This item is the archived peer-reviewed author-version of:

Model-averaging of ab initio spectra for the absolute configuration assignment via vibrational circular dichroism

Reference:

Monaco Guglielmo, Aquino Francesco, Zanasi Riccardo, Herrebout Wouter, Bultinck Patrick, Massa Antonio.- Model-averaging of ab initio spectra for the absolute configuration assignment via vibrational circular dichroism

Physical chemistry, chemical physics / Royal Society of Chemistry [London] - ISSN 1463-9076 - 19:41(2017), p. 28028-28036

Full text (Publisher's DOI): <https://doi.org/10.1039/C7CP05358D>

To cite this reference: <http://hdl.handle.net/10067/1455450151162165141>

Cite this: DOI: 10.1039/xxxxxxxxxx

Model-Averaging of Ab Initio Spectra for the Absolute Configuration Assignment via Vibrational Circular Dichroism[†]

Guglielmo Monaco,^{*a} Francesco Aquino,^a Riccardo Zanasi,^a Wouter Herrebout,^b Patrick Bultinck,^c and Antonio Massa^a

Received Date

Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Vibrational absorption and vibrational circular dichroism spectra (VA and VCD) have been recorded for two hybrid isoindolinone-phthalide stereoisomers dissolved in CDCl₃. Density-functional calculations have been performed to determine their absolute configuration. The comparison of calculated and measured values has been performed using several goodness-of-fit indicators, and also introducing a model-averaging technique, taking into account the variation of calculated spectra with the details of the computational method. The model-averaging technique, preliminarily tested on two VCD spectra already assigned to two diastereomers of tadalafil, gives higher credibility to the ab initio calculations, and should be useful for other molecules with high flexibility and/or more than one stereogenic center.

1 Introduction

The determination of the absolute configuration (AC) is a crucial step in the assessment of pharmacological and toxicological properties of novel drugs.¹ Among the many techniques available for that task, Vibrational Circular Dichroism (VCD) in the mid infrared (IR) region is ever more recognized as a convenient technique; it leads to many signed signals, which, coupled with *ab initio* computations, allow discriminating ACs in an efficient and competitive way.^{2–4} Despite the richness of data, in the presence of many low-energy conformers and/or more than one stereocenter, selection of the best-matching computed spectrum can still require considerable expertise. In fact, not all VCD bands have the same reliability: some of them are non-robust, which means that they can undergo large changes (even of sign) upon modest variations in the computational or experimental method used to obtain them.⁵ Moreover, some patterns, even if well reproduced and easily distinguishable visually, appear shifted on the frequency axis, and thus their acceptable reproduction is hardly taken into account by standard goodness-of-fit indicators (GOFIs).⁶ The iden-

Scheme 1

tification of quantitative and reliable GOFIs has been subject of considerable attention.^{5–10} Most of the methods proposed prescribe some modification of calculated and/or experimental data before the computation of some GOFI. The modifications are intended to circumvent problems with the quality of the calculation and/or with the non-robustness of the bands.

Here, we will propose a new strategy to modify the computed spectra, in order to take into account that no ab initio calculation can be expected to treat all effects perfectly, e.g., solvation, thermal effects, basis set truncation errors etc. Therefore, it is appropriate to endow the calculated values with uncertainties which do not depend on the experimental measures. The strategy will be tested on the spectra of two diastereomers of tadalafil **1**:¹¹ (6*R*,12*aR*)-**1** and (6*S*,12*aR*)-**1**, labeled hereafter (*R,R*)-**1** and (*S,R*)-**1**, for brevity. Moreover, the new strategy will be applied for the AC assignment via VCD spectroscopy of the heterocyclic hybrids isoindolinone-phthalide **2** and **3**, bearing two

^a Department of Chemistry and Biology "A. Zambelli", University of Salerno, Via G. Paolo II, 132, Fisciano 84084, Italy Fax: +39 (0)89 969570; Tel: +39 (0)89 969603; E-mail: gmonaco@unisa.it

^b Department of Chemistry, University of Antwerp, Antwerp, Belgium.

^c Department of Inorganic and Physical Chemistry, Ghent University, Belgium.

[†] Electronic Supplementary Information (ESI) available: Selected values from the conformational analyses, graphical and tabular results of fits of VA and VCD spectra as sum of Lorentzians, with totally free parameters or with parameters taken from ab initio calculations, tables of GOFIs for the plain and MA ab initio models. See DOI: 10.1039/b000000x/

contiguous tertiary and quaternary stereocenters. The results obtained will be compared to those based on statistical methods previously published. Compounds **2** and **3**, recently synthesized under organocatalytic conditions in high yields and with high diastereo- and enantioselectivity,¹² are of high interest because new compounds formed by covalent linking of two known pharmacophores, like the isoindolinone and the isosteric phthalidic heterocyclic ring, can show enhanced therapeutic activity over the isolated parents or even new biological activities.^{13–15} The relative configuration of the synthesized stereoisomer of **2** has been previously determined by NMR spectroscopy,¹² and is (1*S**, 3'*R**), i.e. we have either (1*S*, 3'*R*)-**2** or (1*R*, 3'*S*)-**2**, to be shortened as (1*S**, 3'*R**)-**2** in the following, while no stereochemical information is available for synthesized stereoisomer of **3**, which will be then referred to as (ξ, ξ')-**3**.[‡] The set of four molecules studied covers then 2 cases of known AC, one case of known relative configuration, and a case of totally unknown configuration.

2 Selection of Models for VA and VCD spectra

2.1 Models of the Spectra

Quantum chemical calculations are a valuable tool for the interpretation of vibrational absorption (VA) spectra, and even more for the VCD spectra, where they are necessary to establish the correct AC.[§] With use of fast Fourier transform spectrometers, a spectrum consists in a set of n values $y_i(\tilde{\nu}_i)$ which are functions of absorption at wavenumbers $\tilde{\nu}_i$, where $i = 1, \dots, n$. For VA spectra, of the possible choices for the y_i (transmittance, absorption, molar extinction coefficient), we will take as $y_i(\tilde{\nu}_i)$ the molar extinction coefficients ϵ_i . Analogously, for VCD spectra we will take as $y_i(\tilde{\nu}_i)$ the differential molar extinction coefficients $\Delta\epsilon_i$. These choices imply that the measured (differential) absorbances have been corrected for nominal molar concentration and path-length. No further scaling of the y_i values will be applied in this paper. Pioneering applications of quantum chemical calculations have been performed by i) fitting the n experimental values y_i by the sum of L Lorentzians, with $3L$ adjustable parameters A_k , $\tilde{\nu}_{0,k}$ and γ_k , which are the area, the central wavenumber and the half-width at half-maximum of the isolated k -th Lorentzian; for the typically sharp peaks of the mid-IR region ($\gamma_k \ll \nu_{0,k}$), the model can be written as⁴

$$y_{i,\text{SumLor}}(\tilde{\nu}_i) = \sum_{k=1}^L \frac{A_k}{\pi} \frac{\gamma_k}{\gamma_k^2 + (\tilde{\nu}_i - \tilde{\nu}_{0,k})^2},$$

ii) computing by quantum mechanical methods the areas and harmonic central wavenumbers expected for the conformers of

the stereoisomer $\mathcal{M}$, and weighing these values by the thermal population of the conformers; iii) comparing numerical tables of fitted central wavenumbers and Lorentzian areas with those computed quantum mechanically, iv) computing the spectra $y_{i,\mathcal{M}}(\tilde{\nu}_i; \{A_k, \tilde{\nu}_{0,k}^{\text{harm}}, T, \gamma\})$ by assuming a single γ value for all Lorentzian peaks, whose areas have been weighted by the conformer populations at temperature T ; and eventually v) comparing visually the experimental spectra $y_i(\tilde{\nu}_i)$ with those computed, $y_{i,\mathcal{M}}(\tilde{\nu}_i; \{A_k, \tilde{\nu}_{0,k}^{\text{harm}}, \gamma, T\})$.³ The visual comparison of the spectra can be significantly improved introducing a frequency scaling factor λ to correct overall the computed harmonic frequencies for anharmonicity: $\tilde{\nu}_{0,k}^{\text{corrected}} = \lambda \tilde{\nu}_{0,k}^{\text{harm}}$. Reasonable simulations of VA or VCD spectra can be obtained adding to the quantum chemical values of areas and central wavenumbers just the two parameters γ and λ . If γ and λ parameters are assumed to be the same in VA and VCD spectra, and they are estimated from the VA spectra, VCD spectra can then be simulated with no adjustable parameter, in any case with a number of parameters which is negligible with respect to the number of measurements. In the following, we will refer to the model spectra obtained by thermal averages of quantum chemical spectra (with the addition of 2 fitted or fixed parameters γ and λ) as plain ab initio spectra $y_{i,\mathcal{M}}(\tilde{\nu}_i; \mathbf{P}) \equiv y_{i,\mathcal{M}}(\tilde{\nu}_i; \{A_k, \tilde{\nu}_{0,k}^{\text{harm}}, \gamma, T, \lambda\})$, where we have introduced a vector $\mathbf{P}$ to collect all model parameters. It will be useful to consider vector $\mathbf{P}$ as the union of two vectors $\mathbf{p}$ and $\mathbf{k}$ of p and k elements, containing parameters which are either changed during the data analysis or held fixed, respectively. As previously said, in case of more than two competing ACs, i.e. a molecule with more than one chiral center, selection of the best model among the plain ab initio spectra can be a delicate step, requiring visual scrutiny and band-by-band assignment. Indeed, most plain ab initio spectra are statistically bad models, in the sense that residuals are significantly larger than experimental errors and they have non-normal distributions. In this situation GOFs should be purposely chosen, as the standard ones will generally perform poorly.[¶]

2.2 The GOFs

The simplest possible GOF is likely the sum of squares for error (SSE)¹⁷

$$\text{SSE} := \sum_{i=1}^n (y_i(\tilde{\nu}_i) - y_{i,\mathcal{M}}(\tilde{\nu}_i; \mathbf{P}))^2, \quad (1)$$

which has indeed been used in the selection of best-fitting models in vibrational Raman optical activity spectra.¹⁸ Use of eqn (1) must take into account that its value depends on the scale of the individual values y_i , and also on the number n of elements in the summation. A widely used GOF, which has not the above limita-

[‡] In the recommended IUPAC notation¹⁶ an unknown configuration of a stereogenic center is indicated by the Greek letter ξ. Here we will indicate the configuration of the second stereogenic center by ξ' to stress that it is unknown and bears no relationship with the configuration of the first stereogenic center.

[§] VA spectra are most frequently named by their wavelength, i.e. IR spectra. This practice originates from the fact that VCD spectra have become known after VA spectra. It could be appropriate to distinguish the common name of the wavelength region from the physical phenomenon (absorption or dichroism) causing the spectrum. A recent reference textbook has just opted for this choice.⁴

[¶] All GOFs in this paper are functions of the stereoisomeric model $\mathcal{M}$. To simplify notation, we will add a subscript to the GOF, e.g. $\text{MSE}_{\mathcal{M}}$, only when the identification of the stereoisomeric model can be non-obvious.

tion, is the reduced chi-squared¹⁹

$$\chi_q^2 := \frac{1}{q} \sum_{i=1}^n \frac{(y_i(\tilde{\mathbf{v}}_i) - y_{i,\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{P}))^2}{\sigma_i^2} = \frac{1}{q} \sum_{i=1}^n |z_i(\tilde{\mathbf{v}}_i) - z_{i,\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{P})|^2 =: \text{MSE}, \quad (2)$$

which coincides with the mean squared error MSE¹⁷ in terms of error-normalized dimensionless values $z_i = y_i/\sigma_i$ and $z_{i,\mathcal{M}} = y_{i,\mathcal{M}}/\sigma_i$, where σ_i is the standard deviation of the i -th difference $y_i(\tilde{\mathbf{v}}_i) - y_{i,\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{P})$. Typically, σ_i is considered as an indication of errors due to the experimental setup only. However, in the following we will also consider the role of errors caused by the computational model, that is to say, what is the impact of the basis set choice, functional and more generally all parameters $\mathbf{p}$ that influence the generation of the model spectrum. In eqn (2), q is the number of degrees of freedom. For models which are linear in p independent variables, with the p slopes determined from the experimental data by least squares, $q = n - p$ (in the limit case of a parameter-free model, $q = n$), and the expected value for MSE is 1. Values of MSE either significantly larger or smaller than 1 are considered unacceptable, indicating a bad model and/or the adoption of incorrect variances.¹⁹

Although selection of good models can be performed considering their MSEs (typically via F -tests), this is not the case for bad models, which are strongly affected by outliers; non-normal residuals prevent the use of standard statistical tests and the numerical values of MSEs can be uninformative. We will now consider as possible workarounds the use of more robust GOFs, or the consideration of a different statistic for residuals, which, in its simplest form, corresponds to different variances σ_i^2 .

A more robust GOF is the mean absolute error (MAE), which can be written, in terms of error-normalized values, as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |z_i - z_{i,\mathcal{M}}|. \quad (3)$$

Comparison MAE and MSE should, of course, consider the different powers in the sums: MAE should best be compared with RMSE = $\sqrt{\text{MSE}}$. In presence of important contributions from outliers, one expects $\text{MAE} \ll \text{RMSE}$, while for residuals extracted from a normal distribution one expects similar values of MAE and RMSE, indeed in that case $\text{MAE} = \sqrt{\frac{2\text{MSE}}{\pi}} = 0.7979 \text{ RMSE}$.²⁰ Another possibility to cope with outliers is the cosine-similarity, which is known in the field of information retrieval as a criterion to rank a document for a query.²¹ Considering the values z_i and $z_{i,\mathcal{M}}$ as elements of n -dimensional vectors $\mathbf{z}$ and $\mathbf{z}_{\mathcal{M}}$, eqn (2) is basically an Euclidean distance squared, and can be rewritten as

$$\text{MSE} = \frac{1}{v} \left(\|\mathbf{z}\|^2 + \|\mathbf{z}_{\mathcal{M}}\|^2 - 2\|\mathbf{z}\| \|\mathbf{z}_{\mathcal{M}}\| \cos \theta \right), \quad (4)$$

which shows that a low value of MSE is associated with similar norms of vectors $\mathbf{z}$ and $\mathbf{z}_{\mathcal{M}}$, and with a large cosine-similarity $\cos \theta$ (a small angle θ between the two vectors). As a trivial example of the usefulness of $\cos \theta$, as compared to the MSE, one can consider a simple uniform scaling of experimental data: in this case, the MSE changes significantly, while the cosine-similarity

$$\cos \theta = \frac{\mathbf{z} \cdot \mathbf{z}_{\mathcal{M}}}{\|\mathbf{z}\| \|\mathbf{z}_{\mathcal{M}}\|} \quad (5)$$

is unchanged.

In the case of VCD spectra, which are signed, besides the individual MSE values, it is interesting to consider the relationship between the MSE values of pairs of enantiomers. The computed spectrum for the enantiomer $\text{ent-}\mathcal{M}$ of a stereoisomer $\mathcal{M}$ is just its negative: $z_{i,\text{ent-}\mathcal{M}} = -z_{i,\mathcal{M}}$. Choosing then among any two enantiomers, the stereoisomer $\mathcal{M}$ which gives the better agreement ($\text{MSE}_{\mathcal{M}} \leq \text{MSE}_{\text{ent-}\mathcal{M}}$), and using eqn (4), we have

$$\text{MSE}_{\text{ent-}\mathcal{M}} - \text{MSE}_{\mathcal{M}} = \frac{4\|\mathbf{z}\| \|\mathbf{z}_{\mathcal{M}}\| \cos \theta_{\mathcal{M}}}{q}, \quad (6)$$

which shows that a large distance between the MSEs of two enantiomers parallels a small θ angle between the vectors $\mathbf{z}$ and $\mathbf{z}_{\mathcal{M}}$, whose elements are the measured and the calculated values. Differently stated, this means that for a pair of diastereomers $\mathcal{M}$ and $\mathcal{M}'$, with cosine similarities $\cos \theta_{\mathcal{M}}$ and $\cos \theta_{\mathcal{M}'}$, the preferred stereoisomer - the one with larger cosine-similarity - will undergo the larger variation of the VCD spectrum upon consideration of its enantiomer (we are actually assuming substantial equality of the norms of the computed spectra $\|\mathbf{z}_{\mathcal{M}}\|$ and $\|\mathbf{z}_{\mathcal{M}'}\|$). To analyze the GOF of VCD spectra, two modifications of the cosine-similarity have been proposed in literature: the similarity index (SI) and the enantiomeric similarity index (ESI).⁶ Even when applied to signed spectra, the latter indices turn out bounded in the $[0, 1]$ interval rather than in the $[-1, 1]$ interval, as happens to the simpler cosine-similarity. The mathematical definitions of SI and ESI can be found in literature,^{6,9} and are not repeated here (for ESI, the original definition of ref. 6 has one more absolute value with respect to the definition of ref. 9; the original definition will be used in this paper).

2.3 The MA Technique

Coming now to the possibility of considering a different statistic for the residuals, it can be imagined to build a population of residuals $(y_i - y_{i,\mathcal{M}})$, where the model itself has an uncertainty which does not depend on the experimental setup. A possible approach to give an uncertainty to a model $\mathcal{M}$ of a spectrum recognizes the model as a function of a vector $\mathbf{p}$ of p parameters (frequencies, magnetic and electric transition dipole moments, angles between the magnetic and electric transition dipole moments), assumes a distribution $f_j(p_j)$ for each parameter, and then, upon assumption of independent variation of the parameters, computes a model-averaged (MA) spectrum as

$$y_{i,\text{MA-}\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{k}) = \int_{p_1, p_2, \dots, p_p} y_{i,\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{p}; \mathbf{k}) \prod_{j=1}^p f_j(p_j) dp_j, \quad (7)$$

with variance

$$\sigma_{i,\text{MA-}\mathcal{M}}^2(\tilde{\mathbf{v}}_i; \mathbf{k}) = \int_{p_1, p_2, \dots, p_p} [y_{i,\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{p}; \mathbf{k}) - y_{i,\text{MA-}\mathcal{M}}(\tilde{\mathbf{v}}_i; \mathbf{k})]^2 \prod_{j=1}^p f_j(p_j) dp_j. \quad (8)$$

A simple Gaussian form will be assumed in the following for each distribution, i.e.

$$f_j(p_j) = \frac{1}{\sqrt{2\pi}\sigma_{p_j}} \exp\left[-\frac{(p_j - p_{0,j})^2}{2\sigma_{p_j}^2}\right], \quad (9)$$

requiring to estimate one variance $\sigma_{p_j}^2$ for each parameter $p_{0,j}$ in order to compute the MA spectrum, eqn (7), and its error bound obtained from eqn (8).

We are not aware of any previous computation of model-averaged VA or VCD spectra through eqns 7 and 8.

It should be considered that the introduction of a distribution for the central wavenumber $\nu_{0,k}$ corresponds to considering a family of peaks close to the peak of the plain calculation, and thus leads to broadened spectra, just as happens in the neighborhood similarity approach.^{22,23}

Collections of ab initio calculations can be used to estimate the variances of the relevant spectral parameters (see below). Once averages 7 and 8 have been estimated, the above GOFs can be computed using $y_{i,MA-\mathcal{M}}$ in place of $y_{i,\mathcal{M}}$, and

$$\sigma_i^2 = \sigma_{i,\text{exp}}^2 + \sigma_{i,MA-\mathcal{M}}^2, \quad (10)$$

in place of the usual experiment-only variance $\sigma_i^2 = \sigma_{i,\text{exp}}^2$.

3 Methods

3.1 Synthesis

The synthesis of **2** and **3** has been performed as described in Ref. 12.

The two diastereoisomers of tadalafil were received from RK Tech, Hungary, with purity $\geq 99.8\%$ and were used without further purification. (6*R*, 12*aR*)-tadalafil is the active ingredient of Cialis and Adcirca.

3.2 Measurement of Spectra

Vibrational Absorption (VA) and VCD were recorded on a BioTools dual-PEM ChiralIR-2X spectrometer at room temperature. The PEMs were optimized for 1400 cm^{-1} , and a resolution of 4 cm^{-1} was used throughout. For all experiments on **2** and **3**, solutions were prepared by dissolving 4.2 mg in 135 μL of CDCl_3 . All spectra were recorded using BaF_2 windows and a spacer of 100 μm . For both samples, the solution spectra were averaged over 20 000 scans. Baseline corrections were introduced by subtracting the spectra of the solvent. Acquisition of the spectra of tadalafil **1** has been described elsewhere.¹¹

3.3 Empirical fits of the spectra

According to common practice, we have fitted VA and VCD spectra by sums of Lorentzians, using the non-linear least squares method.³ Plots reported in the Supporting Information (Figs. S2-S5, ESI†) show a good visual agreement of the Lorentzian models $y_{i,\text{SumLor}}$ with the experimental data y_i , as can also be inferred from the cosine-similarities which are hardly different from 1 (Tabs. S4-S5, ESI†). Rounded mean values of γ parameters, also reported in the Tables, are always equal to 4 cm^{-1} for VCD spectra and only somewhat larger for the VA spectra of (*S**,*R**)-**2** and

(ξ, ξ')-**3**. The residuals of the fits, also shown in Figs. S2-S5 in the ESI†, suggest some frequency-dependent trend of the variances of the experimental data $\sigma_{i,\text{exp}}^2$. However, the residuals turn out roughly an order of magnitude smaller than those typically obtained by ab initio methods, and they can be anticipated to play a minor role in model-averaging. Then, we will simply neglect their change with frequency in the spectral range examined, and, assuming small errors on the parameters of the Lorentzian model, we will estimate the single variance of experimental data¹⁹

$$\sigma_{\text{exp}}^2 = \frac{\sum_{i=1}^n (y_i - y_{i,\text{SumLor}})^2}{q}, \quad (11)$$

with $q = n - 3L$. Introduction of the above estimation of variance instead of σ_i^2 in eqn (2) renders useless the RMSE value, which turns out identically equal to 1. However, that estimation has been used instead of σ_i^2 to compute eqn (10), and then the RMSE values for the plain and model-averaged quantum mechanical models.

3.4 Quantum chemical calculations

All quantum chemical calculations have been performed with Gaussian '09.²⁴

The conformational analysis for **1** has been performed, building starting geometries with different values for the two dihedral angles $\phi_1 \equiv \text{H-C6-C6'-C1'}$ and $\phi_2 \equiv \text{C3'-C4'-O-C7'}$ (Scheme 1). Standard optimization at the level used in ref. 11 (SCRF- $\omega\text{B97XD}/6\text{-}31\text{G}^*$) gave 6 low-energy conformers for (*R,R*)-**1** and 4 low-energy conformers for (*R,S*)-**1**, with good but incomplete overlap with previous results.¹¹ Using tight criteria of optimization and ultrafine grids, 2 more low-energy minima could be found (Tab. S1, ESI†). In the conformers found the dihedral angles ϕ_1 and ϕ_2 are close to 180° or 0° but have a small positive or negative deviation from those values. Thus the possible rotational states for each of the dihedral angles are T^+ , T^- , C^+ , C^- .²⁵ Angle ϕ_2 occurs only as C^- and C^+ , angle ϕ_1 has all four possibilities in (*R,R*)-**1**, but only the two possibilities T^- and C^+ in (*R,S*)-**1**. Consistently, $2 \times 2 = 4$ conformers have been found for (*R,S*)-**1**, but only 7 conformers have been found for (*R,R*)-**1**; several attempts to find the missing conformation $\text{C}^- \text{C}^+$ for (*R,R*)-**1** failed. In any case, the potential is very flat, and the conformers are almost degenerate. The 7+4 conformers found in this work with the tight criteria of convergence have been used in the rest of the analysis. VA and VCD spectra have been computed at the lowest level considered in ref. 11: SCRF- $\omega\text{B97XD}/6\text{-}31\text{G}^*$, with a dielectric constant for chloroform of $\epsilon = 4.71$.²⁶

Aiming at showing that MA can effectively help in circumventing inadequacies of the computational protocol, the computations of the VCD spectra **2** and **3** have been performed with the basic approach of a standard conformational search, followed by a B3LYP/6-31G* optimization of the conformers, and then by VA and VCD calculations at the same B3LYP/6-31G* level. 18 conformers have been previously found by a Spartan '04²⁷ MMFF conformational search for (*S,R*)-**2** and (*S,S*)-**2**, and they have been used to determine the relative configuration (*S**,*R**)-**2** by NMR spectroscopy.¹² The conformers differ in the values of three

dihedral angles ϕ_1 , ϕ_3 and ϕ_4 : one on the ethoxyl chain, and two flanking the quaternary carbon (Fig. S1, ESI†). As the bromine atom in (S,R)-**3**, and (S,S)-**3** is far from the atoms forming the just cited dihedral angles, the same conformers are expected for (S,R)-**3**, and (S,S)-**3**. We have therefore optimized at B3LYP/6-31G* level the conformations for all four species, using the 18 dihedral angles found in previous study on (S,R)-**2** and (S,S)-**2**. Relevant dihedral angles obtained by DFT optimization are collected in Tables S2 and S3 in the ESI†. All relevant dihedrals of **2** remained almost unchanged upon bromination and successive optimization. Also, the energy distribution of the conformers remains pretty much the same upon bromination, which is what we expected on the ground that the bromine atom is far from the flexible portion of the molecule.

The spectra of the conformers must be weighted by their populations, which are determined by their relative energies according to the Boltzmann distribution. Seminal studies have either evaluated the relative energies neglecting nuclear motion (i.e. considering only the relative energies on the potential energy hypersurface), or considering it within the rigid-rotor harmonic-oscillator approximation.³ Although the latter approach (to be called GP) is conceptually superior to the former (to be called EP),²⁸ the errors introduced by the approximate thermal corrections can be important, and EP calculations are sometimes preferred in literature. As an intermediate choice, populations can be derived from standard enthalpies (HP populations), thus neglecting only the entropic corrections.^{29–31} We have checked the relevance of the choice of the thermal averaging scheme, adopting all 3 weightings EP, HP and GP.

Frequency scaling factors λ have been determined optimizing the agreement between the experimental VA spectra and those computed by EP weightings and the single γ value of 4 cm^{-1} (irrelevant variations have been found for the HP and GP weighting schemes). Optimal values are reported in Tab. S6 in the ESI†.

3.5 Computation of model-averaged spectra and variances

In order to obtain the MSE with model-averaged computed spectra (Eqs. 2, 7 and 10), an estimate of the uncertainties of the *ab initio* calculations is needed. To this end we have considered the 7+4 conformers of the two diastereomers of **1**, and we have optimized them at 8 different level of theory, which have been used in recent works on VCD spectroscopy of small organic molecules: B3LYP/TZ2P,²⁸ B3LYP/cc-PVTZ,^{6,28,32} B3LYP/6-31G*,¹¹ B3PW91/TZ2P,²⁸ B3PW91/cc-PVTZ,^{28,32} B97D/TZ2P,³³ ω -B97XD/6-31G*,¹¹ ω -B97XD/6-311++G**,³⁴ using in all cases the PCM model,³⁵ to model the effect of chloroform.

In many cases (R,R)-**1** conformers found at the SCRF- ω B97XD/6-31G* level did not correspond to minima on the potential energy surfaces at different levels, as already reported.¹¹ For each conformer we focused on the 800-1800 cm^{-1} spectral window and we estimated mean and standard deviation of spectral parameters, such as central wavenumbers $\nu_{0,k}$, electric dipole transition moments μ_k , magnetic dipole transition moments m_k computed setting as origin the center of mass, and angles ξ_k between the

μ_k and m_k vectors. Standard deviations are of the following orders of magnitude: $\sigma(\tilde{\nu}_0) \sim 20\text{ cm}^{-1}$, $\sigma(\mu) \sim 4 \cdot 10^{-20}\text{ esu cm}$, $\sigma(m) \sim 3 \cdot 10^{-24}\text{ esu cm}$, $\sigma(\xi) \sim 20^\circ$. These values can be expected to be overestimates, because the sequence of the signals in the infrared is not necessarily preserved changing the level of computation. Indeed, it has been reported that, restricting the analysis to normal modes that are roughly preserved upon change of the level of calculation, $\sigma(\xi) \sim 10^\circ$.⁵ We have therefore used halved errors for generating the MA spectra, Eqs. 7 and 8: $\sigma(\tilde{\nu}_0) = 10\text{ cm}^{-1}$, $\sigma(\mu) = 2 \cdot 10^{-20}\text{ esu cm}$, $\sigma(m) = 1.5 \cdot 10^{-24}\text{ esu cm}$, $\sigma(\xi) = 10^\circ$. This choice turns out capable of obtaining values of MA-MSE close to one (Tables 1, S7, S8, S9, ESI†). A further source of error lies in the population of conformers. Using the set of 8 different calculation methods performed on the 7+4 conformers of **3**, we have estimated standard deviations of energies, enthalpies and free enthalpies as 0.05, 0.10 and 0.5 kcal mol^{-1} , showing that the estimates of free enthalpies are affected by a larger error. These errors have been used in calculating MA-EP, MA-HP and MA-GP spectra.

4 Results

Comparison of experimental and computed VCD spectra in terms of the 5 GOFIs presented above (RMSE, MAE, cosine-similarity, SI, ESI) should allow selection of the correct stereoisomer. To avoid as much as possible a selection based on calculated values which are not reliable, the procedure of model-averaging has been introduced. Thus in addition to the 3 plain *ab initio* spectra EP, HP and GP, obtained with the thermal averages discussed above, we have also considered the 3 model-averaged spectra MA-EP, MA-HP and MA-GP, obtained with the same thermal averages, for a total of $2 \times 3 \times 5 = 30$ combinations of method/GOFI to select the best AC. We have first tested this strategy on the spectra of (R,R)-**1** and (S,R)-**1**.¹¹

4.1 Benchmark on (R,R)-**1** and (S,R)-**1**

The 311 spectral data contained in the 1000-1600 cm^{-1} window, previously used in ref. 11 for both VA and VCD spectra of these two stereoisomers of **1**, have been compared with those computed for all 4 possible stereoisomers. Inspection of graphical comparisons between experimental and calculated values (Figs S6 and S7) numerical results for the GOFIs (Table S7 and S8, ESI†) shows that each of the 30 combinations method/GOFI leads to the assignment of the correct AC. Therefore, in this case, the selection performed with the plain *ab initio* spectra remains unaltered upon model averaging: the assignment is not dependent on uncertain features of the calculation. If we consider the VA spectrum, instead, (Figures S10-S11 and Tables S7-S8, ESI†) RMSE and MAE are useless for model selection using plain *ab initio* spectra, while upon model averaging, both these GOFIs point to the correct relative configuration. Interestingly, the cosine-similarity always suggests the correct relative configurations, even in absence of model averaging.

4.2 Hybrid (S^*, R^*)-2

The 311 spectral data contained in the 950-1550 cm^{-1} window, for both VA and VCD spectra of this stereoisomer of **2**, have been compared with those computed for all 4 possible stereoisomers. Graphical comparisons between experimental and calculated values are given in Fig. 1. Numerical results for the GOFIs are given in Tab. 1.

Table 1 GOFIs for the VCD spectrum of (S^*, R^*)-**2** in chloroform, calculated using the four possible ACs, conformer populations obtained by energy, enthalpy or free-enthalpy, and either plain B3LYP/6-31G* calculations or their model-averaged version. To facilitate the reading of the table, for each method and for each GOFI, the extremal GOFI value which hints at the best-fitting AC, has been underlined.

Method	AC	RMSE	MAE	$\cos \theta$	SI	ESI
EP	(<i>R,R</i>)	16.40	9.33	0.34	0.49	0.43
	(<i>S,S</i>)	20.42	11.74	-0.34	0.17	0.14
	(<i>S,R</i>)	<u>15.90</u>	<u>8.58</u>	<u>0.53</u>	<u>0.64</u>	<u>0.61</u>
	(<i>R,S</i>)	22.76	12.44	-0.53	0.11	0.05
HP	(<i>R,R</i>)	15.91	9.16	0.33	0.48	0.42
	(<i>S,S</i>)	19.88	11.54	-0.33	0.17	0.13
	(<i>S,R</i>)	<u>15.70</u>	<u>8.55</u>	<u>0.54</u>	<u>0.65</u>	<u>0.62</u>
	(<i>R,S</i>)	22.68	12.44	-0.54	0.11	0.06
GP	(<i>R,R</i>)	<u>15.49</u>	<u>9.05</u>	0.31	0.45	0.39
	(<i>S,S</i>)	19.15	11.23	-0.31	0.16	0.12
	(<i>S,R</i>)	17.92	9.25	<u>0.55</u>	<u>0.65</u>	<u>0.62</u>
	(<i>R,S</i>)	25.16	13.40	-0.55	0.10	0.07
MA-EP	(<i>R,R</i>)	1.34	0.98	0.07	0.33	0.12
	(<i>S,S</i>)	1.19	1.17	-0.07	0.38	0.33
	(<i>S,R</i>)	0.96	<u>0.81</u>	<u>0.46</u>	<u>0.57</u>	<u>0.56</u>
	(<i>R,S</i>)	1.62	1.28	-0.46	0.08	0.02
MA-HP	(<i>R,R</i>)	1.35	0.99	0.07	0.33	0.12
	(<i>S,S</i>)	1.22	1.19	-0.07	0.38	0.33
	(<i>S,R</i>)	<u>0.96</u>	<u>0.82</u>	<u>0.47</u>	<u>0.58</u>	<u>0.57</u>
	(<i>R,S</i>)	1.65	1.30	-0.47	0.08	0.02
MA-GP	(<i>R,R</i>)	1.12	0.83	0.09	0.33	0.13
	(<i>S,S</i>)	1.03	1.00	-0.09	0.36	0.32
	(<i>S,R</i>)	<u>0.83</u>	<u>0.71</u>	<u>0.51</u>	<u>0.60</u>	<u>0.59</u>
	(<i>R,S</i>)	1.47	1.17	-0.51	0.08	0.01

As can be seen from Tab. 1, 28 combinations of method/GOFI lead to select the AC (*S,R*) $\equiv$ (*1S,3'R*), from the VCD spectrum; the exceptions are the GP calculation of RMSE and MAE, which points to an (*R,R*) AC. This finding could be an indication of the lower accuracy of free-enthalpy calculations, as advocated in literature.²⁹ It is particularly interesting to note that the two discordant conclusions disappear upon model averaging. The performance of the cosine-similarity is pretty much similar to that of its modifications SI and ESI.

GOFIs calculated for the VA spectrum, reported in Tab. S12 of the ESI†, are all close each other and no reliable determination of the relative configuration can be performed from the VA spectrum, at least at this simple level of calculation.

4.3 Hybrid (ξ, ξ')-3

The 311 spectral data contained in the 950-1550 cm^{-1} window, for both VA and VCD spectra of this stereoisomer of **3**, have been compared with those computed for all 4 possible stereoisomers.

Graphical comparisons between experimental and calculated values (Fig. S9) and numerical results for the GOFIs (Tab. S9) can be found in the ESI†.

From the analysis of the VCD spectrum, in absence of model-averaging, the preferred AC is (*S,R*) for all GOFIs, but the RMSE, which points to the (*R,R*) configuration. Upon MA, all GOFIs, RMSE included, point to the (*S,R*) $\equiv$ (*1S,3'R*) configuration, which can be then safely trusted. As just reported for (S^*, R^*)-**2**, GOFIs calculated for the VA spectrum, reported in Tab. S13† are all close each other and no reliable determination of the relative configuration can be performed from the VA spectrum, at least at this simple level of calculation.

Table 2 Overall results of application of the GOFIs to the VCD spectra studied (computed either with plain or model-averaged ab initio calculations as described in the text). Assignments which are not unique using the three population averages EP, HP and GP are marked with a ‡ sign.

		VCD		
		RMSE	MAE	$\cos \theta$
(<i>R,R</i>)- 1	plain	(<i>R,R</i>)	(<i>R,R</i>)	(<i>R,R</i>)
	MA	(<i>R,R</i>)	(<i>R,R</i>)	(<i>R,R</i>)
(<i>S,R</i>)- 1	plain	(<i>S,R</i>)	(<i>S,R</i>)	(<i>S,R</i>)
	MA	(<i>S,R</i>)	(<i>S,R</i>)	(<i>S,R</i>)
(S^*, R^*)- 2	plain	(<i>S,R</i>)‡	(<i>S,R</i>)‡	(<i>S,R</i>)
	MA	(<i>S,R</i>)	(<i>S,R</i>)	(<i>S,R</i>)
(ξ, ξ')- 3	plain	(<i>R,R</i>)	(<i>S,R</i>)	(<i>S,R</i>)
	MA	(<i>S,R</i>)	(<i>S,R</i>)	(<i>S,R</i>)

4.4 Discussion

The results reported above, resumed in Table 2, show that the AC has been determined in a confident manner via VCD, using 5 GOFIs. The use of MA methods gives more credibility to the assignment, which does not depend on uncertainties of parameters of the computation.

The assignment of relative configuration via VA spectra has been more problematic. It could be performed using model-averaging for (*R,R*)-**1** and (*S,R*)-**1**, but not for the other two VA spectra.

Inspection of Fig. 1 leads to conclude that some of the vibrations should be localized on single stereogenic centers. The spectral pattern in the 1250-1400 cm^{-1} window (attention can be given to the strong negative and positive bands at around 1250 and 1400 cm^{-1}) change sign when the chirality of the quaternary stereogenic center 1 changes from (*S*) to (*R*); on the other hand, the bands at 1025 cm^{-1} (negative), 1100 cm^{-1} (positive) change sign when the chirality of the tertiary stereogenic center 3' changes from (*R*) to (*S*). A similar situation occurs for **3** (Fig. S9, ESI†). Analysis of the normal modes of the lowest energy conformer of **2**, consistently reveals that at 1414, 1418, 1297, 1285 cm^{-1} there are vibrations localized on the quaternary stereogenic center 1, while at 1042, 1046 and 1142 cm^{-1} there are C-H bending vibrations.

5 Conclusions

Despite their amazing improvement, quantum chemical calculations still do not lead to quantitative agreement with VCD spectra,

Fig. 1 VCD spectrum of (S^* , R^*)-**2** and its fit with plain and model-averaged B3LYP/6-31G* calculations for the 4 possible ACs. The experimental spectrum is displayed as a blue dotted line. The plain calculation using energies for the conformational average as a continuous black line (EP), the model-averaged calculation as a dashed black line (MA-EP), and the prediction error of the MA-EP calculation as a shaded area.

i.e. one that could be checked by a standard goodness-of-fit indicator (GOFI) as the MSE. In many cases, expert corrections of the calculations and expert procedures of data rejection are used in combination with visual inspection to achieve a confident assignment of the absolute configurations (AC) via VCD spectroscopy, especially in case of flexible molecules with more than a single chiral center. In this paper we have explored the use of alternative GOFIs and the possibility to cope with the variability of the ab initio results as statistical errors. The statistical perspective allows to obtain model-averaged (MA) ab initio spectra, which come together with their uncertainties, and should thus allow a more confident assignment, as compared with the plain ab initio spectra normally used.

MA spectra have been computed and compared with the VA and VCD experimental spectra of 4 stereoisomers with 2 chiral centers: (R,R)-**1**, (S,R)-**1**, (S^* , R^*)-**2** and (ξ, ξ')-**3**. Of these species,

2 had known AC, **1** had known relative configuration only, and **1** had no stereochemical information.

Regarding the GOFIs of the VCD spectra, the performance of the modified cosine-similarity indices SI and ESI⁶ agree with the basic cosine-similarity $\cos \theta$.²¹ All cosine-similarities are in agreement with the MAE, but they occasionally disagree with the RMSE (Table S-S9, ESI†). The latter GOFI, as expected, should be used with caution, while selecting statistically poor models, as the ab initio spectra typically are.

As for the different thermal averages EP, HP and GP, there is some indication of a poorer performance of the GP average, which leads to inconsistent results for the VCD spectrum of (S,R)-**2**. Interestingly all thermal averages lead to the same conclusion when coupled with model-averaging.

MA spectra confirm the assignment made with plain ab initio spectra for the benchmark cases (R,R)-**1** and (S,R)-**1**. In the case

of (S^*, R^*)-2 and (ξ, ξ')-3, almost all GOFs point to the (S, R) configuration, and the few discordant ones reveal the behavior of all other GOFs, upon use of MA. We arrive then to the firm conclusion that the AC is (S, R) $\equiv$ ($1S, 3'R$) (Scheme 1), for (S^*, R^*)-2 and also for (ξ, ξ')-3.

In summary, the model-averaging technique promises to be a useful tool to check the credibility of the AC assignments based on ab initio calculations. The method is anticipated to be generally useful in the AC assignment of molecules with many populated conformations and/or with more than 1 stereogenic center, a task that can still be challenging with standard ab initio calculations of VCD spectra.

The present implementation of the method uses independent Gaussian distributions for model parameters, and includes among the parameters both the magnetic dipole transition moments m_k and the angles ξ_k they form with the electric dipole transition moments μ_k . Future implementations should explore the possibility of more flexible probability distributions for the parameters and/or the adoption of different sets of model parameters, such as the use of dissymmetry factors g_k , which are gauge-invariant in addition of being just half in number as compared to the magnetic dipole transition moments m_k and the angles ξ_k considered here.^{9,36} A joint analysis of VCD and VA spectra in the present scheme could also be performed. Eventually, a set of probability distributions and parameter spaces could be identified and used in conjunction with low level calculations (such as the present B3LYP/6-31G* one), whose usefulness could thus be significantly improved by the adoption of model-averaging. Further work in this direction has been planned, and we hope to disclose it in near future.

Acknowledgment

GM thanks CINECA (IsC37_VCDAC2C project) for the availability of computational resources. Financial support from the MIUR (FARB 2014) is gratefully acknowledged.

References

- With regards to isoindolinones, please see: Kanamitsu, N.; Osaki, T.; Itsuji, Y.; Yoshimura, M.; Tsujimoto, H.; Soga, M. *Chem. Pharm. Bull.* 2007, **55**, 1682-1688; Toyooka, K.; Kanamitsu, N.; Yoshimura, M.; Kuriyama, K. S.; Tamura, H.; Takashi, K. S. Isoindoline derivatives. PCT Patent PCT/JP2003/014986, 2003; Hussein, Z.; Mulford, D. J.; Bopp, B. A.; Granneman, G. R. *Br. J. Clin. Pharmacol.* 1993, **36**, 357-361; Kondo, T.; Yoshida, K.; Yamamoto, M.; Tanayama, S. *Arzneim. Forsch.* 1996, **46**, 11-14; Belliotti, T. R.; Brink, W. A.; Kesten, S. R.; Rubin, J. R.; Wustrow, D. J.; Zoski, K. T.; Whetzel, S. Z.; Corbin, A. E.; Pugsley, T. A.; Heffner, T. G.; Wise, L. D. *Bioorg. Med. Chem. Lett.* 1998, **8**, 1499-1502; A. Di Mola, M. Tiffner, F. Scorzelli, L. Palombi, R. Filosa, P. De Caprariis, M. Waser, A. Massa *Beilstein J. Org. Chem.* 2015, **11**, 2591-2599.
- L. A. Nafie, *Vibrational Optical Activity: Principles and Applications*, John Wiley & Sons Inc., 2011.
- P. J. Stephens, F. J. Devlin and J. R. Cheeseman, *VCD spectroscopy for organic chemists*, CRC Press, Boca Raton, 2012.
- P. L. Polavarapu, *Chiroptical Spectroscopy: Fundamentals and Applications*, CRC Pr Inc, 2015.
- V. P. Nicu and E. J. Baerends, *Phys. Chem. Chem. Phys.*, 2009, **11**, 6107.
- E. Debie, E. De Gussem, R. K. Dukor, W. Herrebout, L. A. Nafie and P. Bultinck, *ChemPhysChem*, 2011, **12**, 1542-1549.
- T. Kuppens, W. Langenaeker, J. P. Tollenaere and P. Bultinck, *J. Phys. Chem. A*, 2003, **107**, 542-553.
- S. Góbi and G. Magyarfalvi, *Phys. Chem. Chem. Phys.*, 2011, **13**, 16130.
- C. L. Covington and P. L. Polavarapu, *J. Phys. Chem. A*, 2013, **117**, 3377-3386.
- J. Vandenbussche, P. Bultinck, A. K. Przybył and W. A. Herrebout, *J. Chem. Theory Comput.*, 2013, **9**, 5504-5512.
- S. Qiu, E. De Gussem, K. Abbaspour Tehrani, S. Sergeyev, P. Bultinck and W. Herrebout, *J. Med. Chem.*, 2013, **56**, 8903-8914.
- A. D. Mola, F. Scorzelli, G. Monaco, L. Palombi and A. Massa, *RSC Adv.*, 2016, **6**, 60780-60786.
- R. Karmakar, P. Pahari and D. Mal, *Chem. Rev.*, 2014, **114**, 6213-6284.
- K. Speck and T. Magauer, *Beilstein J. Org. Chem.*, 2013, **9**, 2048-2078.
- A. D. Mola, L. Palombi and A. Massa, in *Targets in Heterocyclic Systems*, ed. D. S. O. A. Attanasi, R. Noto, Italian Society of Chemistry, 2014, vol. 18, ch. 5, pp. 113-140.
- IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). *Eur. J. Biochem.*, 1984, **138**, 9-37.
- W. Dennis O. Wackerly and R. L. S. III Mendenhall, *Mathematical Statistics with Applications*, Cengage Learning, Inc, 2007.
- B. Simmen, T. Weymuth and M. Reiher, *J. Phys. Chem. A*, 2012, **116**, 5410-5419.
- P. Bevington and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill Book Co, 2002.
- Weisstein, E. W. "Mean Deviation", From MathWorld—A Wolfram Web Resource. <http://mathworld.wolfram.com/MeanDeviation.html>.
- A. Singhal, *IEEE Data Eng. Bull.*, 2001, **24**, 35-43.
- R. de Gelder, R. Wehrens and J. A. Hageman, *J. Comput. Chem.*, 2001, **22**, 273-289.
- T. Kuppens, K. Vandyck, J. Van der Eycken, W. Herrebout, B. J. van der Veken and P. Bultinck, *J. Org. Chem.*, 2005, **70**, 9103-9114.
- Frisch, M. J. et al. Gaussian 09, Revision D.01, 2013, Wallingford CT, Gaussian, Inc.
- International Union of Pure and Applied Chemistry and International Union of Pure and Applied Chemistry, *Compendium of polymer terminology and nomenclature: IUPAC recommendations, 2008*, RSC Pub. : IUPAC, Cambridge, 2009.
- G. Scalmani and M. J. Frisch, *J. Chem. Phys.*, 2010, **132**, 114110.
- SPARTAN '04, Wavefunction Inc. 18401 Von Karman Avenue, Suite 370; Irvine, CA 92612.
- P. J. Stephens, F. J. Devlin and J.-J. Pan, *Chirality*, 2008, **20**,

- 643–663.
- 29 S. Grimme and C. Mück-Lichtenfeld, *Chirality*, 2008, **20**, 1009–1015.
- 30 S. Qiu, K. A. Tehrani, S. Sergeyev, P. Bultinck, W. Herrebout and B. Mathieu, *Chirality*, 2016, **28**, 215–225.
- 31 M. Kohout, J. Vandenbussche, A. Roller, J. Tuma, J. Bogaerts, P. Bultinck, W. Herrebout and W. Lindner, *RSC Adv.*, 2016, **6**, 81461–81465.
- 32 A. Krief, M. Dunkle, M. Bahar, P. Bultinck, W. Herrebout and P. Sandra, *J. Sep. Sci.*, 2015, **38**, 2545–2550.
- 33 A. Massa, P. Rizzo, G. Monaco and R. Zanasi, *Tetrahedron Lett.*, 2013, **54**, 6242–6246.
- 34 F. Cherblanc, Y.-P. Lo, E. De Gussem, L. Alcazar-Fuoli, E. Bignell, Y. He, N. Chapman-Rothe, P. Bultinck, W. A. Herrebout, R. Brown, H. S. Rzepa and M. J. Fuchter, *Chem. Eur. J.*, 2011, **17**, 11868–11875.
- 35 J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999–3094.
- 36 G. Longhi, M. Tommasini, S. Abbate and P. Polavarapu, *Chem. Phys. Lett.*, 2015, **639**, 320–325.