected, until the network stabilizes. Note that the experimental classification of the molecule is not used in determining the winning neuron.

After the training, the CPG NN is able to classify a chiral molecule on input of an object represented by their PAS descriptors—the winning neuron is chosen and the corresponding weight in the output layer is used for prediction (Figure 4).

In the experiments here described, an output value of +1 was given to one class, and a value of −1 to the opposite class. Prediction of the class was based on the sign of the weight at the selected output neuron. Training of the CPG NNs was performed by using a linear decreasing triangular scaling function used with an initial learning rate of 0.1 and an initial learning span half of the network size. The weights were initialized with random numbers that are calculated using the mean and standard deviation of the input data set as parameters. For the selection of the central neuron the minimum Euclidean distance between the input vector and neuron weights has been used. The training was performed typically over 50 cycles, with the learning span and the learning rate linearly decreasing until zero.

Classification Tree.[26] A single classification tree was investigated to make predictions. A classification tree is sequentially constructed, partitioning objects from a parent node into two child nodes. Each node is produced by a logical rule, defined for a single variable, where objects below a certain variable's value fall into one of the two child nodes, and objects above fall into the other child node. The prediction for an object reaching a given terminal node is obtained by majority vote of the objects (in the training set) reaching the same terminal node. The entire procedure comprises three main steps. First an entire tree is constructed by data splitting into smaller nodes; each produced split is evaluated by an impurity function which decreases as long as the new split permits child node's content to be more homogeneous than parent node. Second, a set of smaller, nested trees is obtained by obliteration (pruning) of certain nodes of the tree obtained in the first step. The selection of the weakest branches is based on a cost-complexity measure that decides which subtree, from a set of subtrees with the same number of terminal nodes, has the lowest (within node) error. Finally, from the set of all nested subtrees, the tree giving the lowest value of error in cross-validation (where the set of objects used to grow the tree is different from the prediction set) is selected as the optimal tree. In this study, a classification tree was grown with the R program version


P01
P02
P03
P04
P05
P06
P07
P08
P09
P10
P11
P12
P13
P14 ex
P15
P16
P17
P18
P19
P20
P21
P22
P23
P24
P25
P26
P27
P28
P29
P30
P31
P32
P33
P34
P35
P36
P37 ex
P38 ex
P39
P40
P41
P42
P43
P44
P45
P46

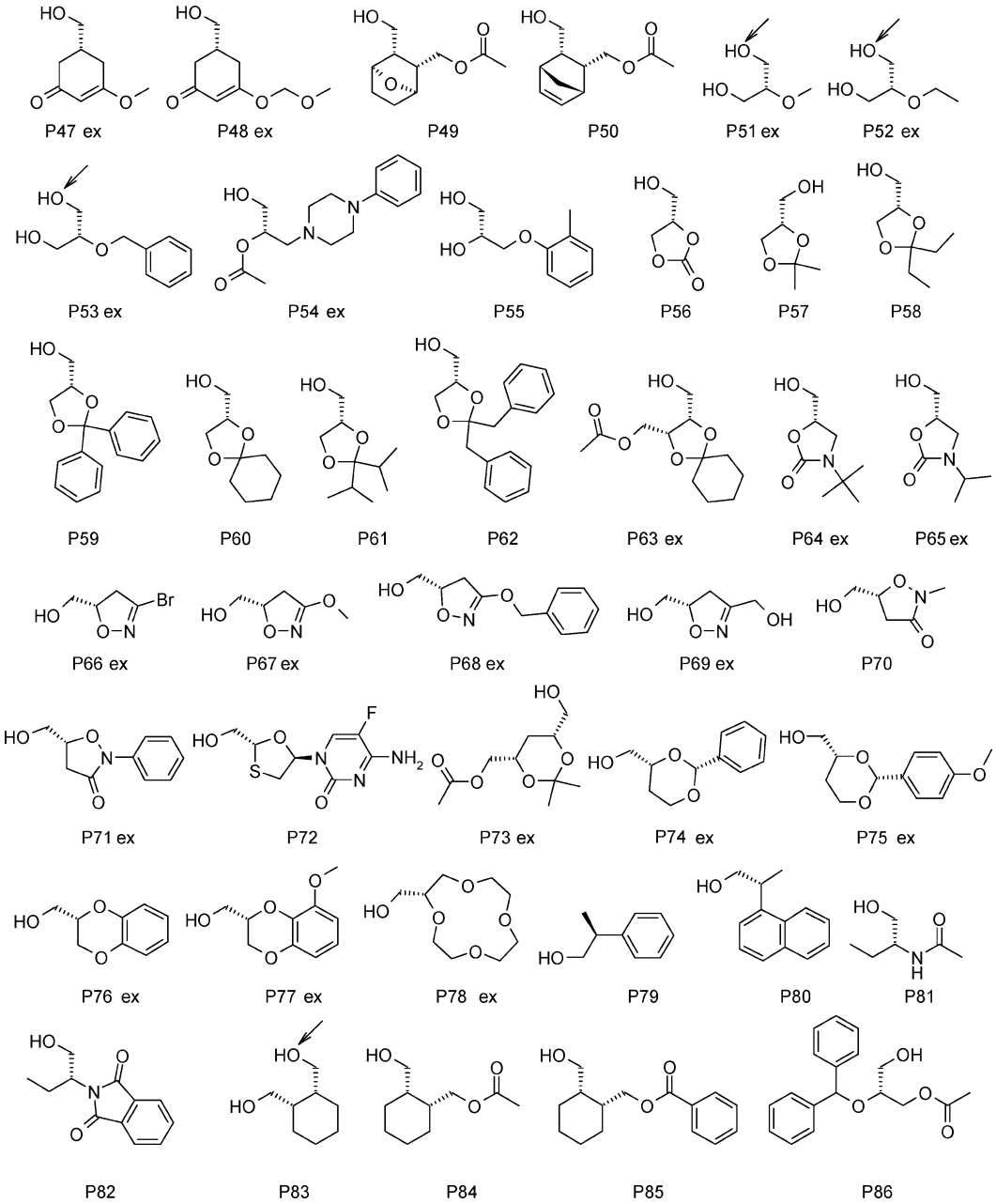


Figure 2. Primary alcohols preferentially produced in hydrolysis or transesterification reactions with lipase from *Pseudomonas cepacia* (PCL) as catalyst. Compounds reacting slower than its opposite enantiomer are labeled as 'ex'.

2.0.1[27] using the RPART library with the default parameters.

Random Forests.[28,29] A random forest is an ensemble of unpruned classification trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node. It is a high-dimensional nonparametric method that works well on large numbers of variables. Prediction is made by majority vote of the individual trees. It has been shown that the method is extremely accurate in a variety of applications.[29] Additionally, performance is internally assessed with the prediction error for the objects left out in the bootstrap procedure. The method quantifies the importance of a variable by the increase in misclassification occurring when the values of the variable are randomly permuted or by the decrease in a node's impurity every time the variable is used for splitting. In this study, random forests were grown with the R program version 2.0.1[27] using the randomForest library.[30] The number of descriptors in each random selection was set to the default values, and 1000 trees were grown for each forest. For the assessment of variable importance, it was checked that the most important variables remained the same after removal of highly intercorrelated variables. The RFs were applied to classify the enantiomers from the PAS code.

RESULTS AND DISCUSSION

Prediction of the Major Enantiomer in the Catalytic Addition of Diethylzinc to Benzaldehyde. Chiral amino alcohols enantioselectively catalyze the addition of diethylzinc to benzaldehyde. The relationship between the features of the chiral centers of the catalysts and the major enantiomer they yield was investigated with PAS descriptors. When there was more than one chiral center in a molecule, the PAS descriptors for all the chiral centers were summed. The ability

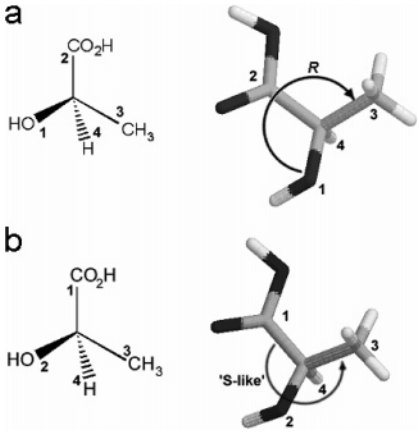


Figure 3. (a) Assignment of CIP descriptor *R* to one enantiomer of lactic acid (CIP priorities: OH > CO_2H > CH_3 > H). (b) Assignment of 'CIP-like' descriptor *S*-like to the same enantiomer of lactic acid, ranking the ligands according to the σ-electronegativity of the atoms bonded to the chiral center.

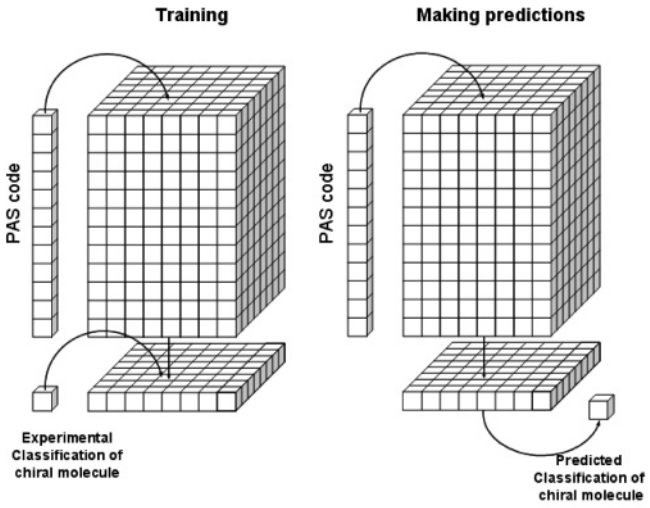


Figure 4. Representation of a counterpropagation neural network (CPG NN). Every small box of the network block represents a weight. The CPG NN is trained by iterative presentation of objects (PAS profiles and the corresponding class of the chiral molecule). After the training, the NN is able to make predictions on input of a PAS profile.

of the global PAS profile to predict the enantioselective outcome of the reaction was first assessed with an unsupervised clustering technique (counterpropagation neural network, CPG NN). CPG networks with a dimension of 9 × 9 neurons were trained with 76 catalysts (38 enantiomeric pairs) and tested with 20 catalysts (10 enantiomeric pairs). This test set includes the test set of ref 9. Figure 5 shows the surface of a CPG NN trained with the PAS 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 catalysts yielding preferentially *S*(−)-**1** (class B) and a clearly distinct region for catalysts yielding the opposite enantiomer (colored with gray). The objects of the test set were mapped onto the same map and were labeled with their reference numbers and experimental classes. It can be seen that 9 out of 10 pairs of catalysts in the test set were correctly classified. The only wrong prediction for the test set were the structures **Z01**. Inspection of the structures from the training set that activate neurons in the same region

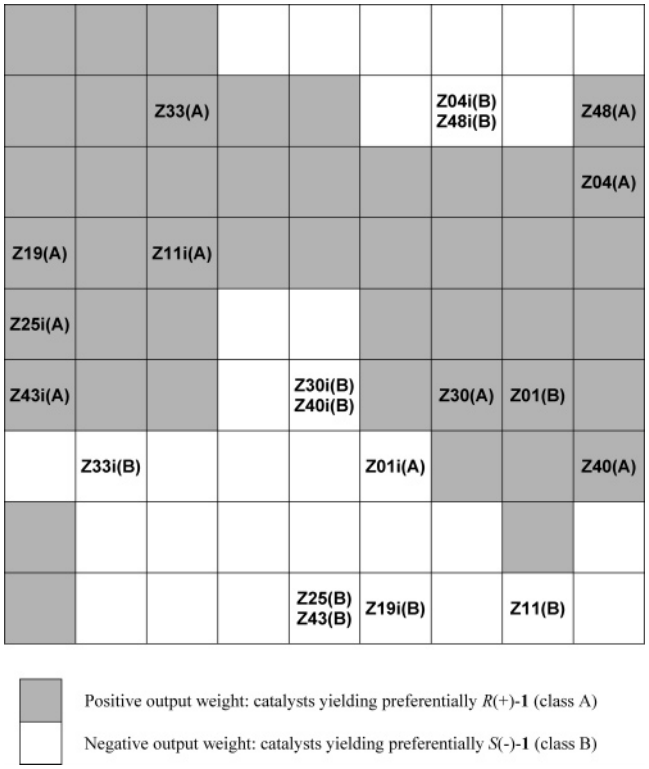


Figure 5. Representation of the output weights of a 9 × 9 CPG NN after the training with 76 catalysts − 38 pairs, represented by PAS codes of size 42 and considering hydrogen atoms. After the training, the 10 pairs of catalysts in the test set (**Z01**, **Z04**, **Z11**, **Z19**, **Z25**, **Z30**, **Z33**, **Z40**, **Z43**, **Z48,** and their enantiomers labeled with 'i') were mapped for classification—catalysts yielding preferentially *R*(+)-**1** were labeled with 'A', while those yielding the opposite enantiomer were labeled with 'B'.

reveals very similar structures such as **Z27**, **Z28**, and **Z29** (all yielding preferentially the opposite enantiomer), which explain the prediction. In CPG NN experiments the entire PAS codes were used with no selection of variables. It is also to note that the distribution of the catalysts on the map is performed in an unsupervised way, i.e., the catalysts are mapped exclusively on the basis of the PAS code, and the information about the class is only used to correct the output neurons.

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 output.

The results obtained with CPG NN, decision tree (CART), and random forest (RF) are presented in Table 1. Experiments were performed with the two coding schemes of PAS descriptors and with/without considering hydrogen atoms not bonded to the chiral center. The two different schemes for the encoding of PAS descriptors gave approximately the same overall results, and inclusion of hydrogen atoms was not relevant as well. A classification tree trained with PAS descriptors of size 42 and with hydrogen atoms correctly predicted 80% of the test set and 92% of the training set. CPG NN performed better, but the best accuracies were observed with random forests, although differences were not dramatic. This technique obtained up to 96% of correct predictions in the out-of-bag estimation using the whole data

Table 1. Prediction of the Major Enantiomer Produced by Amino Alcohol Catalysts, from Their PAS Descriptors

	correct predictions		
	training set (76 cases)	test set (20 cases)	all (OOB estimation)
PAS Descriptors of Size 42, H Atoms Considered			
CPG NN	73 (96%)	18 (90%)	
CART	70 (92%)	16 (80%)	
random forest	74 (97%)[a]	18 (90%)	92 (96%)
PAS Descriptors of Size 42, H Atoms not Considered			
CPG NN	71 (93%)	16 (80%)	
random forest	72 (95%)[a]	18 (90%)	90 (94%)
PAS Descriptors of Size 21, H Atoms Considered			
CPG NN	72 (95%)	18 (90%)	
random forest	73 (96%)[a]	18 (90%)	91 (95%)
PAS Descriptors of Size 21, H Atoms not Considered			
CPG NN	69 (91%)	16 (80%)	
random forest	71 (93%)[a]	18 (90%)	90 (94%)

[a] OOB estimation within the training set.

set and 90% of correct predictions for the independent test set. These results are approximately of the same quality as those previously obtained with chirality codes.[9] The decision tree was built with only two splits: one based on the minimum atomic charge within the third sphere of bonds from the chiral center and the other on the minimum of the ($\sigma + \pi$) residual electronegativity within the third sphere of bonds from the chiral center. The most important variables for the random forest model trained with PAS descriptors of size 42 and H atoms were also based on the minimum atomic charge within the third sphere of bonds from the chiral center and on the maximum $\sigma + \pi$ residual electronegativity within the third sphere of bonds from the chiral center. Since the oxygen atoms generally bear the most negative charge followed by the nitrogen atom, the choice of those properties by the RFs can be grossly interpreted as a way to identify the ligand that includes an oxygen or nitrogen atom.

The methods were further validated with *y*-randomization. Five series of random classes were generated for the objects of the training set (keeping the 1:1 balance between the two classes) and five RFs were built using the random classifications—the highest percentage of correct predictions in the OOB estimation was 55%, 57%, 58%, and 57% for each of the PAS codes in Table 1. When the models trained with the random *y*-values were applied to the test set, the highest percentage of correct predictions were respectively 65%, 70%, 55%, and 65%. Five *y*-randomization experiments performed with CPG NNs on the basis of PAS descriptors of size 42 and with H-atoms yielded 30%, 35%, 65%, 50%, and 30% of correct predictions for the test set. With CART the same experiments yielded 45%, 70%, 50%, 25%, and 50%. All these predictions are considerably inferior to those obtained with the experimental *y*-values.

Prediction of the Preferred Enantiomer of Primary Alcohols in Reactions Catalyzed by *Pseudomonas cepacia* Lipase (PCL). Biocatalytic resolution of racemic mixtures is a well-known method of asymmetric synthesis, although prediction of the enantioselectivity for a given substrate is still a challenge. For reactions catalyzed by the PCL enzyme, a rule was put forward for the prediction of the preferred enantiomer of primary alcohols in esterifications, transesterifications, and hydrolyses. It is based on the sizes of the substituents bonded to the same chiral center as the CH_2OH group involved in the reaction.[17] The rule could account for 54 out of 61 studied primary alcohols. Reliable predictions were however not possible for primary alcohols that have

Table 2. Prediction of the Preferred Enantiomer of Primary Alcohols in Reactions Catalyzed by *Pseudomonas cepacia* Lipase (CPL), from the PAS Descriptors of the Molecular Structures

	correct predictions		
	training set (44 O-bonded, 100 not O-bonded)	test set (14 O-bonded, 14 not O-bonded)	all (OOB estimation)
PAS Descriptors of Size 21, H Atoms Considered			
CPG NN	120 (83%)	19 (68%) O-bonded: 43% not O-bonded: 93%	
GA + CPG NN	118 (82%)	19 (68%) O-bonded: 50% not O-bonded: 86%	
CART	123 (85%)	20 (71%) O-bonded: 57% not O-bonded: 86%	
CART (PAS+properties)[a]	125 (87%)	18 (64%)	
random forest	126 (88%)[b] O-bonded: 91%[b] not O-bonded:86%[b]	18 (64%) O-bonded: 43% not O-bonded: 86%	143 (83%) O-bonded: 81% not O-bonded: 84%
random forest (PAS+properties)[a]	123 (85%)[b] O-bonded: 80%[b] not O-bonded: 88%[b]	25 (89%) O-bonded: 79% not O-bonded: 100%	146 (85%) O-bonded: 72% not O-bonded: 91%
PAS Descriptors of Size 21, H Atoms not Considered			
CPG NN	114 (79%)	18 (64%) O-bonded: 43% not O-bonded: 86%	
GA + CPG NN	125 (87%)	20 (71%) O-bonded: 43% not O-bonded: 100%	
random forest	122 (85%)[b] O-bonded: 82%[b] not O-bonded: 86%[b]	19 (68%) O-bonded: 43% not O-bonded: 93%	135 (78%) O-bonded: 62% not O-bonded: 87%
random forest (PAS+properties)[a]	120 (83%)[b] O-bonded: 70%[b] not O-bonded: 89%[b]	26 (93%) O-bonded: 86% not O-bonded: 100%	147 (85%) O-bonded: 71% not O-bonded: 93%

[a] The CIP-like descriptors were used together with the real properties of the ligands. [b] OOB estimation within the training set.

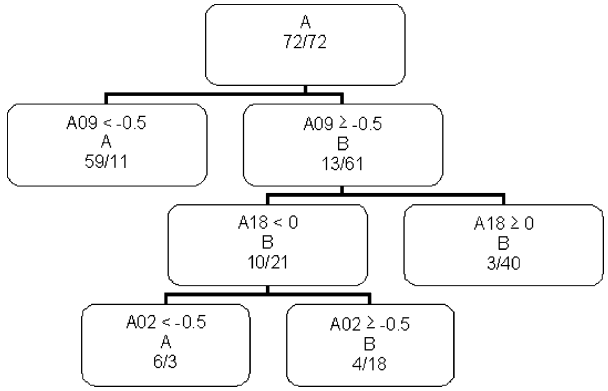


Figure 6. Graphical representation of the classification tree obtained for the prediction of the preferred enantiomer of primary alcohols in reactions catalyzed by PCL. Ovals represent nodes. Inside each node, a condition and the corresponding classification is given for the node as well as the number of major enantiomers/minor enantiomers of the training set falling into it. Descriptor A09 is based on the maximum of the sum of the σ and π residual electronegativity within the third sphere of bonds from the chiral center. Descriptor A18 is based on the maximum resonance stabilization of a positive charge after bond breaking within the third sphere of bonds from the chiral center. Descriptor A02 is based on the number of atoms within the third sphere of bonds from the chiral center.

an oxygen atom attached to the stereocenter.[17] We investigated this data set with the new PAS descriptors, and the results are presented in Table 2. For this particular application, the highest priority in the generation of PAS descriptors was always assigned to the CH_2OH ligand that is involved in the reaction. This allowed for focusing on the reaction center and considering nonchiral products resulting from the hydrolysis of chiral esters (e.g., **P53**). The best results were obtained with random forests, particularly when the real values of the properties were used together with the *R*/*S*-like descriptors. Such models correctly predicted 85% of the cases in the cross-validation test and 93% of the test set. The cases with an oxygen atom bonded to the chiral center could be predicted with accuracy up to 81% and 86% in the cross-validation test and in the independent set test, respectively. All the structures with no oxygen atom bonded to the stereocenter in the test set were correctly predicted as well as 93% in the cross-validation test. When applied to the structures used in this study, the rule available in the literature could only predict 34% of the cases with an O-atom attached to the chiral center and 88% of the others. The best random forest models (using the PAS codes and the real values of the properties) were validated with *y*-randomization in the same way as for the amino alcohol catalysis of the first application. In five experiments with the PAS code including H atoms, the highest percentage of correct predictions was 51% in the OOB estimation with the training set and 61% for the test set. In the experiments without including H atoms, the highest percentages were 55% and 54%, respectively.

CPG NN (with or without selection of variables) and decision trees yielded poor predictions for the structures with an oxygen atom bonded to the chiral center, although reasonable performance was observed for the other cases. The decision tree generated in the experiment with 21 PAS descriptors and hydrogen atoms considered is shown in Figure 6. In the experiment yielding the best overall results (random forests, PAS descriptors + real values of the properties, H atoms not considered) the four most relevant variables for the RF model were the *R*/*S*-like descriptors based on (by order of importance) the following: (a) maximum electronegativity within the third sphere of bonds from the chiral center, (b) number of atoms within the third sphere of bonds from the chiral center, (c) number of atoms in the ligand, and (d) distance (in number of bonds) from the chiral center to the farthest atom in the ligand. The selection of such variables not only highlights the importance of the size of the ligands for the enantioselectivity, which is in accordance with the known rule for primary alcohols with no oxygen atoms bonded to the chiral center, but also suggests that electronic descriptors are crucial for the establishment of models that are able to cover most situations. In any case, no set of simple rules could be established by decision trees that can achieve high accuracy of predictions for the structures with oxygen atoms bonded to the chiral center, in the test set. For these situations, good predictions could only be obtained by the more complex models produced by RFs.

In both applications (enantioselective addition of diethylzinc and biocatalytic reactions), RF experiments were also performed excluding the descriptors more specifically related to reactivity (descriptors 18−21) and yielded essentially the same percentages of correct classifications (differences of 0−2%).

CONCLUSION

CIP-like stereodescriptors based on a series of unambiguously defined physicochemical properties of the ligands were successfully applied to qualitative structure-enantioselectivity relationships in organic reactions and biocatalysis. The series of descriptors for one chiral center can be used as a profile of the chiral center or can be submitted to a selection procedure (e.g. decision trees) for the establishment of rules on the basis of only a few descriptors. In the application concerning addition of diethylzinc to benzaldehyde catalyzed by chiral amino alcohols, the global profile of the chiral centers could establish good models with CPG NN. Improved performance was observed with random forests. In the application to biocatalysis by PCL, the global profiles of the relevant chiral centers allowed for good predictions by CPG NNs only for primary alcohols with no oxygen atom bonded to the chiral center. For the other substrates, accurate predictions were achieved by RFs that employ supervised learning and a complex process for features selection. Although complex, the RF models for the PCL reactions can account for the stereoselectivity toward substrates with or without an oxygen atom at the chiral center, which is not possible with the simple rule available in the literature.

ACKNOWLEDGMENT

The authors 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).

REFERENCES AND NOTES

(1) Aires-de-Sousa, J. Representation of Molecular Chirality. Vol. 3. In *Handbook of Chemoinformatics*; Gasteiger, J., Engel, J., Eds.; Wiley-VCH: New York, 2003; pp 1062−1078.

(2) Cahn, R. S.; Ingold, C.; Prelog, V. Specification of Molecular Chirality. *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 385−419.
(3) Prelog, V.; Helmchen, G. Basic Principles of the CIP-System and Proposals for a Revision. *Angew. Chem., Int. Ed. Engl.* **1982**, *21*, 567−583.
(4) Mata, P.; Lobo, A. M.; Marshall, C.; Johnson, A. P. The Cip Sequence Rules − Analysis and Proposal for a Revision. *Tetrahedron: Asymmetry* **1993**, *4*, 657−668.
(5) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological Organic Chemistry. 9. Graph Theory and Molecular Topological Indices of Stereoisomeric Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 864−870.
(6) Julián-Ortiz, J. V.; Alapont, C. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. Prediction of Properties of Chiral Compounds by Molecular Topology. *J. Mol. Graphics Modell.* **1998**, *16*, 14−18.
(7) Golbraikh, A.; Bonchev, D.; Tropsha, A. Novel Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147−158.
(8) For the proposal of a chirality function on the basis of topological descriptors, see: Lukovits, I.; Linert, W. A Topological Account of Chirality. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1517−1520.
(9) Aires-de-Sousa, J.; Gasteiger, *J.* 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.
(10) 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. Graphics Modell.* **2002**, *20*, 373−388.
(11) Aires-de-Sousa, J.; Gasteiger, J.; Gutman, I.; Vidovic, D. I. Chirality Codes and Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 831−836.
(12) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Feng, J.; Zheng, W.; Tropsha, A. QSAR Modeling of Datasets with Enantioselective Compounds Using Chirality Sensitive Molecular Descriptors. *SAR QSAR Environ. Res.* **2005**, *16* (1−2), 93−102.
(13) Caetano, S.; Aires-de-Sousa, J.; Daszykowski, A.; Heyden, Y. V. Prediction of Enantio Selectivity Using Chirality Codes and Classification and Regression Trees. *Anal. Chim. Acta* **2005**, *544*, 315−326.
(14) Aires-de-Sousa, J.; Gasteiger, J. Prediction of Enantiomeric Excess in a Combinatorial Library of Catalytic Enantioselective Reactions. *J. Comb. Chem*. **2005**, *7*, 298−301.
(15) Zhang, Q. Y.; Carrera, G. A.; Gomes, M. J. S.; Aires-de-Sousa, J. Automatic Assignment of Absolute Configuration from 1D NMR Data. *J. Org. Chem.* **2005**, *70*, 2120−2130.
(16) A similar idea was put forward in a communication to the First Indo−US Workshop On Mathematical Chemistry (Santiniketan, India, January 9−13, 1998) by P. D. Mosier and W. J. Murray, but, to the best of our knowledge, details of such approach were never published.
(17) Weissfloch, A. N. E.; Kazlauskas, R. J. Enantiopreference of Lipase from Pseudomonas Cepacia toward Primary Alcohols. *J. Org. Chem.* **1995**, *60*, 6959−6969.
(18) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, Germany, 1988; pp 119−138.
(19) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(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. Method.* **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: 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.
(26) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, FL, 2000.
(27) R Development Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org (accessed Nov 2004 and Jan 2005).
(28) Breiman, L. RandomForests. *Machine Learning* **2001**, *45*, 5−32.
(29) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1947−1958.
(30) Fortran original by Leo Breiman, Adele Cutler, R port by Andy Liaw and Matthew Wiener. http://www.stat.berkeley.edu/users/breiman/ (accessed Oct 2006).

CI600235W