Organic & Biomolecular Chemistry

RSC Publishing

PAPER

Cite this: *Org. Biomol. Chem.*, 2013, **11**, 4847

Successful combination of computationally inexpensive GIAO ^{13}C NMR calculations and artificial neural network pattern recognition: a new strategy for simple and rapid detection of structural misassignments†

Ariel M. Sarotti*

Received 24th April 2013,
Accepted 28th May 2013

DOI: 10.1039/c3ob40843d

www.rsc.org/obc

GIAO NMR chemical shift calculations coupled with trained artificial neural networks (ANNs) have been shown to provide a powerful strategy for simple, rapid and reliable identification of structural misassignments of organic compounds using only one set of both computational and experimental data. The geometry optimization, usually the most time-consuming step in the overall procedure, was carried out using computationally inexpensive methods (MM+, AM1 or HF/3-21G) and the NMR shielding constants at the affordable mPW1PW91/6-31G(d) level of theory. As low quality NMR prediction is typically obtained with such protocols, the decision making was foreseen as a problem of pattern recognition. Thus, given a set of statistical parameters computed after correlation between experimental and calculated chemical shifts the classification was done using the knowledge derived from trained ANNs. The training process was carried out with a set of 200 molecules chosen to provide a wide array of chemical functionalities and molecular complexity, and the results were validated with a set of 26 natural products that had been incorrectly assigned along with their 26 revised structures. The high prediction effectiveness observed makes this method a suitable test for rapid identification of structural misassignments, preventing not only the publication of wrong structures but also avoiding the consequences of such a mistake.

Introduction

The role of modern NMR techniques in structure elucidation is indisputable, allowing the characterization of complex molecular architectures in milligram quantities. Nevertheless, despite a steady advancement in complex multidimensional NMR experiments and more powerful spectrometers has been achieved over the course of the last few decades, several examples of structural and stereochemical misassignments are still appearing in the literature.[1] As pointed out by Nicolaou, incorrectly assigned natural products complicate the assessment of biosynthetic schemes and can also have profound consequences both in terms of time and money if a research group is willing to venture into their total synthesis.[1a] The accurate calculation of NMR chemical shifts with quantum chemical methods has laid a solid foundation to solve, at least in part, some of these problems.[2–5] The typical approach involves the calculation of the chemical shifts for all the candidate structures to identify the molecular arrangement that best matches the experimental data.[6,7]

This useful technique often has the geometry optimization as the most time-consuming step in the overall NMR calculations procedure, therefore much effort has been made to identify computational methods that afford good geometries for further high quality NMR predictions at minimal computational cost. It has been established that the use of DFT functionals (such as B3LYP or similar) coupled with the 6-31G(d) basis set (or superior) performs reasonably well for common organic compounds.[4] Even though these levels of theory are nowadays affordable for most small to medium sized molecular systems, the CPU time might be found to be prohibitively long for many situations in which quickness is an essential requirement. This is particularly important when dealing with conformationally flexible compounds, from which an extensive conformational search must be done to determine the relative contribution of each conformer by Boltzmann analysis.[4]

Instituto de Química Rosario (CONICET), Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario (2000), Argentina. E-mail: sarotti@iquir-conicet.gov.ar; Fax: +54-341-4370477; Tel: +54-341-4370477

† Electronic supplementary information (ESI) available: (a) Instructions for using the Excel file. Experimental chemical shifts, GIAO isotropic magnetic shielding values and statistical descriptors for all structures. Weight and bias matrices for all trained ANNs. (b) Excel file for automatic chemical shift and ANN calculations (with tutorial). See DOI: 10.1039/c3ob40843d

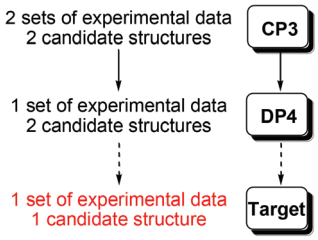

Fig. 1 Target method proposed in this work.

Recently, Smith and Goodman introduced sophisticated statistical methods for the stereochemical assignment of diastereoisomeric compounds using computationally inexpensive molecular mechanics for the geometry optimization step. As depicted in Fig. 1, the CP3 parameter was designed to assign two sets of experimental data to two possible structures,[8] while the DP4 applies to the more difficult case of assigning a pair of isomers with only one set of experimental data.[9]

According to the trend depicted in Fig. 1, having access to a method that could be applied to only one set of both experimental and computational data (situation often found in structural validation, that is, confirm or reject a given putative structure) seemed a valid next step. It would not be trivial to point out that in all current approaches at least two candidate structures are inevitably needed to assess which best matches the experimental data, which is the reason why any method envisaged to solve this problem should require a change in the paradigm.

Although NMR calculations can be employed to aid a structural hypothesis validation,[10] the main drawback comes from the intrinsic absence of any other plausible molecular architecture. As a result, the decision making is a difficult task because it is unclear what level of correlation between experimental and calculated data should be expected for a particular structure to be classified as correct or incorrect. This issue is of critical importance when computationally inexpensive methods are used, where even for a correct match the agreement might be poor.

Bagno pointed out the low probability of two different molecules having the same NMR spectrum.[11] Clearly, this scenario offers an outstanding opportunity to solve the structural validation problem if only the current *ab initio* or DFT methods had the ability to predict NMR data with perfect accuracy. Unfortunately, no such method is available (at least, not with an affordable computational cost). This provides a large limitation considering that different molecules often display similar NMR data (particularly in the cases of stereoisomerism). As a result, assessing the correctness of a putative structure using only one set of inexpensive computational data remains an unsolved task. Fortunately, this restriction cannot prevent the determination of the incorrect nature of a given structural proposal.

The main goal of this work was the development of a simple procedure to be used as a test for rapid and straightforward detection of structural misassignments. The approach herein proposed is to perform the geometry optimization step using computationally inexpensive methods (such as molecular mechanics, semiempirical or low level *ab initio* methods), followed by calculation of the NMR properties at the affordable mPW1PW91/6-31G(d) level of theory. As high quality NMR prediction with such protocols was not expected, the decision making was considered as a problem of pattern recognition. Thus, given a set of statistical parameters computed after correlation between the calculated and experimental ^{13}C values for a candidate structure, the classification might be done using the knowledge derived from analysis of the patterns corresponding to known correct and incorrect structures.[12]

An artificial neural network (ANN) is a mathematical model in which a number of interconnected artificial neurons mimic the behavior of a biological brain. One of the most relevant properties of the ANNs is their ability to learn from the data, which has many applications in pattern recognition, classification, clustering and more,[13] making them ideally suited to be used in this work. The successful combination of quantum chemical calculations with ANNs to tackle different issues has been reported,[14] though this work represents the first application of such an approach for structural validation using NMR chemical shift calculations.

Computational methods

All molecular mechanics calculations were performed using Hyperchem[15] with the MM+ force field[16] and the quantum mechanical calculations were performed using Gaussian 09.[17] In the case of conformationally flexible compounds, the conformational search was done in the gas phase using the MM+ force field, with the number of steps large enough to find all low-energy conformers at least 10 times. All conformers within 5 kcal mol^{-1} of the lowest energy conformer were subjected to further reoptimization at the AM1[18] and HF/3-21G levels of theory. With the most stable conformers in hand (up to 5 kcal mol^{-1} of the lowest energy conformer) the next step was the shielding constants single point calculation using the GIAO (gauge including atomic orbitals) method,[19] with the mPW1PW91 functional[20] (one of the most reliable DFT functionals for NMR calculations)[4,5g,6a,7] and the 6-31G(d) basis set. To simplify the process and reduce the computational cost all calculations were carried out in the gas phase.[7–9]

The NMR shielding constants were subjected to Boltzmann averaging over all conformers according to:[4,7–9]

$$\sigma^x = \frac{\sum_i \sigma_i^x \exp(-E_i/RT)}{\sum_i \exp(-E_i/RT)} \quad (1)$$

where σ^x is the Boltzmann-averaged shielding constant for nucleus x, σ_i^x is the shielding constant for nucleus x in conformer i, R is the molar gas constant (8.3145 J K^{-1} mol^{-1}), T is the temperature (298 K), and E_i is the gas phase single point mPW1PW91/6-31G(d) energy of conformer i (relative to the lowest energy conformer).

Once the shielding constants were computed, the chemical shifts were calculated according to:[7]

$$\delta^{x}_{\mathrm{calc}} = \sigma_{\mathrm{ref}} - \sigma^{x} + \delta_{\mathrm{ref}} \quad (2)$$

where σ_{ref} is the NMR isotropic magnetic shielding value for the reference compound, and δ_{ref} is the experimental chemical shift of the reference compound in deuterated chloroform. In this study two different methods to calculate the NMR chemical shifts were used, namely TMS and MSTD. In the TMS method, all chemical shifts are calculated using TMS as the reference standard (δ_{ref} = 0.00 ppm), while in the multi-standard approach (MSTD), methanol (δ_{ref} = 50.41 ppm) and benzene (δ_{ref} = 128.37) were used as reference standards for sp^3 and sp–sp^2 hybridized carbon atoms, respectively.[21] Sarotti and Pellegrinet have recently found that this simple modification allowed much better accuracy and lower dependence on the theory level employed, both for ^{13}C and ^{1}H NMR shift calculation procedures.[7]

The ANN training was done using the Neural Network Toolbox incorporated in MATLAB 7.0.[22]

Results and discussion

ANNs are computational algorithms that can be used to examine data and develop models to identify patterns in the data (known as the training process), which in turn can be used to make further predictions.[13] Among the two main approaches for network training (supervised and unsupervised), the first is most commonly used for a variety of applications. In supervised training, both the inputs and outputs are given to train the ANN to perform a particular job. The set of data which enables the training is called "training set", which in this study was a set of data derived from known correct and incorrect structures. For the former, 100 small-to-medium sized compounds (Fig. 2) were selected to provide a wide array of chemical functionalities and molecular complexity and also because their ^{13}C NMR spectra are well-known.[23] To create the test set of incorrect structures, some of the compounds shown in Fig. 2 were deliberately modified to create 100 additional molecular architectures (Fig. 3), and the calculated NMR shifts were correlated with the experimental data corresponding to its parent structure.

Once the NMR shifts of the 200 compounds of the training set were computed at the three levels of theory under study, the next step was the calculation of statistical parameters commonly used to quantify the agreement between experimental and computational NMR shifts. To do so, it would be necessary to know which experimental shift corresponds to which calculated shift, which in turn requires having the experimental data fully assigned (that is, having all resonances assigned to all the corresponding nuclei in the proposed structure). However, even with 2D NMR data available, it is not uncommon to misassign at least some of the signals leading to serious consequences when correlating with the calculated data. On the other hand, if the candidate structure is wrong, any assignment would have to be inevitably incorrect. To sort these potential problems, the corresponding δ_{exp} and δ_{calc} were sorted in descending order of size and the resulting values were paired, providing an additional simplification as unassigned data can be used (*e.g.* from the 1D NMR spectrum).

Finally, systematic errors during the shift calculation were removed by empirical scaling according to:[4]

$$\delta_{\mathrm{scaled}} = (\delta_{\mathrm{calc}} - b)/m \quad (3)$$

where m and b are the slope and the intercept, respectively, resulting from a linear regression calculation on a plot of δ_{calc} against δ_{exp}. The scaling factors (b and m) can be determined in two ways: (a) using the data from large databases[4,5b,k,6a] (for example, at http://cheshirenmr.info) or (b) from a plot of δ_{calc} against δ_{exp} for each particular compound under study. This procedure has been extensively used,[4,5i,j,6c,7–9] and was the method of choice in this study. In a perfect correlation, b, m and R^2 (the correlation coefficient) would be 0, 1 and 1, respectively. In addition, six other statistical descriptors were used to quantify the agreement between experimental and calculated data: the mean absolute error (MAE, defined as $\Sigma_n|\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}|/n$), the corrected mean absolute error (CMAE, defined as $\Sigma_n|\delta_{\mathrm{scaled}} - \delta_{\mathrm{exp}}|/n$), the standard deviation (σ, defined as $[\Sigma_n(|\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}| - \mathrm{MAE})^2/(n-1)]^{1/2}$), the corrected standard deviation (Cσ, defined as $[\Sigma_n(|\delta_{\mathrm{scaled}} - \delta_{\mathrm{exp}}| - \mathrm{CMAE})^2/(n-1)]^{1/2}$), the maximum error (MaxErr, defined as $\max|\delta_{\mathrm{calc}} - \delta_{\mathrm{exp}}|$) and the corrected maximum error (CMaxErr, defined as $\max|\delta_{\mathrm{scaled}} - \delta_{\mathrm{exp}}|$). Therefore, for each reference standard used in the chemical shift calculation (TMS and MSTD), a total number of 9 statistical parameters are computed: MAE, CMAE, σ, Cσ, MaxErr, CMaxErr, m, b and R^2. In Table 1 are shown two representative examples to clarify the process.

At this point, it would be important to recall that the main goal of this work was the development of a method that could determine if a putative structure is right or wrong using only one set of both computational and experimental data. For that reason, any attempt to directly compare any pair of correct and incorrect compounds shown in Fig. 2 and 3 must be avoided. For instance, after a simple glance of the data presented in Table 1 for both **1** and **101**, it would be easy to affirm that the last one should be the incorrect one, but the decision making is based on a direct comparison between two candidates. On the other hand, this method was thought to achieve the same conclusion using only the data computed for **101**.

Construction of the network

In this study, two-layer feed-forward networks were used. A schematic representation of such ANNs is given in Fig. 4, in which are highlighted the three main components of the network architecture: the input layer, the hidden layer and the output layer.

These types of ANNs are said to be fully connected, that is, each node is connected to every node in the preceding layer. In the biological neuron, the synapse permits a neuron to transfer an electrical signal to another neuron, allowing the

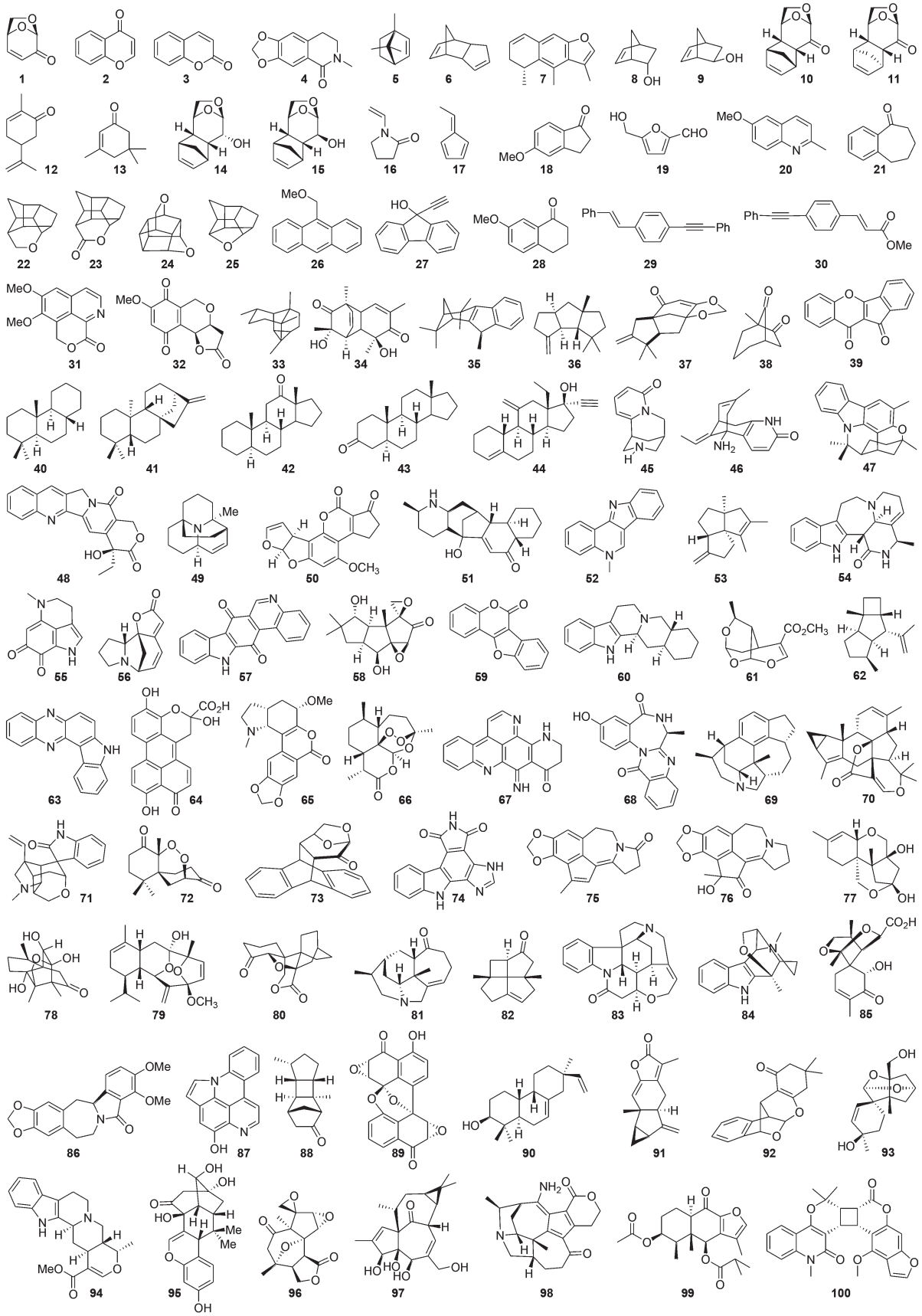

Fig. 2 Molecules used in the correct test set.

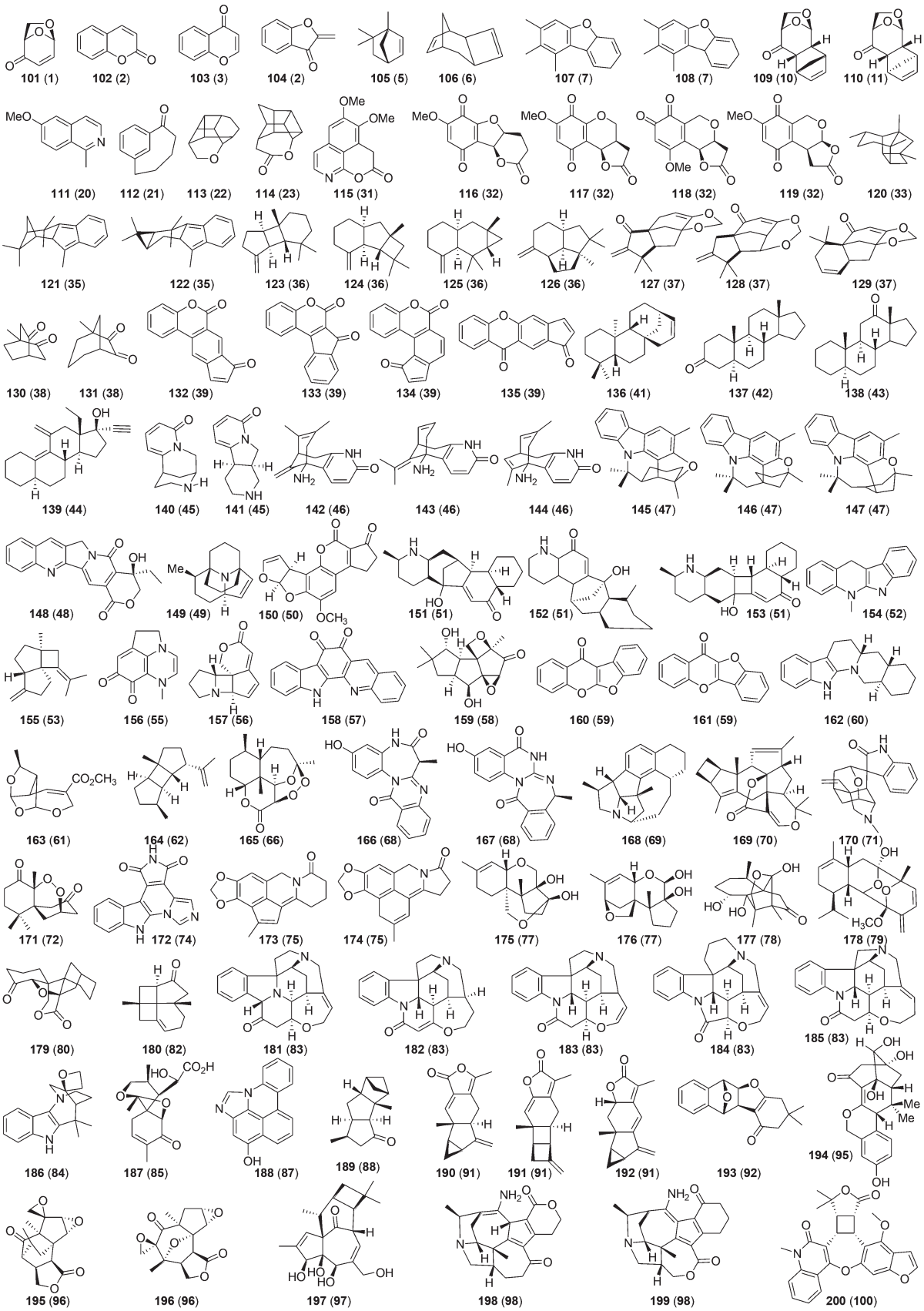

Fig. 3 Molecules used in the incorrect test set. In parentheses are shown the parent compounds from which experimental NMR data were taken.

Table 1 Statistical parameters calculated after matching the calculated chemical shifts for compounds **1** (correct structure) and **101** (incorrect structure) with the experimental chemical shifts of **1** at the mPW1PW91/6-31G(d)//HF/3-21G level of theory

σ^x	δ_{exp}	$\delta_{calc,MSTD}$	$\delta_{scaled,MSTD}$	$\delta_{calc,TMS}$	$\delta_{scaled,TMS}$
Compound **1**					
13.9991	188.6	188.3	189.4	180.8	191.3
57.2862	148.1	145.0	145.3	137.5	143.5
72.7391	126.5	129.6	129.6	122.0	126.5
94.5418	101.4	100.2	99.7	100.2	102.4
121.2435	71.5	73.5	72.6	73.5	73.0
127.9196	66.3	66.8	65.8	66.8	65.6
	MAE	1.70		4.45	
	CMAE	1.65		1.74	
	σ	1.22		4.04	
	Cσ	1.08		1.65	
	MaxErr	3.08		10.62	
	CMaxErr	3.10		4.56	
	m	0.98		0.91	
	b	2.18		7.34	
	R^2	0.9980		0.9971	
Compound **101**					
8.3253	188.6	194.0	190.9	186.4	192.9
53.7939	148.1	148.6	146.3	141.0	144.7
73.7452	126.5	128.6	126.7	121.0	123.5
97.5324	101.4	97.2	96.0	97.2	98.3
114.7059	71.5	80.1	79.2	80.1	80.1
130.9038	66.3	63.9	63.3	63.9	62.9
	MAE	3.83		4.99	
	CMAE	3.40		4.29	
	σ	2.88		2.56	
	Cσ	2.69		2.16	
	MaxErr	8.55		8.56	
	CMaxErr	7.67		8.60	
	m	1.02		0.94	
	b	−0.66		4.51	
	R^2	0.9906		0.9881	

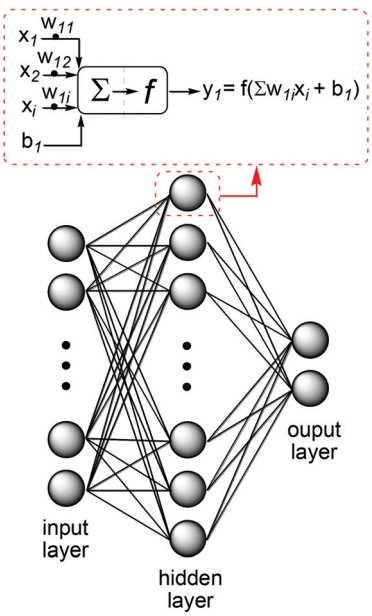

Fig. 4 Schematic illustration of a two-layer feed-forward ANN used in this study.

interconnection between them and providing the strength of the connection. In a similar fashion, in an artificial neuron the synapse is represented by a weight number (w). Each neuron receives the signal from the previous neuron, and each connection carries an assigned weight (that is, the synapse). All inputs are summed altogether and modified by the weights. Finally, an activation function (also known as the transfer function, f) controls the amplitude of the output.[13] Among several transfer functions, the hard-limit (or step), the linear and the sigmoid are the most commonly used for a variety of applications. For example, ANNs with sigmoid transfer functions for the first layer and linear transfer function for the second layer are typically used to fit multi-dimensional problems.[14a,e] In this study, two-layer feed-forward networks, with sigmoid hidden and output neurons, were used because it is known that this net architecture can classify vectors arbitrarily well.[13,22]

The input vector is constituted by the statistical descriptors (MAE, CMAE, MaxErr, *etc.*), denoted as x_i. Since the output values are in the range of 0 and 1, the input values must be scaled in the [−1, 1] region, according to:

$$x_{scaled} = [2(x_i - x_{min})/(x_{max} - x_{min})] - 1 \quad (4)$$

where x_{max} and x_{min} are the maximum and minimum values of the i^{th} input parameter, respectively. Each neuron of the hidden layer receives the scaled values from the input layer and produces the output value according to:

$$y_j = \tanh(net_j)$$
$$net_j = \sum_{i=1}^{n} w_{ji}x_i + b_j \quad (5)$$

where n is the number of neurons in the input layer, w_{ji} is the connection weight of the i^{th} neuron in the input layer to the j^{th} neuron of the hidden layer, x_i is the scaled value of the i^{th} neuron of the input layer, and b_j is the bias value of the j^{th} neuron. The output of the first layer (y_j) is then used as an input value for the second layer (the output layer) according to:

$$z_k = \tanh(net_k)$$
$$net_k = \sum_{j=1}^{m} w_{kj}y_j + b_k \quad (6)$$

where m is the number of neurons in the hidden layer, w_{kj} is the connection weight of the j^{th} neuron in the hidden layer to the k^{th} neuron of the output layer, and y_j is the output from the hidden layer, and b_k is the bias value of the k^{th} neuron. z_k is the output value of the output layer, which is finally scaled in the [0, 1] region according to:

$$z_{scaled} = (z_i + 1)/2 \quad (7)$$

Once the net architecture was established, the question arose on the size of the input, hidden and output layers to afford optimal results. Regarding the size of the input layer, three different sets of statistical descriptors were used as input vectors: (a) the 9 parameters obtained from using TMS as the reference standard, (b) the 9 parameters obtained using MSTD as the reference standard, and (c) the 18 parameters obtained

combining the 9 parameters obtained from using TMS with the 9 parameters obtained from using MSTD (referred to as TMS + MSTD). On the other hand, two categories were chosen for the output layer: category 1 for correct structures and category 2 for incorrect structures. The training pattern for category 1 is (1 0), whereas the training pattern for category 2 is (0 1). Finally, after preliminary investigation the hidden layer was set to a number of 20 neurons since increasing this value did not afford improved results.

In the training process, a scaled conjugate gradient backpropagation learning algorithm was used to determine the optimal values of the weights and bias values for each connection between neurons.[13,22] In this procedure, the error between the calculated and experimental output vector is calculated, and the weights are corrected throughout all the network from the last layer to the first layer, until an acceptable error is reached. Each iteration is called "epoch", and in this work a maximum of 1000 epochs was used to achieve convergence criteria. For the purpose of the training, the input data were randomly divided into three subsets: 80% of the data were used for the training (used to adjust the weights and bias), 15% for the validation (used to decide when to stop the training process, avoiding overfitting) and 5% for testing the net (used to measure the performance of the trained network). The results obtained after training the ANNs are given in Table 2, and all weights and biases corresponding to each trained network can be found in the ESI.†

Table 2 Percentage of correct classification of the nine ANNs used in this study

ANN	Geometry optimization method	Reference standard	No. of neurons in the input layer	% of correct classification
MM-9a	MM+	TMS	9	92%
MM-9b	MM+	MSTD	9	92%
MM-18	MM+	TMS + MSTD	18	94%
AM1-9a	AM1	TMS	9	94%
AM1-9b	AM1	MSTD	9	92%
AM1-18	AM1	TMS + MSTD	18	97%
HF-9a	HF/3-21G	TMS	9	97%
HF-9b	HF/3-21G	MSTD	9	97%
HF-18	HF/3-21G	TMS + MSTD	18	100%

From the data presented in Table 2 it can be seen that all methods performed well in terms of pattern classification (as determined by the percentage of the 200 molecules of the training set that were correctly classified after training of the ANN). Not unexpectedly, the overall performance of the ANNs increased with the computational cost of the method employed in the geometry optimization step (HF/3-21G > AM1 > MM+). Moreover, even though the use of TMS and MSTD as reference standards afforded similar results, the combination of both (18 parameters) slightly increased the ability of the ANNs to differentiate between correct and incorrect structures.

To clarify the outstanding capability of the ANNs to explore and identify patterns in the data, Table 3 shows the minimum and maximum values for the 9 statistical parameters computed using MSTD as the reference standard for the 100 compounds depicted in Fig. 2,[24] along with the percentage of the 100 misassigned compounds (Fig. 3) whose calculated parameters fall within the corresponding range. In view of these results, it can be seen that none of the statistical parameters taken alone can be used to classify the data with the same level of accuracy found for the ANNs after the training process. For example, the R^2 parameter of 89%, 78% and 86% of all incorrect structures (computed at the MM+, AM1 and HF/3-21G levels, respectively) were located within the range corresponding to known correct structures, meaning that only 11%, 22% and 14% of wrong structures could have been correctly classified using only this statistical parameter. Naturally, a more sophisticated algorithm could be found to achieve a better classification process combining some of these parameters, with the ANNs being ideal to tackle this issue in a simple and straightforward fashion.

In a more comprehensive example on the real usefulness of trained ANNs in structural validation, Table 4 shows the statistical parameters that are typically used to quantify the agreement between experimental and computational data (and therefore, to assess whether the proposed structure is wrong or right), of 6 of the 200 compounds shown in Fig. 1 and 2 computed at the mPW1PW91/6-31G(d)//MM+ level of theory using the MSTD approach.[25]

Even with the information provided in Table 3, any assignment on the correct or incorrect nature of compounds **A–F**

Table 3 Minimum and maximum values of the 9 statistical parameters computed using MSTD for the 100 compounds shown in Fig. 2, and percentage of the 100 incorrect compounds shown in Fig. 3 whose calculated parameters fall within the range

	MM+			AM1			HF/3-21G		
	Min	Max	%	Min	Max	%	Min	Max	%
MAE	0.88	7.28	87	1.11	4.24	39	0.84	3.35	20
σ	0.67	3.59	44	0.56	3.25	39	0.51	2.98	31
MaxErr	2.31	15.31	56	1.88	12.64	42	1.71	13.13	53
R^2	0.9494	0.9996	89	0.9719	0.9997	78	0.9592	0.9998	86
m	0.83	1.07	81	0.78	1.08	88	0.89	1.06	82
b	−7.14	23.55	87	−9.82	12.18	82	−6.23	14.73	82
CMAE	0.46	3.77	60	0.45	4.16	63	0.35	3.11	34
Cσ	0.31	3.20	64	0.34	3.17	57	0.24	2.73	47
CMaxErr	1.21	12.45	63	1.21	12.18	63	0.97	10.60	55

Table 4 Selected statistical parameters computed for 6 of the 200 compounds shown in Fig. 1 and 2 at the mPW1PW91/6-31G(d)//MM+ level of theory using the MSTD approach

	Compound					
	A	**B**	**C**	**D**	**E**	**F**
MAE	3.50	3.20	3.59	7.30	3.91	3.82
MaxErr	10.48	5.02	6.88	11.51	11.62	10.65
R^2	0.9979	0.9947	0.9936	0.9972	0.9940	0.9936
CMAE	1.70	2.51	3.52	1.98	2.60	3.05

would not be a trivial task. Remember that the goal here is to assess whether a putative structure is right or wrong based only on the statistical parameters calculated for that particular structure. Because there is not another plausible candidate, any direct comparison is simply impossible. As seen in Table 4, the highest R^2 and lowest CMAE correspond to structure **A**, and the lowest MaxErr are found for structures **B** and **C**. Structure **D** displays high MAE and MaxErr, but the R^2 and CMAE are relatively good. In the case of structures **E** and **F** both MAE, CMAE and MaxErr are high, and the R^2 parameters are not among the highest. However, since all these parameters fall in the range corresponding to known correct structures (Table 3), the decision making remains an obstacle difficult to overcome. Actually, **A**, **B** and **C** are the incorrect compounds **110**, **111** and **118**, respectively (Fig. 2), while **D**, **E** and **F** are the correctly assigned compounds **58**, **89** and **70**, respectively (Fig. 1). All these compounds were successfully classified by the trained ANN-MM-18, proving further utility of the herein proposed methodology.

Testing the trained ANNs

The capability of the trained ANNs to successfully classify the data corresponding to known correct and incorrect structures of the training set indicated that the information provided in the input layer was enough. However, it is important to recall that the ANNs were trained using a "virtual" set of incorrect structures. For that reason, the evaluation of the performance of the ANNs in real cases of structural misassignment was considered a fundamental stage in this study. Accordingly, a set of 26 natural products that were incorrectly assigned based on NMR data (Fig. 5) was selected to test the overall performance of the optimal trained ANNs (ANN-MM-18, ANN-AM1-18 and ANN-HF-18). The collected results are given in Table 5.

Fortunately, even with the large variety of chemical functionalities and molecular complexity of compounds **201–226** the overall performance of the trained ANNs was excellent. As expected from the training process, the ANN-HF-18 could successfully classify all structures in category 2 (incorrect), while the ANN-MM-18 failed in two cases (compounds **216** and **219**) and the ANN-AM1-18 in only one case (compound **219**). These results not only provided further evidence on the ability of the trained ANNs to correctly identify structural misassignments on compounds that were not used in the training set, but also validated the choice of using "virtual" incorrect structures in it.

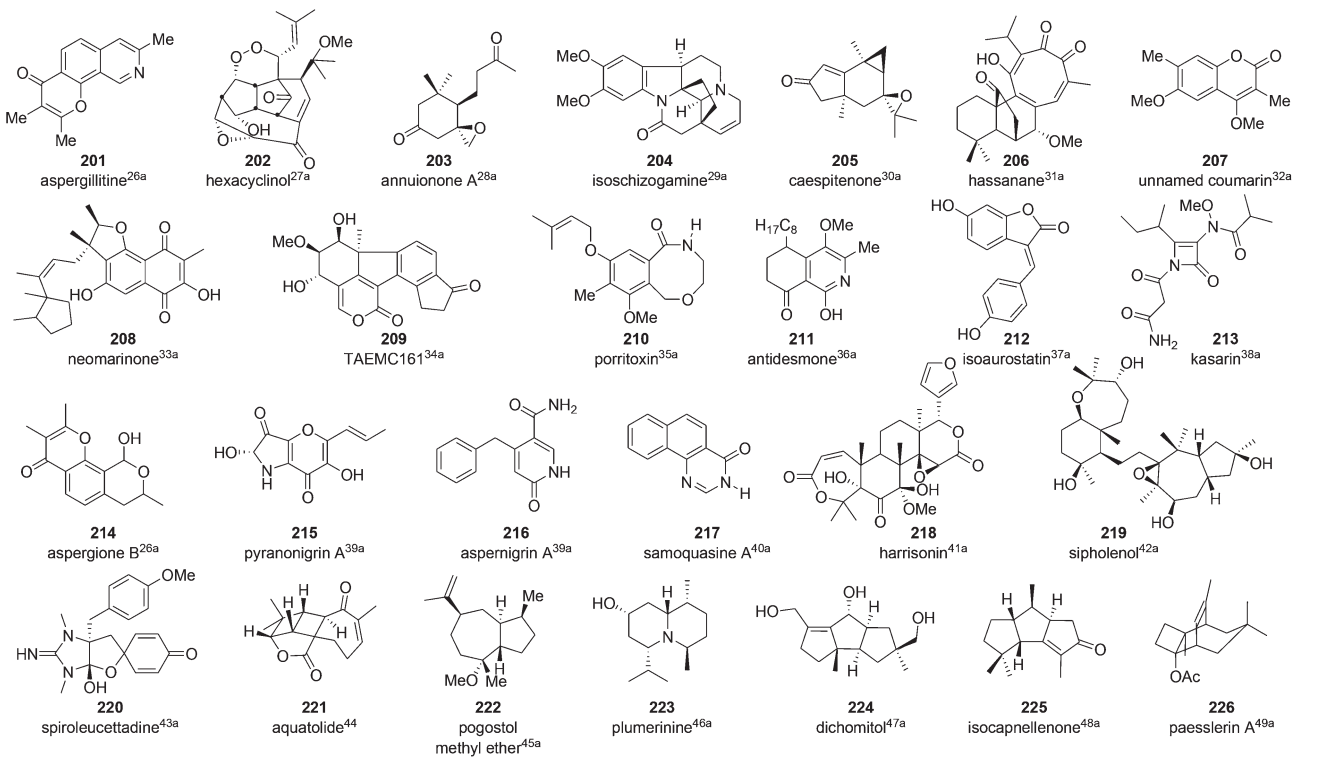

Fig. 5 Natural products that were originally misassigned based on NMR data.

Table 5 Output patterns obtained after testing the optimally trained ANNs with the 26 molecules shown in Fig. 5[a]

Structure	ANN-MM-18	ANN-AM1-18	ANN-HF-18
201	0.0071; 0.9957	0.0016; 0.9992	0.0021; 0.9977
202	0.0208; 0.9851	0.0008; 0.9998	0.0000; 1.0000
203	0.0001; 0.9999	0.0000; 1.0000	0.0000; 1.0000
204	0.0166; 0.9882	0.0000; 1.0000	0.0000; 1.0000
205	0.0002; 0.9998	0.0209; 0.9905	0.0000; 1.0000
206	0.0000; 1.0000	0.0000; 1.0000	0.0000; 1.0000
207	0.0000; 1.0000	0.0000; 1.0000	0.0000; 1.0000
208	0.0329; 0.9784	0.0120; 0.9917	0.0000; 0.9999
209	0.0366; 0.9723	0.0000; 1.0000	0.0000; 1.0000
210	0.2646; 0.7788	0.0019; 0.9992	0.0003; 0.9994
211	0.0000; 1.0000	0.0000; 1.0000	0.0000; 1.0000
212	0.0541; 0.9818	0.0017; 0.9992	0.0000; 1.0000
213	0.0000; 1.0000	0.0000; 1.0000	0.0000; 1.0000
214	0.1161; 0.9299	0.0237; 0.9846	0.0125; 0.9790
215	0.0000; 1.0000	0.0000; 1.0000	0.0000; 1.0000
216	0.7458; 0.2998	0.0733; 0.9103	0.0009; 0.9993
217	0.0003; 0.9998	0.0006; 1.0000	0.0000; 1.0000
218	0.2749; 0.7289	0.0000; 1.0000	0.1221; 0.7964
219	0.6705; 0.3565	0.9886; 0.0063	0.2353; 0.6994
220	0.0431; 0.9700	0.0001; 1.0000	0.1102; 0.8307
221	0.0000; 1.0000	0.0000; 1.0000	0.0000; 1.0000
222	0.4957; 0.5667	0.2523; 0.7674	0.0602; 0.8814
223	0.0004; 0.9999	0.0000; 1.0000	0.0000; 1.0000
224	0.0020; 0.9990	0.0000; 1.0000	0.0000; 1.0000
225	0.0054; 0.9948	0.0000; 1.0000	0.0000; 1.0000
226	0.0607; 0.9591	0.1582; 0.9084	0.0000; 1.0000

[a] The output values of the ANNs were required to assume discrete values 0 or 1 in the training process. However, in general the final output patterns show continuous values between 0 and 1. This is a limitation of the resolution ability of the network, and should not be taken as probability distribution. Nevertheless, an output pattern closer to (1 0) than to (0 1) corresponds to category 1 (correct). Otherwise, corresponds to category 2 (incorrect).

Finally, to reject the possibility of an excessive tendency of the ANNs to classify structures in category 2, the revised structures of the compounds **201–221** (Fig. 6) were next computed (Table 6). On the other hand, compounds **222–226** are natural products whose putative structures were found to be incorrect by total synthesis, even though their correct structures have not been revised yet. In these cases, the calculated chemical shifts were correlated with the experimental chemical shifts of the synthesized structures.[45b,46b,47b,48b,49b]

Here again, the ANN-HF-18 correctly classified all structures in category 1 (correct), while the ANN-MM-18 and ANN-AM1-18 failed in two (compounds **232** and **244**) and three cases (compounds **232**, **242** and **244**), respectively.

These results clearly proved that the ANNs performed well when dealing with correct structures which were not used during the training stage, although a slight tendency of ANN-AM1-18 to overclassify structures in category 2 (incorrect) was observed.

Case study

Finally, to give a further example on the utility of the proposed methodology to solve structural elucidation problems, a recent case study of the structural revision of aquatolide by Shaw, Tantillo and co-workers[6a] is presented. This sesquiterpenoid was originally assigned as compound **221** (Fig. 5)[44] and further revised as **247** (Fig. 6). In one of the key steps to identify the correct structure of aquatolide, the authors made a combination of rational and arbitrary changes of the originally proposed structure **221**, followed by calculation of NMR shifts of all candidate structures considered at the SCRF-B3LYP/6-31+G-(d,p)//B3LYP/6-31G(d) level of theory. After this *in silico*

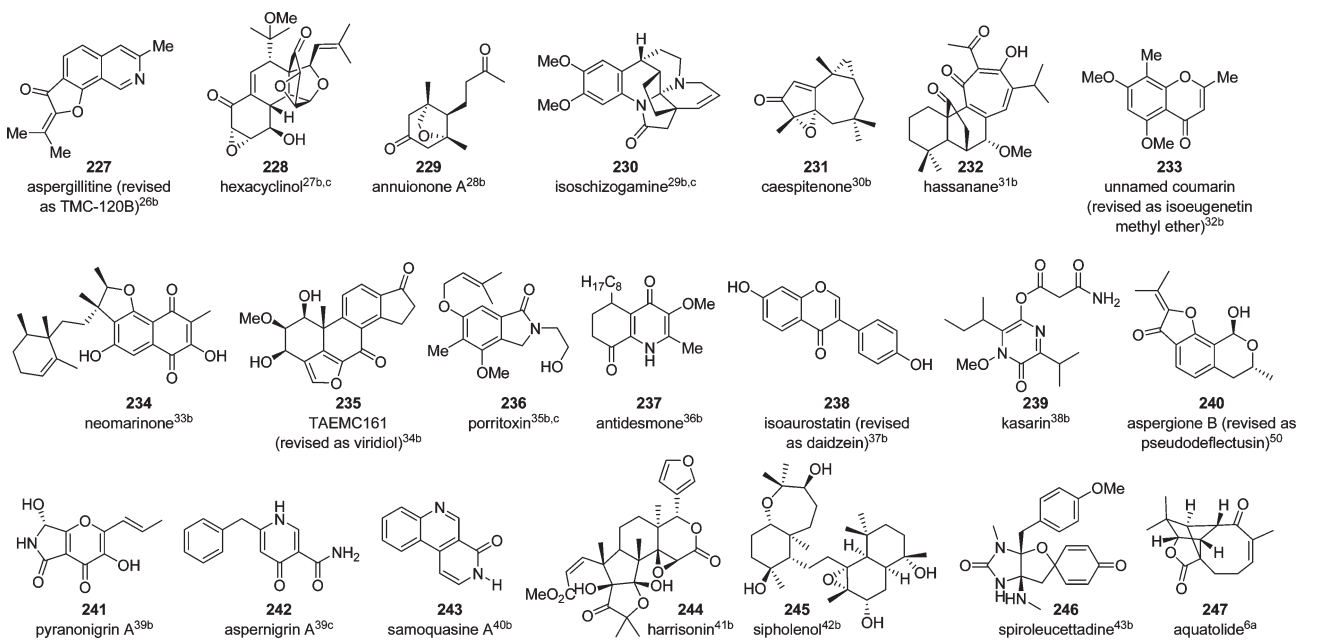

Fig. 6 Revised structures of compounds **201–221**.

Table 6 Output patterns obtained after testing the optimally trained ANNs with the 21 molecules shown in Fig. 6, and with the new NMR data of compounds **222–226**

Structure	ANN-MM-18	ANN-AM1-18	ANN-HF-18
227	0.5330; 0.4721	0.9680; 0.0207	0.9993; 0.0007
228	0.8726; 0.1289	0.6750; 0.3833	0.8903; 0.0805
229	0.9621; 0.0244	0.9913; 0.0064	0.9999; 0.0001
230	0.8338; 0.1589	0.9994; 0.0003	0.9997; 0.0003
231	0.7294; 0.2664	0.9919; 0.0056	0.9739; 0.0221
232	0.0008; 0.9995	0.0944; 0.9182	0.9998; 0.0002
233	0.7543; 0.2759	0.9996; 0.0002	0.9991; 0.0008
234	0.9474; 0.0478	0.9982; 0.0009	0.9990; 0.0008
235	0.9216; 0.0634	0.9286; 0.0652	0.9999; 0.0001
236	0.8853; 0.1313	0.9943; 0.0038	0.8813; 0.0873
237	0.9700; 0.0237	0.9998; 0.0001	0.9998; 0.0002
238	0.6833; 0.4091	0.9857; 0.0074	0.9851; 0.0121
239	0.9165; 0.0839	0.8847; 0.1296	0.9990; 0.0003
240	0.8628; 0.1506	0.9993; 0.0003	0.9976; 0.0021
241	0.9362; 0.0637	0.9672; 0.0230	0.9990; 0.0008
242	0.9766; 0.0173	0.0458; 0.9496	0.9996; 0.0003
243	0.9670; 0.0288	0.9998; 0.0001	1.0000; 0.0000
244	0.0092; 0.9929	0.0044; 0.9981	0.9278; 0.0425
245	0.9503; 0.0408	0.9964; 0.0018	0.9980; 0.0021
246	0.9678; 0.0236	0.9422; 0.0418	0.9990; 0.0007
247	0.7839; 0.2679	0.9788; 0.0171	0.9999; 0.0001
222	0.9890; 0.0064	0.9996; 0.0002	1.0000; 0.0000
223	0.9898; 0.0056	0.9999; 0.0000	1.0000; 0.0000
224	0.8424; 0.1718	0.6040; 0.4574	0.9753; 0.0146
225	0.9795; 0.0115	0.9975; 0.0012	0.9999; 0.0001
226	0.9674; 0.0269	0.9847; 0.0117	0.9997; 0.0004

screening the authors were able to propose an alternative structure (**247**), which was experimentally verified after reisolation, and extensive experimental NMR and X-ray crystallographic analysis.[6a]

To analyze whether the methodology proposed in the present study could have been useful in such computational screening process, 45 of the many alternative (incorrect) structures for aquatolide were randomly selected (see the ESI†). All these compounds, along with **221** and **247**, were optimized at the MM+, AM1 and HF/3-21G levels of theory (after extensive conformational searches), and the NMR shielding tensors calculated at the proposed mPW1PW91/6-31G(d) level. Once the statistical parameters were computed, the performance of the trained ANN-MM-18, ANN-AM1-18 and ANN-HF-18 was evaluated. In all cases, the trained ANNs could successfully classify all candidate structures as incorrect with the only exception of compound **247**, which is precisely the correct structure of aquatolide.

Final considerations

The results described herein clearly supported the initial goal of this work: developing a test method that could detect cases of structural misassignments in a simple, rapid and confident fashion. However, since care must be taken on hasty conclusions, some final considerations must be pointed out. A "category 1" classification means that the proposed structure has calculated NMR shifts that correlate with the experimental data in a similar manner to the correct structures of the training samples, it does not mean that the structure is actually correct. This situation, often found in cases of stereoisomerism, represents the main limitation of this method. As a consequence, the stereoassignment by computational methods needs to be done using procedures based on the comparison between all candidate structures, with the CP3 and DP4 methods being developed by Goodman highly recommended.[8,9] On the other hand, a "category 2" classification indicates that the candidate structure is probably wrong and should be conveniently revised.

Another issue that is important to discuss represents the well-known effect of higher than average errors obtained when calculating NMR shifts of carbon atoms attached to third row or greater elements (known as the heavy-atom effect), which mainly derives from the neglection of spin–orbit contributions from relativistic effects.[4] Since the ANNs were not trained to account for this effect, any carbon atom attached to heavy atoms should be excluded to prevent misclassification.

Practical aspects

This methodology was thought to help NMR spectroscopists in their everyday work. Naturally, the user must have the proper basic skills in computational chemistry but is not supposed to be an expert in ANNs nor in using MATLAB or related software. For that reason, an Excel file is provided as part of the ESI† to considerably simplify the process. Filling the enabled cells with the experimental NMR shifts and with the calculated GIAO shielding tensors, the spreadsheet computes the unscaled and scaled chemical shifts using TMS and MSTD, along with the 18 statistical parameters that are automatically introduced in the trained ANN (also provided in the same file) to obtain the output value.

Conclusions

In this work it has been shown that the NMR shift calculations using low theory levels in the geometry optimization, which constitutes one of the most time-consuming steps, coupled with trained ANNs represent a powerful test for simple, rapid and reliable identification of structural misassignments. From the results obtained, the method based on HF/3-21G geometries proved to be the most reliable in terms of classification ability, and should be used when computational resources are not a problem. On the other hand, the methods based on MM+ and AM1 geometries also displayed very good performances in terms of pattern recognition, having the advantage of being much faster than the former one. For that reason, those methods should be used if rapid and/or preliminary results are needed. Properly used, the methodology presented in this work could detect errors in early stages of structural assignment, preventing the publication of wrong structures and therefore lowering the probability of chasing molecules that were never there.[1a]

Acknowledgements

The author thanks CONICET, Universidad Nacional de Rosario, and ANPCyT for financial support.

Notes and references

1 (*a*) K. C. Nicolaou and S. A. Snyder, *Angew. Chem., Int. Ed.*, 2005, **44**, 1012–1044; (*b*) T. L. Suyama, W. H. Gerwick and K. L. McPhail, *Bioorg. Med. Chem.*, 2011, **19**, 6675–6701.

2 T. Helgaker, M. Jaszunski and K. Ruud, *Chem. Rev.*, 1999, **99**, 293–352.

3 G. Bifulco, P. Dambruoso, L. Gomez-Paloma and R. Riccio, *Chem. Rev.*, 2007, **107**, 3744–3779.

4 M. W. Lodewyk, M. R. Siebert and D. J. Tantillo, *Chem. Rev.*, 2012, **112**, 1839–1862.

5 (*a*) T. Bally and P. R. Rablen, *J. Org. Chem.*, 2011, **76**, 4818–4830; (*b*) R. Jain, T. Bally and P. R. Rablen, *J. Org. Chem.*, 2009, **74**, 4017–4023; (*c*) Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2008, **112**, 6794–6799; (*d*) Z. Wu, Y. Zhang, X. Xu and Y. Yan, *J. Comput. Chem.*, 2007, **28**, 2431–2442; (*e*) A. Bagno, F. Rastrelli and G. Saielli, *Chem.–Eur. J.*, 2006, **12**, 5514–5525; (*f*) K. W. Wiitala, T. R. Hoyle and C. J. Cramer, *J. Chem. Theory Comput.*, 2006, **2**, 1085–1092; (*g*) P. Cimino, L. Gomez-Paloma, D. Duca, R. Riccio and G. Bifulco, *Magn. Reson. Chem.*, 2004, **42**, S26–S33; (*h*) C. F. Tormena and G. V. da Silva, *J. Chem. Phys. Lett.*, 2004, **398**, 466–470; (*i*) G. Barone, L. Gomez-Paloma, D. Duca, A. Silvestri, R. Riccio and G. Bifulco, *Chem.–Eur. J.*, 2002, **8**, 3233–3239; (*j*) G. Barone, D. Duca, A. Silvestri, L. Gomez-Paloma, R. Riccio and G. Bifulco, *Chem.–Eur. J.*, 2002, **8**, 3240–3245; (*k*) P. R. Rablen, S. A. Pearlman and J. Finkbiner, *J. Phys. Chem. A*, 1999, **103**, 7357–7363.

6 For leading references, see: (*a*) M. W. Lodewyk, C. Soldi, P. B. Jones, M. M. Olmstead, J. Rita, J. T. Shaw and D. J. Tantillo, *J. Am. Chem. Soc.*, 2012, **134**, 18550–18553; (*b*) K. W. Quasdorf, A. D. Huters, M. W. Lodewyk, D. J. Tantillo and N. K. Garg, *J. Am. Chem. Soc.*, 2012, **134**, 1396–1399; (*c*) G. Saielli, K. C. Nicolaou, A. Ortiz, H. Zhang and A. Bagno, *J. Am. Chem. Soc.*, 2011, **133**, 6072–6077; (*d*) M. W. Lodewyk and D. J. Tantillo, *J. Nat. Prod.*, 2011, **74**, 1339–1343; (*e*) J. A. Mendoza-Espinoza, F. López-Vallejo, M. Fragoso-Serrano, R. Pereda-Miranda and C. M. Cerda-García-Rojas, *J. Nat. Prod.*, 2009, **72**, 700–708; (*f*) B. Wang, A. T. Dossey, S. S. Walse, A. S. Edison and K. M. Merz, *J. Nat. Prod.*, 2009, **72**, 709–713; (*g*) S. G. Smith, R. S. Paton, J. W. Burton and J. M. Goodman, *J. Org. Chem.*, 2008, **73**, 4053–4062; (*h*) S. M. Koskowich, W. C. Johnson, R. S. Paley and P. R. Rablen, *J. Org. Chem.*, 2008, **73**, 3492–3496; (*i*) G. Hu, K. Liu and L. J. Williams, *Org. Lett.*, 2008, **10**, 5493–5496; (*j*) E. Fattorusso, P. Luciano, A. Romano, O. Taglialatela-Scafati, G. Appendino, M. Borriello and C. Fattorusso, *J. Nat. Prod.*, 2008, **71**, 1988–1992; (*k*) A. M. Belostotskii, *J. Org. Chem.*, 2008, **73**, 5723–5731; (*l*) A. R. Allouche, D. Graveron-Demilly, F. Fauvelle and M. Aubert-Frecon, *Chem. Phys. Lett.*, 2008, **466**, 219–222; (*m*) K. N. White, T. Amagata, A. G. Oliver, K. Tenney, P. J. Wenzel and P. Crews, *J. Org. Chem.*, 2008, **73**, 8719–8722; (*n*) D. C. Braddock and H. S. Rzepa, *J. Nat. Prod.*, 2008, **71**, 728–730; (*o*) A. G. Griesbeck, D. Blunk, T. T. El-Idreesy and A. Raabe, *Angew. Chem., Int. Ed.*, 2007, **46**, 8883–8886; (*p*) C. Bassarello, G. Bifulco, P. Montoro, A. Skhirtladze, E. Kemertelidze, C. Pizza and S. Piacente, *Tetrahedron*, 2007, **63**, 148–154; (*q*) J. X. Pu, S. X. Huang, J. Ren, W. L. Xiao, L. M. Li, R. T. Li, L. B. Li, T. G. Liao, L. G. Lou, H. J. Zhu and H. D. Sun, *J. Nat. Prod.*, 2007, **70**, 1706–1711; (*r*) C. Fattorusso, E. Stendardo, G. Appendino, E. Fattorusso, P. Luciano, A. Romano and O. Taglialatela-Scafati, *Org. Lett.*, 2007, **9**, 2377–2380; (*s*) K. C. Nicolaou and M. O. Frederick, *Angew. Chem., Int. Ed.*, 2007, **46**, 5278–5282; (*t*) G. Rasul, G. A. Olah and G. K. S. Prakash, *J. Phys. Chem. A*, 2006, **110**, 7197–7201; (*u*) S. D. Rychnovsky, *Org. Lett.*, 2006, **8**, 2895–2898; (*v*) G. Bifulco, L. Gomez-Paloma, R. Riccio, C. Gaeta, F. Troisi and P. Neri, *Org. Lett.*, 2005, **7**, 5757–5760; (*w*) A. Aiello, E. Fattorusso, P. Luciano, A. Mangioni and M. Menna, *Eur. J. Org. Chem.*, 2005, 5024–5030.

7 (*a*) A. M. Sarotti and S. C. Pellegrinet, *J. Org. Chem.*, 2009, **74**, 7254–7260; (*b*) A. M. Sarotti and S. C. Pellegrinet, *J. Org. Chem.*, 2012, **77**, 6059–6065.

8 S. G. Smith and J. M. Goodman, *J. Org. Chem.*, 2009, **74**, 4597–4607.

9 S. G. Smith and J. M. Goodman, *J. Am. Chem. Soc.*, 2010, **132**, 12946–12959.

10 (*a*) Y. J. Hong and D. J. Tantillo, *Org. Lett.*, 2011, **13**, 1294–1297; (*b*) M. Weitman, L. Lerman, S. Cohen, A. Nudelman, D. T. Major and H. E. Gottlieb, *Tetrahedron*, 2010, **66**, 1465–1471; (*c*) G. Hu, K. Liu and L. J. Williams, *Org. Lett.*, 2008, **10**, 5493–5496; (*d*) C. Timmons and P. Wipf, *J.Org. Chem.*, 2008, **73**, 9168–9170.

11 G. Saielli and A. Bagno, *Org. Lett.*, 2009, **11**, 1409–1412.

12 This work has been limited to ^{13}C NMR calculations since the chemical shifts are spread over a wide range, they are relatively insensitive to solvent changes and display high influence on steric and electronic factors in the structure. In addition, the low probability of signal overlapping considerably simplifies the entire procedure.

13 J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley VCH, Weinheim, 1999.

14 (*a*) H. M. Le, T. S. Dinh and H. V. Le, *J. Phys. Chem. A*, 2011, **115**, 10862–10870; (*b*) A. T. H. Le, N. H. Vu, T. S. Dinh, T. M. Cao and H. M. Le, *Theor. Chem. Acc.*, 2012, **131**, 1158–1170; (*c*) H. T. T. Nguyen and H. M. Le, *J. Phys. Chem. A*, 2012, **116**, 4629–4638; (*d*) K. Makarova, E. V. Rokhina, E. A. Golovina, H. Van As and J. Virkutyte, *J. Phys. Chem. A*, 2012, **116**, 443–451; (*e*) R. M. Balabin and E. I. Lomakina, *J. Chem. Phys.*, 2009, **131**, 074104; (*f*) R. Vendrame, R. S. Braga, Y. Takahata and D. S. Galvão, *J. Mol. Struct. (THEOCHEM)*, 2001, **539**, 253–265.

15 *Hyperchem Professional Release 7.52*, Hypercube, Inc., 2005.

16 N. L. Allinger, *J. Am. Chem. Soc.*, 1977, **99**, 8127–8134.

17 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09*, Gaussian, Inc., Wallingford CT, 2009.

18 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.

19 (*a*) R. Ditchfield, *J. Chem. Phys.*, 1972, **56**, 5688–5691; (*b*) R. Ditchfield, *Mol. Phys.*, 1974, **27**, 789–807; (*c*) C. M. Rohlfing, L. C. Allen and R. Ditchfield, *Chem. Phys.*, 1984, **87**, 9–15; (*d*) K. Wolinski, J. F. Hinton and P. Pulay, *J. Am. Chem. Soc.*, 1990, **112**, 8251–8260.

20 C. Adamo and V. Barone, *J. Chem. Phys.*, 1998, **108**, 664–675.

21 The chemical shifts of benzene and methanol in $CDCl_3$ were used regardless of the solvent employed to obtain the experimental data, since the chemical shifts of these reference standards in all common deuterated solvents are nearly the same.

22 MATLAB, *MathWorks*, Natick, MA, USA, 2007.

23 Full references of the original papers are provided in the ESI.†

24 Similar observations were made when comparing the statistical parameters derived from TMS.

25 Similar results can be found with AM1 and HF/3-21G geometries, using both TMS and MSTD as reference standards.

26 (*a*) W. H. Lin, G. Brauers, R. Ebel, V. Wray, A. Berg, Surdasono and P. Proksch, *J. Nat. Prod.*, 2003, **66**, 57–61; (*b*) S. O. Simonetti, E. L. Larghi, A. B. J. Bracca and T. S. Kaufman, *Org. Biomol. Chem.*, 2012, **10**, 4124–4134.

27 (*a*) G. Schlegel, A. Hartl, H. M. Dahse, F. A. Gollmick, U. Gräfe, H. Dorfelt and B. Kappes, *J. Antibiot.*, 2002, **55**, 814–817; (*b*) S. D. Rychnovsky, *Org. Lett.*, 2006, **8**, 2895–2898; (*c*) J. A. Porco Jr., S. Su, X. Lei, S. Bardhan and S. D. Rychnovsky, *Angew. Chem., Ind. Ed.*, 2006, **45**, 5790–5792.

28 (*a*) F. A. Macías, R. M. Varela, A. Torres, R. M. Oliva and J. M. G. Molinillo, *Phytochemistry*, 1998, **48**, 631–636; (*b*) H. Takikawa, K. Isono, M. Sasaki and F. A. Macías, *Tetrahedron Lett.*, 2003, **44**, 7023–7025.

29 (*a*) U. Renner and H. Fritz, *Helv. Chim. Acta*, 1965, **48**, 308–317; (*b*) J. Hájícek, J. Taimr and M. Budesínky, *Tetrahedron Lett.*, 1998, **39**, 505–508; (*c*) J. L. Hubbs and C. H. Heathcock, *Org. Lett.*, 1999, **1**, 1315–1317.

30 (*a*) Y. Asakawa, A. Yamamura, T. Waki and T. Takemoto, *Phytochemistry*, 1980, **19**, 603–607; (*b*) M. Tori, K. Nakashima, M. Toyota and Y. Asakawa, *Tetrahedron Lett.*, 1993, **34**, 3751–3752.

31 (*a*) J. G. Luis, E. H. Lahlou and L. S. Andrés, *Tetrahedron*, 1996, **52**, 12309–12312; (*b*) J. Yang, S. X. Huang and Q. S. Zhao, *J. Phys. Chem. A*, 2008, **112**, 12132–12139.

32 (*a*) T. H. Al-Tel, M. H. A. Zarga, S. S. Sabri, M. Feroz, N. Fatima, Z. Shah and Atta-Ur-Rahman, *Phytochemistry*, 1991, **30**, 3081–3085; (*b*) A. V. Kalinin and V. Snieckus, *Tetrahedron Lett.*, 1998, **39**, 4999–5002.

33 (*a*) I. H. Hardt, P. R. Jensen and W. Fenical, *Tetrahedron Lett.*, 2000, **41**, 2073–2076; (*b*) J. A. Kalaitzis, Y. Hamano, G. Nilsen and B. S. Moore, *Org. Lett.*, 2003, **5**, 4449–4452.

34 (*a*) E. Sakuno, K. Yabe, T. Hamasaki and H. Nakajima, *J. Nat. Prod.*, 2000, **63**, 1677–1678; (*b*) P. Wipf and A. D. Kerekes, *J. Nat. Prod.*, 2003, **66**, 716–718.

35 (*a*) R. Suemitsu, K. Ohnishi, M. Horiuchi, A. Kitaguchi and K. Odamura, *Phytochemistry*, 1992, **31**, 2325–2326; (*b*) M. Horiuchi, T. Maoka, N. Iwase and K. Ohnishi, *J. Nat. Prod.*, 2002, **65**, 1204–1205; (*c*) I. Cornella and T. R. Kelly, *J. Org. Chem.*, 2004, **69**, 2191–2193.

36 (*a*) A. Buske, S. Busemann, J. Mühlbacher, J. Schmidt, A. Porzel, G. Bringmann and G. Adam, *Tetrahedron*, 1999, **55**, 1079–1086; (*b*) J. Bringmann, H. Schlauer, H. Rischer, J. Wohlfarth, J. Mühlbacher, A. Buske, A. Porzel, J. Schmidt and G. Adam, *Tetrahedron*, 2000, **56**, 3691–3695.

37 (*a*) K. Suzuki, S. Yahara, M. Kazutomo and M. Uyeda, *J. Nat. Prod.*, 2001, **64**, 204–207; (*b*) S. Vankateswarlu, G. K. Panchagnula, M. B. Guraiah and G. V. Subbaraju, *Tetrahedron*, 2005, **61**, 3013–3017.

38 (*a*) K. Suenaga, S. Aoyama, W. Xi, H. Arimoto, K. Yamaguchi, K. Yamada, T. Tsuji, A. Yamada and D. Uemura, *Heterocycles*, 2000, **52**, 1033–1036; (*b*) M. Kita, R. Miwa, T. Widianti, Y. Ozaki, S. Aoyama, K. Yamada and D. Uemura, *Tetrahedron Lett.*, 2007, **48**, 8628–8631.

39 (*a*) J. Hiort, K. Maksimenka, M. Reichert, S. Perovi-Ottstadt, W. H. Lin, V. Wray, K. Steube, K. Schaumann, H. Weber, P. Proksch, R. Ebel, W. E. G. Müller and G. Bringmann, *J. Nat. Prod.*, 2004, **67**, 1532–1543; (*b*) G. Schlingmann, T. Taniguchi, H. He, R. Bigelis, H. Y. Yang, F. E. Koehn, G. T. Carter and N. Berova, *J. Nat. Prod.*, 2007, **70**, 1180–1187; (*c*) Y. H. Ye, H. L. Zhu, Y. C. Song, J. Y. Liu and R. X. Tan, *J. Nat. Prod.*, 2005, **68**, 1106–1108.

40 (*a*) H. Morita, Y. Sato, K. L. Chan, C. Y. Choo, H. Itokawa, K. Takeya and J. Kobayashi, *J. Nat. Prod.*, 2000, **63**, 1707–1708; (*b*) C. Timmons and P. Wipf, *J. Org. Chem.*, 2008, **73**, 9168–9170.

41 (*a*) I. Kubo, S. P. Tanis, Y. W. Lee, I. Miura, K. Nakanishi and A. Chapya, *Heterocycles*, 1976, **5**, 485–498; (*b*) M. S. Rajab, J. K. Rugutt, F. R. Fronczek and N. H. Fischer, *J. Nat. Prod.*, 1997, **60**, 822–825.

42 (*a*) S. Jain, S. Laphookhieo, Z. Shi, L. W. Fu, S. Akiyama, Z. S. Chen, D. T. A. Youssef, R. W. M. von Soest and K. A. El

Sayed, *J. Nat. Prod.*, 2007, **70**, 928–931; (*b*) S. Jain, I. Abraham, P. Carvalho, Y. H. Kuang, L. Shaala, D. T. A. Youssef, M. A. Avery, Z. S. Chen and K. A. El Sayed, *J. Nat. Prod.*, 2009, **72**, 1291–1298.

43 (*a*) P. Ralifo and P. Crews, *J. Org. Chem.*, 2004, **69**, 9025–9029; (*b*) K. N. White, T. Amagata, A. G. Oliver, K. Tenney, P. J. Wenzel and P. Crews, *J. Org. Chem.*, 2008, **73**, 8719–8722.

44 A. San Feliciano, M. Medarde, J. M. Miguel del Corral, A. Aramburu, M. Gordaliza and A. F. Barrero, *Tetrahedron Lett.*, 1989, **30**, 2851–2854.

45 (*a*) T. C. Fleischer, R. D. Waigh and P. G. Waterman, *J. Nat. Prod.*, 1997, **60**, 1054–1056; (*b*) K. I. Booker-Milburn, H. Jankins, J. P. H. Charmant and P. Mohr, *Org. Lett.*, 2003, **5**, 3309–3312.

46 (*a*) S. N. Kazmi, Z. Ahmed, W. Ahmed and A. Malik, *Heterocycles*, 1989, **29**, 1901–1906; (*b*) D. L. Comins, X. Zheng and R. R. Goehring, *Org. Lett.*, 2002, **4**, 1611–1613.

47 (*a*) Z. Huang, Y. Dan, Y. Huang, L. Lin, T. Li, W. Ye and X. Wei, *J. Nat. Prod.*, 2004, **67**, 2121–2123; (*b*) G. Mehta and K. Pallavi, *Tetrahedron Lett.*, 2006, **47**, 8355–8360.

48 (*a*) A. Romo de Vivar, D. A. Nieto, R. Gaviño and A. L. C. Pérez, *Phytochemistry*, 1995, **40**, 167–170; (*b*) G. Mehta, S. K. Murthy and J. D. Umarye, *Tetrahedron Lett.*, 2002, **43**, 8301–8305.

49 (*a*) M. F. Rodríguez Brasco, A. M. Seldes and J. A. Palermo, *Org. Lett.*, 2001, **3**, 1415–1417; (*b*) K. Inanaga, K. Takasu and M. Ihara, *J. Am. Chem. Soc.*, 2004, **126**, 1352–1353.

50 F. Saito, K. Kuramochi, A. Nakazaki, Y. Mizushina, F. Sugawara and S. Kobayashi, *Eur. J. Org. Chem.*, 2006, 4796–4799.