UNIVERSITY OF CAMBRIDGE

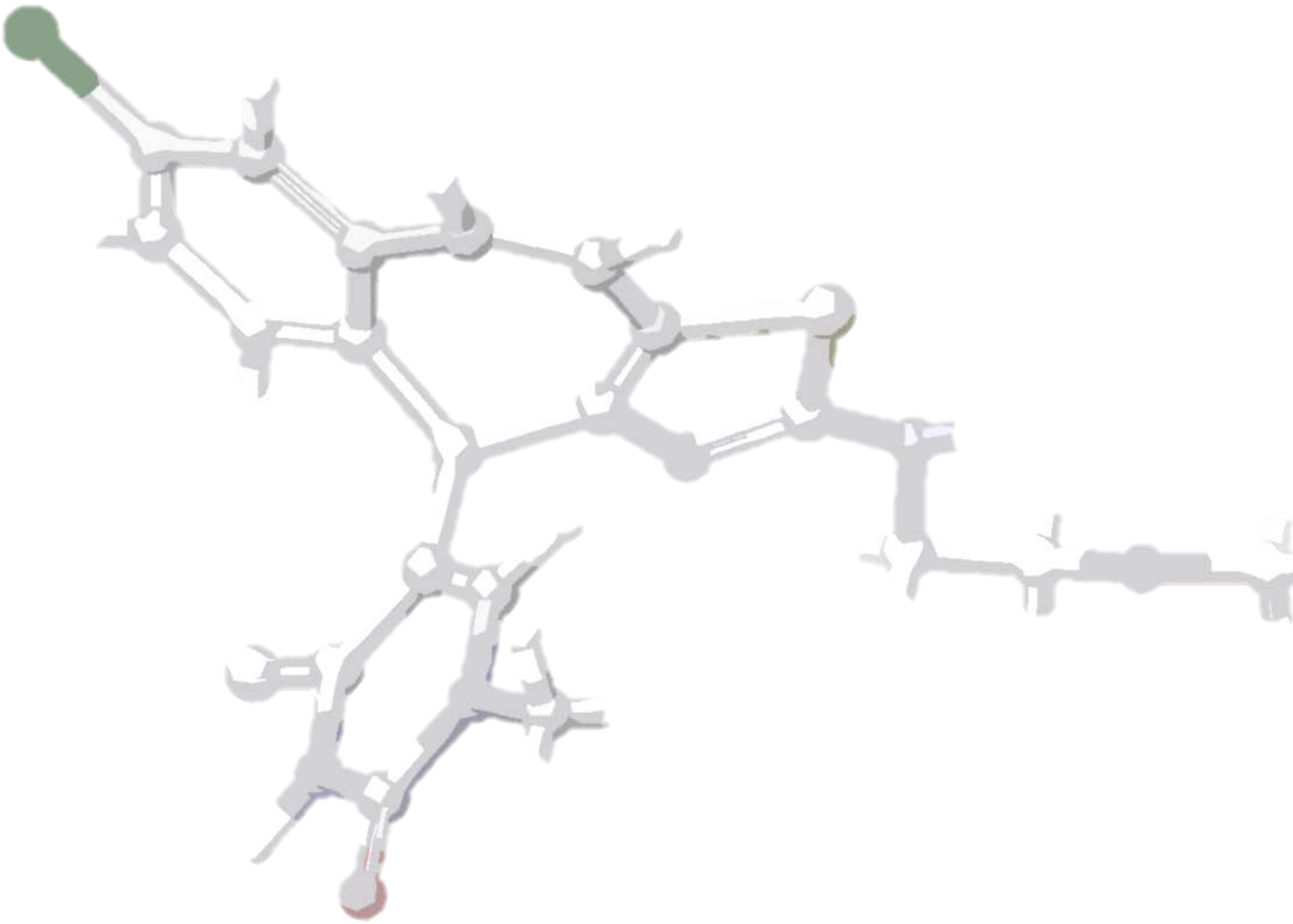
CLARE COLLEGE

Stereostructure Determination using Vibrational Circular Dichroism

Jonathan Lam
Department of Chemistry

September 2018

This thesis is submitted for the degree of Doctor of Philosophy



To the memory of Andrew Pang

Except as indicated by conventional references, this dissertation is the result of my own work and except where indicated is not the result of collaboration. It has not been submitted, either in whole or in part, for a degree or other qualification at another university. In accordance with the regulations, this dissertation does not exceed 60 000 words in length.

Jonathan Lam

September 2018

Acknowledgements

I wish to thank my supervisor, Professor Jonathan Goodman, for the interesting and thought-provoking project and for his support and insight throughout the duration of the work. My thanks also go to Dr Richard Lewis, Dr Marie Rydén Landergren and Dr Maria Saxin for their kind advice and assistance, as well as to AstraZeneca Sweden for funding my research. Finally, I wish to thank all members of the Goodman and Paterson groups for their stimulating discussion and ideas all of which have proven invaluable to the progress of my project.

Special thanks must also go to my parents for their support throughout the years, as well as all my friends in Cambridge for making the journey worthwhile. I wish in particular to thank Stephanie Ho for her help in the cover design, as well as Fiona and Henry Tai for technical support and a very patient introduction to LaTeX.

Stereostructure Determination using Vibrational Circular Dichroism

Jonathan Lam

Abstract

The application of chiroptical methods such as vibrational circular dichroism (VCD) spectroscopy is valuable in the assignment of absolute chirality. Assignment of chirality traditionally involves visual comparison of calculated and experimental spectra which can be subjective and time consuming.

Experimental VCD spectra for 40 compounds were acquired on different VCD spectrometers over the course of this work. To these data were added several compounds from published literature for which VCD data was already available. By successive trials over a dataset containing 60 compounds and their VCD spectra, various automated methods for interpretation of VCD spectra are developed and investigated, with the aim of achieving a confidence-level assignment algorithm for absolute chirality. The algorithm functions by comparing baseline-corrected VCD spectra of a compound with theoretically predicted spectra for each enantiomeric form.

The prediction method for VCD spectra involves conformational searching at the molecular mechanics level to find stable conformational minima, followed by quantum calculations using density functional theory (DFT) to find the vibrational modes. DFT calculations are performed using the B3LYP and B3PW91 functionals, in conjunction with the 6-31G(d,p) and cc-pVTZ basis sets.

Comparison between calculated and observed data is performed using a novel multiplicative percentage scoring method, scanning through a range of scale factors between 0.95 and 1.00. The degree of similarity is given a score ranging from -100% to +100%, with percentage values given such that uncertain cases need not be ignored or overconfidently assigned, but can be realistically evaluated according to existing spectral data.

This analysis is optimised using a database of 30 pairs of small-molecule organic compounds, including many drug-like and drug precursor molecules. Through these and further computational methods we present a reliable and time-efficient method for determination of absolute chirality, with the aim of developing into a standardised procedure for small-molecule organic compounds.

Table of contents

Nomenclature

Acronyms / Abbreviations

AC Absolute Chirality

B3LYP Becke, three-parameter, Lee, Yang, Parr

B3PW91 Becke, three-parameter, Perdew-Wang 91

cc-pVTZ correlation-consistent polarised valence-only triple-zeta

CD circular dichroism

COSMO conductor-like screening model

CPCM conductor-like polarisable continuum model

CPL circular polarised light

DCMTIQA 3-(6,8-dichloro-2-methyl-1,2,3,4-tetrahydroisoquinolin-4-yl)aniline

DFT density functional theory

DMSO dimethyl sulfoxide

ECD electronic circular dichroism

EM electromagnetic

FTIR Fourier Transform infrared

FWHM full width at half maximum

GIAO gauge-invariant atomic orbital

IEF-PCM integral equation formalism polarisable continuum model

IR infrared

LCP left circular polarised

LIA lock-in amplifier

MCPOPC methyl 2-(4-chlorophenyl)-4-oxocyclopentane-1-carboxylate

MM molecular mechanics

NMR nuclear magnetic resonance

PEM photoelastic modulator

QM/MM quantum mechanics/molecular mechanics

RCP right circular polarised

RMSD root mean square deviation

ROA Raman optical activity

SF scale factor

TBPTA *tert*-butyl(phenyl)thiophosphoric acid

UV/vis ultraviolet/visible

VCD vibrational circular dichroism

Chapter 1

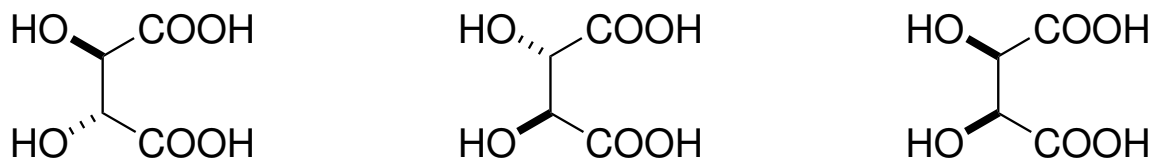
Introduction

1.1 Chirality and stereochemistry

Most organic molecules are not flat. A carbon atom with four single bonds possesses tetrahedral geometry, with two bonds pointing in and out of the plane of the remaining two. This gives rise to a form of asymmetry crucial to organic chemistry: a simple molecule with a central C atom attached to four different substituents is non-identical to its mirror image. Such a property is termed *chirality*, from the Greek word χεῖρ meaning "hand". It is easily illustrated by the non-superimposable nature of left and right human hands, i.e. a left hand will not fit in a right-handed glove.

Carbon atoms bonded to four different substituents are most commonly, but not exclusively, the source of chirality in organic chemistry. Such carbon atoms are termed stereogenic centres. Ever since Pasteur's first "optical resolution" separating the D- and L-isomers of tartaric acid in 1848[1], the study of stereochemistry has been central to organic synthesis and to the behaviour of organic compounds.

Absolute stereochemistry is defined as the complete three-dimensional arrangement of atoms in a molecule. This distinguishes it from relative stereochemistry, which does not differentiate between enantiomeric compounds, chiral molecules which are mirror images of each other. For molecules containing chiral carbon atoms, the enantiomeric structure can be generated by inverting all the chiral centres in the molecule. Note that the presence of stereogenic centres does not guarantee that the molecule is chiral, as in the case of mesotartaric acid **1c**.



1a **1b** **1c**

Scheme 1.1 Stereoisomers of tartaric acid. **1a** and **1b** are enantiomers of each other; both are diastereomers of **1c**.

1.2 Stereochemistry in medicine

Among molecules of biological interest, the vast majority possess chirality. All chiral amino acids incorporated by organisms into proteins during translation exist as the L-form. Sugars and nucleic acids are likewise homochiral, with one enantiomer occurring in biological samples in vast excess of the other. As a result, care must be taken when administering pharmacologically active chiral compounds as the effect of one enantiomer can differ drastically from the other [2] [3].

One example in which this difference was made painfully clear is the case of the drug Thalidomide[4], which in the late 1950s was used to treat morning sickness in pregnant women. While the pharmaceutically active (*R*)-isomer had the desired sedative and antiemetic properties, the (*S*)-isomer presented adverse effects resulting in infants born with phocomelia and other severe teratogenic birth defects.

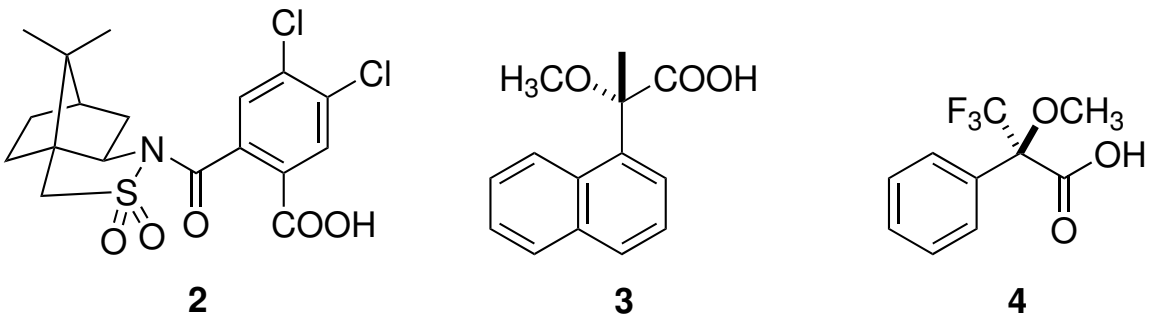
In reaction to this tragedy and to increasingly stringent regulations, the pharmaceutical industry as a whole chose to direct much of its attention away from chiral molecules during the following decades, focusing instead on aromatic and achiral compounds. Recent years, however, have seen renewed interest in single-enantiomer chiral compounds used for medicinal applications.

The determination of absolute stereochemistry in organic compounds of biological interest has been a persistent challenge in spite of an arsenal of modern spectroscopic methods at the scientist's disposal. This is especially relevant to the present day when entire fields of study are devoted to ensuring the stereoselectivity of catalysts and installing the correct stereogenic centres in organic synthesis, and the creation of single-enantiomer drugs is of paramount importance in both the pharmaceutical industry and its corresponding regulatory bodies. The determination and discrimination of absolute stereochemistry in chemical synthesis and structural analysis is therefore of utmost significance.

1.3 Methods used in the determination of absolute stereochemistry

NMR and IR spectroscopy are unable to distinguish enantiomeric compounds directly. However, there remain techniques through which NMR can be a useful probe for absolute stereochemistry. Esterification of compounds containing a hydroxy group with chiral derivatising agents such as (-)-camphorsultam dichlorophthalic (CSDP) acid **2** [5], (S)-(+)-2-methoxy-2-(1-naphthyl)propionic (MαNP) acid **3** or Mosher's acid **4** [6] shown below can generate pairs of diastereomers from enantiomeric alcohols, allowing separation of racemates via high-performance liquid chromatography (HPLC) on silica gel. In some cases, crystallisation has also been achieved with MαNP esters, enabling stereostructure determination by X-ray crystallography which is mentioned below.

Since the esters produced possess a carboxylate moiety of known stereochemistry and exhibit different chemical shifts under NMR, careful analysis of NMR shifts and coupling constants can yield the correct assignment of absolute stereochemistry. Unfortunately, such an analysis is not straightforward for larger and more complex molecules and the differences in NMR between diastereomers can be subtle, leading to ambiguity in such an approach.



Scheme 1.2 Chiral derivatising agents used in separation and NMR analysis of enantiomeric alcohols. From left to right: (-)-CSDP acid **2**, (*S*)-MαNP acid **3** and (*R*)-Mosher's acid **4**.

Another widely used method is single-crystal X-ray diffraction, in which a monochromatic beam of X-rays is directed at a single crystal of the compound under investigation. The angles and intensities of diffracted beams allow for the mapping of electron density in three dimensions. This gives the precise location of all the atoms in the molecule besides hydrogen, which lacks the sufficient electron density. This does not usually pose a significant barrier to determination of stereochemistry, since the positions of hydrogen atoms can be inferred from those of other atoms. However, the method of predicting absolute stereochemistry via X-ray

diffraction, making use of anomalous dispersion, depends on the presence of heavy atoms such as S, P, Cl and beyond. This achieves an assignment whose reliability can be quantified via the Flack parameter.[7] Though there exist natural products[8] [9] and pharmaceuticals containing such atoms, they are absent in most examples. Light atoms (C, N, O etc.) which have a weak anomalous scattering can lead to uncertainty in the assignment of absolute stereochemistry, or give a physically meaningless Flack parameter. Finally, sufficiently large and regular crystals are required for single-crystal X-ray diffraction to be possible, and when crystals of sufficient quality are unobtainable then other methods must be considered.

The analytical methods most commonly associated with determination of absolute stereochemistry are termed the chiroptical techniques [10], which utilise the behaviour of polarised light transmitted through a sample. The first to be discovered among these was optical rotation, a phenomenon in which the plane of plane polarised light is rotated as it travels through a chiral material. The phenomenon of optical rotation has given rise to the (+) and (-) designations of stereoisomeric forms of the same compound, formerly known as “optical isomers”. However, optical rotation often shows a degree of temperature dependence and cannot always be measured reliably.

One possibility for overcoming this problem lies in the Cotton effect [11]: near the absorption wavelength of a given substance, the optical rotation exhibits a characteristic sinusoidal variation. The Cotton effect is termed “positive” if the optical rotation first increases with decreasing wavelength and “negative” if it first decreases. This allows for the measurement of optical rotatory dispersion, the variation in optical rotation as a function of wavelength, as a spectroscopic approach to determination of absolute stereochemistry.

In plane polarised light, the electric and magnetic field vectors oscillate in orthogonal planes, with the direction of propagation being the intersection between these planes. In addition to plane polarisation, light can also exist in a circular polarised form, with the electric and magnetic field vectors of the electromagnetic (EM) wave maintaining a constant magnitude while rotating around the centre. Two forms of circular polarised light (CPL) exist: in right circular polarised (RCP) light the electric and magnetic field vectors trace out a right-handed helix, and in left circular polarised (LCP) light they trace out a left-handed helix. The two types of CPL are thus mirror images of each other.

The chiral nature of CPL means that each of its two forms will interact with a sample of chiral molecules in a different way. This phenomenon has also been utilised in chiroptical spectroscopy. For example, Raman optical activity (ROA) [12] is a technique which relies on the differing intensities of LCP and RCP light after Raman scattering through chiral media.

This allows for an extension of Raman spectroscopy into the investigation of molecular chirality, much as circular dichroism methods (as will be discussed) build on the basis of UV/vis and IR spectroscopy. One notable feature of Raman optical activity is that the same number of bands appear in its spectra as in those of circular dichroism (CD), allowing these techniques to complement each other.

Finally, absorption spectra can be measured using CPL in order to differentiate chiral compounds. Circular dichroism (CD) [13] of a sample is defined as the difference in absorption of RCP and LCP light; specifically, the differential absorption ΔA is calculated by the following equation:

$$\Delta A = A_{\mathrm{L}} - A_{\mathrm{R}} \tag{1.1}$$

Equation 1.1: Derivation of the differential absorption value used in circular dichroism.

where A_L and A_R refer to the sample's absorbance of LCP and RCP light respectively. Hence, CD is positive if $A_L > A_R$ and negative if $A_L < A_R$. Note that this convention is the reverse of that used in ROA.

CD is measured as a function of wavenumber, generating a spectrum with bands for different frequencies of absorption. In the ultraviolet/visible (UV/vis) region of the electromagnetic spectrum, where the majority of absorption bands are due to excitation of electrons between quantised states, this is referred to as electronic circular dichroism (ECD); in the IR range, where the absorption signals arise from the excitation of bond vibrations, it is termed vibrational circular dichroism (VCD). It is VCD spectroscopy which will be the main focus of this work.

1.4 Vibrational Circular Dichroism

1.4.1 Theory of VCD

The theoretical background of VCD measurement is very similar to that of IR spectroscopy [14] [15]. When placed in the path of a beam of IR radiation, a sample containing organic molecules will attenuate the light by absorbing photons of a given frequency. These frequencies correspond to specific vibrational modes of the molecules in the sample, and the energy carried by the photons is used in the excitation of these vibrational modes.

The relationship between the light absorbed and the path length of the beam through a sample is given by Beer's law, also known as the Beer-Lambert law:

$$A = \varepsilon \cdot c \cdot l \tag{1.2}$$

Equation 1.2: Beer's law relating the absorption of a sample to its pathlength, l. The concentration of the sample (with units of mol dm^{-3}) is represented by c and the molar extinction coefficient by ε.

An IR spectrum is a recording of these absorptions over a range of wavenumbers. Since bond lengths and bond angles are preserved under reflection in a mirror plane, the vibrational frequencies of two enantiomers are likewise the same, giving them identical vibrational absorption spectra when measured using unpolarised IR radiation. This is provided that the molecules in the sample are randomly oriented, as is the case with liquid and gaseous solutions, as well as pure liquids and gases. For this reason, unpolarised IR spectroscopy is unable to distinguish enantiomeric compounds.

However, as the enantiomeric molecules are mirror images of each other, so are the dipole transition moments involved in the vibrational modes under IR excitation [16]. These are the dipole moments associated with changes in the electrical and magnetic fields induced by the bond vibrations in the molecule. Since each type of CPL is chiral, and carries its own electric and magnetic field vectors, RCP radiation will interact with a single enantiomer in a different manner from LCP. This results in a difference in absorption between RCP and LCP, which is then measured as ΔA(CD).

By symmetry, a chiral molecule should interact with LCP light in an analogous way to its enantiomer's interaction with RCP light and vice versa. Replacing a compound under study with its opposite enantiomer ought, therefore, to cause the band shapes of circular dichroism spectra to invert in sign, theoretically producing a mirror image of the original spectrum.

1.4.2 Measurement of VCD Spectra

CPL cannot be generated directly from a beam of unpolarised IR radiation. Having directed the beam through a linear polariser, however, the resulting plane polarised IR radiation may be converted into CPL by passing it through an optical device called a quarter-wave plate, creating a source of CPL in the IR region of the electromagnetic spectrum.

The measurement of VCD spectra is in principle similar to that of Fourier transform IR spectroscopy. However, since the quantity being measured is the difference in absorbance of circular polarised light and not the absorbance itself (as in IR spectroscopy), the signal strength for VCD is relatively weak, about four or five orders of magnitude smaller than that of a typical IR measurement.[17] Due to a low signal-to-noise ratio, therefore, it is not practical to acquire a VCD spectrum simply by taking absorbance spectra for left- and right-handed circular polarised light and subtracting one from the other, as background noise may vary from experiment to experiment giving rise to misleading results.

Instead, a VCD spectrometer applies what is termed dual polarisation modulation. [18] This makes use of a vibrating piezoelectric material, the photoelastic modulator (PEM), to convert incoming linearly polarised light into alternating left- and right-handed circular polarised light, such that the difference between the two may be measured and recorded "on the fly" and minimising the effect of any variations in background. The PEM effectively acts as an alternating quarter-wave plate to convert plane polarised IR into RCP and LCP light at a controlled frequency.

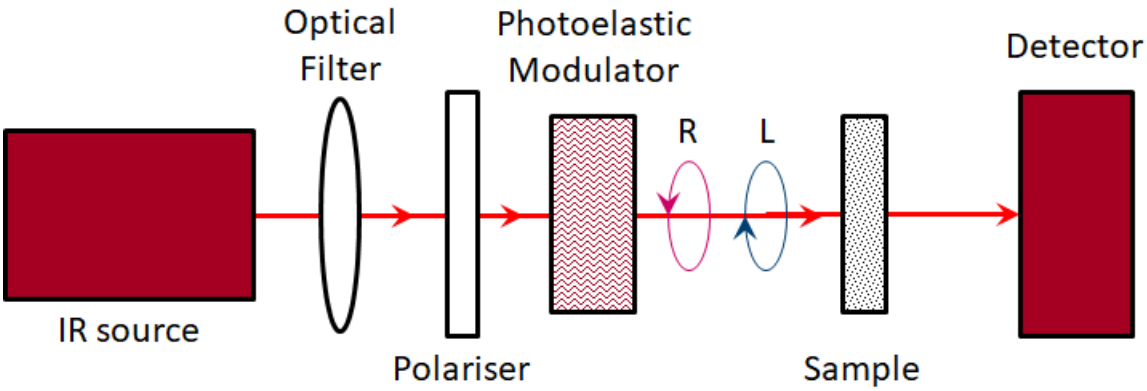


Fig. 1.1 Schematic representation of the components of a basic VCD spectrometer. An FT-IR spectrometer is often used as the IR source.

Similarly to IR spectroscopy, VCD analysis may use solution-state [19] or solid-state [20] samples. Solution-state spectra are typically obtained using a liquid cell with KBr windows, which are transparent to infrared radiation. For solid-state spectra, samples in KBr pellets may be used. It is also possible, though costly, to perform VCD experiments on samples in which the molecule of interest is isolated in an argon matrix, precluding any oligomerisation of the sample or unwanted solute-solvent interactions. Spectra acquired from the same compound in different physical states may differ significantly for reasons outlined below.

1.4.3 Computational prediction of VCD Spectra

Chemical informatics and computational methods [21] lend themselves conveniently to the calculation of VCD spectra. The absorption of circular polarised light in VCD spectroscopy is due to the excitation of vibrational transitions in the molecule, in a similar manner to those in IR spectroscopy. The same vibrational modes of each moiety are possible; however, the strength of each band depends not only on dipole strength (as in IR spectroscopy) but also on rotational strength R, a value which can be derived from the electric and magnetic dipole transition moments as shown below. These can be predicted by performing gauge-invariant atomic orbital (GIAO) calculations on accurately modelled conformers of a molecule [22] [23].

$$R = \frac{1}{2mc} Im \int \Psi_g \hat{M}_{(elect.dipole)} \Psi_e d\tau \cdot \int \Psi_g \hat{M}_{(mag.dipole)} \Psi_e d\tau \tag{1.3}$$

Equation 1.3: The derivation of a theoretical value for rotational strength. The symbols $\hat{M}_{(elect.dipole)}$ and $\hat{M}_{(mag.dipole)}$ refer to the electric and magnetic dipole moment operators respectively.

The elucidation of the relationship between chemical structure and the position and intensity of VCD bands enables the *in silico* calculation of theoretical VCD spectra for a given structure. This presents a valuable tool for the determination of stereostructure by combining cheminformatics with analysis of experimentally derived VCD spectra. A typical computational VCD analysis, and the procedure followed in this work, is given as follows:

1. For a given molecule under investigation, a VCD spectrum is recorded and a list of candidate structures is drawn up. Depending on the amount of information already available regarding the stereochemistry of the molecule, the number of candidate structures may range from two (a single pair of enantiomers) to the total number of unique combinations of configurations of chiral centres in the molecule.

2. The input conformation for DFT calculations is crucial to VCD prediction [24]. Therefore, a conformational search is carried out on each isomer to find its most stable conformers. To save computational effort, this process is usually performed by molecular mechanics packages which treat atoms as single particles and chemical bonds as Hookean springs. In order that conformational space is sampled as thoroughly as possible, a suitable number of steps is chosen for the conformational search. The search involves repeated rotation of single bonds at random to generate new conformations, whose energies are then calculated and

optimised for the lowest energy. This gives rise to a series of low-energy conformers which present themselves as conformational minima on the potential energy surface.

3. For each of the conformational minima within a given energy window, density functional theory (DFT) calculations are applied to generate a series of vibrational modes, their frequencies and rotational strengths [25]. There exists an extensive list of computational models for this process, in the form of numerous possible functionals and basis sets; however, there does not appear as yet to be a clear consensus on the most accurate method to use. Among the more commonly used are the B3LYP hybrid functional and the 6-31G(d,p) and cc-pVTZ basis sets. These calculations often involve DFT-based optimisation of conformational geometries, though this was not always performed in this particular work. This generates a list of predicted VCD peaks for each individual conformer; however, the calculated spectra for all relevant conformers must be merged before they can be compared to the observed spectrum.

4. A broadening algorithm is applied to the discrete signals calculated for each conformer to produce a continuous line shape resembling an experimentally obtained spectrum.

5. The Boltzmann weight for each conformer is calculated according to DFT-generated conformational energies. The calculated conformational spectra, each multiplied by their corresponding weight, are summed together to produce a final Boltzmann averaged spectrum.

6. The calculated spectrum is compared to the observed VCD spectrum of the molecule under investigation. This comparison may be done visually, or via one of many computational algorithms. If one of the predicted structures matches the molecule of interest, its calculated spectrum ought to give VCD signals with the same sign and similar relative intensities to those of the experimental spectrum.

The final step, comparison between observed and calculated spectra, may appear straightforward but is often a complex process. Whether due to issues with the experimental data (such as the presence of noise in the baseline) or inaccuracies in the computational process, discrepancies between the two spectra often arise creating ambiguity regarding the assignment of a particular structure. An investigation aimed at creating, applying and optimising an algorithm to overcome these difficulties is given in the following chapters.

The chirality predicted need not be limited to stereogenic centres. Studies [26] have shown that atropisomeric forms of a compound, arising from restricted rotation around a single bond, can likewise be analysed *via* VCD. Inherent chirality at the supramolecular level, such as that of helicenes [27] and calixarenes[28], can also be studied using chiroptical spectroscopies.

In addition, even when stereogenic centres are present, the groups around such a centre may mask the VCD signal or hide it entirely, making interpretation difficult [29].

1.4.4 Causes of Discrepancies between Calculated and Experimental Spectra

As shown in the previous section, quantum chemical calculations such as those applying DFT are now able to predict VCD spectra of simple rigid molecules.[30] The visual similarity between a calculated spectrum and one acquired from a VCD experiment does not always translate into certainty of the proposed structure, or more commonly, the calculated spectrum of the correct structure differs significantly from the experimentally observed one. Baseline noise in the experimental spectra and anharmonicity of certain VCD transitions (deviations from the behaviour of a simple harmonic oscillator) are two possible causes of this.

Additionally, these calculations generally ignore the effects of solute-solute and solute-solvent interactions. In many cases, the effects of solvation on the observed VCD spectrum can be easily estimated using simple solvent models, such as the integral equation formalism of the polarisable continuum model (IEF-PCM) [31] whereby the solvent is treated as a continuum dielectric environment and no explicit intermolecular interactions are taken into account. Other similar solvent models include the conductor-like polarisable continuum model (CPCM) [32] and the conductor-like screening model (COSMO) [33].

The success of these solvent models notwithstanding, they are unsatisfactory in dealing with cases where intermolecular interactions interfere strongly with non-robust VCD transitions; that is, transitions with a near -90° angle between their electric and magnetic dipole transition moments.[34] This leads to a change in sign of the corresponding VCD band and as a result, equivalent bands between observed and calculated spectra cannot always be matched trivially. The concept of robust vibrational modes is further explained in the following section.

Furthermore, there exist examples where solution-state spectra can only be interpreted in terms of hydrogen bonded complexes. Nicu and Baerends[34] conducted a study of several small-molecule carboxylic acids and found that despite discrepancies between the observed spectra and those calculated for single-molecule vibrational modes, they were able to accurately predict experimental spectra for molecules known to form such complexes by explicitly considering the dimers, oligomers or solute-solvent complexes in question together with their relative populations, thus creating Boltzmann-weighted VCD spectra

that accurately modelled the solution state. They also noted that while it was possible to acquire VCD spectra for the acid monomers, this was generally only possible from very dilute solutions. For such compounds, therefore, the solution-state VCD spectrum is dependent on concentration of the solute under investigation. This adds difficulty to the task of determining the spectrum for the monomer: a high concentration causes hydrogen bonding to become an issue while a low one hinders the acquisition process owing to a poor signal-to-noise ratio. It is possible, however, to remove the effects of hydrogen bonding entirely via matrix isolation of the compound of interest.

Another result of hydrogen bonding is the chirality transfer effect, observed in aqueous solution by Losada and Xu.[35] Upon dissolution of enantiopure methyl lactate in water, the VCD spectrum of the resulting solution exhibited broad signals in the 1600-1700 cm^{-1} region. These were found to correspond with the H-O-H bending bands of water, a solvent with no inherent chirality. This lends further weight to the idea that the effects of solvents upon VCD spectra, though complex and sometimes undesirable, are not in fact unpredictable and that they can occasionally (as with chirality transfer to solvent) be used as an aid [36] to stereostructure determination.

1.5 Difficulties in VCD interpretation

Various challenges present themselves in the course of determining AC using prediction of VCD spectra. One of the challenges in the interpretation of VCD spectra lies in the poor signal-to-noise ratio, as mentioned in Section 1.4. As a result, the correction of baseline noise is a significant and often-used step to clean up experimental VCD data before comparison with calculations.

In terms of sample preparation for experimental VCD measurement, it may be possible to remove some of the causes of baseline artefacts. By taking background measurements, for example using a flow cell filled with pure solvent, the effects of noise can be subtracted from the VCD spectrum while the sample measurement is taking place. The advantages of the various background subtraction techniques are summarised in Table 1.1.

Orientation of the cell can affect the polarisation of the beam passing through, and birefringence artefacts may vary from once cell to another. Path lengths minute differences can introduce errors. Therefore, using the same flow cell is recommended when measuring VCD

	Causes of artefacts
A	Instrument
B	Cell window, orientation, birefringence etc.
C	Solvent impurities or properties
D	Sample impurities or properties

Baseline correction	Issues resolved			
	A	B	C	D
Racemic mixture	✓	✓	✓	✓
Solvent-filled cell	✓	✓	✓	
Empty instrument	✓			

Table 1.1 Some possible causes of baseline artefacts in VCD and methods of background subtraction which may help to remove these sources of error.

spectra of related compounds, such as two enantiomers, and ensuring that the cell has a consistent position in the sample holder when taking measurements is highly recommended.

The degree to which each individual VCD signal is diagnostic of a chirality assignment is dependent on various factors. While the strength of a VCD signal does not always increase its reliability, a strong peak is easier to calculate and predict. As a vibrational spectroscopic method, VCD is highly sensitive to conformational changes, and under some calculations, it can be found that a portion of VCD signals appear to support one assignment of AC, while the remainder of signals support the opposite assignment.

Paul Nicu *et al.* outlined the theory of robustness[34] in a molecule's vibrational modes as the tendency of a VCD signal to undergo a sign change as a result of small perturbations in the conformation; in other words, more robust modes are likely to be correct even when the calculated conformation of the molecule under investigation is not absolutely accurate.

1.6 Outline of the rest of the Dissertation

The remainder of this dissertation is structured is as follows:

Chapter 2 outlines the procedure of measuring VCD spectra and the computational methods used in their prediction.

Chapter 3 provides a list of structures for the compounds studied in this work.

Chapter 4 discusses a few early model algorithms tested and the issues investigated which led up to the development of a new scoring method, which is detailed in Chapter 5.

Chapter 6 applies this method to solving the AC of previously unassigned molecules and further investigates some of the weighting and scaling methods used.

Chapter 7 explores the compilation of VCD spectra into an online database and outlines the requirements such a database would have to enable users worldwide to access and submit data.

Finally, Chapter 8 draws up the important conclusions from the dissertation and provides an outlook for future investigations.

Chapter 2

Methods

2.1 Acquisition of Experimental spectra

In the development of the comparison algorithms in the following chapters, trials were carried out on pairs of commonly available organic enantiomers, as well as several novel compounds with previously unassigned stereochemistry. Compounds for the analysis were selected with varying degrees of conformational flexibility, ranging from structures with only one stable conformation to those with many low-energy conformers.

Unless otherwise stated, the compounds were dissolved in chloroform-d_1, at a concentration adequate for measurement of VCD spectra. For a vibrational mode to be reliable in the VCD spectrum, the corresponding IR signal should lie in the range between 0.2-0.8 absorbance units [37]. This was taken into account when deciding the concentration of the solute, while avoiding the formation of dimer aggregates as in the case of carboxylic acid groups.

In order to minimise the absorption of the transmitted IR beam by the solvent, the path length of the beam through the sample should be kept low. The concentration must also be sufficiently low such that Beer's law is obeyed; that is, the level of solute-solute interactions is at a level not measurable by the spectrometer, though likewise concentrated enough for the solute VCD signals to be detectable.

The dissolved compounds were analysed using NMR spectroscopy to ensure their purity before further measurements. IR and VCD spectra for the compounds were recorded on a BioTools ChiralIR-2X Spectrometer and a Bruker TENSOR FTIR spectrometer with a

PMA50 module for polarisation modulated measurements. Solutions of the samples were held in a BaF2 transmission cell with a path length of 100 micron.

Before measurement of the VCD spectra, the detector must be cooled with liquid N_2 for approximately 20 minutes. Since the IR spectra were acquired separately from the VCD spectra on the Bruker instrument, necessitating a change of settings on the spectrometer, a new calibration curve must be acquired before every VCD measurement using a CdS multiple wavelength plate and polariser.

Both IR and VCD spectra were recorded at a spectral resolution of $4cm^{-1}$ by accumulating 18,000 scans in blocks of 3,000 over an accumulation time of approximately 6 hours. For the VCD spectra, baseline correction was performed by subtraction of the spectra of the opposite enantiomers.

In the following, *via* thorough investigation of a range of compounds by DFT calculation, we evaluate the effects of weighting and variation of scale factors on some methods used in calculating VCD spectra, and report on their applicability to a standard chirality prediction procedure.

In the development of the comparison algorithms in the following chapters, trials were carried out on pairs of commonly available organic enantiomers whose experimental VCD data were provided by AstraZeneca Gothenburg or found in published literature.

2.2 Computational methods

The technique applied to each pair of enantiomers is as follows:

Each enantiomeric form was drawn using the Maestro[38] molecular modelling interface (Version 9.3). After geometry cleanup using the MM2 Force Field,[39] a Macromodel[40](Version 9.9) conformational search was set up. All conformational searches were performed using the Mixed torsional/Low-mode sampling method,[41][42] with the torsional sampling option set to “Extended”, and the Merck Molecular Force Field (MMFF)[43][44] to determine the energy of each conformer.

The environment of the conformational search aimed to model that of the VCD experiment as closely as possible. Hence the $CHCl_3$ solvent model was applied in the vast majority of searches. However, owing to the few available solvent models for MMFF, the $CHCl_3$ solvent

model was also applied in certain cases where VCD data had been taken in other common organic solvents, such as dimethyl sulfoxide (DMSO).

A suitable number of steps was chosen for each conformational search such that each conformation within 10 kJ mol^{-1} above the ground state was found five times or more. This has been shown[45] to be a reasonable indication for a thorough search of the conformational landscape.

For each conformer found in the 10 kJ mol^{-1} window, GIAO DFT calculations were used to calculate the position and intensity of VCD rotational transitions. This was achieved through the use of Jaguar[46] software (Version 7.9) to apply the widely-used B3LYP functional[47][48][49][25] and the 6-31G(d,p) basis set,[50] as well as the cc-pVTZ basis set[51], this has been stated accordingly in the corresponding chapter. Using the Jaguar[46] software package, geometry optimisations and frequency calculations were performed for both enantiomers of each compound at the following levels of theory:

B3LYP/6-31G(d,p)

B3LYP/cc-pVTZ

B3PW91/6-31G(d,p)

B3PW91/cc-pVTZ

This yielded a list of VCD transitions and their rotational strengths for each conformer, which were then subjected to analysis as detailed in the following chapters.

Chapter 3

Molecules Studied

This chapter lists the various compounds studied in the following chapters together with some background information regarding specific compounds, grouped according to the source of experimental VCD data and the studies in which they appear.

3.1 Test set

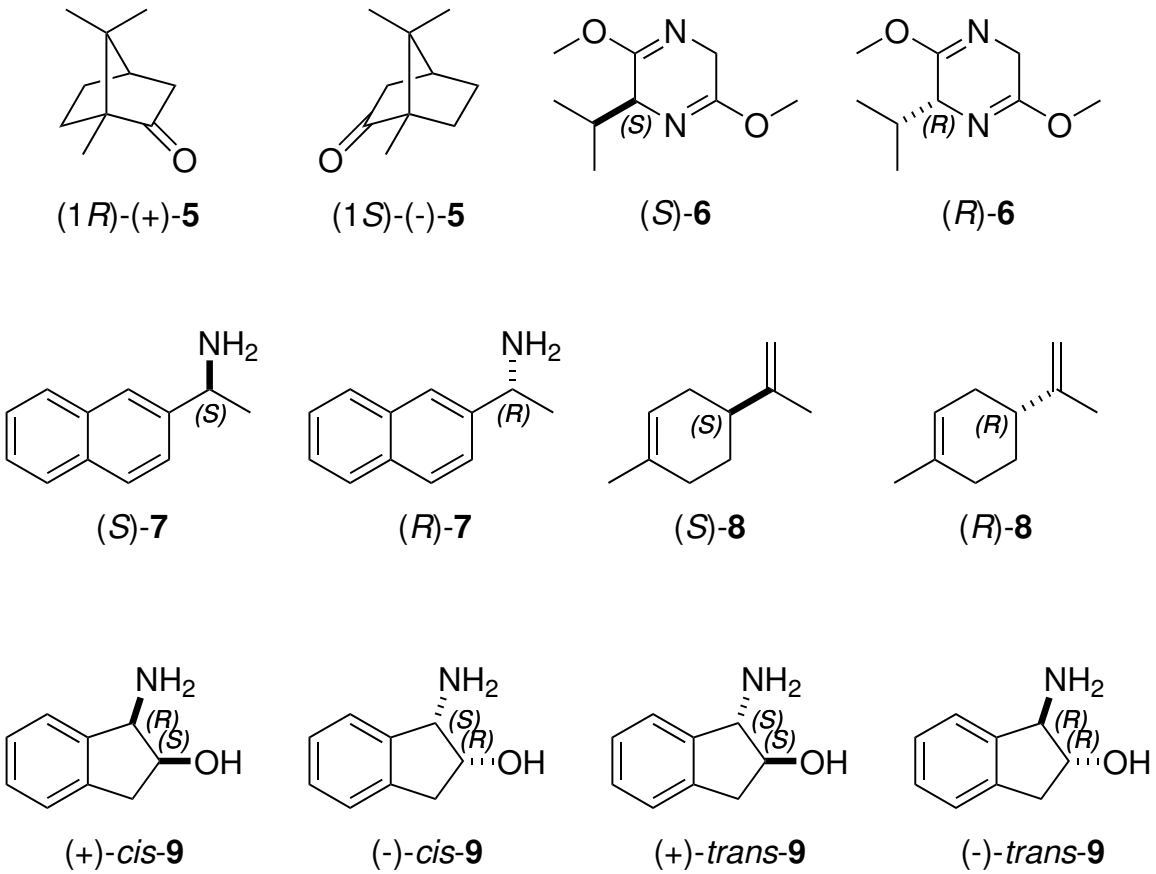
For the development of a comparison algorithm, a test set of twelve compounds, comprising six enantiomeric pairs, was used. Worth noting is the diastereomeric relationship between the *cis*- and *trans*-1-amino-2-indanols **9**, which were also investigated to ascertain if the diastereomers could be differentiated by VCD.

As the solubility of the 1-amino-2-indanols in $CDCl_3$ was poor, initial studies of these compounds used VCD spectra of the compounds dissolved in d_6-DMSO. These spectra proved difficult to assign analytically, with parts of the VCD-relevant wavenumber region covered by solvent-related noise below 1100 cm^{-1}. Later measurements were carried out using CD_2Cl_2 as a solvent for the *cis*-diastereomers and $CDCl_3$ for the *trans*-diastereomers, allowing the VCD bands within the 1000 - 1800 cm^{-1} wavenumber region to be fully measured.

Experimental data were acquired on a BioTools ChiralIR-2X Spectrometer. Solutions of the samples were held in a BaF_2 transmission cell with a path length of 100 μm. Both IR and

VCD spectra were recorded at a spectral resolution of 4cm^{-1} by accumulating 18,000 scans in blocks of 3,000 over 6 hours.

After conformational searching, the conformational vibrational modes and VCD rotational strengths for each compound were calculated and compiled. Since these molecules were small and relatively rigid, both enantiomers were calculated at little additional cost to computational time.



Scheme 3.1 Compounds used during the studies in Chapter 4.

3.2 Additional compounds tested

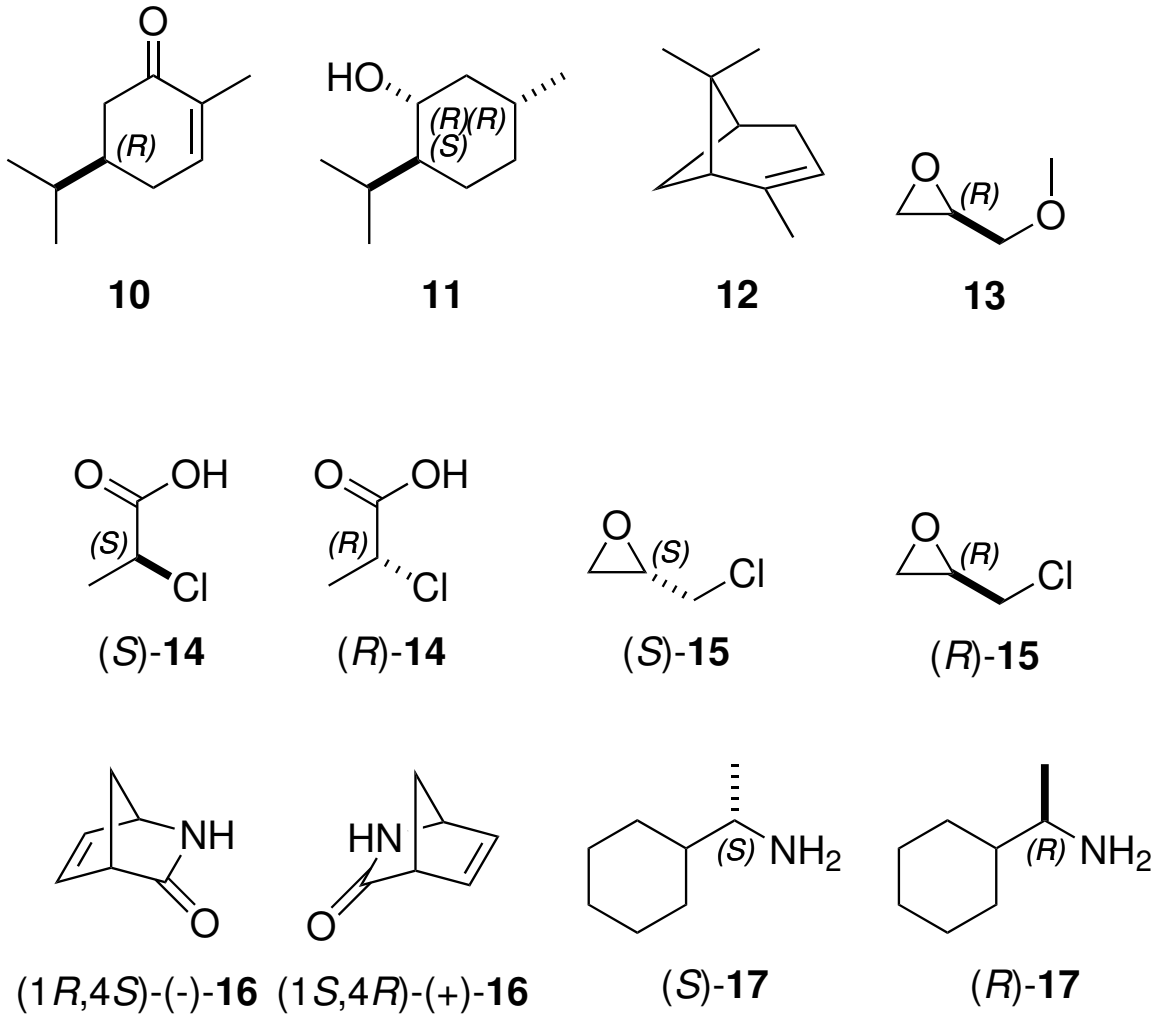
During the course of algorithm development, several additional compounds were measured and these were added to the analysis. Not all of these compounds were available as enantiomeric pairs, and therefore the baseline correction method of subtracting enantiomers was not always available. Similarly to the compounds in Scheme 3.1, VCD data were acquired on a BioTools instrument in AstraZeneca, Sweden. Some further measurements were made with a BioTools ChiralIR-2X spectrometer in the Molecular Spectroscopy Group at the University of Antwerp.

Solvent effects were also studied, for example in the VCD spectra for 2-chloropropionic acid **14**. Previous studies [52] had demonstrated for this compound that owing to the presence or absence of solute-solvent H-bonding in different solvent systems, the VCD signals would be changed to a different wavenumber depending on the solvent used. Spectra of 2-chloropropionic acid were therefore acquired in CD_2Cl_2, as well as CCl_4, in order to compare the positions of the VCD signals and to determine which, if any, the gas-phase calculations DFT calculations would correspond to the best. The expectation was that spectra acquired in CCl_4, lacking any H-bonding effects, would be the most similar to the calculated VCD spectrum.

3.3 AstraZeneca Compounds

Following the conception and development of the comparison procedure detailed in Chapter 5, IR and VCD spectra for a variety of chiral compounds owned by AstraZeneca were acquired. In addition to measurements on the BioTools ChiralIR-2X instrument, a selection of spectra were also taken on a Bruker TENSOR FTIR spectrometer with an attached PMA50 module for polarisation modulated measurements.

Compared to previous examples, most of these compounds are larger in size, and many of them possess a greater degree of conformational flexibility. While performing computational analyses of these compounds, therefore, it was expected that a confident assignment of AC would be more difficult to achieve than with the previous sets. This is despite spectra of both enantiomeric forms being available for these compounds in all the examples given, rendering baseline correction by subtraction of enantiomers an available option.

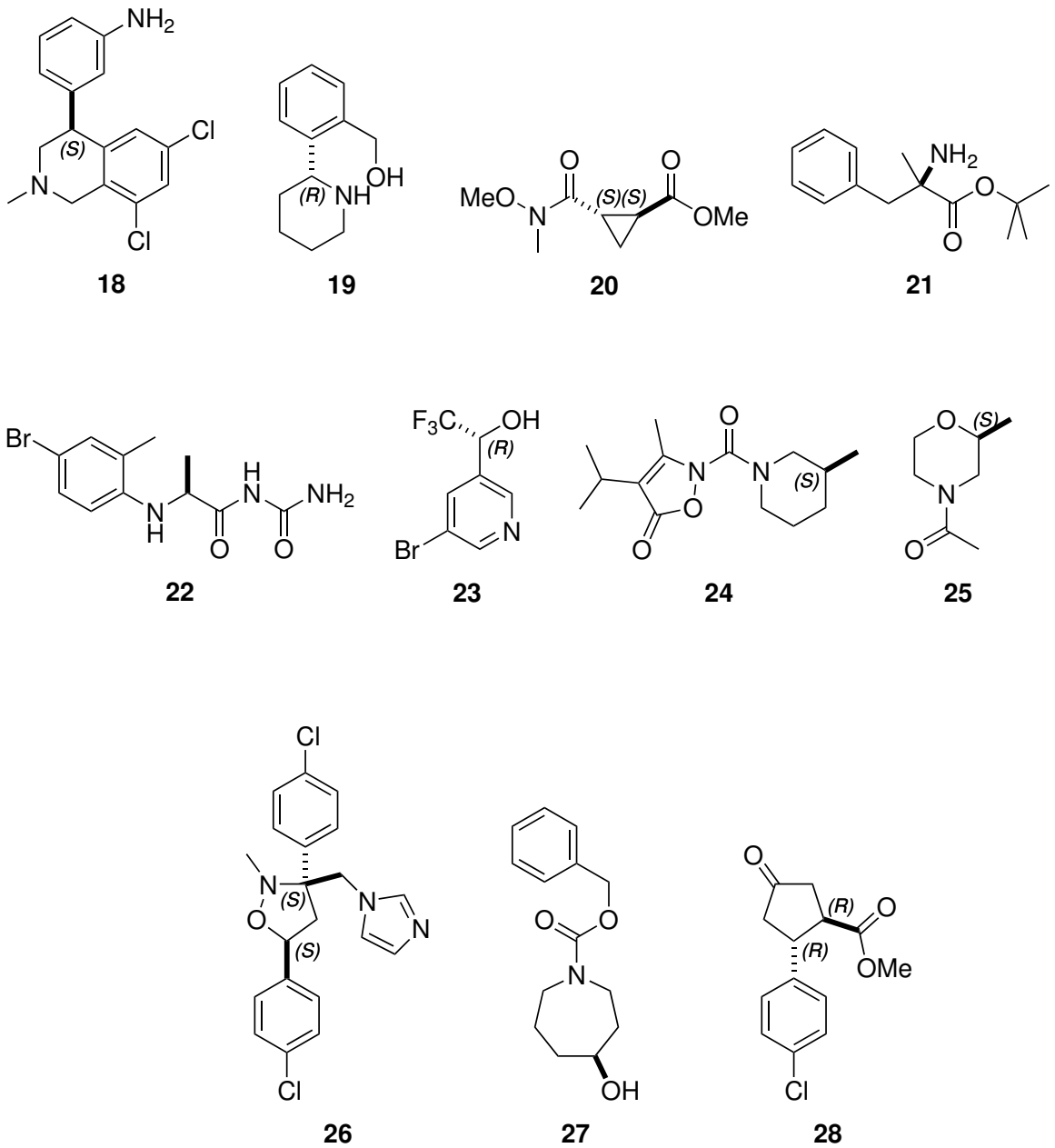


Scheme 3.2 Additional compounds used in Chapter 4; some were available for VCD measurement only as single enantiomers.

A few of these compounds proved to be too large and flexible for a full conformational search; particularly noteworthy is the case of compound **27**. Even after 300,000 steps, a full exploration of the potential energy surface of the molecule appeared to be incomplete, with 584 conformations found within the 10 kJ mol^{-1} energetic window, among which 257 were found fewer than five times and 76 were poorly converged in energy. A conformational search of the opposite enantiomer likewise found 588 conformers within the same energy range.

The highly flexible seven-membered ring system is likely to be a major contributing factor in this case, a difficulty exacerbated by the location of the stereogenic centre on the 3-position of the ring. As a result of this, the hydroxy group has no neighbouring groups in the immediate vicinity with which H-bonding interactions can be made, and its placement along the carbon chain is also far away from the remainder of the molecule (though some long range H-bonding conformations have not been ruled out). While conformational searches to find all the conformational minima have not been successful, it appears likely that the major contributions to the experimental VCD are made by the vibrational modes of the hydroxyl group in isolation, adding to the difficulty in assigning AC to a conformationally flexible molecule.

Compounds for which a full sampling of the conformational landscape could be performed had their individual conformers catalogued and submitted for DFT calculation. Once again, an energetic window of 10 kJ mol^{-1} was used to select the conformations for which DFT calculations were performed; however, the increased flexibility of these compounds leads to a greater number of conformers within the energetic window. In addition to this, the increased size of the molecules leads to a greater number of nuclei for which DFT must be applied as well as an increased number of permutations in conformation available. As a result, the DFT calculations for this set of compounds (especially the geometry optimisations) are significantly more time-consuming than the previous sets.



Scheme 3.3 Compounds provided by AstraZeneca.

3.4 Compounds added for library analysis

Together with a few of the compounds provided by AstraZeneca, the finalised comparison algorithm was tested on a further 16 compounds which were added to the library for analysis in Chapter 5.

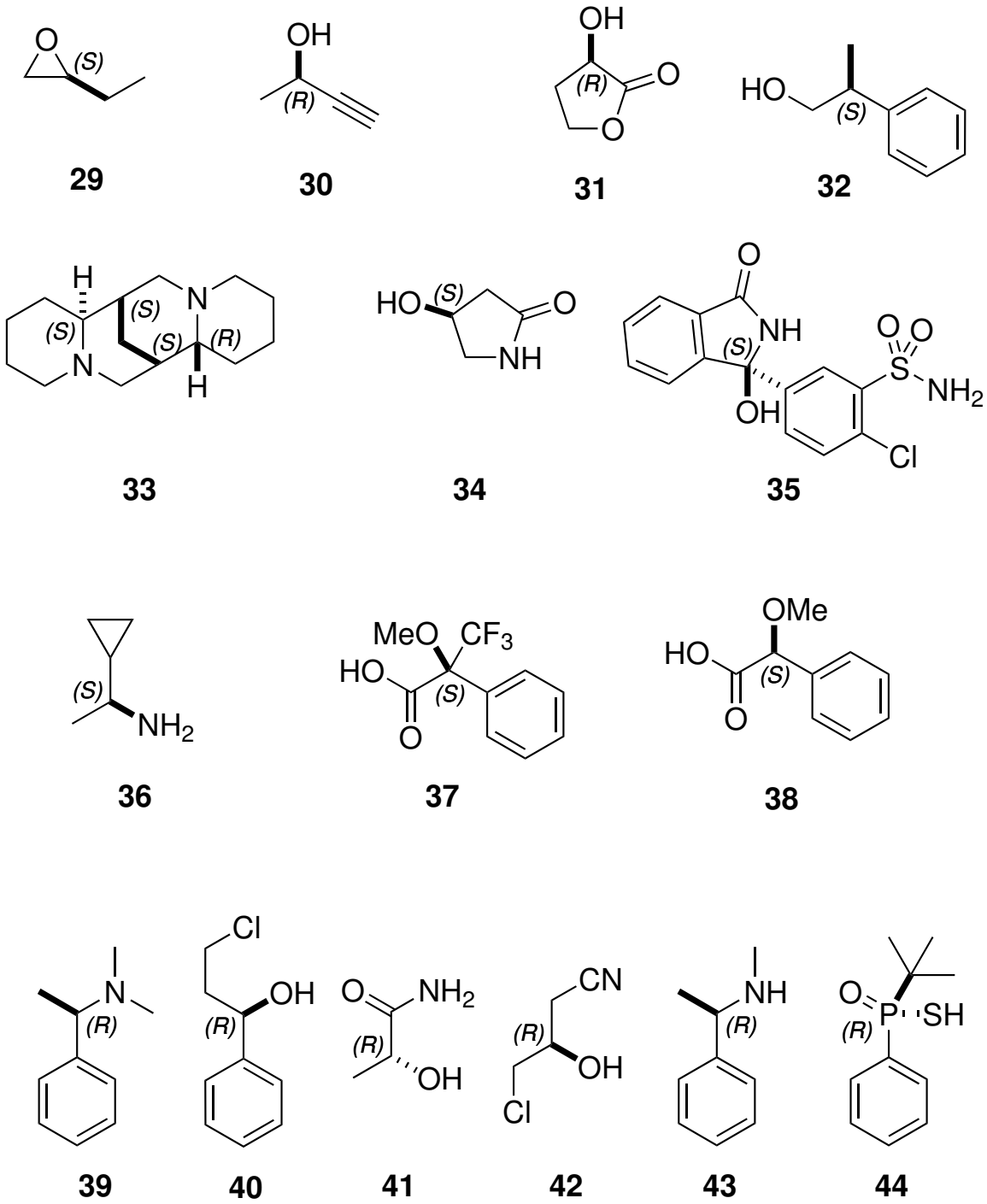
Along with adding to the degree conformational flexibility, this library also contained various pharmaceutically relevant molecules. Chlortalidone **35** is a drug usually administered as a racemic mixture, which for the purposes of this test was separated into both enantiomeric forms. Several other compounds on this list are chiral building blocks commonly used in organic synthesis and drug development.

VCD measurements for these compounds were likewise taken on both BioTools and Bruker instruments, and the relevant data is referenced in Chapter 5 where appropriate.

The shape of the baseline noise and the artefacts present in spectra measured using the Bruker spectrometer differ significantly from those taken from the BioTools instrument.

Spectra measured on the Bruker instrument were also used in the cases of 4-chloro-3-hydroxybutanenitrile **42** and *tert*-butylphenylthiophosphoric acid **44**. For each pair of compounds, spectra of both enantiomers are available from the Bruker instrument, but only a single enantiomeric form has been measured on the BioTools spectrometer.

A full list of experimental details can be found in Table 3.1. The #enant BioT. and #enant Bruk. columns denote the number of enantiomers measured on each instrument type, while the conformers generated from the Monte Carlo conformational search are shown in the rightmost column. A full list of spectra obtained from these experiments can be found in Appendix C.



Scheme 3.4 Compounds added to the library for investigation in Chapter 5.

Compounds	#enant BioT.	#enant Bruk.	Solvent	Concn. / mol dm^{-3}	Calcd. conf.s
5	2	0	$CDCl_3$	1.0	1
6	2	0	$CDCl_3$	1.0	5
7	2	0	$CDCl_3$	1.0	6
8	2	2	$CDCl_3$	1.0	5
cis-**9**	2	0	DMSO	0.6	5
trans-**9**	2	0	DMSO	0.6	7
cis-**9**	2	0	CD_2Cl_2	1.0	5
trans-**9**	2	0	$CDCl_3$	1.0	7
10	1	0	$CDCl_3$	1.0	6
11	1	0	$CDCl_3$	1.0	7
12	1	0	-	neat	1
13	2	0	$CDCl_3$	1.0	6
14	2	0	CD_2Cl_2	0.2	5
14	2	2	$CDCl_3$	0.2	5
14	2	0	CCl_4	0.2	5
15	2	1	-	neat	4
16	1	0	$CDCl_3$	1.0	1
17	2	0	$CDCl_3$	1.0	4
18	2	0	$CDCl_3$	0.5	4
19	2	0	$CDCl_3$	1.0	4
20	2	0	$CDCl_3$	1.0	64
21	1	0	$CDCl_3$	1.0	14
22	2	0	DMSO	0.5	4
23	2	0	DMSO	1.0	11
24	2	0	$CDCl_3$	1.0	14
25	2	0	$CDCl_3$	1.0	4
26	2	0	$CDCl_3$	0.5	4
27	2	0	$CDCl_3$	1.0	588
28	2	0	$CDCl_3$	1.0	7
29	2	2	-	neat	3
30	2	2	$CDCl_3$	1.0	3
31	2	2	$CDCl_3$	0.8	2
32	2	2	$CDCl_3$	0.7	7
33	0	2	$CDCl_3$	0.4	3
34	2	0	$CDCl_3$	0.8	3
35	0	2	DMSO	0.9	9
36	1	0	$CDCl_3$	1.0	2
37	2	0	$CDCl_3$	0.9	2
38	2	0	$CDCl_3$	0.9	1
39	0	2	$CDCl_3$	0.9	3
40	0	2	$CDCl_3$	1.2	8
41	0	2	$CDCl_3$	1.2	3
42	1	2	$CDCl_3$	1.3	6
43	0	2	$CDCl_3$	0.9	4
44	1	2	$CDCl_3$	0.4	1

Table 3.1 Experimental details for the VCD measurements taken in this work.

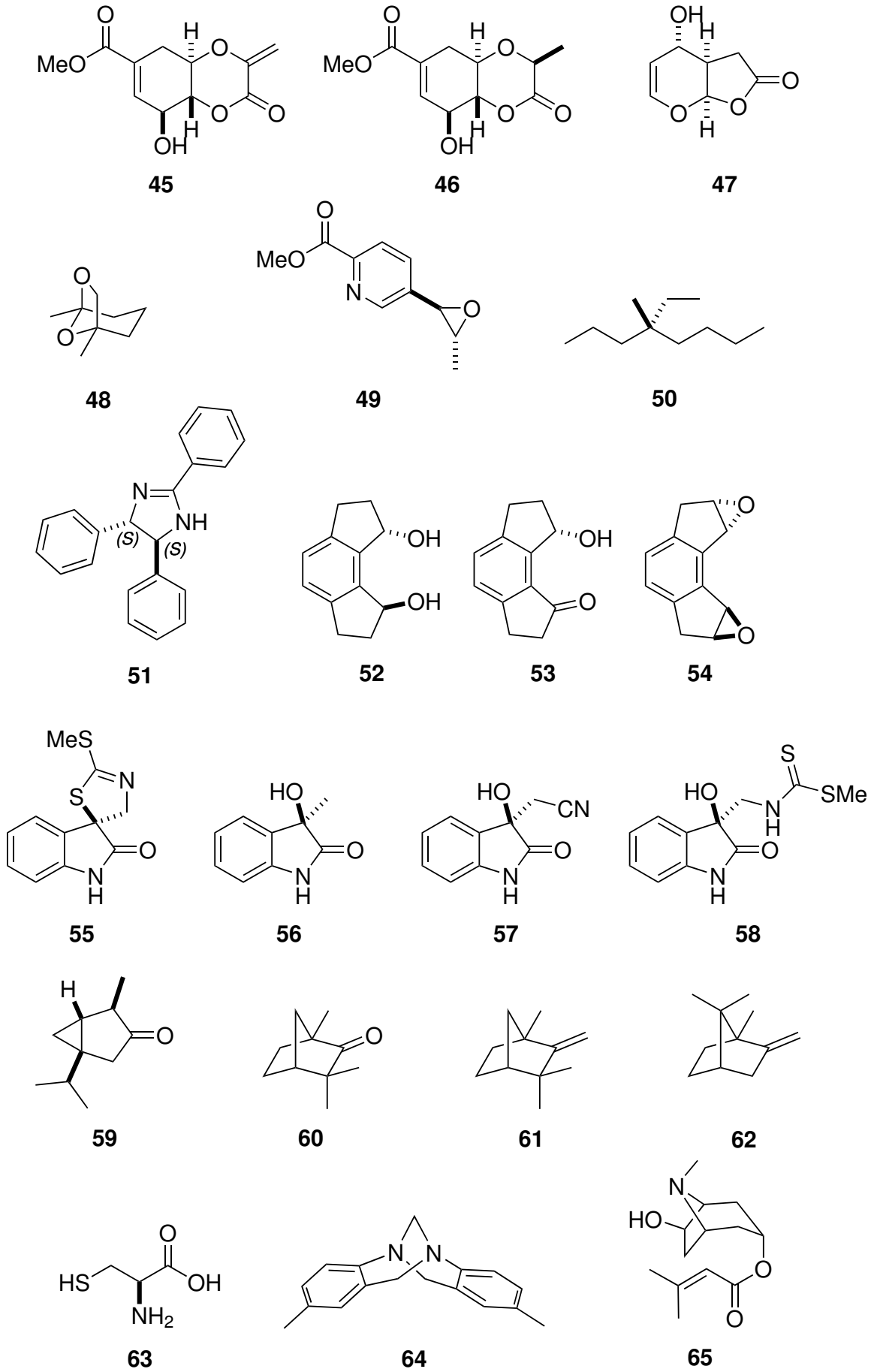
3.5 Compounds from Published Literature

Finally, a further set of compounds from previous publications [24] [53] [54] [55] [56] [57] [58] [29] [59] [60] [61] [62] [63] [64] [65] was added to the analysis in order to compare the performance of the multiplicative scoring method with other algorithms. These are compounds investigated by VCD for which experimental data is available, though the experimental details (solvent, instrument etc.) may vary between compounds. A more detailed account of the VCD measurements is included in Chapter 6.

Because of their inclusion in studies of congeners or similarly related compounds, many of these molecules have structures that are comparable with each other or serve as successive steps of a reaction scheme. This may serve to illustrate the effect of altering certain functionalities on the chiral vibrational modes of a compound. In the case of fenchone **60**, 2-methylenefenchone **61** and 2-methylenecamphor **62**, these compounds were studied alongside camphor **5** in a study [60], demonstrating the contribution of the carbonyl group to the VCD spectrum in contrast to the C=C bonded methylene moiety.

Also of interest is the compound (*S*)-4-ethyl-4-methyloctane **50** [56] [66]. This molecule, lacking any significant functionality, represents the simplest chiral alkane with a quarternary carbon atom. This makes it a cryptochiral compound, so termed due to its very small optical rotation, and leads to increased difficulty in assigning AC *via* chiroptical methods.

Many of these naturally occurring products are included only as single enantiomers. This makes baseline correction more difficult, since no mirror-image form is available with which to compare the VCD signals (and thereby separate the physical signals from background noise). However, an advantage of using the published data is that in most cases, the baseline correction has already been performed, thus cleaning up the data for analysis.



Scheme 3.5 Compounds from published data used in Chapter 6.

Chapter 4

Construction of a Comparison Algorithm

4.1 Preamble

This chapter explores the various elements necessary to construct a confidence-based AC prediction method. Several components are characteristic to a procedure for evaluating the similarity between experimental and calculated VCD spectra (and hence making an assignment of AC):

1. Firstly, the low signal-to-noise ratio of VCD spectroscopy necessitates a method of suppressing the baseline noise. This ensures that the VCD signals being analysed are a product of the solute molecule, rather than due to sources of noise inherent to the instrument, flow cell, solvent etc.

2. A method for combining the individual spectra for each conformer is necessary to generate a final calculated spectrum, which will be compared to the experimental data. This procedure may be accompanied by a weighting method, such as used in Boltzmann weighting, in order to account for the different contributions of each conformational minimum to the overall VCD spectrum.

3. Finally, a scoring step for quantitatively comparing the similarity between spectra is essential to any evaluation procedure.

The various methods applied to create each component are assessed using various datasets from the previous chapter and their performances compared.

4.2 Baseline Correction

The removal of baseline noise is critical to the quality of the results in the procedure. Before the VCD transitions and their rotational strengths can be matched to the experimental spectra, the positions of peaks must be separated from the baseline noise. It was noted that the baseline in the experimental spectrum usually appears to curve. Therefore, the baseline error for these spectra cannot be corrected by simple subtraction of a constant value.

Based on the principle that VCD signals of mirror-image compounds undergo a change of sign from one enantiomer to another, a correction method was therefore considered by taking the average of the enantiomeric spectra and using the resulting values as a baseline to subtract from the experimental data. Assuming identical solute concentrations, the baseline should thus follow the mean value between the spectra for each enantiomer on the vertical scale. This is known as the half-sum spectrum of the compound. [67]

Note that the half-sum spectrum is not representative of the baseline at all points. For example, in the spectrum of camphor (Fig. 4.1), the carbonyl C=O stretch appears as a signal at 1730 cm^{-1}. The C=O stretches in the enantiomeric forms produce VCD peaks of opposite sign as expected; however, the entire region from 1720 - 1750 wavenumbers appears to have a positive rotational strength for both enantiomers. This is due to the IR signal exceeding the bounds for ideal VCD measurement, causing the VCD signal to become saturated. It is ideal to keep the IR signal strength within the range of 0.2-0.8 absorbance units, in order for the measured VCD signal to be proportional to A_L - A_R. [37]

Subtraction of the half-sum spectrum is mathematically identical to taking the difference between both enantiomeric spectra and dividing this by a factor of two, i.e. obtaining the half-difference spectrum. This method is used later in Chapter 5 when applying the multiplicative scoring method to a library of 30 enantiomeric pairs of compounds.

Taking the half-difference spectrum removes from the analysis signals which have the same sign in the VCD across both enantiomers. While these signals are often instrument artefacts and rarely contain useful information to the assignment of AC, they can provide clues as to what instrument the spectrum was taken on and (by the relative strengths of the artefacts and the solute signals) the concentration of the solute under investigation. They are also occasionally pertinent to the molecular structure of the compound, as in the case of camphor where they arise from the insufficient filtering of the IR signal (Fig. 4.1). Therefore, while by no means intended to replace the original spectrum, using the half-difference spectrum as a

baseline-corrected version of the compund's VCD represents a simple and convenient way to provide a cleaned up source of VCD signals.

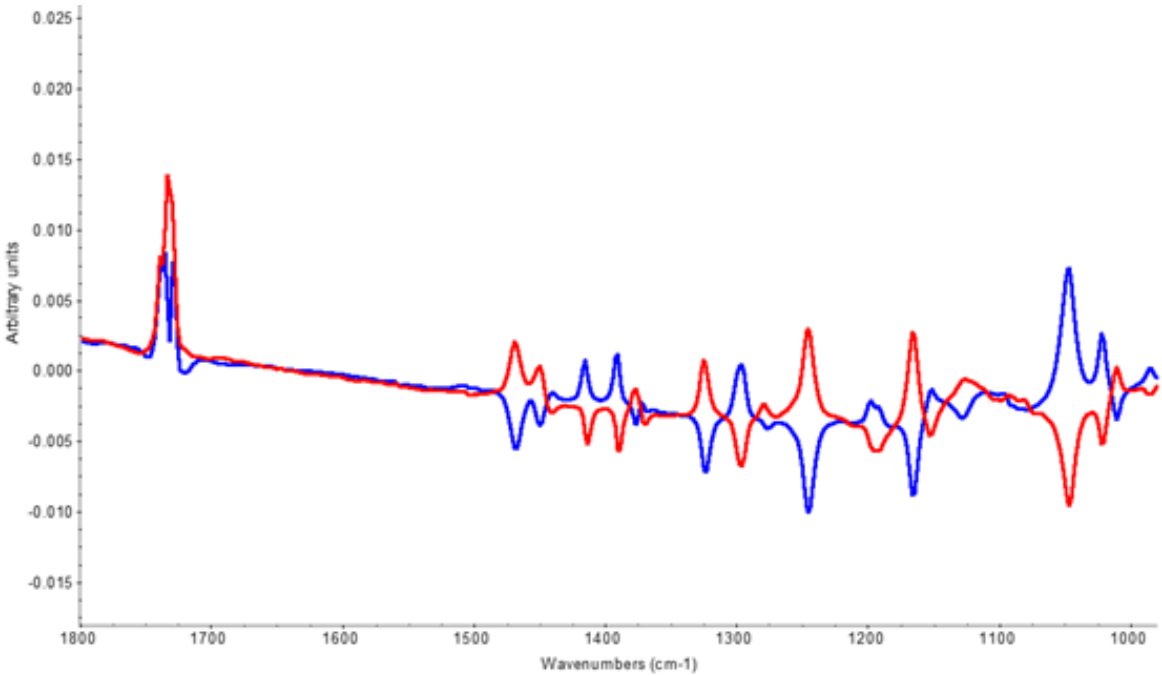


Fig. 4.1 VCD spectra of the D- and L-isomers of camphor, in red and blue respectively. The signals given by the D-isomer are inverted in the spectrum of the L-isomer; however, the two spectra do not mirror each other completely due to imperfect screening of the baseline.

4.2.1 Baseline correction for single enantiomers

When only a single stereoisomer of a compound is present for measurement, however, the half-sum and half-difference spectra are no longer available. As an alternative, the possibility of using a fixed function was explored for these cases.

The fixed function originates in the enantiomeric measurements of the initial test set. It was noted that for several compounds provided by AstraZeneca, measured using the BioTools instrument, the half-sum spectra that were being used as a baseline correction method were very similar, including similar artefact peaks as well as a distinctive curved outline within the 1000-2000 cm^{-1} region. These compounds include camphor, limonene, 2,5-dihydro-3,6-dimethoxy-2-isopropylpyrazine and the 1-amino-2-indanols, whose raw VCD data is shown in Fig. 4.2.

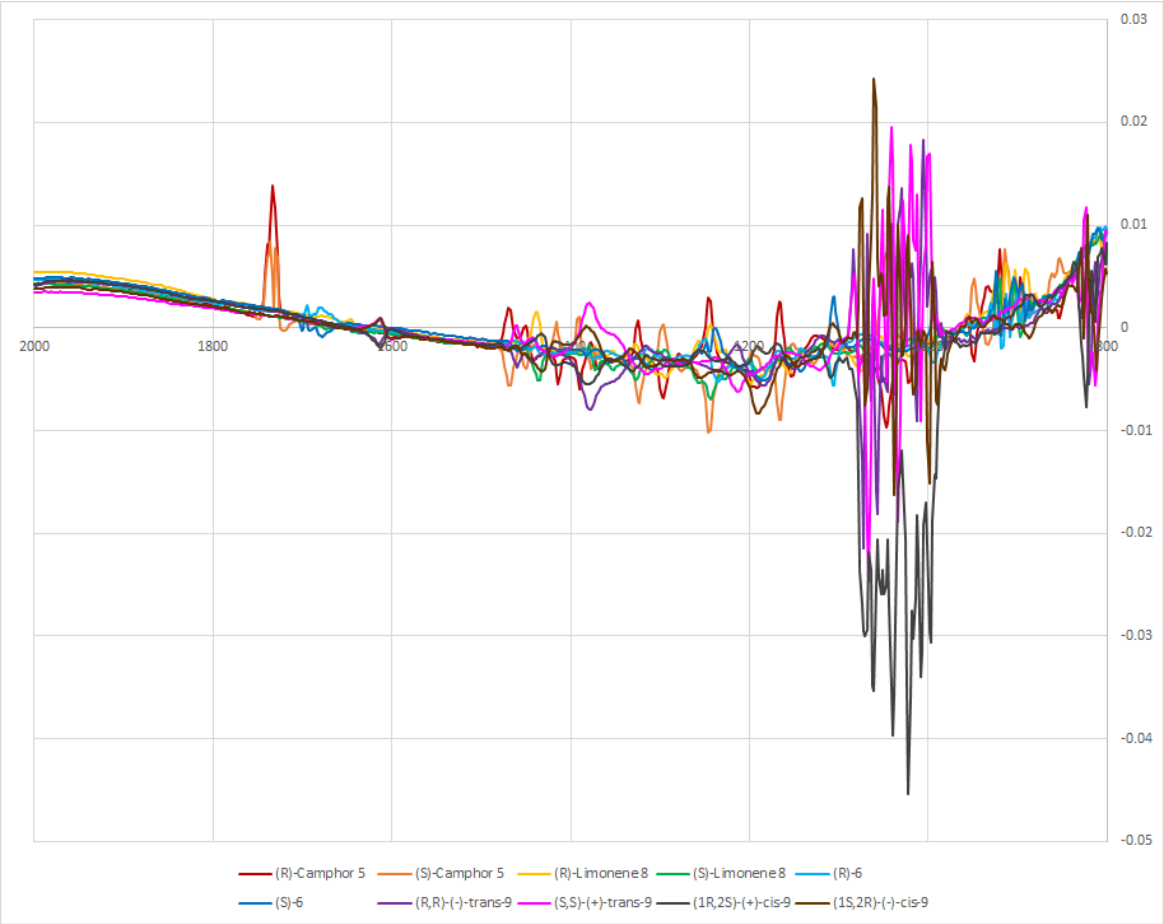


Fig. 4.2 VCD spectra for both enantiomers of camphor **5**, limonene, 2,5-dihydro-3,6-dimethoxy-2-isopropylpyrazine and the 1-amino-2-indanols.

While both diastereomers of the *trans*-1-amino-2-indanols were found to fit this baseline curve, there was a high level of random noise in the VCD spectra of these compounds in the 970-1100 cm^{-1} region. This was due to the use of DMSO as a solvent, which gives strong absorptions in this wavenumber range. The *cis*-diastereomeric forms of 1-amino-2-indanol were, therefore, not included the analysis, and the noisy between 970-1100 cm^{-1} from the *trans*-diastereomer were likewise omitted in creating the baseline curve.

For the remaining four enantiomeric pairs, the half-sum spectra were calculated, then averaged to give an overall mean curve.This gives a shape that can be roughly approximated by a curved line, as shown in Fig. 4.3. Since the curve lacked a vertical axis of reflection, thus precluding a quadratic approximation, and had only one minimum value (whereas a cubic function would have either two or none), the simplest polynomial which could be obtained to act as an empirical baseline was determined to be a fourth-order polynomial. This polynomial was calculated and fitted to the mean curve using least-squares regression fitting.

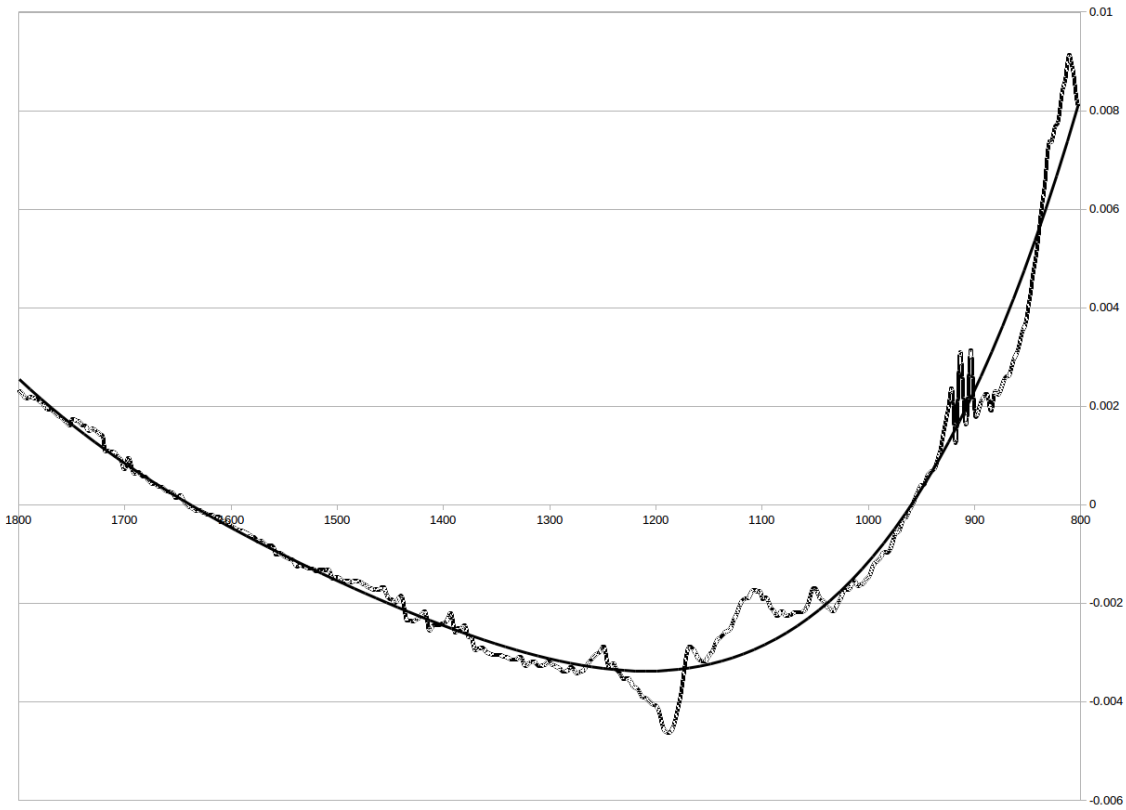


Fig. 4.3 The overall mean curve approximating baseline values for camphor, limonene **8**, 2,5-dihydro-3,6-dimethoxy-2-isopropylpyrazine **6** and the *trans*-indanols **9**, with its best fit fourth-order polynomial superimposed.

The curve is shown to fit the region above 1250 cm^{-1} closely, while there are still some rough areas unaccounted for in the 800-1200 cm^{-1} region. However, due to the majority of useful spectroscopic information being found in the region above 1000 cm^{-1}, the use of this fourth-order polynomial as a baseline correction was considered sufficient for the purposes of stereostructure determination.

4.3 Combining and weighting of conformers

4.3.1 "Ground state only" method

One possibility is to simply choose the lowest energy (ground state) conformation and use it as the final calculated VCD. This option works well for the rigid test set compounds in which only a few low energy conformers have been found by the conformational search, and the ground state conformation therefore takes up a large majority of the population. Since the bands for different conformers of the same structure generally lie at similar wavelengths, this method is applied in the one-dimensional error counting method detailed in the next section.

While convenient for making a first approximation, this method does not provide a theoretically exact prediction of the molecule's VCD, and there remains the possibility of individual signals differing from the experimental spectrum. For example, if a signal is not present in the ground state VCD spectrum but predicted to be the same sign for all other conformers, it may have a high enough intensity to be significant in the final calculated spectrum. By using the ground state only, the omission of such signals would thus prevent the calculated spectrum from matching all of the signals in the experimental VCD.

In addition, while it remains true that the lowest energy conformer dominates the outcome of the calculated spectrum summed across all conformers, there are cases where the energetic ordering of conformers as predicted by molecular mechanics differ significantly from the order of energies found by DFT calculations. This means that finding the most appropriate conformer to use may often not be known until the VCD spectra of all conformers are computed, which may be missed if the only conformer used is the ground state predicted by molecular mechanics.

4.3.2 Lineshape calculation

Before the spectra of individual conformers can be combined or weighting methods can be applied, it is important to note that for the same vibrational transition in different conformers, there is often a slight difference in frequency of the absorbed radiation (due to conformational changes causing different frequencies of vibration) and thus the signal appears at different wavenumbers. For this reason, and in order to match the broadening of signals found in the experimental VCD, it is necessary for the algorithm to be able to predict spectral line shapes.

For a species experiencing homogeneous broadening, as is the case with VCD spectroscopy, computational modelling packages give the line shape of each signal in VCD spectra as a Lorentzian distribution. Unfortunately, since there are no mathematical methods to analytically find the maximum of a sum of Lorentzian distributions, the peak absorption value of several summed conformational spectra must be approximated by:

1. Calculating the absorption value at a given wavelength for each conformer;

2. Scaling the absorption value via multiplication by the Boltzmann weight;

3. Summing the fractional combination of each predicted signal to give the overall absorption value at that particular wavelength.

4.3.3 Boltzmann weighting

Boltzmann weighting averages the VCD spectra of individual conformations according to the calculated energy of each. The relative populations of the conformers are described by the Boltzmann weight, a value derived from the conformational energy E_i by the equation

$$Population = \frac{(E_i \times e^{\frac{-E_i}{kT}})}{S(E)}$$

where S(E) is the sum of conformational energies, T is the absolute temperature (in Kelvin) and k is the Boltzmann constant, which has an approximate value of 1.38×10^{-23} J K^{-1}. In this work, room temperature is estimated at 298.15 K (that is, about 25 °C) for the purposes of VCD measurements.

Subsequently to the one-dimensional technique (of Subsection 4.5.1), this method of Boltzmann weighting was applied consistently to compounds for which the conformational search found more than one conformational minimum.

By implementing Boltzmann averaging for individual conformers, the contribution of all conformational minima within the energetic window of 10 kJ mol^{-1} can be considered. While the ground state conformer is still expected to dominate the spectrum overall, combining the weighted contributions of all the low-energy conformers towards the final spectrum will give a fuller picture of the VCD signals present, improving the accuracy of the prediction.

4.4 Scale Factor

The DFT-calculated frequencies of the vibrational modes for each molecule are not entirely accurate. Systematic errors can be included in the calculation of vibrational frequencies, in part due to the anharmonicity of molecular vibrations which are not accounted for in the quantum calculations. This is likewise true for the set of compounds included in this study, such as in the case of camphor.

Upon visual comparison it was clear that the calculated peaks at high wavenumbers (1400 cm^{-1} and above) lay at a position of noticeably higher wavenumber than the experimental ones. This tendency towards a slight exaggeration of the wavenumber axis was also confirmed in literature, and can be corrected with the use of an appropriate scale factor (SF). In order to achieve the best match, therefore, the calculated spectrum was scaled down to fit. Upon visual evaluation, a value of 0.97 was chosen as a preliminary SF, though this value was refined over the course of algorithm development.

After trialing the algorithm on enantiomeric pairs within the test set, a method was developed for obtaining the value of the wavenumber SF *via* scanning through a range of values. By default, this was done over a range from 0.90 to 1.00 in increments of 0.0001.

The first criterion used in obtaining the correct SF was to maximise the number of paired signals between the observed and calculated spectra, a process which is discussed in further detail below. Since this usually generated a range of values a second criterion was applied: the value which gave the smallest errors between paired signals was selected as the SF.

Two versions of this were trialled: one with the minimisation of the mean error and the other using the maximum error. It was expected that, since the errors follow a roughly uniform

distribution over a similar range up to the maximum error, the values given by each method would not differ greatly. The results of this study are shown in Table 4.1.

Enantiomeric pair	SF minimising err_{max}	SF minimising err_{mean}
5	0.9687	0.9703
6	0.9767	0.9768
8	0.9775	0.9788
trans-**9**	0.9087	0.9077
cis-**9**	0.9770	0.9784

Table 4.1 SF values obtained via scanning, with scanning criteria to minimise maximal error and mean error respectively.

Most of the values generated by this procedure lie within the range of 0.975 ± 0.007, with the exception of the values for the *cis*-1-amino-2-indanols. However, as the spectra for all stereoisomers of 1-amino-2-indanol had been measured in DMSO for this analysis, and owing to the difficulties encountered in making the correct assignment of AC in these compounds, both pairs of enantiomers were excluded from the eventual determination of SF, resulting in a final value of 0.9748 using the values from camphor, limonene and 2,5-dihydro-3,6-dimethoxy-2-isopropylpyrazine.

After the development of the multiplicative percentage score (Chapter 5), a more comprehensive study was performed using a library of 30 enantiomeric pairs of molecules. To spare computational time, a larger incremental value of 0.005 was used in this study, and the range to be sampled was halved by bringing the lower limit up from 0.90 to 0.95. In accordance with expectations, the SF giving a best fit overall between observed and calculated VCD spectra lies within the range of 0.97 – 0.98, with an optimal value of 0.975 for the whole dataset, though this may vary between individual compounds.

A more in-depth analysis of the different levels of theory in DFT calculation and their impact on SF values can be found in Chapter 5.

4.5 Scoring algorithms

4.5.1 One-dimensional error minimisation

The experimental VCD spectrum records, for each wavenumber value, the overall rotational strength measured inside the flow cell. Besides the molar absorptivity of the species being

studied, this value may also depend on other factors including but not limited to solvent, solute concentration, and path length traversed through the cell.

Without the aid of a solvent model or any input regarding the solute concentration, the calculated spectrum is based only on properties of the solute molecule. Thus, while theoretically proportional to the experimental spectra, the calculated spectra may differ in from it in terms of absolute value, especially when small inaccuracies in the quoted concentration accumulate from weighing errors and other measurements.

One simple way of circumventing this problem is by setting aside the intensities of the peaks and using only the sign of the signals. This gives rise to a one-dimensional method in which both the experimental and calculated spectra are simplified to a series of positive and negative signals. While information such as the signals' intensities is lost in this process, this procedure is unambiguous and simple to implement in a computational algorithm, providing a quick numerical output to any pair of experimental and calculated spectra.

The procedure performed by the algorithm is as follows: After superimposing the reduced spectra on top of each other, the closest signals from different datasets within an upper bound (for example, 100 cm^{-1}) are paired together. The value of this upper bound is intended to define the largest possible error between an experimental VCD signal and its calculated wavenumber, beyond which two signals may no longer be considered for pairing. After they have been paired to a signal from the corresponding dataset, VCD peaks are removed from the available pool, preventing more than one calculated signal from being matched with one experimental peak and vice versa.

Having removed the closest pair of experimental and calculated peaks from the pool of available VCD signals, the procedure repeats until the distance between the closest available neighbours exceeds the maximum wavenumber limit. Orphaned signals, those without a closest available neighbour from the other dataset, are ignored. Thus, for each combination of calculated and experimental spectra, a list of paired signals and the differences in wavenumber between them can be produced.

From this procedure, three values can be recorded from each pair of experimental and calculated VCD spectra:

1. The number of pairs of corresponding peaks in calculated and experimental spectra

2. The mean difference in wavenumber between corresponding peaks

3. The maximal difference in wavenumber between corresponding peaks

The baseline noise in the experimental spectrum becomes a significant issue, even after applying the baseline correction. Without a cutoff point with which to filter out small errors in the baseline, every local minimum within the negative region and every local maximum in the positive region is treated by the algorithm as an experimental signal, resulting in the perceived number of experimental peaks far exceeding the calculated ones. A threshold value is therefore required in order to separate VCD signals, which will be included in the analysis, from baseline artefacts, which can be ignored.

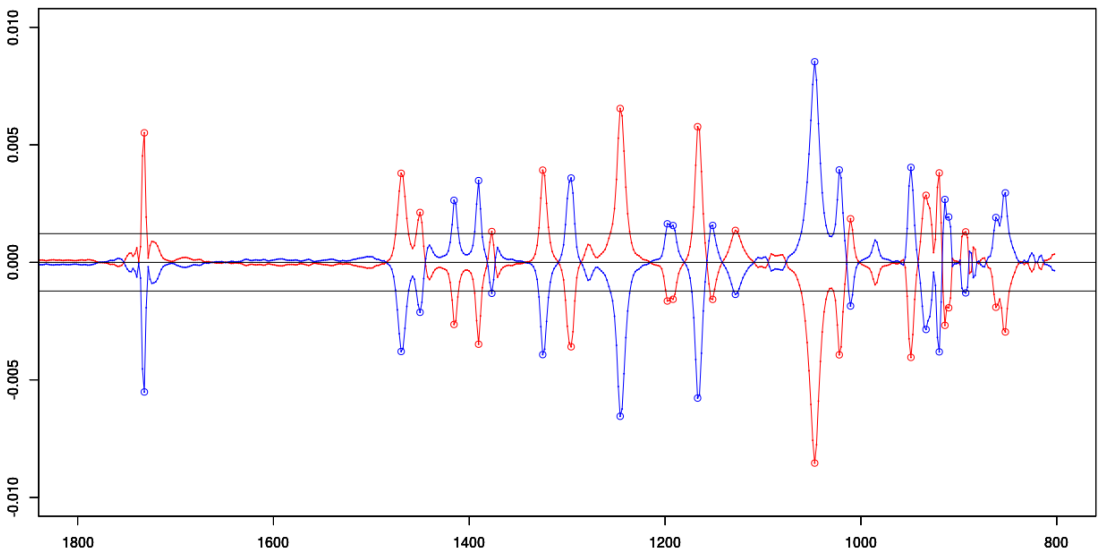


Fig. 4.4 Baseline-corrected spectrum for D- and L-camphor with positions of experimental peaks circled, after filtering out signals of low intensity using a threshold value of ±0.012.

As the strength of VCD signals is dependent on solute concentration, the cutoff value must also be set accordingly rather than using a fixed value. Using camphor as an example (Fig. 4.4), the threshold value is set at ±0.012, corresponding to one-seventh of the intensity of the strongest peak present in the 800-1800 cm^{-1} range. The selected peaks are circled on the red trace in Fig. 4.5; these appear as local maxima in the positive region and local minima in the negative region. Peaks with a magnitude of less than 0.012 are not circled and are excluded from the analysis, whether or not they are real signals from the experimental VCD.

For the upper bound of the error, used to define pairs of signals, initial values of 100 cm^{-1} and 20 cm^{-1} were considered. The value of the upper bound must neither be so stringent as to exclude corresponding peaks, nor so lenient as to include unrelated signals.

The method of one-dimensional error counting was trialled on compounds in the test set including camphor, limonene, 2,5-dihydro-3,6-methoxy-2-isopropylpyrazine and the *cis* and *trans*-1-amino-2-indanols. Early on in the analysis, even after implementation of the threshold value, it was noticed that there was a mismatch between the number of experimental peaks and those predicted by DFT calculations. While certain peaks in the experimental spectrum appear not to correspond to any predicted signals, the converse is likewise true in that there exist calculated signals without a corresponding experimental peak. This is illustrated in the case of camphor (Fig. 4.5):

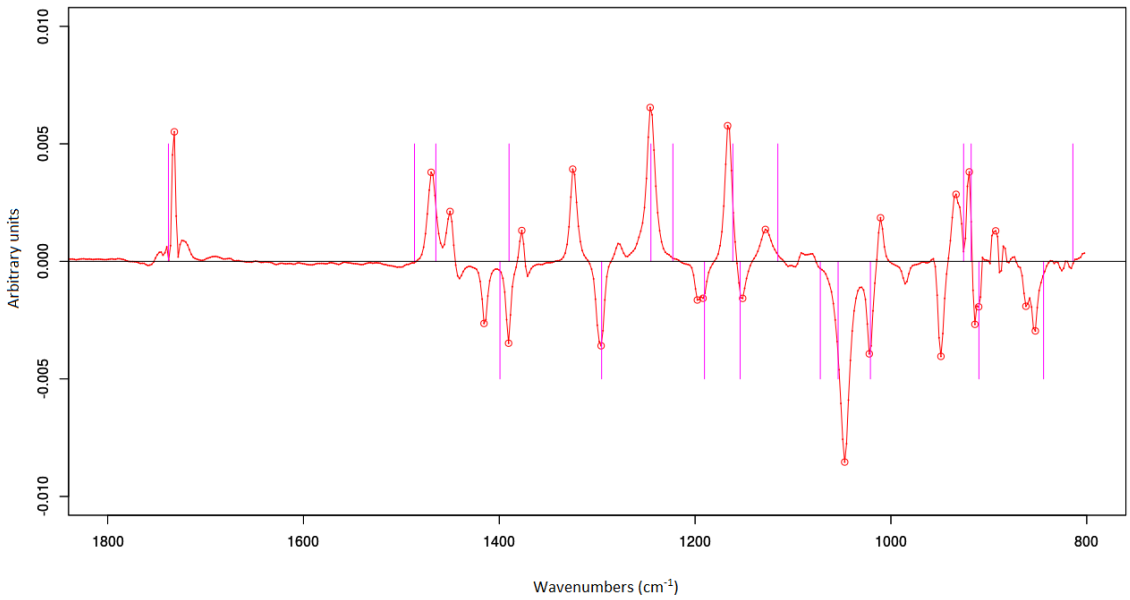


Fig. 4.5 Baseline-corrected spectrum for D-camphor with positions of calculated peaks superimposed (calculated intensities are not shown).

A direct one-to-one mapping of VCD signals is, therefore, not a straightforward task, even with the simplest and most conformationally rigid compounds. Nevertheless, it was reasoned that the greatest number of matched signals ought to come from the matched cases (where both experimental and calculated molecules had the same AC). It was further expected that the largest error values would be those originating from the mismatched cases, where the experiment and calculation had opposite AC.

4.5.2 Cumulative probabilities

In order for the fit score between two spectra to better reflect a probabilistic value, a method using probability distributions was attempted. This technique is designed to apply to cases where two configurations must be assigned to two compounds for which both experimental spectra are available, one way round or the other.

The assignment of two possible ACs to two structures, and the level of confidence with which this assignment is made, can be calculated based on the quality of the match obtained when making four comparisons among these spectra:

1) Experimental spectrum #1 with calculated spectrum #1;

2) Experimental spectrum #1 with calculated spectrum #2;

3) Experimental spectrum #2 with calculated spectrum #1; and

4) Experimental spectrum #2 with calculated spectrum #2.

The AC assignment made based on these comparisons has either two outcomes: either the first experimental spectrum corresponds to the first calculated spectrum, implying that the both the second spectra must also correspond ("EXPT1-CALC1 and EXPT2-CALC2"), or the opposite assignment is more likely ("EXPT1-CALC2 and EXPT2-CALC1").

For each combination of experimental and calculated spectra, a series of two-dimensional normal probability distributions (a "scoring distribution") is generated for each set of experimental data, centred on the location of each experimental peak. Two phenomena are accounted for in generating the scoring distribution:

1. In order that peaks of higher intensity are given greater significance, the amplitude of the scoring distribution is calculated proportionally to the intensity of the corresponding peak.

2. To accommodate the greater spread of intensity values (and thus greater variance) for stronger signals in the calculated spectra, the scoring distribution for a high-intensity peak is adjusted (also proportionally to the signal intensity) in the vertical axis to allow more lenient scoring.

In other words, the scoring distribution for a peak of greater intensity will have a greater amplitude as well as appearing stretched in the vertical axis in comparison to the scoring distribution for a lower-intensity peak.

As there is no preceding data for the standard deviation of DFT-predicted VCD wavelengths and intensities, various values ranging from as large as 5.00cm-1 to as little as 0.05cm-1 were trialed (see below). These two extreme cases respectively correspond to a broad distribution with wide tails and a narrow, sharp distribution which only awards a significant score to very close predictions.

Each predicted peak, appearing as a coordinate point overlaid on the experimental spectrum, is assigned the highest score obtained from neighbouring scoring distributions, after which the scoring distribution used is no longer available to score other predicted peaks. The scoring system begins with the highest-scoring predicted peak and proceeds to award successively lower-scoring predictions until all the predictions have been assigned a score. The sum of these scores is then calculated to estimate the quality of the match between the predicted and experimental spectra.

The program scans through a range of horizontal and vertical scale factors to find the combination of scale factors which affords the largest possible score in the matched case (Expt. #1–Calc. #1, Expt. #2–Calc. #2). This is then repeated for the mismatched case (Expt. #1–Calc. #2, Expt. #2–Calc. #1).

This method accounts for the varying size of VCD signals and affords a corresponding weighting to each match, obviating the need for a cutoff value as in the previous method. It should be noted that while these cases are referred to as “matched” and “mismatched”, there is no discernment of a “correct way round” when applying the algorithm, in order to prevent the introduction of bias into the scoring procedure. Instead, the algorithm reports the greatest value for the matched case, **M**, and likewise denotes the greatest value for the mismatched case, **m**. A numerical representation of the confidence in the assignment can then be given, dependent on these maximal scores and their relative magnitude, as detailed in the following subsection.

4.5.3 Logarithmic scoring method

Using the interpretation of the scoring distribution as a probability distribution, the scores **M** and **m** may be treated roughly as probability values. By this interpretation, the ratio of M:m represents the relative likelihood, according to the scoring algorithm, of the matched and mismatched cases being correct.

The relative sizes of these "probabilities" can be expressed as the fraction $\frac{\mathbf{M}}{\mathbf{m}}$, giving an estimate of the factor by which the matched case is more likely to be correct than the mismatched case. However, one disadvantage of using the fraction $\frac{\mathbf{M}}{\mathbf{m}}$ to express this ratio is the potential of numerical magnitudes to mislead the reader.

This situation is best illustrated by considering the extreme cases where either assignment is strongly preferred. When **M** is much larger than **m**, even a small increase in the matched score **M** can cause the fraction $\frac{\mathbf{M}}{\mathbf{m}}$ to grow without bound. On the other hand, if the mismatched case is strongly preferred, $\frac{\mathbf{M}}{\mathbf{m}}$ becomes vanishingly small, but being a fractional value between two positive numbers, it can never decrease past zero. This gives the appearance of asymmetry between both situations, exaggerating matched cases and understating mismatched ones.

This issue may be resolved by instead taking the natural logarithm of the fraction. The $\frac{\mathbf{M}}{\mathbf{m}}$ score reflects the level of confidence in the assignment "EXPT1-CALC1 and EXPT2-CALC2", with the magnitude of the score increasing with relative confidence. A large positive number indicates high confidence in the match, a large negative number indicates high likelihood of the mismatched case (i.e. the assignment is the wrong way around), and a number close to or equalling zero indicates that both cases have a similar score and that an assignment cannot be made with certainty. As an example, a result of $\ln(\frac{\mathbf{M}}{\mathbf{m}}) = 3$ would mean that the matched case is more likely than the by a factor of approximately 20 (that is, e^3, where e is the natural constant). However, since the properties for each signal are additive and not multiplicative, this interpretation only serves as a rough guideline and is not to be taken as a literal probability.

4.6 Results of each evaluation

4.6.1 One-dimensional error minimisation

Predicting the VCD spectra using the lowest energy conformation of each stereoisomer, the one-dimensional method was trialled on camphor **5**, limonene **8** and the *cis*- and *trans*-1-amino-2-indanols **9**, finding the maximal errors remaining after all possible calculated VCD signals had been paired to those in the observed spectrum. The values found are shown in Table 4.2.

In the cases of camphor (**5**) and the 1-amino-2-indanols (**9**), it is clear that pairing each experimental spectrum with its corresponding calculated form returns error values much

SF = 0.970	same enant. err/cm^{-1}	opp enant. err/cm^{-1}
(+)-**5**	13.03	56.60
(-)-**5**	16.80	49.59
(+)-**8**	27.95	23.15
(-)-**8**	28.00	22.99
(+)-*trans*-**9**	23.09	57.90
(-)-*trans*-**9**	23.12	57.91
(+)-*cis*-**9**	31.79	83.90
(-)-*cis*-**9**	31.70	83.97

Table 4.2 Maximal errors (wavenumbers) obtained for various test compounds.

smaller than those of the mismatched cases. This test also establishes the horizontal scaling factor of 0.970 as a viable one.

However, when the same technique is applied to limonene **8**, larger errors were yielded for the matched case (approximately 28 cm^{-1}) than for the mismatched case (approximately 23 cm^{-1}). Further analysis reveals that while the maximal errors were indeed larger, 16 signals were successfully paired between like enantiomers in limonene, whereas only 9 were paired in each case of unlike enantiomers. It is clear from the large disparity in the number of paired signals (by almost a factor of 2) that the chirality assignment is correct. As the number of paired signals is central to the evaluation of the fit, this value was added to the output for further studies, as shown in Table 4.3.

Experiment	match/cm^{-1}	SF	same #match	same err/cm^{-1}	opp #match	opp err/cm^{-1}
(+)-**6**	100	0.970	6	15.53	8	60.19
	20	0.970	7	7.75	2	3.77
(-)-**6**	100	0.970	6	15.50	8	60.03
	20	0.970	7	7.72	2	3.83
(+)-**8**	100	0.970	16	27.95	9	23.15
	100	0.965	15	22.05	9	54.19
	100	0.960	12	21.37	10	46.77
(-)-**8**	100	0.970	16	28.00	9	22.99
	100	0.965	15	22.16	9	54.07
	100	0.960	12	21.44	10	46.65

Table 4.3 Number of matches and maximal errors (wavenumbers) obtained for 2,5-dihydro-3,6-dimethoxy-2-isopropylpyrazine **6** and limonene **8** under various parameters.

Within the range of 0.960 – 0.970 (Table 4.3), the quality of the match for limonene increases with SF; that is, a SF value of 0.970 performs better than 0.960 or 0.965. At SF = 0.970, it is clear from the large disparity in the number of paired signals (by almost a factor of 2) that the calculated (+)-isomer and the experimental data for (+)-limonene are a good match.

The results for the updated method show a peculiarity in the spectra of 2,5-dihydro-3,6-dimethoxy-2-isopropylpyrazine **6**: when comparing maximal errors only, initial results for this pair of compounds seem to indicate successful assignment of stereostructure. However, the number of matches found adds ambiguity to this outcome. Given the large error between the last matched pair (60 cm^{-1}), it is likely that the paired signals from different datasets are not physically related, but rather paired together simply due to lack of other signals in between. This error can be removed by imposing a more stringent limit on the definition of matching pairs, to within 20 cm^{-1} of each other instead of the original 100 cm^{-1}, as shown in the top lines of Table 4.3.

Redefining the limit in this way gives the matched case a decisively larger number of matched pairs than the mismatched case. Despite the matched case now having the largest maximal errors, it is clear that its number of paired signals are conclusive in giving the correct assignment of stereostructure. Even so, the limit for defining a matched pair (100 cm^{-1}, 20 cm^{-1} etc.) in either case is an arbitrarily chosen figure. An ideal comparison algorithm would therefore have to find a means of justifying this value, or removing the need for such a definition entirely.

For the analysis of more complex molecules, it is not adequate to rely only on the number of matched peaks, or the mean and maximal error values between paired signals. These values do not ultimately convey meaningful statistical information, or give a good probabilistic indication of the AC assignment's likelihood of being correct.

Furthermore, despite the original intention for the one-dimensional method to be minimalistic and simple to implement, its application depends on several pre-defined quantities. On the vertical scale, it requires threshold or cutoff values to define experimental VCD peaks, and on the horizontal scale, upper bounds for errors (to constrain the pairing of peaks) and scale factors (for adjustment of the wavenumber scale). These quantities do not have an agreed literature value, and have needed repeated adjustment while testing the one-dimensional model on a set of simple and rigid compounds. Reliance on these arbitrary definitions thus renders this method problematic for application to more complex structures, and unable to give a quantifiable level of confidence to AC assignments based on experimental data.

Overall, the one-dimensional algorithm for comparison of VCD spectra is inadequate to meet the needs of stereostructure assignment. By making use of important information such as signal intensities, it is possible to extract more information from the spectra and avoid matching of low-intensity signals to high-intensity ones.

4.6.2 Logarithmic scoring method

Persistent discrepancies between calculated and experimental spectra, such as in the case of the 1-amino-2-indanols, led to the need for a more accurate ab initio prediction method for VCD spectra. One candidate was the correlation-consistent polarised valence triple-zeta (cc-pVTZ) basis set developed by Dunning and coworkers[51] for DFT calculations on molecules containing first- and second-period atoms (from H to Ne). This basis set, when used in conjunction with the B3LYP functional, has been shown to yield accurate predictions of VCD spectra in a study by Kuppens *et al.*[68] Furthermore, in cases where calculations using the 6-31G* basis set failed to converge towards some conformational minima in the potential energy surface, the cc-pVTZ basis set was yet able to describe the conformational landscape uniformly. Given the great sensitivity of VCD spectra to molecular conformation, and because this method was expected to give the most accurate picture of molecular conformations, it was therefore considered an ideal alternative to the 6-31G(d,p) basis set previously used.

DFT-based geometry optimisation was required in order for the calculations to describe the potential energy surface, rather than simply calculate the energy of given conformations with fixed geometry as had previously been done in single point energy calculations. This would require multiple iterations of DFT calculation and thus take up more computational time than the previous method; however, as the intention of the new prediction procedure was aimed at maximising accuracy, this was considered worthwhile.

Other than the changes to the DFT, the procedure was kept identical to the previous method. All the conformers generated by the conformational search were submitted to DFT calculation to predict individual conformational spectra, with optimisation of geometry beforehand. Variations were present upon comparing the conformational energies generated to those obtained from the previous calculations, although no dramatic changes were observed. After Boltzmann weighting and summation of spectra, the same cumulative probability algorithm was applied to generate a $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value indicating confidence in the correct assignment. The

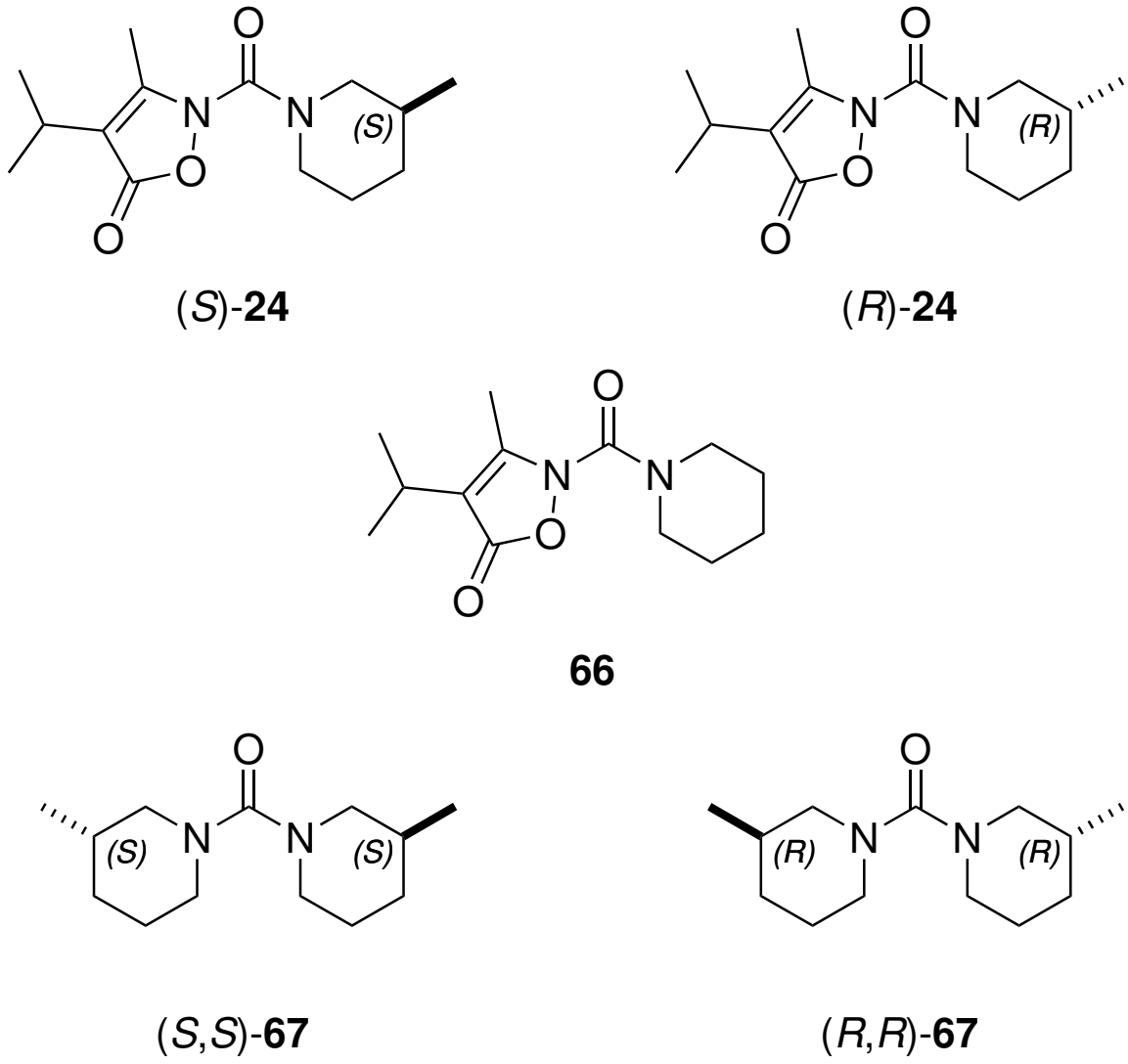
resulting values are shown, alongside their corresponding values from the previous analysis, in Table 4.4.

For four pairs of enantiomers shown, the cumulative probability algorithm returns a positive $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value using the polynomial baseline correction method. This reflects a higher confidence in the matched case than in the mismatched case. However, while analysing the 1-amino-2-indanols, the algorithm returned a negative $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value under various calculation methods, indicating a higher degree of confidence in the mismatched case than in the matched case.

In the case of the *trans*-diastereomer, the correct assignment is made under treatment by the average baseline but not the polynomial baseline. While the incorrect assignments may be attributed to the solvation of the compounds in DMSO as mentioned in the previous chapter, it is also worth noting that the comparison made under the average baseline correction method performs significantly better than that using the polynomial baseline in both cases. Since the *cis*-1-amino-2-indanols did not fit the mean curve used to derive the polynomial baseline, this outcome was at least partly to be expected; however, the fact that this correction method fails on both isomers highlights its shortcomings and points towards the need for a method that is better able to adapt itself to the experimental baseline than before.

It is worth noting that while application of the cc-pVTZ basis set significantly improves the accuracy of prediction in both cases of the 1-amino-2-indanols, even reversing the incorrect assignment in the case of the *cis*-isomer, there are also several cases in which the $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value decreases under this treatment, among which the value for camphor is especially dramatic. While the calculated spectrum from the cc-pVTZ method is expected to be the more accurate one, it is possible that the program may return a high fit score for a mismatched pair of spectra while near the lower or upper bounds of the scanning range for horizontal scale factors. Accurate scaling of predicted spectra is therefore a crucial requirement for future quantitative comparison algorithms. The use of wavenumber scaling is investigated further in the following chapter, with the multiplicative percentage score being tested over a range of SFs.

Before assessing the behaviour of the cumulative probability comparison algorithm when applied to highly complex molecules, a further trial was carried out to determine the algorithm's capability of determining AC for a compound slightly larger than the others studied, with an intermediate degree of flexibility. Compound **24** was one such structure, possessing an isoxazolone and a piperidine ring, joined by a bridging urea moiety.



Scheme 4.1 Compound **24** and the various calculated structures used in modelling it.

B3LYP	6-31G(d,p), s. p.		cc-pVTZ, opt.	
	Polynomial	Half-diff.	Polynomial	Half-diff.
5	1.3	1.8	0.1	0.1
6	0.7	0.2	0.3	0.1
7	0.1	0.3	0.1	0.5
8	0.2	0.1	0.4	0.0
trans-**9**	-0.6	0.2	0.5	0.9
cis-**9**	-4.8	-2.2	-1.1	0.1
24	-	-0.73	-	1.11
24, mod.	-	-0.54	-	0.34

Table 4.4 $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ values returned for various test compounds under two DFT calculation protocols, with the results of varying baseline correction methods shown. Compound **24** was tested against a further set of calculated test structures, for which the results are shown under "**24**, mod.".

A conformational search was performed on compound **24**, both enantiomers of which had experimental VCD data available. These spectra were used to assess the ability of the algorithm to assign the single unknown stereocentre on the methyl-bearing position of the piperidine ring, whose configuration is marked on the structure. A total of 14 conformers was generated for each enantiomer. These were then used to generate individual conformational VCD spectra, which were then Boltzmann averaged to give the overall calculated spectrum.

Despite being measured with the same VCD spectrometer as those from which the polynomial baseline had been derived, the average baseline correction method was applied to this case. This was due to the fact that the mean curve as seen in previous cases was visibly a poor fit for the experimental data.

Since the chirality of each sample was not originally known when taking the experimental spectra, the certainty of the algorithm's assignment must be treated with caution. However, from a visual comparison, an initial guess was made matching the strong positive signal near 1700 cm^{-1} (blue trace, Figure 4.9) with structure (*S*)-**24** and the negative signal with structure (*R*)-**24**. This preliminary assignment was used only to determine the "matched" and "mismatched" combinations and would not lend any bias to the outcome of the algorithm.

DFT calculations were run using single-point energies with the B3LYP functional and the 6-31G(d,p) basis set, and separately using the cc-pVTZ basis set with geometry optimisation. The resulting $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ values are given in Table 4.4.

While the disagreement between the two DFT methods leads to uncertainty in the final assignment, it can be seen that the cc-pVTZ calculation gives a result largely in agreement

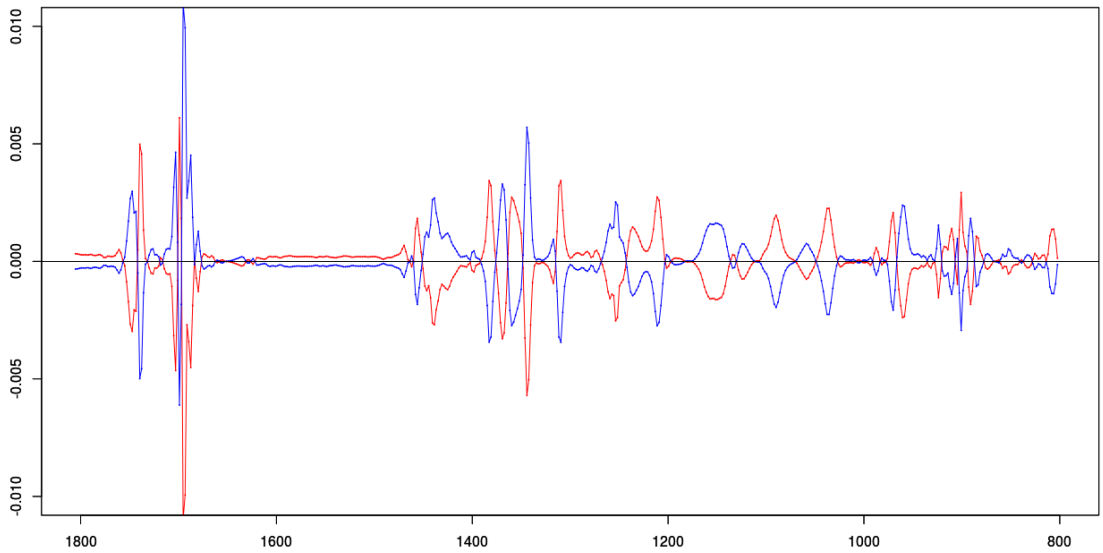


Fig. 4.6 Baseline-corrected experimental spectra for the two enantiomers of **24**; the blue trace is initially guessed to be that for the (*S*)-form and the red trace for the (*R*)-form.

with the initial guessed combination. Given the greater degree of accuracy in the cc-pVTZ calculation, and the greater magnitude of the $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value, it is thought that the initial assignment is therefore the correct one.

Regarding the reasons for the disagreement, it is possible that the highly conformation-sensitive VCD prediction may be affected by the orientation of the plane of the isoxazole ring relative to the urea subunit, as well as to the piperidine methyl group. One way to isolate the VCD prediction from the effects of the isoxazolone unit was by calculating predicted VCD signals for the theoretical structure **67** (Scheme 4.1), for which the methyl-carrying piperidine was duplicated onto the other side of the urea bridging unit.

The conformational search for (*S*,*S*)-**67** and (*R*,*R*)-**67** generated 13 stable conformers each, whose Boltzmann averaged spectra were compared to the experimental spectra for (*R*)- and (*S*)-**24**. As in the previous section, two different DFT calculations were run and the $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ values resulting from each are shown (Table 4.4), under the heading of "**24**, mod."

The results for the B3LYP/6-31G(d,p) single point calculation give a $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ of -0.54, while the B3LYP/cc-pVTZ optimisation returns +0.34. Although the level of confidence is reduced (moves closer to zero) in both cases, which is to be expected due to the absence of the isoxazolone unit in structure 12, it is observed that the replacement does not resolve the

disagreement between the two DFT methods. Also, the $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value in the cc-pVTZ case is significantly reduced, such that the 6-31G(d,p) method now has greater confidence in the opposite assignment. Thus, due to the disagreement between the two DFT methods, the assignment of stereostructure remains inconclusive.

A further factor affecting the calculated spectra may be the orientation of the piperidine methyl group relative to the urea carbonyl group. One proposed solution would thus be an analysis of the individual conformers of compound **66** (Figure 4.10). This may be achieved by using the calculated conformational VCD spectra separately, rather than applying Boltzmann averaging. One advantage this affords over using structure **67** is that it represents only a slight change from the parent compound **24**, so that the predicted VCD spectra are more likely to closely resemble those of (*R*)- and (*S*)-**24**.

Finally, in order for the comparison algorithm to be successful, it would have to give accurate stereostructure assignments beyond the rigid structures previously investigated and perform well on larger complex molecular systems. To test this, a list of compounds with published VCD data was compiled from various literature sources. Predicted VCD spectra for these compounds were calculated and compared to the published ones in the literature. Since these data were acquired from literature values, no baseline correction was applied to the experimental VCD signals. Where VCD data were only available for a single enantiomer, the "experimental" spectrum for the opposite enantiomer was created by reflecting the original spectrum in the horizontal axis.

The first group contained mostly small molecules used as standard compounds for VCD experiments. Some of the compounds investigated were chlorine-containing molecules, as seen in the examples of epichlorohydrin and 2-chloropropanoic acid (2-CP). Use of the cc-pVTZ basis set is not ideal when calculating spectra for molecules containing Cl atoms, which would require further functions to be accurately described. Nevertheless, it was trialed as a possible alternative DFT calculation method to the previous B3LYP/6-31G(d,p) combination, in order to ascertain whether it was able to provide more accurate stereostructure assignments.

While the 6-31G(d,p) method makes several incorrect assignments, these are corrected to some extent by the cc-pVTZ basis set calculation. Similarly to the first set of compounds, the cc-pVTZ calculated set seems to give lower confidence in its assignments, even those that are correct. Remarkably, the 2-chloropropanoic acid in which dimerisation due to hydrogen bonding was expected to give anomalous results gave the best $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value among this group.

Compound	6-31G(d,p), sp.	cc-pVTZ, m.
12	-0.4	-0.2
10	-0.3	0.1
11	0.8	0.1
14	1.7	0.5
15	-0.8	0.0

Table 4.5 $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ scores for various simple compounds from literature data.

Further trials were carried out on a range of larger literature compounds, **45**-**50**. These structures were mostly natural products for which the stereochemistry had been determined using a variety of techniques including VCD spectroscopy. Owing to the complex structure of these compounds, and therefore the inefficiency of the geometry optimisation method, only 6-31G(d,p) calculations were available to use in predicting VCD spectra.

Compound	Score
45	23.5
46	-0.2
47	0.0
48	-5.5
49	11.3
50	-26.9

Table 4.6 $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ scores for larger compounds from literature data.

As seen in Table 4.6, the results vary drastically across compounds, even between phyllostin and scytolide where the structures only differ in one double bond. In comparing the DFT calculated spectra with the published experimental data, the algorithm makes three incorrect assignments for six pairs of structures, the algorithm lacks reliability in providing the correct assignment. Thus, these compounds likely represent the limit to the ability of the cumulative probability model to predict absolute stereostructure, at least under DFT calculations using the 6-31G(d,p) basis set.

The $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ values given in a few cases are very large (23.5, 11.3, -26.9). These likely arise when the values of either **M** or **m** are very close to zero, leading to an unrealistically large logarithmic value—for example, a $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value of 23 would indicate that the cumulative probabilities from one case are larger than the other by approximately $e^{23} = 9.7 \times 10^9$ times. Such values can be misleading and do not reflect the physical reality. The use of $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ is therefore not a good scoring system for the degree of confidence in assignment in these cases.

As expected, the cumulative probability algorithm performs better than the simpler one-dimensional algorithm on small, rigid structures. It also obviates the need for an approximate probability value to account for missing peaks. However, it remains lacking when applied to more complex structures with a greater number of atoms or containing multiple rings, for example.

Furthermore, despite the new polynomial correction method, baseline noise remains a problem — as can be seen in the case of the 2-amino-1-indanols. Possible improvements that may be considered for a later correction method include an interactive baseline which is generated individually according to the experimental input. Given the anomalous behaviour of the algorithm in cases where either the matched or mismatched combinations yields a very small fit score (M or m close to zero), a future algorithm will possibly require a warning system to alert the user when such cases arise. The scanning method to obtain correct scaling factors may also be replaced with a direct calculation using the predicted IR signals (which have the same frequency and thus the same wavenumber as those for calculated VCD, but different amplitudes). This would improve the speed of the algorithm and avoid cases where incorrect scale factors lead to incorrect predictions.

A major flaw in using the $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ values was that when either the cumulative score for the matched case **M** or the mismatched case **m** was close to zero, the resulting ln score would be a positive or negative number of very large magnitude. This would often be the case even when neither cumulative probability (**M** or **m**) was large.

Chapter 5

The Multiplicative Percentage Scoring Value

5.1 Overview of the Multiplicative Percentage Score

This Chapter details the rationale behind the usage of the Multiplicative Percentage Scoring value as an indicator of similarity between calculated and observed VCD spectra, and its application to a library of VCD measurements from available compounds.

The multiplicative percentage score is designed with an aim to provide an intuitive measure of spectral similarity as well as to avoid the misleading numerical outputs of previously discussed procedures are evaluating the fit score, such as the $\ln(\frac{\mathbf{M}}{\mathbf{m}})$ value, where an experimental spectrum gave a poor fit for the calculated spectra of either or both of the possible configurations.

While the concept of cumulative scoring is useful and worth retaining, it is also necessary for the presentation of the data to be changed in order to more closely resemble a probabilistic model. With this in mind, the scoring value is no longer designed to apply solely to a two-to-two assignment, though this is often the case for many of the AC assignments in the library.

A requirement for this method would be that a perfect match between calculated and observed spectra would return a maximum score, +100%, while a complete inversion of peaks between the two would similarly return -100%. In addition to this, the score should be close to zero

when neither configuration's calculated VCD spectrum gives a good match for the observed one. In this respect, the new method is similar to integral-based scores already in use [30] [69] [70]. A solution to fulfil these requirements comes in the form of a multiplicative scoring method which first multiplies together individual values along the wavenumber axis, then sums the resulting products.

The procedure followed for calculating the score is as follows:

1. The calculated and observed spectra are superimposed.

2. For each data point along the experimental wavenumber axis, the calculated and observed rotational strengths at the given wavenumber value are multiplied. The product is positive if both rotational strengths have the same sign and negative otherwise.

3. Take the sum of these products over the entire spectrum, Σ_{CO}, where C and O refer to the calculated and observed spectra respectively.

4. By superimposing each spectrum over itself, repeat to find the values of Σ_{CC} and Σ_{OO}.

5. The quality of the match is evaluated using the expression:

$$\text{Fit score} = \frac{\Sigma_{CO}}{\sqrt{\Sigma_{CC}\Sigma_{OO}}} \times 100\%$$

This is a modification of the SimVCD integral as proposed by Shen *et al.* in 2010:

$$\frac{I_{CO}}{I_{CC} + I_{OO} - |I_{CO}|}$$

A distinction between the multiplicative fit score and Shen *et al.*'s SimVCD integral is that in the multiplicative scoring method, scaling of the rotational strengths is included in the calculation of the percentage score. As a result, the multiplicative percentage score is not dependent on absolute rotational strength of the experimental VCD absorptions, but only on their relative intensities. This further allows the baseline correction *via* enantiomer subtraction to be performed without affecting the quality of the match between calculated and observed spectra.

Among the methods used in this dissertation, the multiplicative percentage score is the most heavily biased towards high-intensity VCD signals. Given that high-intensity peaks can nevertheless be non-robust and therefore give the incorrect assignment of AC, a level of

uncertainty remains in this method. Even so, the strong absorption of these signals in the VCD makes them easy to separate from the background noise and artefacts. They are thus more reliable than other signals from an experimental point of view.

5.2 Trial against a range of compounds

For the library of compounds used to evaluate the multiplicative percentage scoring method, a set of compounds used was chosen such that varying degrees of conformational flexibility were present, ranging from structures with only one stable conformation to those with many low-energy conformers. These were obtained *via* AstraZeneca, Sweden and used without further purification.

As several of these were taken from AstraZeneca stores, many drug compounds and drug precursor molecules were included in this study. Various compounds with less-common functionalities were also used, such as the sulfonamide group present in chlortalidone **35**, as well as the chiral phosphorus atom in TBPTA **44**.

5.2.1 Experimental details

Spectra acquired in this study were all measured in $CDCl_3$ solution, at a concentration chosen such that the absorption bands lay within the 0.2 - 0.8 absorption range in the IR. While the samples must be sufficiently concentrated for the detection of VCD signals, they could not be made too concentrated in order to avoid the formation of dimer aggregates as in the case of carboxylic acid groups. The concentrations of each sample are listed in Chapter 3.

IR and VCD spectra for the compounds were recorded on a BioTools ChiralIR-2X spectrometer and a Bruker TENSOR FTIR spectrometer with a PMA50 module for polarisation modulated measurements, allowing the comparison of data acquired on different instruments. Solutions of the samples were held in a transmission cell with BaF_2 windows at a fixed path length of 100 μm. Both IR and VCD spectra were recorded at a spectral resolution of 4 cm^{-1} by accumulating approximately 24,000 scans over an accumulation time of 6 hours.

5.2.2 VCD calculation

Vibrational line broadening was simulated by assigning a Lorentzian band shape with a HWHM of 5 cm^{-1} to the calculated dipole and rotational strength. Using the `scipy.optimize` function in Python [71], the scaled calculated spectra were shifted along the wavenumber scale in order to maximise the absolute value (whether positive or negative) of the multiplicative percentage score between the two spectra. Once again, the sign of the output score was neglected when optimising the fit score in order to prevent bias from entering the algorithm.

Boltzmann weighting was applied to combine the calculated spectra for individual conformers into a single calculated spectrum. Because accuracy of the VCD signals was important in this study, baseline correction was performed by subtraction of opposite enantiomers. Pairs of enantiomers were measured under identical conditions, and the resulting spectral traces subtracted from one another to cancel out baseline noise and artefacts.

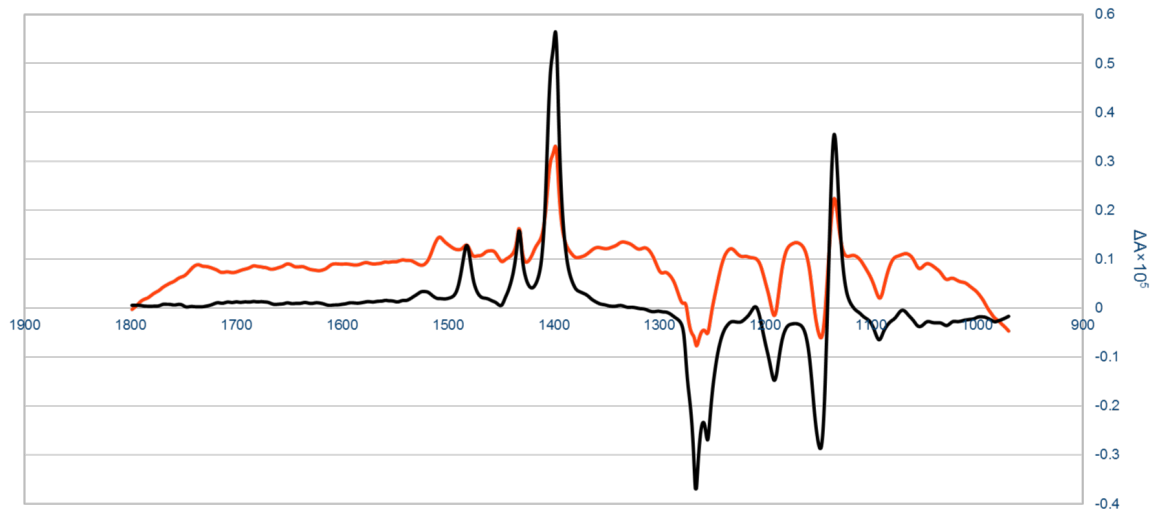


Fig. 5.1 Demonstration of the baseline correction method used. The (*R*)-enantiomer (red) is overlaid on the enantiomer-subtracted spectrum (black), comprising the data of the (*R*)-form with the (*S*)-form subtracted.

Due to the resulting spectra being calculated by subtraction of the opposite enantiomers, the intensity of the signals in the baseline-corrected spectra is doubled when compared to a single-enantiomer spectrum. Having suppressed the baseline noise, however, this increase in signal intensity is tolerated by the multiplicative scoring method, which is only dependent on relative intensities of the VCD signals.

This multiplicative percentage score is applied to a set of 30 enantiomeric pairs comprising compounds from previous analyses as well as several new additions. The outcome of the

analysis is presented in heat map format, in Fig. 5.2 - Fig. 5.5. Negative scores are shown in red and positive results in blue, with magnitude of the score denoted by colour intensity.

5.3 Optimisation of the Method

Given the high dependency of the comparison outcome on the SF used, a range of SFs was investigated in order to examine the scaling profiles of the multiplicative percentage method as well as confirm the optimal range of SFs. Under previous analyses, optimisation algorithms most often converged on values between 0.970 and 0.980 for the best fitting between observed and calculated spectra. In order to verify this, and to ensure that values over a suitable range were sampled, SFs from 0.950 to 1.000 were tested in increments of 0.005, giving a list of 11 SFs in total.

The compounds in the heat maps (Fig. 5.2 - Fig. 5.5) are ranked in order of average fit score, between 0.965 and 0.985. Within this range, as can be noted from the diagrams, the highest-scoring compounds reach a maximum in their fit score.

As expected, horizontal SFs in the centre of the range 0.950 to 1.000 perform better than those on either end. Among the variables tested, the largest increase in match score is achieved by applying geometry optimisation to the structures in DFT.

On average, the match score is also improved by respective application of the cc-pVTZ basis set and the B3PW91 functional. Accordingly, the highest average score is reached by geometry optimisation at the B3PW91/cc-pVTZ level of theory. This result is within expectations, owing to cc-pVTZ being a triple-zeta basis set, as opposed to 6-31G(d,p) which is double-zeta. The match scores under various methods of calculation are compared in Fig. 5.11, in which the B3PW91/cc-pVTZ level of theory is shown to have the highest average score.

Examining the optimised calculation under B3PW91/cc-pVTZ in further detail (Fig. 5.6 - Fig. 5.8, plotting match score against SF also reveals differences particular to the spectra of certain compounds. Scanning across the range of horizontal SFs, many compounds increase to a maximum value within the range of 0.970 – 0.980 before decreasing again near 1.000. The compounds presented in the heat maps are listed in decreasing order of the averaged match score between the SF values of 0.965 and 0.985.

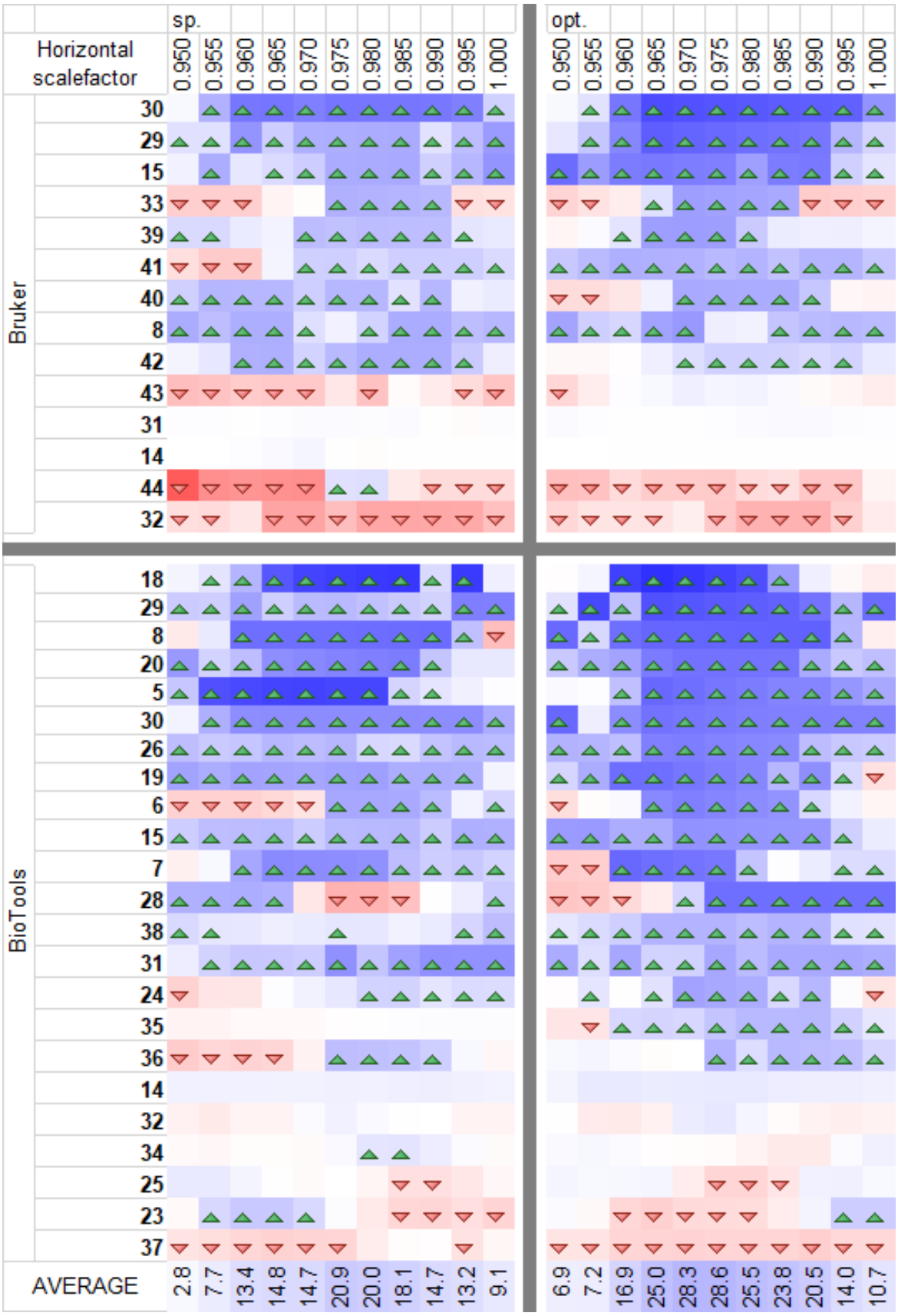


Fig. 5.2 Heatmap of multiplicative percentage scores obtained from single point (left) and geometry optimised calculations (right) at the B3LYP/6-31G(d,p) level of theory. Scores with a magnitude greater than 10% are marked with upward and downward facing arrows for positive and negative results respectively.

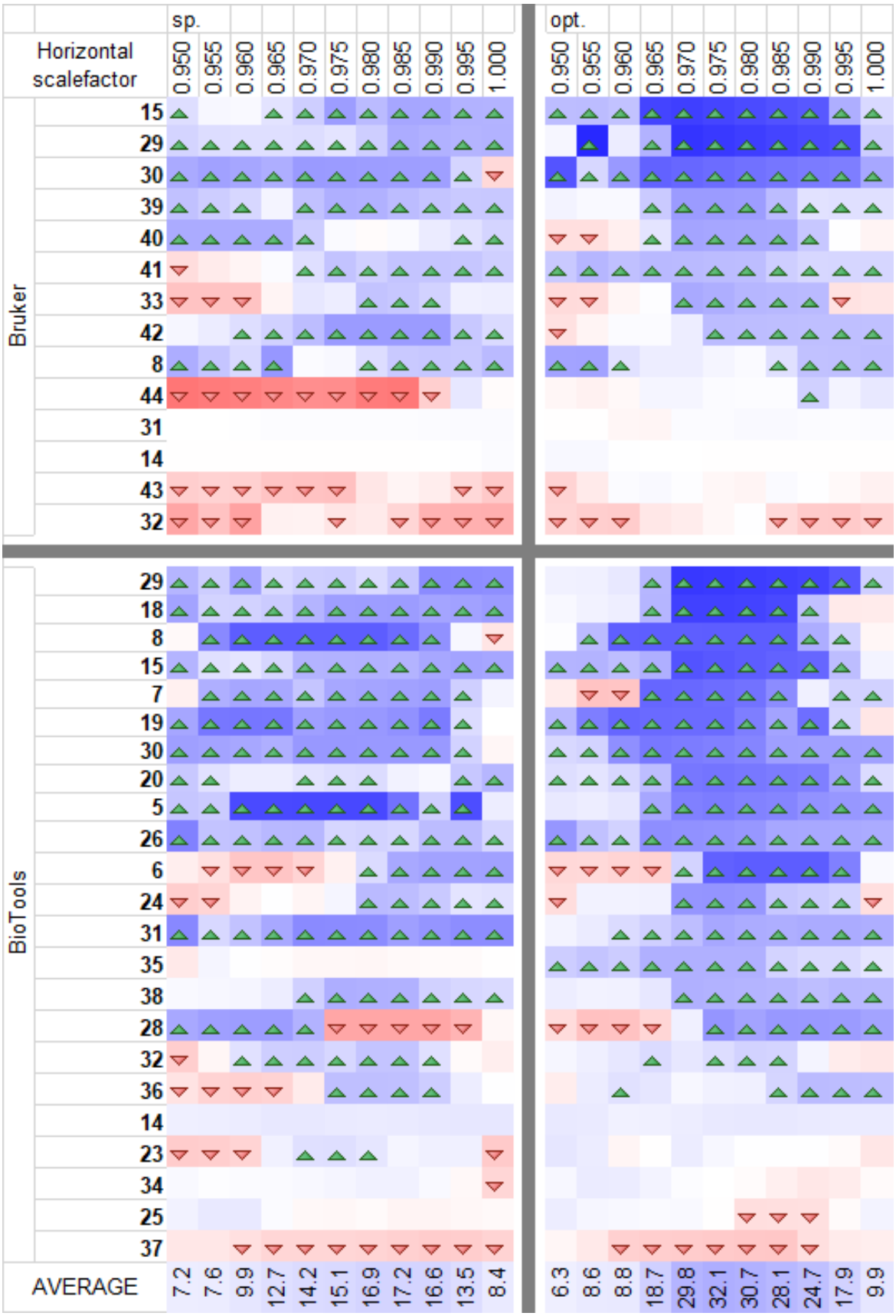


Fig. 5.3 Heatmap of multiplicative percentage scores obtained from single point (left) and geometry optimised calculations (right) at the B3LYP/cc-pVTZ level of theory. Scores with a magnitude greater than 10% are marked with upward and downward facing arrows for positive and negative results respectively.

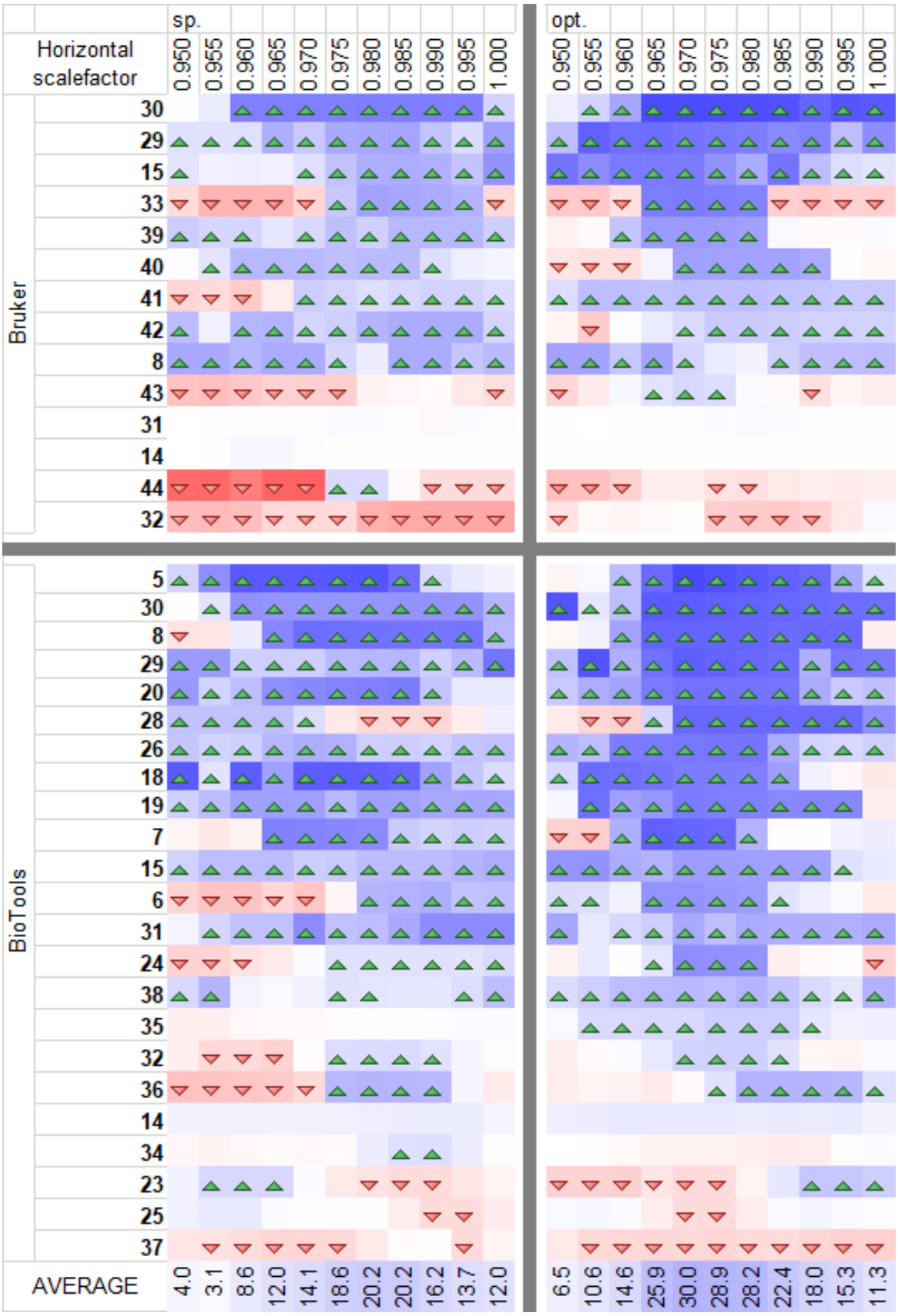


Fig. 5.4 Heatmap of multiplicative percentage scores obtained from single point (left) and geometry optimised calculations (right) at the B3PW91/6-31G(d,p) level of theory. Scores with a magnitude greater than 10% are marked with upward and downward facing arrows for positive and negative results respectively.

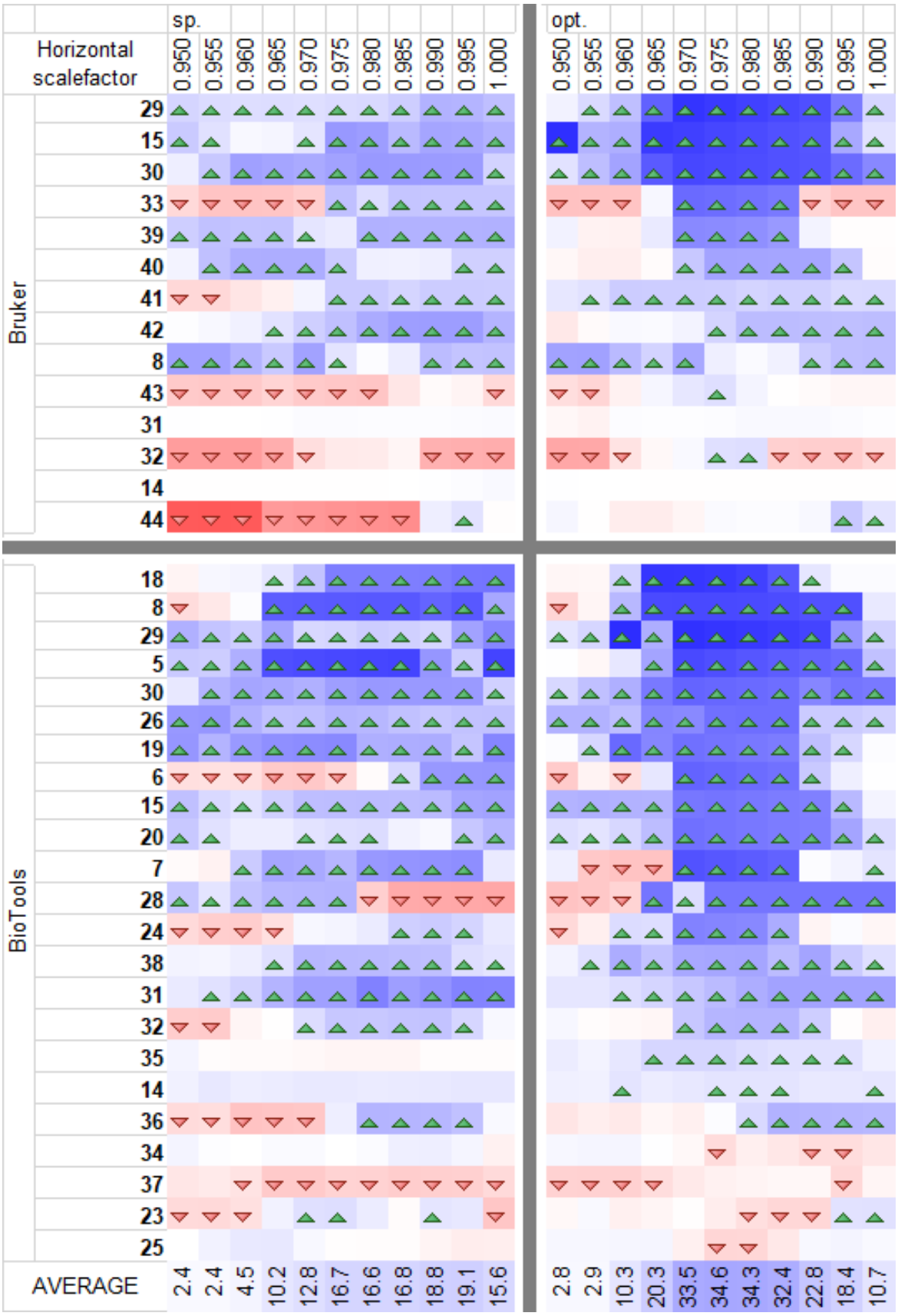


Fig. 5.5 Heatmap of multiplicative percentage scores obtained from single point (left) and geometry optimised calculations (right) at the B3PW91/cc-pVTZ level of theory. Scores with a magnitude greater than 10% are marked with upward and downward facing arrows for positive and negative results respectively.

The quality of spectra acquired also has an impact on the match score. This is evidenced by the spectra of the compounds limonene **8** and 2-phenylpropan-1-ol **32**: for these two compounds, the spectra acquired on the Bruker instrument consistently return near-zero match scores, not sufficiently positive or negative to make an assignment. However, spectra given by the BioTools instrument still return sufficiently positive results within the SF range of 0.965 – 0.985. These appear as the dark green and yellow traces in Fig. 5.6 and Fig. 5.7. The plateau curve is clearly seen in the case of the BioTools spectra but not in the Bruker. This disparity in scaling profile between the BioTools and Bruker spectra shows the quality of the spectra acquired on the BioTools instrument to be higher for these two compounds. For this reason, the data for these two compounds acquired on the BioTools spectrometer are used in the following analysis.

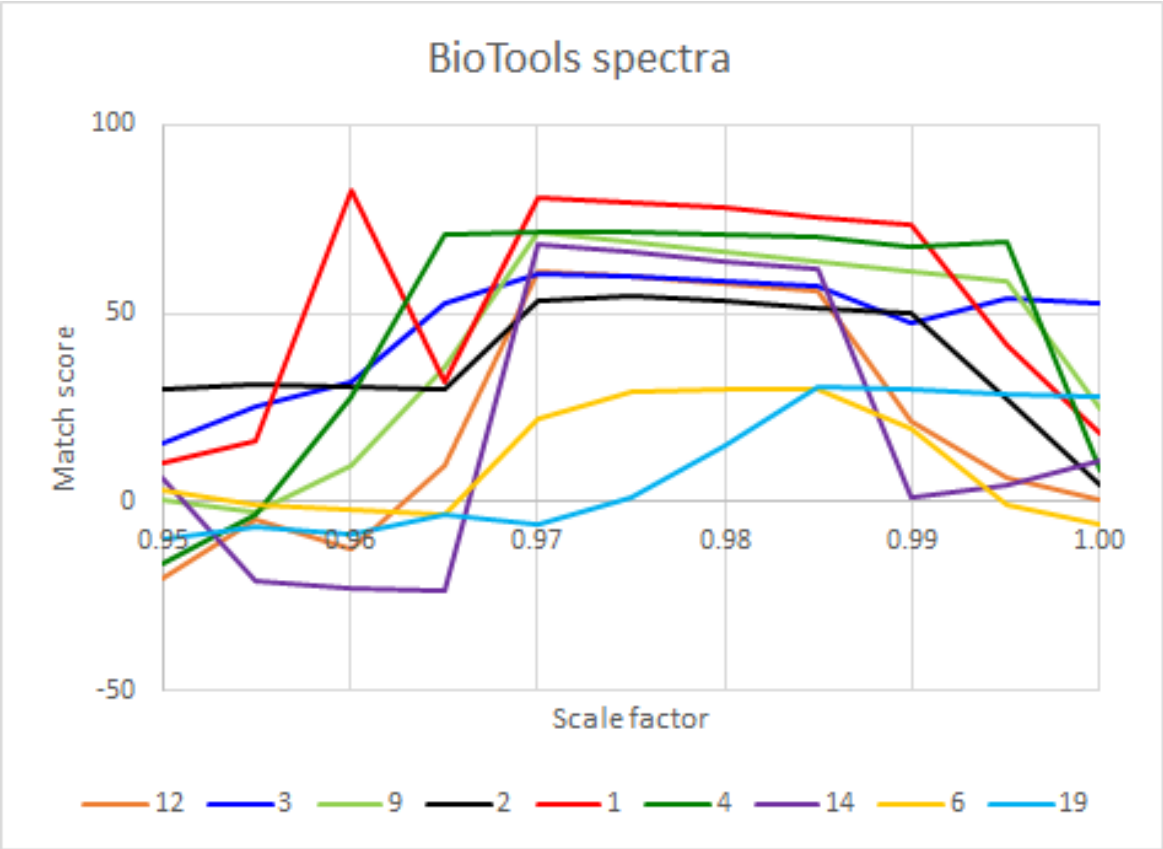


Fig. 5.6 Scaling profile of BioTools spectra of well-behaved compounds under B3PW91/cc-pVTZ calculation.

Under the B3PW91/cc-pVTZ analysis, the match scores between the SF values of 0.965 and 0.985 were averaged. Those with an average score above +25.0 (including limonene) comprise 60% of the dataset. When plotted against SF, the match score of these compounds

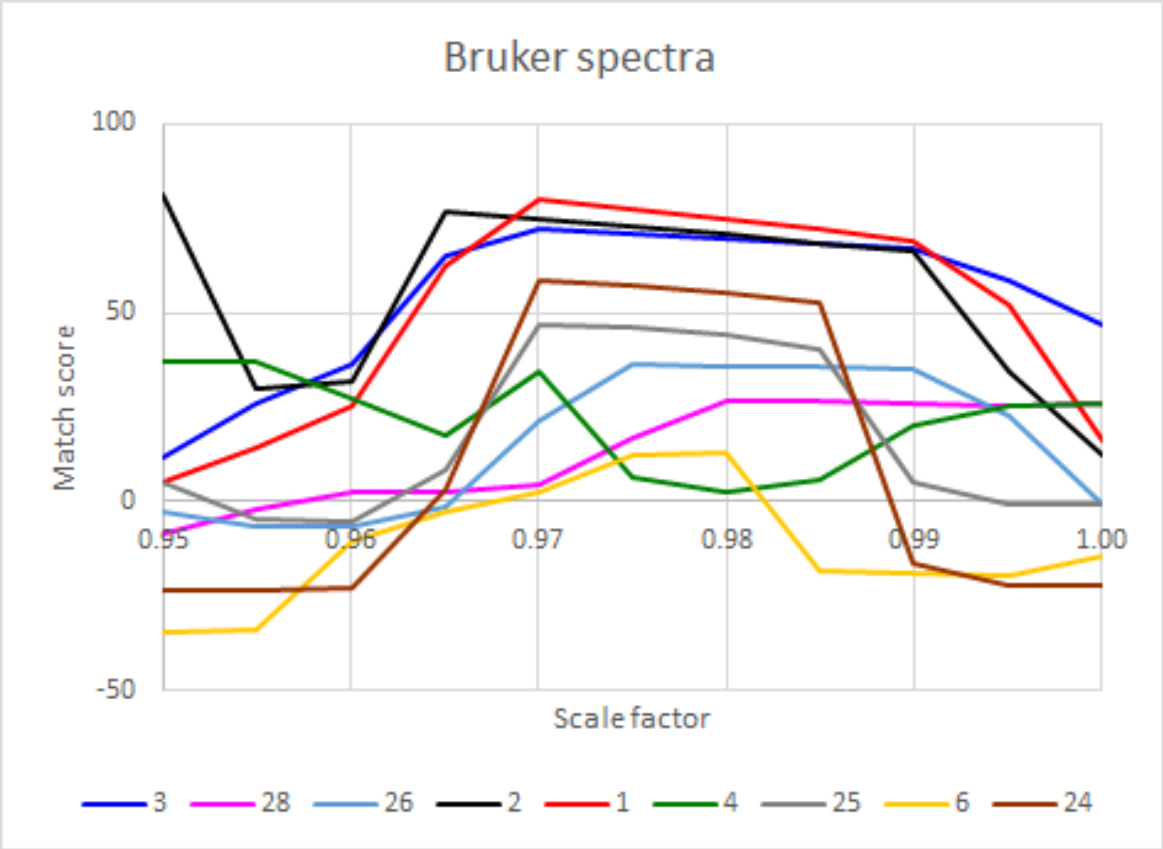


Fig. 5.7 Scaling profile of Bruker spectra of well-behaved compounds, under B3PW91/cc-pVTZ calculation. The match scores of limonene **11** and 2-phenylpropan-1-ol **17** are significantly reduced when compared with the BioTools spectra.

gives a curve which reaches a maximum between the values of 0.970 and 0.980. For these well-behaved compounds, between the SF values of 0.970 and 0.980, the graph tends to remain near its maximum value, giving rise to a plateau region over this range. The scaling profiles of these compounds give rise to the optimal SF for the various calculation methods, which all lie between 0.970 and 0.980 for the geometry optimised cases.

The typical behaviour of these compounds is also reflected in the average values obtained from geometry optimised calculations. When averaged across all compounds studied, the multiplicative percentage scores for each of these methods likewise reach a maximum at SF values between 0.970 and 0.980. The performance of various DFT methods as a function of SF is further explored in the following section.

A number of compounds such as epichlorohydrin **15** and epoxybutane **29** likewise display a plateau curve but with a sudden increase in score outside of this region. Two examples of these are visible in the B3PW91/cc-pVTZ calculation: in the case of epichlorohydrin this occurs at a SF of 0.95 in the Bruker spectrum; for epoxybutane this is found at SF 0.96 in the BioTools spectrum. These "islands" are also found in other calculation methods, as can be clearly seen in the heat maps above. Their presence is possibly due to vibrational bands in specific regions calculated as being closer together than in the experiment. However, they may also be caused by a small number of strong signals dominating the multiplicative score, in which case the SF value at which the islands occur is not a true optimal value but an anomaly.

For the remainder of the compounds, the match score is not sufficiently confident to make an informed assignment of AC. While the correct chirality is usually assigned within the SF range of 0.970 – 0.980, the trend in the scaling profile is not significant and thus cannot be used for a reliable prediction.

In keeping with typical results for the compounds studied, the SF giving a best fit overall between observed and calculated VCD spectra lies within the range of 0.970 - 0.980. This observation holds true for geometry-optimised calculations under all the methods studied.

5.4 Hybrid DFT Method

Having generated multiplicative percentage scores for the various levels of theory for DFT calculation (B3LYP/6-31G**, B3LYP/cc-pVTZ, B3PW91/6-31G**, B3PW91/cc-pVTZ), as

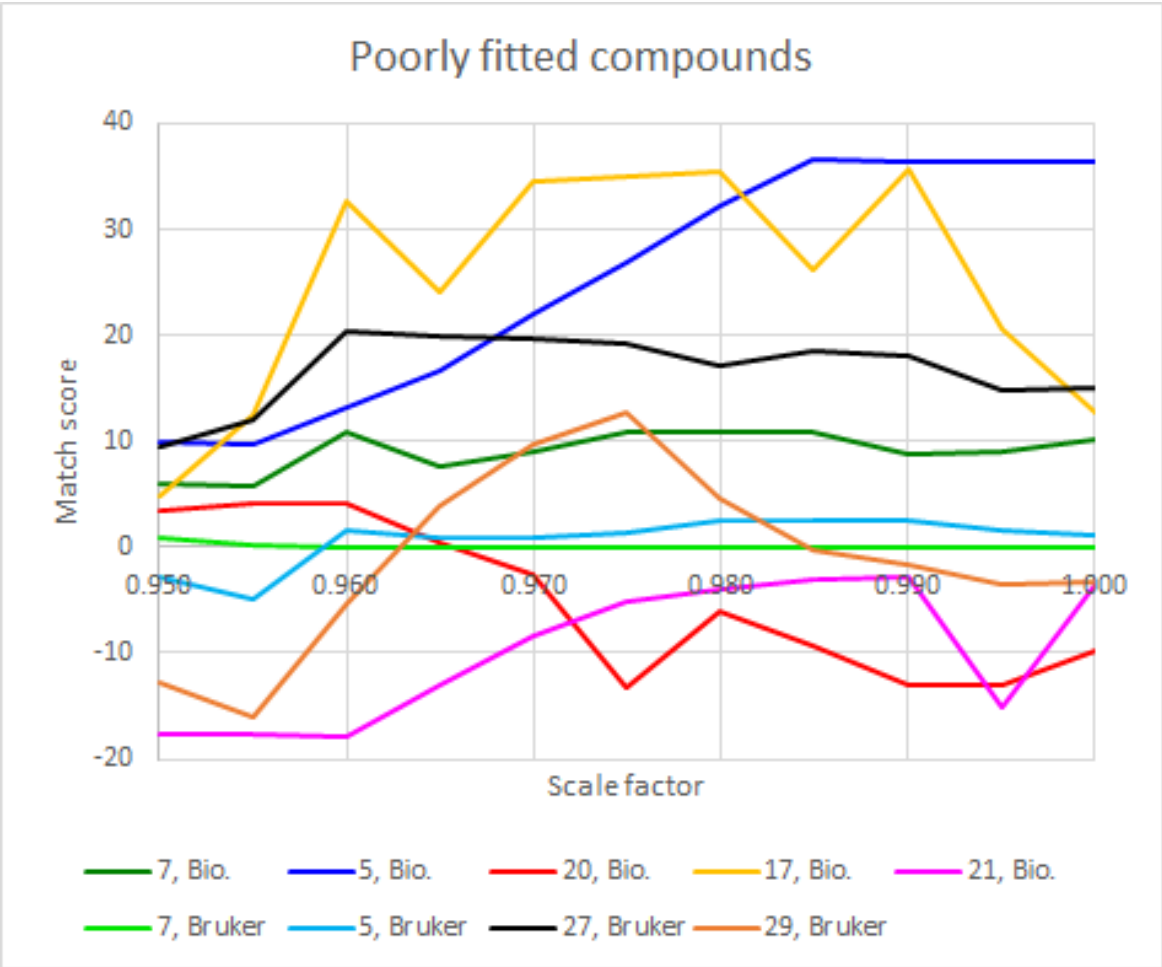


Fig. 5.8 Scaling profile of the remaining compounds from the initial analysis under B3PW91/cc-pVTZ calculation.

well as the single-point energy and geometry-optimised scores, the discrepancies in scoring values between the different methods were taken into account.

Whereas use of the cc-pVTZ basis set significantly improves results over those of the 6-31G(d,p) basis set, its use is also accompanied by a sizable increase in computational time. The reason for the improved performance of the B3PW91 functional over B3LYP is less clear. Although it is accompanied with a slight increase in computational time, this not considered significant. Furthermore, the improvement in score between B3PW91 and B3LYP increases from 0.2 to 2.7 when switching basis sets from 6-31G(d,p) to cc-pVTZ, suggesting a synergistic effect between the B3PW91 functional and the cc-pVTZ basis set. Further studies were carried out in order to investigate the greatest cause of the discrepancies between these scores.

Taking the heatmaps of each calculation method and subtracting the corresponding scores from one another, the differences between the scores generated using the B3LYP and B3PW91 were calculated. Five compounds were determined to have the largest discrepancy in calculated scores under treatment with the two functionals, namely:

limonene **8**,

1,2-epoxybutane **29**,

1-(1-naphthyl)ethylamine (NEA) **7**,

methyl 2-(4-chlorophenyl)-4-oxocyclopentane-1-carboxylate (MCPOPC) **28** and

3-(6,8-dichloro-2-methyl-1,2,3,4-tetrahydroisoquinolin-4-yl)aniline (DCMTIQA) **19**.

For each of these five compounds, an analysis was carried out to examine the optimal geometries obtained under different levels of theory for DFT calculations. Using the Maestro[38] molecular modelling package, the geometry-optimised conformations were superimposed and the RMSD values of atomic positions calculated for each conformer. This analysis was first performed by keeping basis set constant and varying the functional, then *vice versa*.

From the data above Table 5.1 and Table 5.2, when varying functional between B3LYP and B3PW91, the majority of conformations vary by an RMSD of 0.1 or less, with only one conformer differing by more than 0.4, which was later found to be due to a plane rotation of a the chlorophenyl group of MCPOPC. In addition, having excluded this data point, none of the maximal deviation values exceed 1.0, showing that the geometries generated by the B3LYP

Conf.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.
6					0.03	0.05	0.02	0.04		
5	0.02	0.03			0.01	0.03	0.07	0.19		
4	0.01	0.02			0.10	0.26	0.05	0.10	0.04	0.09
3	0.02	0.04	0.01	0.01	0.31	0.62	0.08	0.18	0.03	0.06
2	0.01	0.02	0.01	0.01	0.01	0.02	0.04	0.06	0.04	0.08
1	0.01	0.03	0.01	0.01	0.01	0.02	0.04	0.07	0.02	0.05
	Limonene **8**		Epoxybutane **29**		NEA **7**		MCPOPC **28**		DCMTIQA **19**	
1	0.01	0.02	0.01	0.01	0.01	0.02	0.04	0.08	0.02	0.05
2	0.01	0.02	0.01	0.01	0.01	0.02	0.04	0.09	0.03	0.08
3	0.02	0.04	0.01	0.01	0.01	0.03	0.12	0.27	0.02	0.05
4	0.01	0.02			0.01	0.03	1.08	2.55	0.00	0.08
5	0.02	0.03			0.01	0.02	0.03	0.05		
6					0.03	0.05	0.13	0.37		
7							0.02	0.04		

Table 5.1 RMSD (left) and maximal error (right) values comparing optimised geometries under B3LYP/6-31G(d,p) against B3PW91/6-31G(d,p) for the compounds giving the largest multiplicative score differences.

Conf.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.
6					0.01	0.02	0.04	0.12		
5	0.02	0.02			0.03	0.06	0.03	0.05		
4	0.02	0.04			0.02	0.04	0.07	0.17	0.03	0.08
3	0.02	0.04	0.01	0.01	0.02	0.03	0.08	0.18	0.10	0.22
2	0.01	0.02	0.01	0.01	0.03	0.06	0.06	0.10	0.08	0.19
1	0.02	0.04	0.01	0.01	0.03	0.05	0.23	0.63	0.02	0.04
	Limonene **8**		Epoxybutane **29**		NEA **7**		MCPOPC **28**		DCMTIQA **19**	
1	0.02	0.06	0.01	0.01	0.02	0.05	0.12	0.25	0.02	0.05
2	0.01	0.02	0.01	0.01	0.03	0.06	0.03	0.06	0.08	0.19
3	0.02	0.04	0.01	0.01	0.02	0.03	0.05	0.15	0.13	0.32
4	0.04	0.09			0.02	0.04	0.02	0.04	0.13	0.33
5	0.02	0.03			0.03	0.06	0.02	0.02		
6					0.01	0.02	0.09	0.17		
7							0.02	0.05		

Table 5.2 RMSD (left) and maximal error (right) values comparing optimised geometries under B3LYP/cc-pVTZ against B3PW91/cc-pVTZ for the compounds giving the largest multiplicative score differences.

and B3PW91 functionals are very similar. This similarity holds irrespective of whether the 6-31G(d,p) or the cc-pVTZ basis set is being used.

Conf.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.
6					0.06	0.13	0.46	1.37		
5	0.03	0.07			0.14	0.28	0.86	2.19		
4	0.03	0.07			0.06	0.18	1.74	2.93	0.37	0.97
3	0.03	0.06	0.03	0.03	0.05	0.10	0.39	0.75	0.61	1.47
2	0.03	0.05	0.04	0.06	0.04	0.08	0.21	0.35	0.44	1.07
1	0.04	0.10	0.02	0.04	0.08	0.16	1.63	2.85	0.07	0.13
	Limonene **8**		Epoxybutane **29**		NEA **7**		MCPOPC **28**		DCMTIQA **19**	
1	0.04	0.09	0.02	0.04	0.08	0.15	0.31	0.49	0.07	0.14
2	0.03	0.05	0.03	0.05	0.04	0.08	0.22	0.37	0.45	1.08
3	0.06	0.15	0.02	0.03	0.06	0.10	0.18	0.40	0.59	1.42
4	0.03	0.06			0.07	0.13	1.61	2.85	0.39	1.01
5	0.03	0.06			0.14	0.28	1.64	2.86		
6					0.05	0.11	0.85	2.26		
7							0.45	1.37		

Table 5.3 RMSD (left) and maximal error (right) values comparing optimised geometries under B3LYP/6-31G(d,p) against B3LYP/cc-pVTZ for the compounds giving the largest multiplicative score differences.

In contrast, when the difference between geometries generated by 6-31G(d,p) and cc-pVTZ were measured, notable differences were found in the conformations of DCMTIQA and MCPOPC. While the RMSDs between conformers of simpler compounds such as epoxybutane, limonene and 1-(1-naphthyl)ethylamine remained low (less than 1.0), this is likely due to the rigidity of the molecules rather than the similarity of the calculation methods. As such, while optimisation with the B3LYP and B3PW91 functionals return comparable geometries, this is not the case with the 6-31G(d,p) and the cc-pVTZ basis sets. This result lies within reasonable expectations, since both of the aforementioned are among Becke's three-parameter functionals.

The two basis sets used have less in common: 6-31G(d,p) belongs to the family of Pople's split-valence basis sets [72]. cc-pVTZ, on the other hand, is a correlation-consistent basis set developed by Dunning *et al.* [51] Moreover, cc-pVTZ is a triple-zeta basis set, while 6-31G(d,p) is only double-zeta, further increasing the difference between the two.

Given the overall similarity between geometries obtained *via* calculations using the B3LYP and B3PW91 functionals, as well as a small decrease in computation times when changing from B3PW91 to B3LYP, the possibility of a hybrid method was raised. In this proposal,

Conf.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.	RMSD	max.
6					0.09	0.18	0.42	1.24		
5	0.03	0.06			0.12	0.24	0.83	2.04		
4	0.04	0.09			0.06	0.12	1.75	2.98	0.37	0.97
3	0.02	0.05	0.02	0.03	0.07	0.13	0.40	0.78	0.67	1.62
2	0.03	0.06	0.04	0.06	0.02	0.04	0.19	0.33	0.48	1.17
1	0.04	0.08	0.02	0.03	0.07	0.13	1.76	2.98	0.08	0.15
	Limonene **8**		Epoxybutane **29**		NEA **7**		MCPOPC **28**		DCMTIQA **19**	
1	0.03	0.06	0.02	0.03	0.07	0.14	0.25	0.40	0.08	0.15
2	0.03	0.05	0.03	0.05	0.02	0.04	0.22	0.36	0.49	1.19
3	0.06	0.16	0.02	0.03	0.06	0.11	0.23	0.42	0.69	1.68
4	0.06	0.14			0.06	0.12	1.37	2.70	0.48	1.26
5	0.03	0.07			0.12	0.25	1.66	2.90		
6					0.07	0.14	0.75	1.94		
7							0.44	1.33		

Table 5.4 RMSD (left) and maximal error (right) values comparing optimised geometries under B3PW91/6-31G(d,p) against B3PW91/cc-pVTZ for the compounds giving the largest multiplicative score differences.

geometries would be predicted using the less time-consuming B3LYP functional, with the more accurate B3PW91 functional then being used to calculate vibrational frequencies and rotational strengths.

Initial studies were performed on DCMTIQA, limonene and both sets of spectra of epoxybutane, using the previously obtained geometries in the B3LYP case. As the cc-pVTZ basis set had been shown to return superior results, this was also the basis set chosen for the hybrid method.

Since the geometries obtained from B3LYP and B3PW91 optimisations were similar, the expected result for the multiplicative percentage score would be an average value of those obtained using B3LYP only and B3PW91 only, or a value in between both scores. While this was the case for DCMTIQA and limonene, the scores obtained under the hybrid method in both sets of spectra for epoxybutane surpassed those obtained using either B3LYP or B3PW91 alone. While these results may simply be the result of chance, the B3LYP functional has the added advantage of being optimised for organic compounds, which may be the cause of increased accuracy in the calculation of vibrational frequencies under B3PW91.

Due to the promising results obtained from the initial study, the hybrid method was extended to the entire library previously used for testing the multiplicative percentage scoring technique, the results of which are shown in Fig. 5.10.

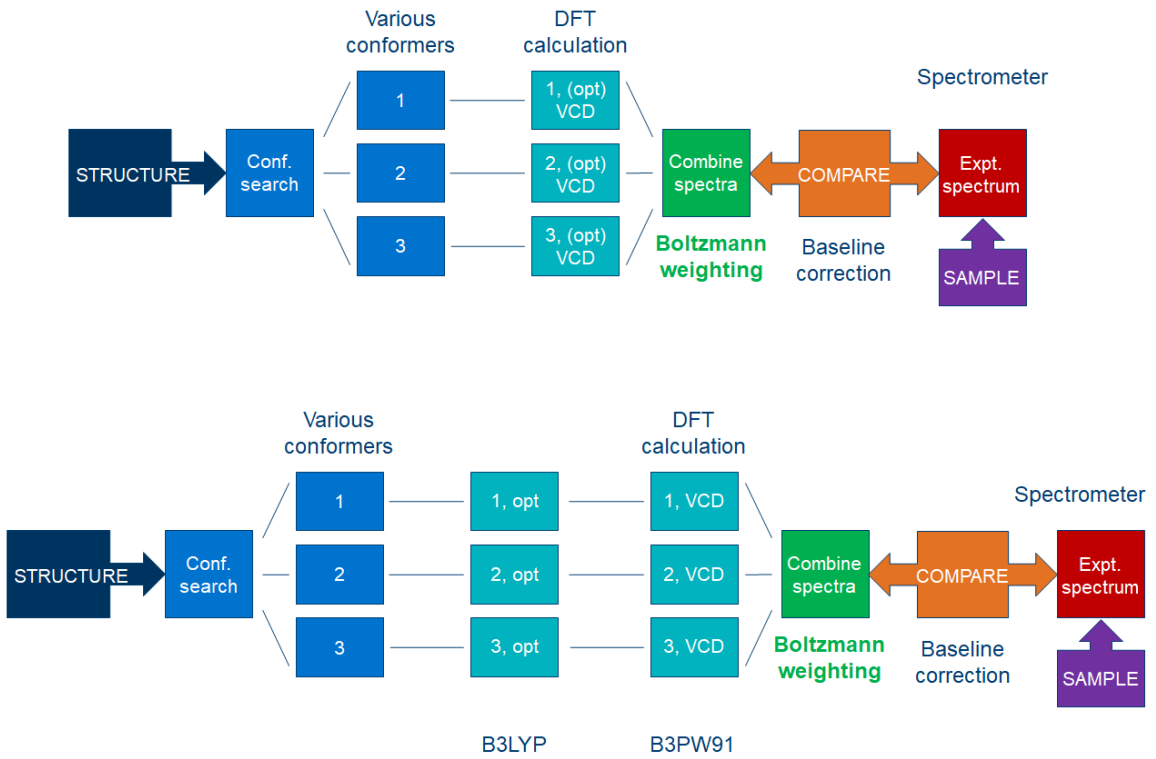


Fig. 5.9 Comparison of the original (top) and hybrid (bottom) methods of calculating predicted VCD spectra.

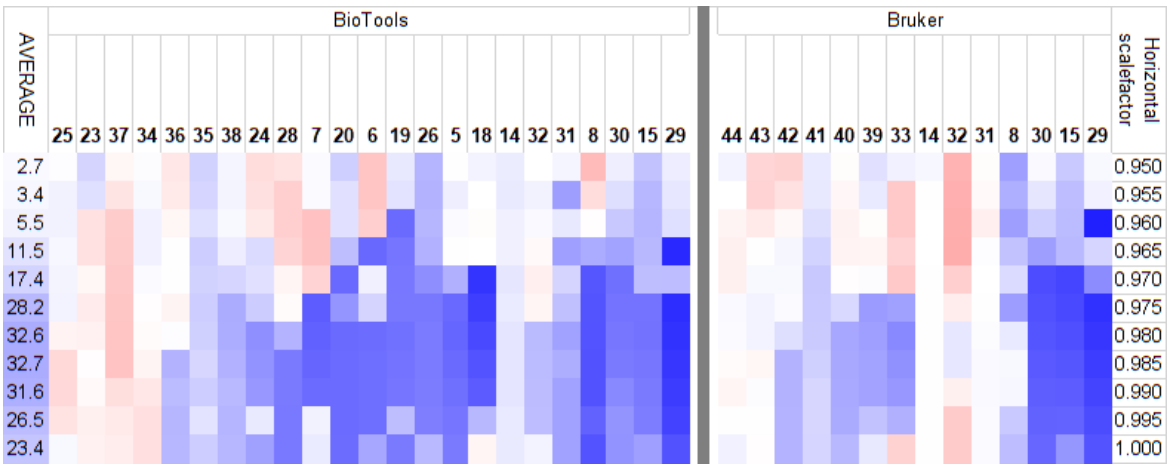


Fig. 5.10 Heatmap showing the results of the hybrid method.

Having tested the hybrid method on the library of compounds, the scores obtained were similar in nature to those of the optimisation under B3PW91/cc-pVTZ. Taking an average over all the sets of spectra compared, a maximum score of +35.6% was attained for the SF value of 0.980.

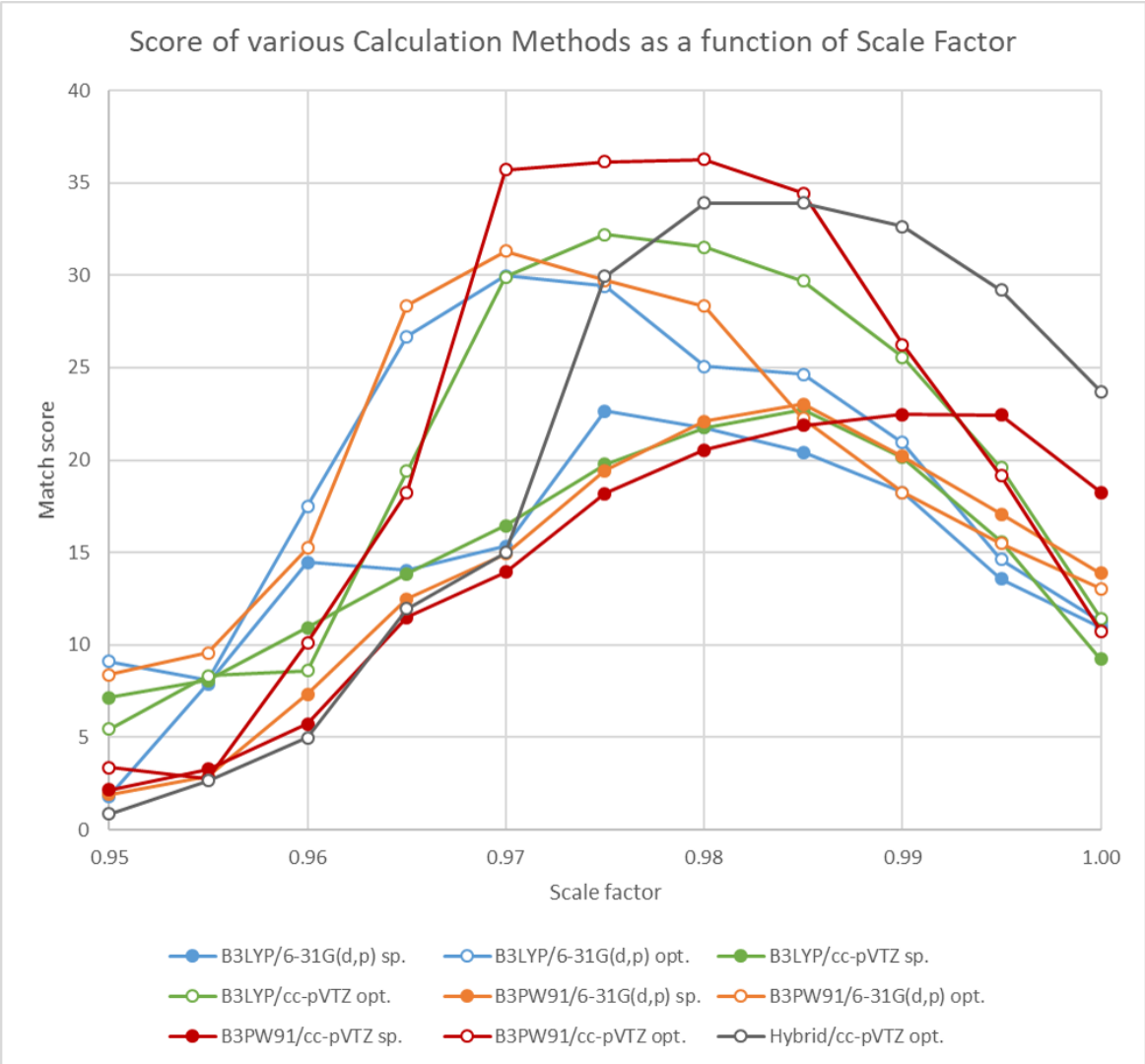


Fig. 5.11 Averaged scaling profiles of each calculation method. Filled circles represent single-point calculations; open circles represent optimisations.

Given the fact that the hybrid method did not perform as well as B3PW91 calculation alone, and since the small gain in computational time was further offset by the time required to prepare the conformers obtained using B3LYP for calculation with B3PW91, it was ultimately decided not to continue testing the hybrid calculation procedure.

5.5 Comparison of various DFT calculation methods

Together with the hybrid DFT calculation method, the various combinations of functionals and basis sets give multiplicative percentage scores which grow more positive when moving from lower to higher levels of theory. These results demonstrate that multiplicative scoring is a reliable and efficient method for comparing experimental and calculated VCD spectra.

As is expected from the DFT calculations, geometry optimisation significantly outperforms single point calculations in terms of the multiplicative percentage score attained. The cc-pVTZ basis set performs better than 6-31G(d,p) in all cases. Likewise, use of the B3PW91 functional gives better performance than B3LYP; however, the difference in score between the two is greatly enhanced when the cc-pVTZ basis set is applied. Some synergistic effects may therefore be present in these calculations.

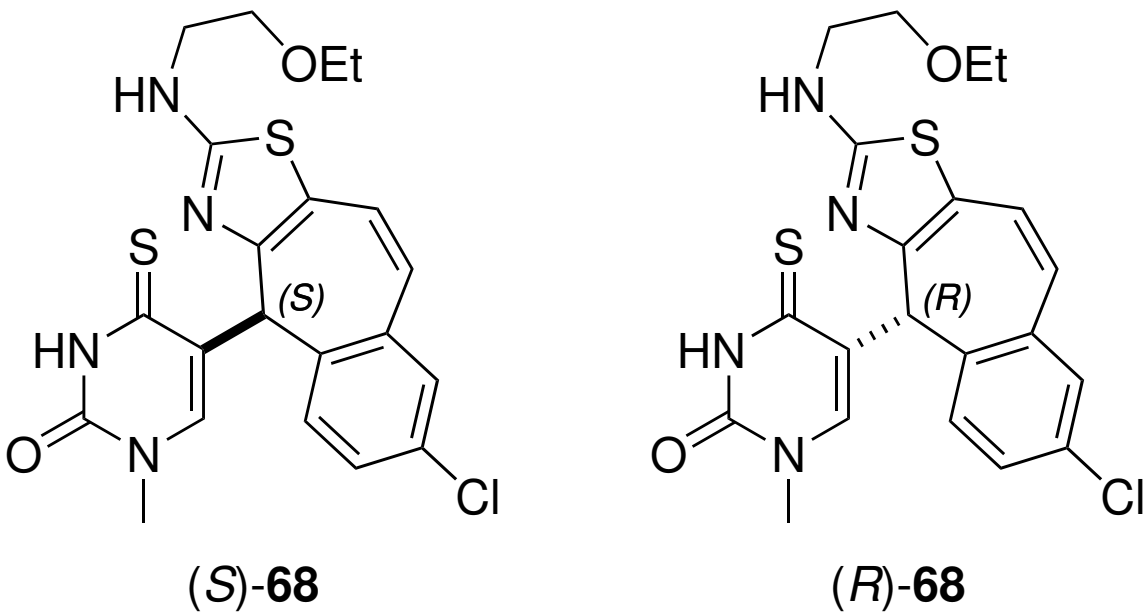
For geometry-optimised calculations, a horizontal SF of 0.980 gives the best fit overall for the compounds studied. Worth noting is the observation that the single point energy calculations tend to reach their optimal score at a higher SF value on average. It appears that the process of geometry optimisation, while giving a significant increase in the match score, also decreases the optimal SF for fitting the calculated spectrum to the observed data. This is unexpected since decreasing the molecular energy level is thought to lower vibrational frequencies, which should necessitate a higher SF in order to match the real experimental signal.

Based on the above data and their analysis, multiplicative scoring is shown to be a simple and efficient weighted method for comparing calculated and observed VCD data. A multiplicative score of +25% or above demonstrates a good level of confidence in the assignment under B3PW91/cc-pVTZ calculation.

Chapter 6

Configuration Assignment of a New Compound

6.1 Compound to be tested



Scheme 6.1 Structures of the enantiomeric compounds studied in this chapter.

Compound **68** was obtained from outside the dataset of compounds previously analysed. While the compound was supplied as separate enantiomers, labelled as MSG-257 and MSG-258, the AC of each remained unknown.

Samples of MSG-257 and MSG-258 were each dissolved in 120 μl of chloroform-d_1. Using a BioTools ChiralIR-2X instrument, IR and VCD spectra were acquired at 20,000 scans each over 6 hours, with a spectral resolution of 4 cm^{-1}.

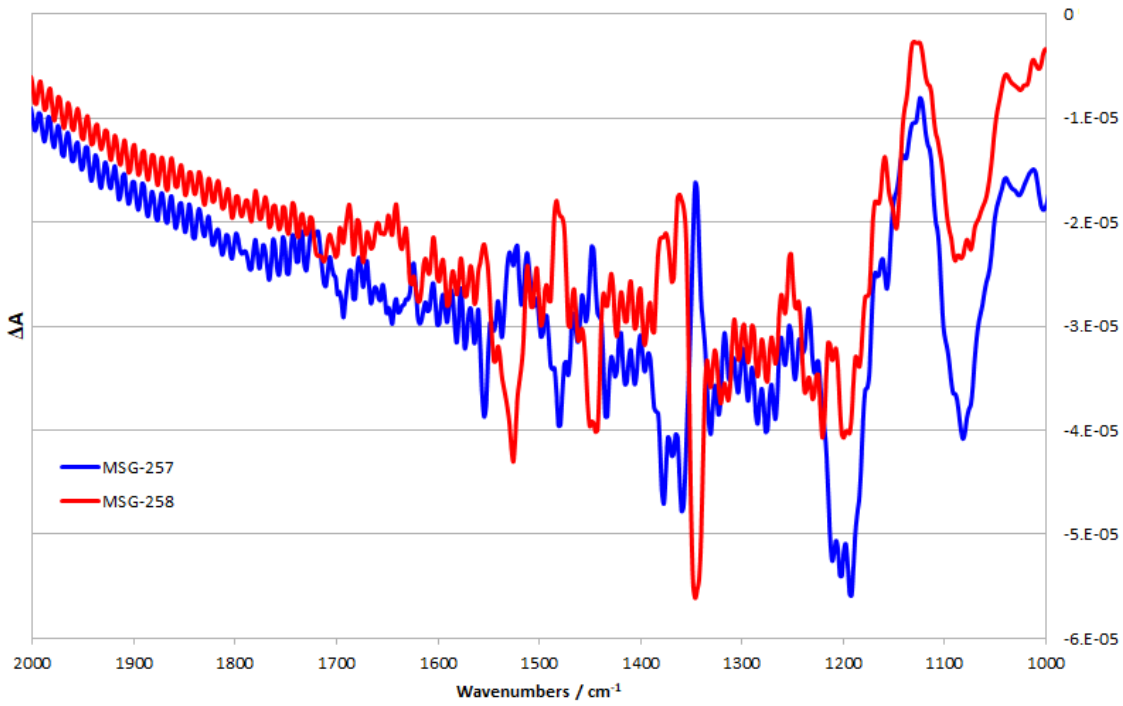


Fig. 6.1 The raw data obtained for samples MSG-257 and MSG-258.

It is noted that in addition to the characteristic curve previously noted in measurements using this BioTools instrument, there is also an oscillating signal present in the baseline, likely due to an alignment problem with the spectrometer. Although the error causing the baseline ripple was later corrected in the instrument, both enantiomers of **68** were no longer available shortly after the measurement and as such, no futher spectra could be recorded. In spite of the baseline noise, therefore, an assignment of AC was attempted using this set of experimental data.

6.2 Computational procedure

A Monte Carlo conformational search was set up for both enantiomers, each sampling the conformational landscape over 100,000 steps. Similar results were obtained for each case, with the (*R*)-enantiomer returning 35 conformations and the (*S*)-enantiomer returning 34 conformations within the 10 kJ mol^{-1} energy window.

These conformers were submitted for geometry optimisation under DFT at the B3PW91/cc-pVTZ level. However, upon conclusion of the DFT calculations only nine conformers of the (*S*)-form had converged to a stable energy conformation, whereas fourteen conformers were found to converge to a stable minimum in the case of the (*R*)-form. These were labelled (*R*)-**68**a - (*R*)-**68**p (the letters j and l were omitted for clarity), following the order of the newly calculated DFT energies, and are shown in Table 6.1.

While it may appear that the failure of a large number of MM-generated conformers to converge may impair the accuracy of a VCD prediction, the result obtained was considered acceptable as the majority of these conformers were from the higher energy portion of the energetic window (conformer 21 and onwards), and therefore unlikely to contribute significantly to the final calculated spectrum. The failure of the four lowest energy conformers to converge on an energetic minimum in both the (*R*)- and (*S*)-enantiomers, however, warrants concern. Further DFT optimisations were therefore carried out in order to ensure that no low-energy conformers were omitted from the analysis, as have been detailed in the following paragraphs. A full justification of the VCD prediction is provided at the end of the chapter.

The resulting conformations generated using the conformational search of the (*R*)-form are used from this point onwards.

Three of the conformers generated, namely (*R*)-**68**n, (*R*)-**68**o and (*R*)-**68**p, have DFT energies more than 15 kJ mol^{-1} above the ground state and thus gave no significant contribution to the final VCD spectrum under Boltzmann weighting.

Within a limit of the first 100 steps, the four conformers with the lowest MM-estimated energies failed to converge under DFT calculation. These were labelled (*R*)-**68**w, (*R*)-**68**x, (*R*)-**68**y and (*R*)-**68**z. Due to the fact that they occupied the lowest energy conformational minima under MM, and in order to ensure that no low-energy conformations would be inadvertently omitted from the analysis, they were each submitted for further DFT optimisation, upon which the following results were obtained:

Boltzmann weight / %	E_{rel} / kJ mol^{-1}	(*S*)	(*R*)	E_{rel} / kJ mol^{-1}	Boltzmann weight / %	Conformer
		1	1		-	w
		2	2		-	x
		3	3		-	y
		4	4		-	z
14.9	0.392	5	5	0.408	8.8	h
		6	6	0.547	8.3	k
		7	7	0	10.4	a
17.5	0	8	8	0.244	9.4	d
17.4	0.0124	9	9	0.593	8.2	m
17.1	0.0489	10	10	0.241	9.4	c
		11	11	0.21	9.5	b
		12	12			
		13	13	15.8	0	o
		14	14	0.285	9.2	e
		15	15			
		16	16			
		17	17	17.5	0	p
		18	18	0.493	8.5	i
		19	19			
15.9	0.237	20	20	0.331	9.1	g
		21	21			
1.7	5.75	22	22			
0.4	9.27	23	23	0.311	9.1	f
15.2	0.343	24	24			
		25	25			
		26	26			
		27	27			
		28	28			
0	22.9	29	29			
		30	30	15.3	0	n
		31	31			
		32	32			
		33	33			
		34	34			
			35			

Table 6.1 Energetic minima found by Monte Carlo conformational searching for the (*S*)- and (*R*)-forms of **68**, alongside energies of the DFT-optimised conformers where found. Alphabetical labels are provided for conformers of the (*R*)-form.

(*R*)-**68**w, the lowest energy conformer predicted by MM, continued to increase in energy under DFT calculation. After failing to converge to a stable conformation after 100 steps of DFT optimisation had elapsed, the calculation was terminated.

(*R*)-**68**x and (*R*)-**68**y both converged after one additional set of DFT geometry optimisation calculations. (*R*)-**68**x was found to be 8.91 kJ mol^{-1} above the ground state, and (*R*)-**68**y was 9.95 kJ mol^{-1} above the ground state. As the Boltzmann weights of conformers at these energies would not contribute significantly to the final calculated spectrum, these were excluded from the final analysis.

In the case of (*R*)-**68**z, after repeated attempts at geometry optimisation, the geometry of the conformation remained stuck at a high energy, more than 10 kJ mol^{-1} above the ground state. Therefore, this conformer was likewise excluded.

While the higher energy conformations remained unaccounted for, all of these conformers have an energy above 5 kJ mol^{-1} according to MM estimates. These are therefore unlikely to have a significant effect on the final calculated VCD spectrum, and were excluded from the analysis.

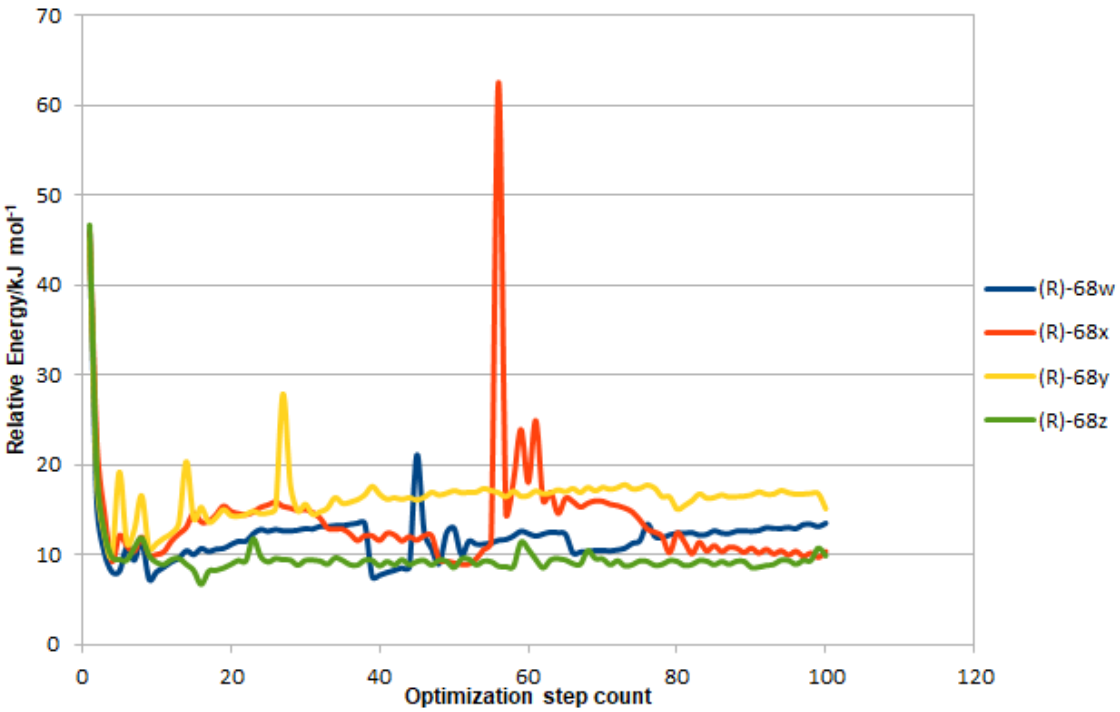


Fig. 6.2 DFT-calculated energies of the conformers (*R*)-**68**w, (*R*)-**68**x, (*R*)-**68**y and (*R*)-**68**z over the course of geometry optimisation. Energetic values are shown relative to the ground state, conformer (*R*)-**68**a (not shown).

6.3 Conformational analysis of the (*R*)-form

For the conformers which converged successfully, the greatest difference was found to be the orientation of the $–NH_2(CH_2)_2OEt$ moiety. Among higher energy conformers, the twist of the 4-thiouracil ring relative to the tricyclic core was also found to be a source of variation among the low-energy conformations. However, since it was confirmed that they were unable to converge to an energy less than 10 kJ mol^{-1} above the ground state and thus would not have made a significant contribution towards the final VCD spectrum, these were not included in the prediction of the calculated VCD.

Crucially to the assignment of AC, for nine of these low-energy conformers((*R*)-**68**b - (*R*)-**68**i, (*R*)-**68**m), the predicted VCD spectra for the individual conformations were nearly identical, with all of the higher intensity VCD signals sharing the same sign, position and rotational strengths. As a result, even though conformer **68**a occupies the ground state and thus contributes the most individually to the final calculated VCD spectrum under Boltzmann averaging, the final spectrum bears a closer resemblance to the conformational spectra of these nine conformers than it does to that of conformer **68**a.

The conformer (*R*)-**68**k likewise differs from the remaining ten conformers, both in conformation and in the shape of its VCD signals (appearing as the deep blue trace in Fig. 6.5).

Besides the ground state (*R*)-**68**a and the conformer (*R*)-**68**k, the nine VCD-similar conformers cover a range from 0.210 to 0.593 kJ mol^{-1}. Examining these in more detail, they are shown to have very similar geometries,

When superimposed, conformers (*R*)-**68**c and (*R*)-**68**d have a RMSD value of 0.1543, and a maximal deviation of 0.4277. Similarly, conformers (*R*)-**68**f and (*R*)-**68**g have a RMSD of 0.1858 and a maximal deviation of 0.4760.

Based on their geometries, two clusters of conformational minima can be described: one including (*R*)-**68**c, (*R*)-**68**d and (*R*)-**68**h, and the other including (*R*)-**68**b, (*R*)-**68**e, (*R*)-**68**f, (*R*)-**68**g, (*R*)-**68**i and (*R*)-**68**m. The geometries within each of these clusters are shown in Fig. 6.6 and Fig. 6.7.

The question of how dissimilar two conformations need to be before they are considered separately is not straightforward, but is important to the calculation of predicted VCD spectra. For example, if two similar conformations found by DFT calculation are in fact duplicated

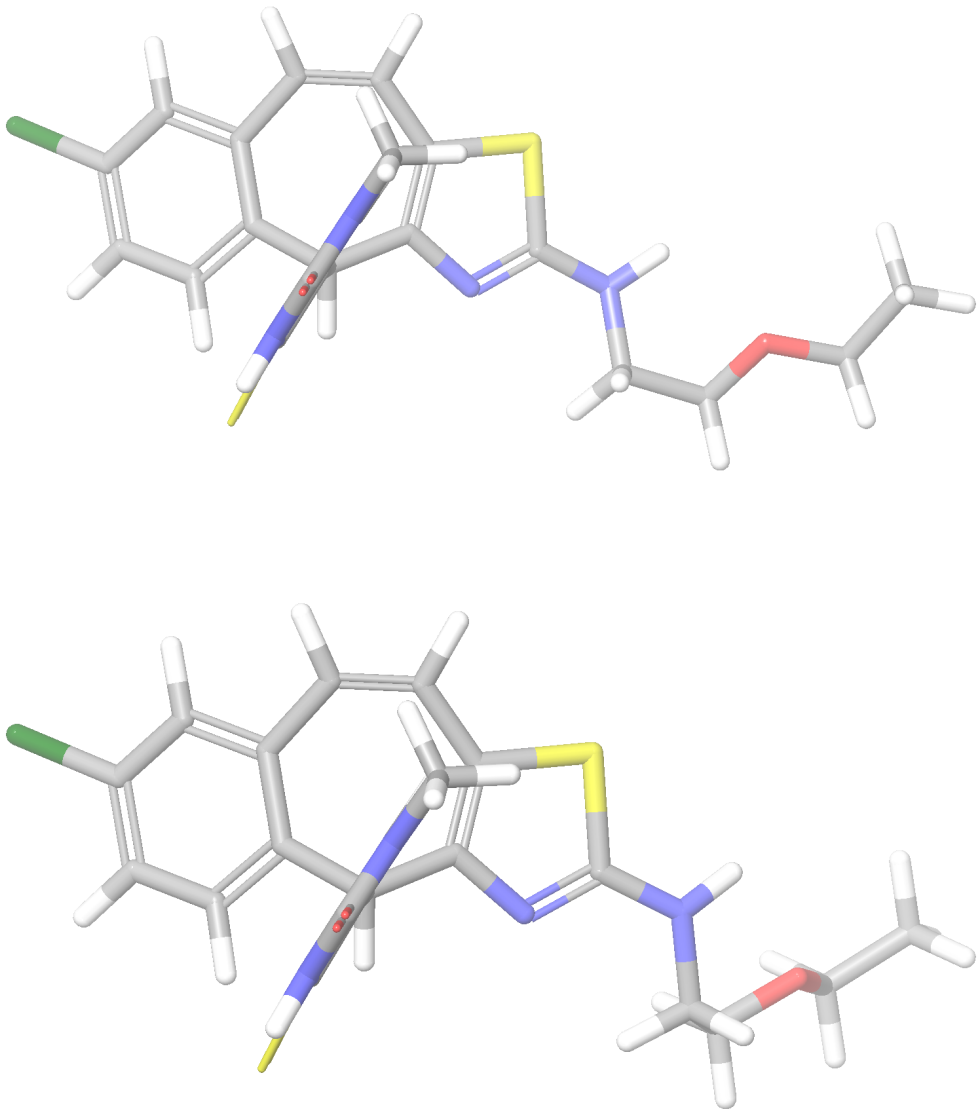


Fig. 6.3 Conformations **68**a (top) and **68**k (bottom) for the (*R*)-form of the compound.

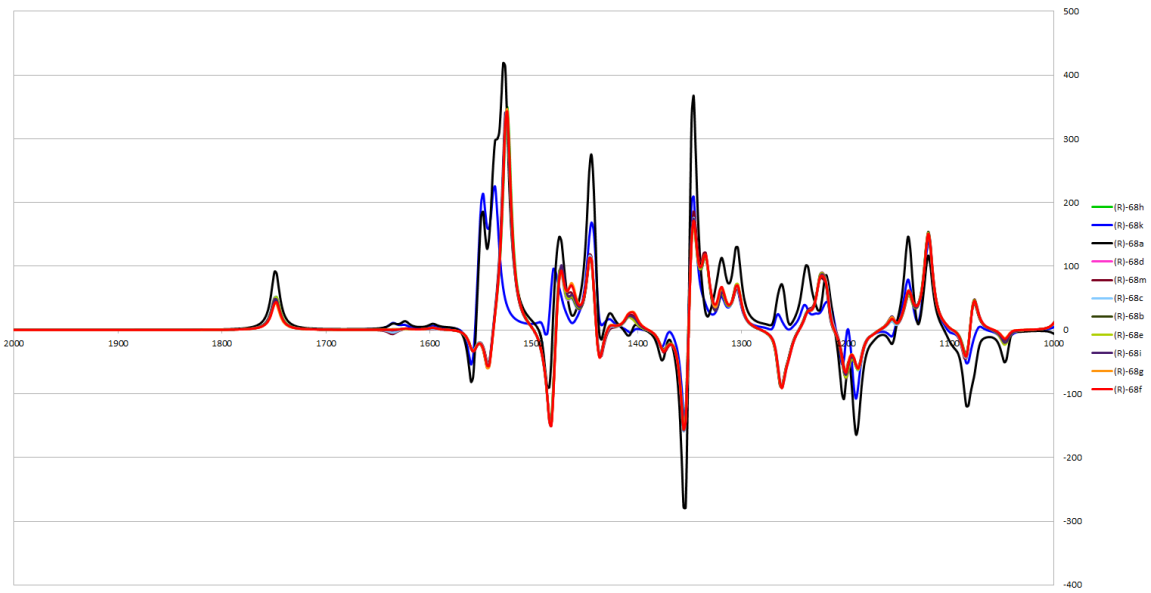


Fig. 6.4 Calculated spectra of the 11 low-energy conformers found under DFT at the B3PW91/cc-pVTZ level.

from a single local minimum in the physical potential energy surface, then including both of them in the analysis would significantly increase the Boltzmann weight of this conformer. This in turn affects the final predicted spectrum, by awarding a disproportionately large weight to the conformational VCD of the duplicated conformer.

In the case of these nine conformers ((*R*)-**68**b - (*R*)-**68**i, (*R*)-**68**m), three options are available:

1) merge all nine into a single conformational minimum

2) merge each of the two clusters, resulting in two minima

3) treat them as nine distinct conformers

However, in cases 1) and 2), where merging of different DFT-optimised conformations is required, the presence of variations in the relative energies of the conformations to be merged poses a problem: whether to take the average of each cluster's energies as the overall energy for that cluster, or to use that of the lowest energy conformer.

Proceeding under the assumption that the lowest energy conformer within each cluster was the closest to that local energetic minimum, and therefore representative of all the conformers found within that cluster, a predicted VCD spectrum can be produced for each of the aforementioned options can be produced using the following conformers:

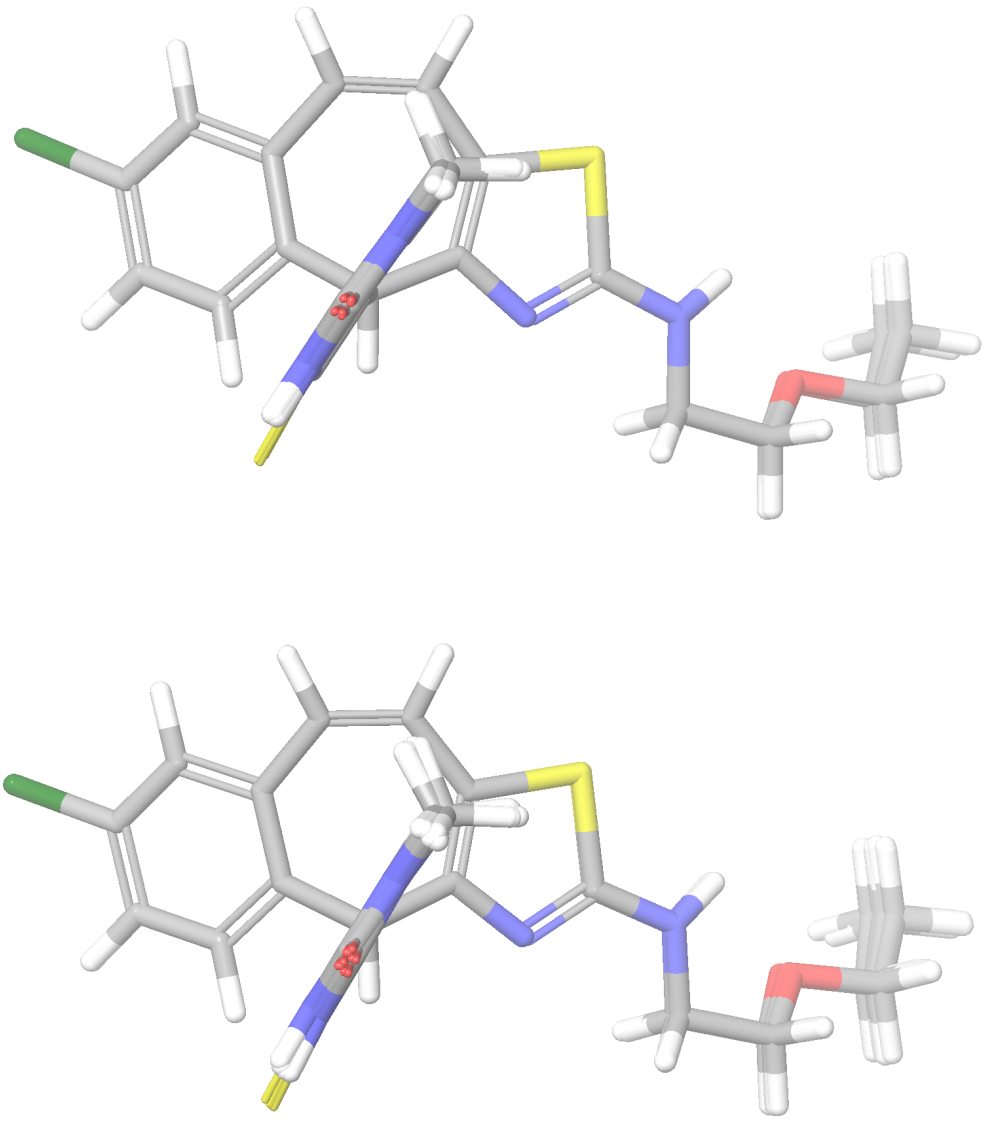


Fig. 6.5 Top: Superimposed geometries of the conformational cluster including (*R*)-**68**c, (*R*)-**68**d and (*R*)-**68**h. Bottom: Superimposed geometries of the conformational cluster including (*R*)-**68**b, (*R*)-**68**e, (*R*)-**68**f, (*R*)-**68**g, (*R*)-**68**i and (*R*)-**68**m.

1) (*R*)-**68**a, (*R*)-**68**b, (*R*)-**68**k

2) (*R*)-**68**a, (*R*)-**68**b, (*R*)-**68**c, (*R*)-**68**k

3) (*R*)-**68**a - (*R*)-**68**i, (*R*)-**68**m

A predicted VCD spectrum was then calculated from each of these combinations, for which the results can be seen in the next section.

6.4 Multiplicative scoring against experimental spectra

Making use of the optimisation of the multiplicative percentage score from the previous chapter, the finalised procedure was applied to each of the combinations of conformers mentioned in the previous section. The results obtained are set out in Table 6.2.

Method used	Conformers included	Mult. % score
1	a, b, k	47.74
2	a, b, c, k	46.98
3	a - i, m	44.15
Reduced energetic window	a, h, k	48.49

Table 6.2 Various combinations of generating the final calculated spectrum, using the lowest energy conformer for each cluster.

Interestingly, the highest score obtained when calculating the multiplicative percentage score against the observed VCD was arrived at by setting the energy window to 5 kJ mol^{-1}. In so doing, only three converged conformers are now included in the analysis: (*R*)-**68**h, (*R*)-**68**k and (*R*)-**68**a. By applying Boltzmann weighting to these 3 conformers, a calculated VCD spectrum was generated and compared to the experimental spectrum.

Despite only using 3 conformers in this calculation for (*R*)-**68**, this nevertheless slightly improved the match with the half-difference spectrum. The multiplicative percentage score achieved using these 3 conformers with the experimental data was +48%, likewise assigning MSG-257 to the (R)-enantiomer and MSG-258 to the (S)-enantiomer. This suggests that the contribution towards the real VCD spectrum by the nine redundant conformers, (*R*)-**68**b - (*R*)-**68**i and (*R*)-**68**m, may be overstated by the DFT calculation method, as (*R*)-**68**h alone from among these nine is sufficient to give an accurate match.

After Boltzmann averaging of all the DFT-converged conformers, the calculated spectra were scaled horizontally by a factor of 0.975 and Lorentzian line broadening was applied with a FWHM of 5cm^{-1}. The resulting overlap with the baseline-corrected experimental spectrum gave a multiplicative match score of +44%, assigning the spectrum of MSG-257 to the (R)-enantiomer and MSG-258 to the (S)-enantiomer.

The major diagnostic signals, which were calculated to be the same sign for all DFT-converged conformers of the (*R*)-enantiomer, lie at approximately 1345 cm^{-1} (+), 1360 cm^{-1} (-) and 1450 cm^{-1} (+). Each of these signals has a high intensity in the experimental VCD spectrum, and while they do not comprise all the highest intensity signals, they are shown to have the same sign among all the converged conformers, increasing the reliability of the assignment.

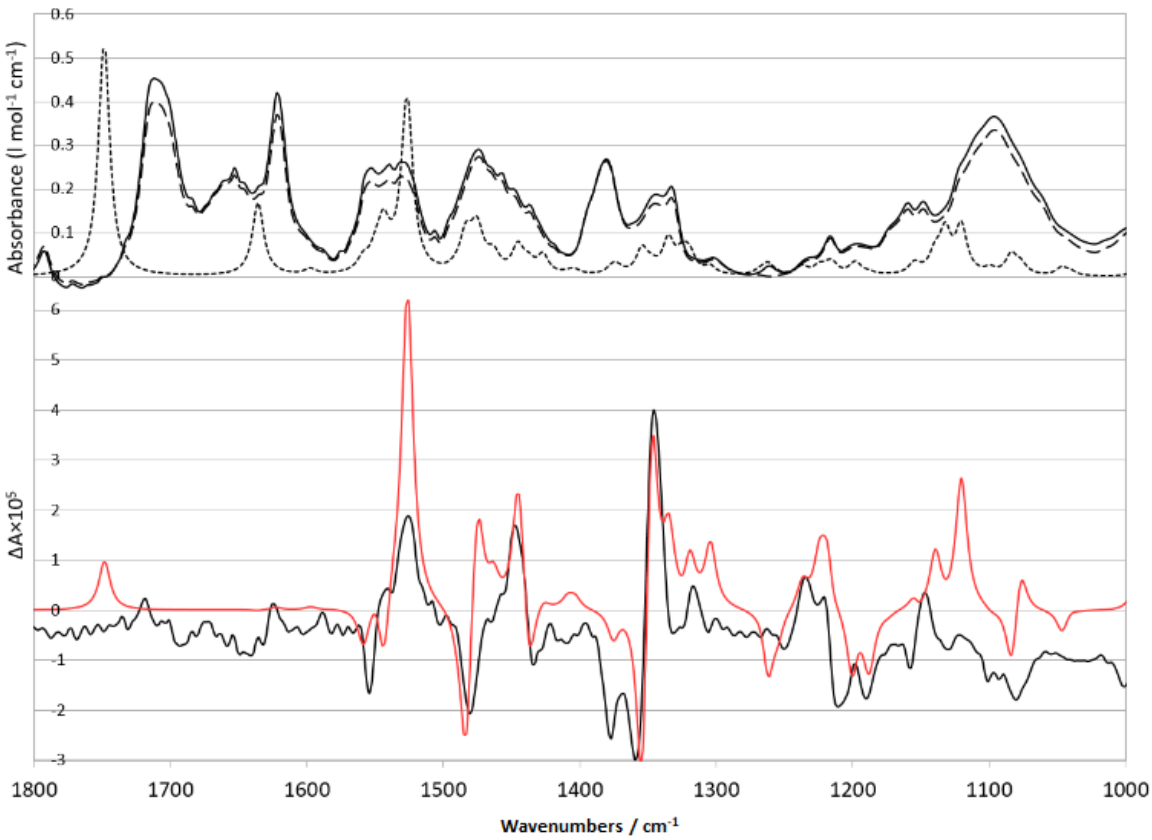


Fig. 6.6 Calculated vibrational and VCD spectra for (*R*)-**68** (fine dashed IR; red VCD) and the observed spectra from MSG-257 (long dashed IR) and MSG-258 (solid IR). The experimental VCD trace (black) is obtained by subtraction as MSG-257 minus MSG-258.

Apart from making an assignment of AC for compound **68**, the optimised multiplicative percentage score is also an indicator of the quality of the SF used when superimposing the

calculated and observed spectra. The value of the SF used is worth further exploration when the IR and VCD spectra are viewed together (Fig. 6.8).

When the calculated and observed IR spectra are overlaid, the SF value of 0.975 appears to be an overestimate. This can be inferred from the fact that the strong IR absorptions at 1630 and 1750 cm^{-1} in the calculated spectrum appear to correspond to calculated signals near 1620 and 1710 cm^{-1} respectively; furthermore, the weaker calculated absorptions at 1120 and 1130 cm^{-1} appear to correspond to the experimental peaks at 1150 and 1160 cm^{-1}, causing the calculated spectrum to appear stretched overall in comparison. By applying a smaller SF value, it would be possible to shrink the wavenumber scale sufficiently to bring these calculated signals into coincidence with the observed peaks in experiment.

Nevertheless, these signals lie on the outside ends of the diagnostic region for the VCD signals, and do not possess a strong rotational strength in the VCD. As a result, this change in SF is unlikely to have a significant impact on the multiplicative percentage score when evaluating the match between the calculated and experimental spectra in the case of compound **68**.

Two major factors justify the assignment of AC made by a limited number of converged conformations:

Firstly, the level of baseline noise in the experimental spectra is a greater source of error than can be caused by differences in minor conformers. This is especially noticeable in the 1000 - 1200 cm^{-1} region of the experimental VCD spectrum, where the baseline continually and significantly deviates in value from zero. In spite of this source of error, a confident multiplicative percentage score is attained for the assignment. The level of baseline noise is likely a greater source of error than the omission of the higher energy conformations, whose contribution to the final calculation is unlikely to exceed 8% (the highest energy included conformer, (*R*)-**68**m, lies only 0.6 kJ mol^{-1} above the ground state and has a Boltzmann weight of 8.2%). In case a higher quality set of experimental spectra can be provided, removing the baseline error, the omission of the higher energy conformers may become a relatively greater source of error, but the assignment of AC is unlikely to be overturned as a result.

Secondly, and more importantly, the repetitive nature of nine of the conformations, namely (*R*)-**68**a - (*R*)-**68**i and (*R*)-**68**m, in producing nearly identical calculated VCD spectra cannot be overlooked. This redundancy is evidence for the robustness of the VCD-active vibrational modes, indicating that they are not likely to change sign under small conformational pertur-

bations. A final calculated spectrum that predicted the opposite assignment of AC would possess negative overlap with all nine of these low energy conformers, a highly unlikely result.

While not the ideal SF value for **68**, the SF of 0.975 is sufficient to make a confident AC assignment for the compound, and thus MSG-257 is assigned to the (R)-enantiomer and MSG-258 to the (S)-enantiomer.

Chapter 7

An Online Database for VCD Spectra

This chapter discusses studies and steps taken towards developing an online database for VCD spectra, together with possible components such a database would require.

In order for new VCD prediction and comparison methods to be tested, a large repository of data is indispensable. The larger the dataset, the more thoroughly an evaluation method can be trialled and improved. VCD spectra already appear in numerous publications, either as direct proof or supporting evidence for an assignment of AC. In addition to this, VCD is approved by the United States' Food and Drug Administration [73] as a technique by which chiral drugs can be characterised.

While there are commonly available spectra of many compounds in published literature and on the Internet, a large amount of data from VCD measurements is acquired privately, such as for drug companies to assign AC to lead compounds for pharmaceutical targets.

In order to facilitate the process of model development for VCD prediction, one solution is to have a database of VCD spectra available online, where all users may submit and view data. Together with user feedback on the data and contributions from researchers worldwide, such a database can facilitate development of improved VCD comparison methods, and in doing so, help computational chemists to refine calculation methods for the prediction of vibrational spectra.

7.1 Calculated spectra

Calculated spectra are presented in line-broadened form, with a FWHM of 5 cm^{-1}. While the DFT-calculated conformational energies are included alongside the conformational spectra, Boltzmann weights have not been presented here. This intention behind this presentation is that in case additional conformations are located, or two or more conformations already present are found to be redundant, these can be added or removed by the original user with ease.

Individual tables of signals and intensities (without line broadening) are also available for those wishing to apply different broadening algorithms.

7.2 Experimental VCD spectra

Collection of experimental VCD spectra is likely to be the main aim of a VCD database in its early stages. Precedents for databases of chemical and spectral information abound and can be found online. The National Institute of Advanced Industrial Science and Technology in Tokyo, Japan, maintains a Spectral Database for Organic Compounds [74] which may act as a reference for a repository of experimental VCD spectra.

Besides spectral data, it is crucial for any database to include relevant details such that the VCD measurement is reproducible. As an example, the following lists some of the useful information that may be included in a database for VCD spectra:

7.2.1 Compound structure

The structure of the molecule under investigation is one of the first items of information to be included in each database entry. Crucially for VCD spectroscopy, stereochemical data must be fully assigned with the compound structure, or, where the stereochemical assignment is uncertain, stereogenic centres of unknown configuration must be clearly marked. However, besides providing the molecular structure for each compound under investigation, chemical structures should be searchable online via an applet, allowing users to compare the VCD spectra of related compounds and congeners.

7.2.2 Instrument

As has been noted in previous chapters, experimental VCD data can vary drastically from one instrument to another in terms of baseline noise and artefact peaks in the spectrum.

Even with information available regarding instrument vendor and type, detailed knowledge of the instrument's construction is often beneficial. For example, several different generations of the BioTools ChiralIR-2X model have been constructed, and while major features such as the dual-PEM setup have been conserved, spectra acquired on newer instruments will differ in quality from those of older measurements. As a result, reporting the details regarding the instrument used may aid in ranking spectra by value or usefulness.

7.2.3 Solvent or physical state

The solvent used in measuring the experimental VCD spectrum is useful information to include, in order that the solvent absorption bands might be excluded from comparison algorithms. Users who wish to make use of solvent models in the MM or DFT phase of calculations may also refer to the solvent used in acquiring the spectra. Additionally, different solvents can have a significant effect on the frequency of VCD absorptions via H-bonding, necessitating the solvent used to be reported for these special cases.

Besides solution-state measurements, VCD spectra can be acquired for neat liquids, as well as solid-state samples [20]. When exceptionally high quality spectra are required, without the effects of any external neighbouring interactions (solvent or otherwise), VCD spectra can be performed in isolation in an argon matrix, which should likewise be noted.

7.2.4 Concentration

Adding the concentration of the solute under investigation allows users to easily reproduce the VCD spectra at the same intensity as the quoted spectra. While the absolute value of the VCD absorptions is not always necessary in order for an accurate assignment of AC to be made, having an ideal concentration of solute available - one which ensures sufficient intensity of peaks while still preventing solute-solute interactions, thus obeying Beer's law - will save researchers the effort of having to attempt various concentrations before arriving at the same conclusion.

7.2.5 Sample and experimental details

Details enabling reproduction of the experimental data (path length, spectral resolution, acquisition time) should be included in each submission to the database. Additionally, the wavenumber range over which data is available may be set such that comparison methods have sufficient data with which to operate.

The supplier of the sample compounds is occasionally relevant to VCD measurement as well. As different suppliers sometimes use different purification methods, which may affect the identity of impurities present as well as their relative amounts. These in turn give rise to different anomalies and impurity signals in the measured spectrum, which can be accounted for in cases of H-bonding and other solute interactions.

7.2.6 IR and VCD traces

As is the accepted standard for publication of VCD spectra, IR spectra are submitted alongside the VCD; doing so ensures the purity of the compound under investigation as well as clearly showing the positions of the vibrational signals along the wavenumber axis.

7.2.7 References in literature

Publications which reference the experimental VCD spectrum of the compound used should be included in database entries. This will enable easy referencing of the initial VCD experiment, as well as ensuring that credit is given for those performing the measurements.

7.3 User manipulation of experimental spectra

As with any online database, user interaction with the spectra is important. An initial version of the database should enable users to annotate and manipulate the experimental spectra, as well as applying scaling and fitting algorithms as appropriate. Further steps may be taken to enable the application of user-defined SF values, and once the user has chosen an ideal overlap, to automatically calculate similarity scores such as the ones described in this work.

Chapter 8

Conclusions

This chapter summarises the major conclusions from this work, as well as extrapolating possible directions for future investigation from these conclusions.

We put forward the use of multiplicative percentage scoring as a simple and efficient weighted method for comparing calculated and observed VCD data. The procedure is straightforward to implement and has led to the automated interpretation of thirty-one sets of enantiomeric spectra, as well as the assignment of AC for five pairs of previously unassigned or uncertain compounds.

Use of the B3PW91 functional and cc-pVTZ basis set generally increase the accuracy of VCD predictions. The improvement in multiplicative percentage score associated with changing to the B3PW91 functional increases when using the cc-pVTZ basis set, and synergistic effects may be at play when applying this combination.

Wavenumber scale factors between 0.970 and 0.980 are recommended for the purpose of comparing calculated and observed spectra. Under B3PW91/cc-pVTZ optimisation, at a SF of 0.975, the multiplicative score of the assignment varied from -15% to +77%, with a mean value of +36%.

The optimised method is applied to a compound from outside the original dataset and returns a confident assignment of absolute configuration based on the match between calculated and experimental VCD spectra.

For the newly assigned compound **68**, nine of the conformational minima found by the Monte Carlo method return highly similar conformational VCD spectra, as well as broadly

overlapping geometries. It is likely that several iterations of the same conformational minimum have been treated as distinct conformers by the conformational search. This may also be true of other compounds tested, and investigation into this phenomenon is included in the suggestions for future work.

The calculated VCD signals for all low energy conformers of the compounds analysed are available for download in XY format from the data repository of the University of Cambridge. The script used in implementing the procedure is likewise included in the accessible data.

8.1 Future work

Various categories exist for further investigation in regards to this work. Firstly, the baseline noise in both BioTools and Bruker instruments used can be corrected to some degree by enantiomer subtraction, but for cases where single enantiomer compounds are to be investigated, the baseline correction methods in use are unable to remove all the instances of artefact peaks or background noise. Further improvements may be possible to the baseline correction algorithm, enabling confident assignment of AC to compunds with only one stereoisomer available, such as is the case with natural products.

Secondly, while this work has explored the use of the B3LYP and B3PW91 functionals, as well as the 6-31G(d,p) and cc-pVTZ basis sets, many additional levels of theory are available for DFT calculation of VCD spectra. With the systematic application of other functionals and basis sets, it is possible to take steps towards finding better levels of theory by which VCD spectra can be predicted *ab initio*. A general application of solvent models in DFT, such as IEF-PCM or COSMO, may also be applied to improve the fit between calculated and observed VCD spectra, as well as to ascertain the accuracy of these models.

By applying the multiplicative percentage scoring method to more sets of calculated VCD data, the database can be further extended, simultaneously allowing evaluation of the scoring method. In addition to this, while concepts of solute-solvent H-bonding and carboxylic acid solute-solute dimerisation are well documented, the broader relationship between molecular structure and the accuracy of VCD prediction has yet to be elucidated. By adding newer compounds to this database, more light may be shed on the relationship between molecular structure and accuracy of VCD calculations.

Under Monte Carlo searching methods, the mis-identification of what should be a single conformational minimum as several different conformers is an issue that requires investigation.

Repeated inclusion of the same conformer in a Boltzmann weighted calculation can skew the predicted weights away from the physical reality; as a result, the identification and correction of these errors is an essential undertaking for accurate prediction not only of VCD but also of other spectroscopic methods.

Finally, development of the data acquired in this work into an online database where VCD spectra can be uploaded and compared, together with the availability of spectral annotation and manipulation, various methods of similarity scoring can be compared. It is hoped that with a large enough repository, in addition to saving researchers time and effort in predicting VCD spectra, this database can also contribute to refining calculation methods for the prediction of vibrational spectra.

References

[1] Pasteur, L. "Recherches sur les propriétés spécifiques des deux acides qui composent l'acide racémique." *Impr. Bachelier*, **1850**.

[2] N. Berova, K. Nakanishi and R. W. Woody. "Circular dichroism: principles and applications." John Wiley & Sons, **2000**.

[3] M. Quack. "How important is parity violation for molecular and biomolecular chirality?" *Angew. Chem. Int. Ed.*, 41(24):4618–4630, **2002**.

[4] W. G. McBride *et al*. "Thalidomide and congenital abnormalities." *Lancet*, 2(1358):90927–8, **1961**.

[5] N. Harada. "Chiral auxiliaries powerful for both enantiomer resolution and determination of absolute configuration by x-ray crystallography." *Top. Stereochem.*, 25:177, **2001**.

[6] J. A. Dale, D. L. Dull and H. S. Mosher. "α-methoxy-α-trifluoromethylphenylacetic acid, a versatile reagent for the determination of enantiomeric composition of alcohols and amines." *J. Org. Chem.*, 34(9):2543–2549, **1969**.

[7] H. Flack and G. Bernardinelli. "Absolute structure and absolute configuration." *Acta Crystallogr. B*, 55(5):908–915, **1999**.

[8] K. Mori. "Bioactive natural products and chirality." *Chirality*, 23(6):449–462, **2011**.

[9] P. Krastel, F. Petersen, S. Roggo, E. Schmitt and A. Schuffenhauer. "Aspects of chirality in natural products drug discovery." *Chirality in Drug Research*, 33:67–94, **2006**.

[10] L. A. Nafie. "Vibrational optical activity: principles and applications." John Wiley & Sons, **2011**.

[11] E. L. Eliel, S. H. Wilen and L. N. Mandel. "Stereochemistry of organic compounds." *Rec. Trav. Chim. Pays-Bas*, 114(8):378, **1995**.

[12] L. Barron, M. Bogaard and A. Buckingham. "Raman scattering of circularly polarized light by optically active molecules." *J. Am Chem. Soc.*, 95(2):603–605, **1973**.
J. Phys. Chem., 98(45):11623–11627, **1994**.

[13] L. A. Nafie. "Circular polarization spectroscopy of chiral molecules." *J. Mol. Struct.*, 347:83–100, **1995**.

[14] P. J. Stephens. "Theory of vibrational circular dichroism." *J.Phys.Chem.*, 89(5):748–752, **1985**.

[15] E. Jiang. "Advanced FT-IR spectroscopy: Principles, experiments and applications." Thermo Fisher Scientific Inc., Verona Road, Madison, **2003**.

[16] J. Cheeseman, M. Frisch, F. Devlin and P. J. Stephens. "*Abinitio* calculation of atomic axial tensors and vibrational rotational strengths using density functional theory." *Chem. Phys. Lett*, 252(3-4):211–220, **1996**.

[17] A. D. Buckingham. "Introductory lecture the theoretical background to vibrational optical activity." In *Optical,ElectricandMagneticPropertiesofMolecules*, pp. 291–302. Elsevier, **1997**.

[18] L. A. Nafie. "Dual polarization modulation: a real-time, spectral-multiplex separation of circular dichroism from linear birefringence spectral intensities." *Appl. Spectrosc.*, 54(11):1634–1645, **2000**.

[19] T. B. Freedman, X. Cao, R. K. Dukor and L. A. Nafie. "Absolute configuration determination of chiral molecules in the solution state using vibrational circular dichroism." *Chirality*, 15(9):743–758, **2003**.

[20] E. Castiglioni, P. Biscarini and S. Abbate. "Experimental aspects of solid state circular dichroism." *Chirality*, 21(1E):E28–E36, **2009**.

[21] J. M. Goodman *etal*. "Chemical applications of molecular modelling." Royal Society of Chemistry, **1998**.

[22] P. J. Stephens and F. J. Devlin. "Determination of the structure of chiral molecules using *ab initio* vibrational circular dichroism spectroscopy." *Chirality*, 12(4):172–179, **2000**.

[23] P. J. Stephens, F. J. Devlin and J-J. Pan. "The determination of the absolute configurations of chiral molecules using vibrational circular dichroism (VCD) spectroscopy." *Chirality*, 20(5):643–663, **2008**.

[24] G. Mazzeo, E. Santoro, A. Andolfi, A. Cimmino, P. Troselj, A. G. Petrovic, S. Superchi, A. Evidente, and N. Berova. "Absolute configurations of fungal and plant metabolites by chiroptical methods. ORD, ECD, and VCD studies on phyllostin, scytolide, and oxysporone." *J. Nat. Prod.*76(4):588–599, **2013**.

[25] P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch. "*Abinitio* calculation of vibrational absorption and circular dichroism spectra using density functional force fields." *J. Phys. Chem.*, 98(45):11623–11627, **1994**.

[26] M. Urbanova, V. Setnicka, F. J. Devlin and P. J. Stephens. "Determination of molecular structure in solution using vibrational circular dichroism spectroscopy: The supramolecular tetramer of *S*-2,2'-dimethyl-biphenyl-6,6'-dicarboxylic acid." *J. Am. Chem. Soc.,* 127(18):6700–6711, **2005**.

[27] V. P. Nicu, J. Neugebauer, S. K. Wolff and E. J. Baerends. "A vibrational circular dichroism implementation within a slater-type-orbital based density functional framework and its application to hexa-and hepta-helicenes." *Theor. Chem. Acc.*, 119(1-3):245–263, **2008**.

[28] C. Talotta, C. Gaeta, F. Troisi, G. Monaco, R. Zanasi, G. Mazzeo, C. Rosini and P. Neri. "Absolute configuration assignment of inherently chiral calix[4]arenes using DFT calculations of chiroptical properties." *Org. Lett.*, 12(13):2912–2915, **2010**.

[29] K. Monde, T. Taniguchi, N. Miura, S.-I. Nishimura, N. Harada, R. K. Dukor and L. A. Nafie. "Preparation of cruciferous phytoalexin related metabolites, (-)-dioxibrassinin and (-)-3-cyanomethyl-3-hydroxyoxindole, and determination of their absolute configurations by vibrational circular dichroism (VCD)." *Tetrahedron Lett.*, 44(32):6017–6020, **2003**.

[30] J. Shen, C. Zhu, S. Reiling and R. Vaz. "A novel computational method for comparing vibrational circular dichroism spectra." *Spectrochim. Acta A*, 76(3-4):418–422, **2010**.

[31] J. Tomasi, B. Mennucci and E. Cances. "The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM *ab initio* level." *J. Mol. Struct. THEOCHEM*, 464(1-3):211–226, **1999**.

[32] Y. Takano and K. Houk. "Benchmarking the conductor-like polarizable continuum model (cpcm) for aqueous solvation free energies of neutral and ionic organic molecules." *J. Chem. Theory Comput.*, 1(1):70–77, **2005**.

[33] A. Klamt. "Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena." *J. Phys. Chem.*, 99(7):2224–2235, **1995**.

[34] V. P. Nicu and E. J. Baerends. "Robust normal modes in vibrational circular dichroism spectra." *Phys. Chem. Chem. Phys.*, 11(29):6107–6118, **2009**.

[35] M. Losada and Y. Xu. "Chirality transfer through hydrogen-bonding: experimental and *ab initio* analyses of vibrational circular dichroism spectra of methyl lactate in water." *Phys. Chem. Chem. Phys.*, 9(24):3127–3135, **2007**.

[36] R. Schweitzer-Stenner, F. Eker, K. Griebenow, X. Cao and L. A. Nafie. "The conformation of tetraalanine in water determined by polarized Taman, FT-IR, and VCD spectroscopy." *J. Am. Chem. Soc.*, 126(9):2768–2776, **2004**.

[37] R. Day Jr. and A. Underwood. "Quantitative analysis, 4th edition." Prentice-Hall, **1982**.

[38] Maestro, Schrödinger, LLC, New York, NY, **2012**.

[39] N. L. Allinger. "Conformational analysis 130. MM2: A hydrocarbon force field utilizing V_1 and V_2 torsional terms." *J. Am. Chem. Soc.*, 99(25):8127– 8134, **1977**.

[40] F. Mohamadi, N. G. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson and W. C. Still. "Macromodel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics." *J. Comput. Chem.*, 11(4):440–467, **1990**.

[41] G. Chang, W. C. Guida and W. C. Still. "An internal-coordinate monte carlo method for searching conformational space." *J. Am. Chem. Soc.*, 111(12):4379–4386, **1989**.

[42] I. Kolossváry and W. C. Guida. "Low mode search. an efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides." *J. Am. Chem. Soc.*, 118(21):5011–5019, **1996**.

[43] T. A. Halgren. "MMFF VI. MMFF94s option for energy minimization studies." *J. Comput. Chem.*, 20(7):720–729, **1999**.

[44] T. A. Halgren. "MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries." *J. Comput. Chem.*, 20(7):730–748, **1999**.

[45] S. G. Smith, Ph.D. dissertation, University of Cambridge, **2010**.

[46] Jaguar, Schrödinger, LLC, New York, NY, **2012**.

[47] A. D. Becke. "Density-functional exchange-energy approximation with correct asymptotic behavior." *Phys. Rev. A*, 38(6):3098–3100, **1988**.

[48] C. Lee, W. Yang and R. G. Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density." *Phys. Rev. B*, 37(2):785, **1988**.

[49] A. D. Becke. "Density-functional thermochemistry III: The role of exact exchange." *J. Chem. Phys.*, 98(7):5648–5652, **1993**.

[50] J. A. Pople, P. von R. Schleyer and W.J. Hehre. "*Ab Initio* Molecular Orbital Theory." Wiley, New York, **1986**.

[51] T. H. Dunning Jr. "Gaussian basis sets for use in correlated molecular calculations I: The atoms boron through neon and hydrogen." *J. Chem. Phys.*, 90(2):1007–1023, **1989**.

[52] S. Góbi, E. Vass, G. Magyarfalvi and G. Tarczay. "Effects of strong and weak hydrogen bond formation on VCD spectra: a case study of 2-chloropropionic acid." *Phys. Chem. Chem. Phys.*, 13(31):13972–13984, **2011**.

[53] C. S. Ashvar, P. J. Stephens, T. Eggimann and H. Wieser. "Vibrational circular dichroism spectroscopy of chiral pheromones: frontalin (1, 5-dimethyl-6, 8-dioxabicyclo[3.2.1]octane)." *Tetrahedron: Asymmetry*, 9(7):1107–1110, **1998**.

[54] C. S. Ashvar, F. J. Devlin, and P. J. Stephens. "Molecular structure in solution: an *ab initio* vibrational spectroscopy study of phenyloxirane." *J. Am. Chem. Soc.*, 121(12):2836–2849, **1999**.

[55] L. Andernach and T. Opatz. "Assignment of the absolute configuration and total synthesis of (+)-caripyrin." *Eur. J. Org. Chem.*, **2014**(22):4780–4784.

[56] T. Fujita, K. Obata, S. Kuwahara, A. Nakahashi, K. Monde, J. Decatur and N. Harada. "(*R*)-(+)-[VCD-(–)-984]-4-ethyl-4-methyloctane: A cryptochiral hydrocarbon with a quaternary chiral center. (1) Synthesis of the enantiopure compound and unambiguous determination of absolute configuration." *Eur. J. Org. Chem.*, **2010**(33):6372–6384.

[57] S. Ma, C. A. Busacca, K. R. Fandrick, T. Bartholomeyzik, N. Haddad, S. Shen, H. Lee, A. Saha, N. Yee, C. Senanayake *et al*. "Directly probing the racemization of imidazolines by vibrational circular dichroism: Kinetics and mechanism." *Org. Lett.*, 12(12):2782–2785, **2010**.

[58] T. Kuppens, K. Vandyck, J. van der Eycken, W. Herrebout, B. van der Veken and P. Bultinck. "Determination of the Absolute Configuration of Three as-Hydrindacene compounds by vibrational circular dichroism." *J. Org. Chem.*, 70(23):9103–9114, **2005**.

[59] M. E.-A. Said, P. Vanloot, I. Bombarda, J-V. Naubron, A. Aamouche, M. Jean, N. Vanthuyne, N. Dupuy, C. Roussel *et al*. "Analysis of the major chiral compounds of *Artemisia herba-alba* essential oils (EOs) using reconstructed vibrational circular dichroism (VCD) spectra: En route to a VCD chiral signature of EOs." *Anal. Chim. Acta*, 903:121–130, **2016**.

[60] F. J. Devlin, P. J. Stephens, J. Cheeseman and M. Frisch. "Prediction of vibrational circular dichroism spectra using density functional theory: camphor and fenchone." *J. Am. Chem. Soc.*, 118(26):6327–6328, **1996**.

[61] J. C. Dobrowolski, M. H. Jamróz, R. Kołos, J. E. Rode and J. Sadlej. "Theoretical prediction and the first IR matrix observation of several L-cysteine molecule conformers." *ChemPhysChem*, 8(7):1085–1094, **2007**.

[62] A. Aamouche, F. J. Devlin, and P. J. Stephens. "Structure, vibrational absorption and circular dichroism spectra, and absolute configuration of Tröger's base." *J. Am. Chem. Soc.*, 122(10):2346–2354, **2000**.

[63] M. Reina, E. Burgueño-Tapia, M. A. Bucio and P. Joseph-Nathan. "Absolute configuration of tropane alkaloids from schizanthus species by vibrational circular dichroism." *Phytochemistry*, 71(7):810–815, **2010**.

[64] P. Mobian, C. Nicolas, E. Francotte, T. Burgi and J. Lacour. "Synthesis, resolution, and VCD analysis of an enantiopure diazaoxatricornan derivative." *J. Am. Chem. Soc.*, 130(20):6507–6514, **2008**.

[65] T. Buffeteau, D. Cavagnat, J. Bisson, A. Marchal, G. D. Kapche, I. Battistini, G. Da Costa, A. Badoc, J-P. Monti, J-M. Merillon *et al*. "Unambiguous determination of the absolute configuration of dimeric stilbene glucosides from the rhizomes of *Gnetum africanum*." *J. Nat. Prod.*, 77(8):1981–1985, **2014**.

[66] S. Kuwahara, K. Obata, T. Fujita, N. Miura, A. Nakahashi, K. Monde and N. Harada. "(*R*)-(+)-[VCD-(–)-984]-4-ethyl-4-methyloctane: A cryptochiral hydrocarbon with a quaternary chiral center. (2) Vibrational CD spectra of both enantiomers and absolute configurational assignment." *Eur. J. Org. Chem.*, **2010**(33):6385–6392.

[67] P. J. Stephens, F. J. Devlin and J. R. Cheeseman. "VCD spectroscopy for organic chemists." CRC Press, **2012**.

[68] T. Kuppens, K. Vandyck, J. van der Eycken, W. Herrebout, B. van der Veken and P. Bultinck. "A DFT conformational analysis and VCD study on methyl tetrahydrofuran-2-carboxylate." *Spectrochim. Acta A*, 67(2):402–411, **2007**.

[69] E. Debie, E. De Gussem, R. K. Dukor, W. Herrebout, L. A. Nafie and P. Bultinck. "A confidence level algorithm for the determination of absolute configuration using vibrational circular dichroism or Raman optical activity." *ChemPhysChem*, 12(8):1542–1549, **2011**.

[70] P. L. Polavarapu. "Chiroptical spectroscopy: fundamentals and applications." CRC Press, **2016**.

[71] Python, Python Software Foundation, **2015**.

[72] W. J. Hehre, R. Ditchfield and J. A. Pople. "Self-consistent molecular orbital methods XI:. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules." *J. Chem. Phys.*, 56(5):2257–2261, **1972**.

[73] USFDA. "FDA's policy statement for the development of new stereoisomeric drugs." Fed. Reg. 22, 249, **1992**.

[74] T. Yamaji, T. Saito, K. Hayamizu, M. Yanagisawa, and O. Yamamoto. "Spectral Database for Organic Compounds SDBS." National Institute of Advanced Industrial Science and Technology (AIST), Japan.

Appendix A

Bruker VCD Spectrometer instructions

To acquire VCD spectra using Bruker instrumentation, the TENSOR FTIR spectrometer is fitted with a PMA50 module for polarisation modulated measurements.

In order to ensure that the solute sample is at an appropriate concentration, an IR spectrum is usually acquired before taking a VCD measurement. This changes the settings for the instrument, requiring the spectrometer to be recalibrated before each VCD measurement.

The following instructions, prepared by Richard J. Lewis at AstraZeneca, are applicable to the Bruker instrumentation used in this work and provide a stepwise procedure to acquiring an IR spectrum on the spectrometer followed by a calibration and VCD measurement.

Acquiring Spectra on the Bruker VCD Spectrometer

Filling the sample cell

The sample cell takes around 70-80 μl solvent. For many organic compounds, around 6 mg compound in this volume gives a reasonable spectrum.

Bubbles can be a problem. The best way to get rid of them seems to be as follows:

1. Fill the cell from one side such that the filling port is horizontal. Capillary attraction will fill the cell, but leave a bubble at the top.

2. Plug the port used for filling but leave the other one unplugged. Rotate the cell such that the unplugged port and bubble are at the top.

3. Gently partially unplug the lower port. The bubble should exit the flow cell. Sometimes repeating the plugging and unplugging a few times is necessary.

Acquire a regular IR spectrum

This is measured in the room temperature detector (left hand side compartment of the instrument).

1. Click the test tube icon at top of screen to left of VCD.

2. Load experiment MIR_Transmission.xpm.

3. Check detector setting (Optic panel) is at RT_DLaTGS (internal).

4. Under the "Basic" tab, enter the sample description (this will be the file name). Additional notes about the sample can be entered under "sample form".

5. Run a background scan.

6. Insert sample (left hand side room temperature compartment, left hand side of sample holder).

7. Press "Sample single channel".

8. Save data in JCAMP format.

9. To save data as XY file, choose "Data point table" under "Mode" tab.

10. Remove the sample cell.

The absorbance of signals of interest should be between 0.3 and 0.8. If necessary the sample should be diluted before running a VCD spectrum.

Acquire a VCD spectrum

This is measured in the cooled detector (right hand side). A calibration should be made for each sample (this must be done if you have previously changed the settings to acquire an IR spectrum).

1. Cool the detector with liquid N_2 and wait for about 20 minutes.

Calibration

2. **Important!** Ensure that the sample cell has not been left in the IR compartment. If the sample is in the IR compartment when performing the calibration, the result will not be valid.

3. Load the VCD calibration experiment "PMA50-VCD-Calibration.xpm".

4. Reduce the sensitivity of the LIA to 1V (strong signal)

5. Insert the CdS multiple wavelength pate, ensuring that it is the correct way up. This should be on the left hand side of the sample holder.

6. Insert the polariser on the right hand side of the sample holder. The axis must be set at 0°.

7. In the "Optic" tab, check that the detector setting is LN-MCT Narrow 24h [PMA module].

8. Go to "Check signal" tab and confirm that the interferogram shows a single burst in the centre of the spectrum.

9. In the "Optic" tab, change the detector setting to DC-IN [PMA module].

10. Go to "Check signal" tab and confirm that the interferogram shows a double burst centred in the spectrum.

11. Press autophase on the LIA. This sometimes needs to be repeated several times. **Important. Autophase MUST complete correctly in order to get correct results and correct phase of spectrum.**

12. If the autophase does not work (the LIA displays "bad phase" then increase the time constant by 2-3 steps (top left hand buttons). Try again. When locked, reduce the time constant back again to its original value (30 μs).

13. In the "Optic" tab, change the detector setting to LN-MCT Narrow 24h AC+DC_IN [PMA module].

14. Press "Sample single channel" (on the "Basic" tab).

15. After measurement, remove the CdS plate and polariser.

Sample measurement

16. Insert sample into left hand side of the sample holder.

17. Load the VCD experiment "PMA50-VCD-Measurement.xpm".

18. In the "Optic" tab, change the detector setting to LN-MCT Narrow 24h [PMA module].

19. Optimise the aperture setting (on the "Optic" tab; 6mm maximum seems most used). Go to "Check signal" tab and confirm that the interferogram shows a single burst in the centre of the spectrum. Note the ADC counts. (ADC counts rarely reaches optimum level with maximum aperture.)

20. Change the sensitivity of the LIA back to 1mV.

21. In the "Optic" tab, change the detector setting to LN-MCT Narrow 24h AC+DC_IN [PMA module].

22. Adjust experiment timing (on the "Advanced" tab). 4000 scans takes 1 hour.

23. Press "Sample single channel" (on the "Basic" tab) to start sample measurement.

Calculation

24. Press the VCD button on the top menu. Drag and drop calibration and data files to the relevant window.

25. Under the "Adjust parameter" tab, press "Normalization" and make sure the figures in "Calibration" and "Sample" are both 1000.

26. Press "Calculate".

27. To manipulate the curves, right-click in the window and select "Shift curve", "Whole curve" or "Top" to move or expand as appropriate. Then click and drag the curve.

28. Save data in JCAMP format.

29. To save data as XY file, choose "Data point table" under "Mode" tab.

Appendix B

Python scripts

These are the Python scripts developed and used in this work. The optimised method is shown in the form of `vcd_tomte.py` and is the final version of the script written for use by AstraZeneca's VCD analyses. The various baseline correction and weighting parameters detailed in Chapter 5 appear in `modvcd.py`, a longer script with modifiable sections of code which can be switched on and off for ease of use.

vcd_tomte.py

```
#created by Jonathan Lam for AstraZeneca
#!/user/bin/python

import os, numpy as np
from operator import sub
from scipy import interp, optimize

#a simple function for checking if input is interpreted as number
def is_number(isn_s):
  try:
    float(isn_s)
    return True
  except ValueError:
    return False

#a function for extracting data from .CSV files
def get_data(dimension, spec_file, wav_min, wav_max):
  lst_axis = []
  for line in open(spec_file):
    check_value = line.split( )[0].split(',')[0]
    if (is_number(check_value) and
          wav_min < float(check_value) < wav_max):
      array_targt = line.split( )[0].split(',')[dimension]
      lst_axis.append(float(array_targt))
  return lst_axis

#
def my_interp(mark, horz_lo, horz_hi, vert_lo, vert_hi):
  vert_diff = vert_hi - vert_lo
  horz_diff = horz_hi - horz_lo
  m = vert_diff / float(horz_diff)
  return m *(mark - horz_lo) + vert_lo
```

```
#function to interpolate graphs ('fine')
#onto a different horizontal scale ('cors')
def interpolate(lst_horz_fine, lst_vert_fine, lst_horz_cors):
  if len(lst_horz_fine) != len(lst_vert_fine):
    print('ERROR: Interpolating different length templates!')
    exit()
  lst_vert_cors = []
  for cors_mark in lst_horz_cors:
    if  (cors_mark < lst_horz_fine[0] or
          cors_mark > lst_horz_fine[-1]):
      lst_vert_cors.append(0)
    elif cors_mark in lst_horz_fine:
      ind = lst_horz_fine.index(cors_mark)
      lst_vert_cors.append(lst_vert_fine[ind])
    else:
      ind_hi = 0
      while cors_mark > lst_horz_fine[ind_hi]:
        ind_hi += 1
      ind_lo = ind_hi - 1
      extp = lst_vert_fine[ind_lo]
      lst_vert_cors.append(my_interp(
      cors_mark,
      lst_horz_fine[ind_lo],lst_horz_fine[ind_hi],
      lst_vert_fine[ind_lo],lst_vert_fine[ind_hi]))
  return lst_vert_cors

#
def sq_scale(lst_one, lst_two):
  one_sqsum = sum([n**2 for n in lst_one])
  two_sqsum = sum([n**2 for n in lst_two])
  sq_sf = np.sqrt(two_sqsum / one_sqsum)
  return [sq_sf * i for i in lst_one]

#
def score(axis, expt, calc):
  #scale expt signals to range of calc signals
```

```
  scaled_expt = sq_scale(expt, calc)
  #calculate multiplicative score
  multiplied = []
  for i in range(len(axis)):
    multiplied.append(scaled_expt[i] * calc[i])
  return sum(multiplied) / sum([n**2 for n in calc])

#a function to be minimized by scipy.optimize.
#Calculates fit between spectra
def fit(shift, calc_wav, calc_sig, expt_wav, expt_sig):
  shift_calc_wav = [shift + i for i in calc_wav]
  intp_calc_sig = interpolate(shift_calc_wav, calc_sig, expt_wav)
  return -abs(score(expt_wav, expt_sig, intp_calc_sig))

#main
#get experimental data
for item in sorted(os.listdir('.')):
  print(item)
alfa_expt_file = input('Experimental File A: ')
beta_expt_file = input('Experimental File B: ')
expt_wavenums = get_data(0, alfa_expt_file, 1000, 2000)

#check both experimental files have same wavenumber scale
if expt_wavenums != get_data(0, beta_expt_file, 1000, 2000):
  print('ERROR: Different wavenumber scale! Exiting.')
  exit()
alfa_expt_signals = get_data(1, alfa_expt_file, 1000, 2000)
beta_expt_signals = get_data(1, beta_expt_file, 1000, 2000)

#subtractive baseline correction: generates corr_expt_signals
corr_expt_signals = list(map(sub,
alfa_expt_signals,
beta_expt_signals))

#get calculated data
alfa_calc_file = input('Calculated File A: ')
```

```
alfa_calc_wavenums = get_data(0, alfa_calc_file, 1000, 2000)
alfa_calc_signals = get_data(1, alfa_calc_file, 1000, 2000)

beta_calc_file = input('Calculated File B: ')
beta_calc_wavenums = get_data(0, alfa_calc_file, 1000, 2000)
beta_calc_signals = get_data(1, beta_calc_file, 1000, 2000)

#fitting
print('Fitting Calculation A...')
alfa_fit = optimize.minimize(fit, 0,
(alfa_calc_wavenums, alfa_calc_signals,
expt_wavenums, corr_expt_signals))
alfa_shift = alfa_fit.x[0]
print('Calculation A fitted with shift of ' +
str(alfa_shift) +
' wavenumbers')
afit_calc_wavenums = [alfa_shift + i for i in alfa_calc_wavenums]

print('Fitting Calculation B...')
beta_fit = optimize.minimize(fit, alfa_shift,
(beta_calc_wavenums, beta_calc_signals,
expt_wavenums, corr_expt_signals))
beta_shift = beta_fit.x[0]
print('Calculation B fitted with shift of ' +
str(alfa_shift) +
' wavenumbers')
bfit_calc_wavenums = [beta_shift + i for i in beta_calc_wavenums]

#interpolation of calculated signals to expt wavenumber scale
aext_calc_signals = interpolate(
afit_calc_wavenums, alfa_calc_signals, expt_wavenums)
bext_calc_signals = interpolate(
bfit_calc_wavenums, beta_calc_signals, expt_wavenums)

#comparison with XYZ
a_final = score(expt_wavenums,
```

```
corr_expt_signals,
aext_calc_signals)
b_final = score(expt_wavenums,
corr_expt_signals,
bext_calc_signals)
print("Match score between Calc'd file A and Corrected Exp't A:")
print(a_final)
print("Match score between Calc'd file B and Corrected Exp't A:")
print(b_final)

#decision and confidence
confid = str(round(50 * abs(a_final - b_final), 1))

if a_final > b_final:
  print('Input assignment is SUPPORTED with ' +
confid +
'% confidence.')
else:
  print('Input assignment is REVERSED with ' +
confid +
'% confidence.')
```

modvcd.py

```
#created by Jonathan Lam for model testing
#and optimisation
#!/usr/bin/python

import os, re, glob, math, numpy as np
import matplotlib.mlab as mlab
from operator import add, sub
from scipy import optimize
from pdb import set_trace as bp

os.chdir('VCD/Database/')
#
def is_number(isn_s):
  try:
    float(isn_s)
    return True
  except ValueError:
    return False
#
def check_empty(lst_one,lst_two):
  if len(lst_one) * len(lst_two) != 0:
    print('ERROR: Different types of spectra present in folder')
    exit()
#
def check_two(lst):
  if len(lst) != 2:
    print('ERROR: Wrong number of spectra')
    exit()
#
def check_same(lst_one, lst_two):
  if lst_one != lst_two:
    print('ERROR: Non-matching x-Axis for spectra')
    exit()
```

```
#
def check_equal_length(lst_one, lst_two):
  if len(lst_one) != len(lst_two):
    print('ERROR: lists are unequal length')
    exit()
#
def sq_scale(lst_one, lst_two):
  one_sqsum = sum([n**2 for n in lst_one])
  two_sqsum = sum([n**2 for n in lst_two])
  sq_sf = np.sqrt(two_sqsum / one_sqsum)
  return [sq_sf * i for i in lst_one]
#
def get_biot_xdata(biot_spec_file, wav_min, wav_max):
  lst_wav_x_biot = []
  for line in open(biot_spec_file):
    check_value = line.split( )[0]
    if (is_number(check_value) and
wav_min < float(check_value) < wav_max):
      array_targt = line.split( )[0]
      lst_wav_x_biot.append(float(array_targt))
  return lst_wav_x_biot
#
def get_biot_ydata(biot_spec_file, wav_min, wav_max):
  lst_vib_y_biot = []
  for line in open(biot_spec_file):
    check_value = line.split( )[0]
    if (is_number(check_value) and
wav_min < float(check_value) < wav_max):
      array_targt = line.split( )[1]
      lst_vib_y_biot.append(float(array_targt))
  return lst_vib_y_biot
#
def get_bruk_xdata(bruk_spec_file, wav_min, wav_max):
  lst_wav_x_bruk = []
  for line in reversed(list(open(bruk_spec_file))):
    check_value = line.split( )[0]
```

```
    if (is_number(check_value) and
wav_min < float(check_value) < wav_max):
      array_targt = line.split( )[0]
      lst_wav_x_bruk.append(float(array_targt))
  return lst_wav_x_bruk
#
def get_bruk_ydata(bruk_spec_file, wav_min, wav_max):
  lst_vib_y_bruk = []
  for line in reversed(list(open(bruk_spec_file))):
    check_value = line.split( )[0]
    if (is_number(check_value) and
wav_min < float(check_value) < wav_max):
      array_targt = line.split( )[1]
      lst_vib_y_bruk.append(float(array_targt))
  return lst_vib_y_bruk
#
BIOxData = get_biot_xdata('BiotBaseLine.txt',
wav_min = 1000,
wav_max = 1800)
BIOyData = get_biot_ydata('BiotBaseLine.txt',
wav_min = 1000,
wav_max = 1800)
BIOmean = np.mean([abs(item) for item in BIOyData])
BRUxData = get_bruk_xdata('BrukBaseLine.txt',
wav_min = 950,
wav_max = 1800)
BRUyData = get_bruk_ydata('BrukBaseLine.txt',
wav_min = 950,
wav_max = 1800)
BRUmean = np.mean([abs(item) for item in BRUyData])
#
def get_biot_val(VALx):
  if VALx in BIOxData:
    ind = BIOxData.index(VALx)
    return BIOyData[ind]
  elif VALx < BIOxData[0]:
```

```
    return -0.0278530778
  elif VALx > BIOxData[-1]:
    return 0.0002336204 * VALx - 0.382694828
  else:
    ind = 0
    while VALx > BIOxData[ind]:
      ind += 1
    BIOyData[ind] + BIOyData[ind + 1]
    BIOnume = BIOyData[ind] + (BIOyData[ind + 1] - BIOyData[ind])
    BIOdeno = BIOxData[ind + 1] - BIOxData[ind]
    BIOdenb = VALx - BIOxData[ind]
    return BIOnume / (BIOdena * BIOdenb)
#
def get_bruk_val(VALx):
  if VALx in BRUxData:
    ind = BRUxData.index(VALx)
    return BRUyData[ind]
  elif VALx < BRUxData[0]:
    return 0.0015218237 * VALx - 1.4742669374
  elif VALx > BRUxData[-1]:
    return -0.002578096 * VALx + 4.6538036161
  else:
    ind = 0
    while VALx > BRUxData[ind]:
      ind += 1
    BRUyData[ind] + BRUyData[ind + 1]
    BRUnume = BRUyData[ind] + (BRUyData[ind + 1] - BRUyData[ind])
    BRUdena = BRUxData[ind + 1] - BRUxData[ind]
    BRUdenb = VALx - BRUxData[ind]
    return BRUnume / (BRUdena * BRUdenb)
#
def fitpoly(pfit, var):
  tot = 0
  pwr = len(pfit) - 1
  for coeff in pfit:
    tot += coeff*(var**pwr)
```

```
    pwr -= 1
  return tot
#
def get_enants(index):
  return sorted(list(
set([item[:-index] for item in glob.glob('*.out')])
))
#
def get_confs(enant):
  return sorted(
[item.split('_vcd.spm')[0] for item in glob.glob(
enant + '*_vcd.spm'
)]
)
#
def get_calpeak(conf, wav_min, wav_max):
  lst_cal_peak = []
  opt_echo = -1
  for line in open(conf + '_vcd.spm'):
    if opt_echo == 1:
      if ':::' in line:
        break
      elif wav_min <= float(
re.sub(r'\s+', ' ', line
).split( )[1]) <= wav_max:
        lst_cal_peak.append(
re.sub(r'\s+', ' ', line
).split( ))
    if opt_echo == 0 and ':::' in line:
      opt_echo = 1
    if opt_echo == -1 and 's_j_Symmetry\n' in line:
      opt_echo = 0
  return lst_cal_peak
#
def loren(lor_xi, lor_Ap, lor_xo, lor_gm):
  lor_diff = (lor_xi - lor_xo) / lor_gm
```

```
  return lor_Ap / (1 + lor_diff **2)
#
def gauss(gau_xi, gau_Ap, gau_mu, gau_SD):
  return gau_Ap * np.exp(-(gau_xi - gau_mu)**2/(2.* gau_SD**2))
#
Phy_RootTwoPi = np.sqrt(2 * math.pi)
def gamus(gau_xi, gau_Ap, gau_mu, gau_SD):
  return gau_Ap * gau_SD * Phy_RootTwoPi * mlab.normpdf(
gau_xi, gau_mu, gau_SD)
#
def squarewave(sqa_xi, sqa_Ap, sqa_mu, sqa_SD):
  if abs(sqa_xi - sqa_mu) < sqa_SD:
    return sqa_Ap
  else:
    return 0
#
def sanit(lst, peakfn, template_EXPx):
  lst_FITdist = []
  lst_vib_ExSani = [0 for item in lst]
  num_iter = 0
  while True:
    try:
      yAbs = [abs(item) for item in lst]
      iPeak = np.argmax(yAbs)
      Peak = lst[iPeak]
      iMin = iPeak
      while abs(lst[iMin]) > abs(Peak/2):
        iMin -= 1
      iMax = iPeak
      while abs(lst[iMax]) > abs(Peak/2):
        iMax += 1
      EXPFitSolution = optimize.curve_fit(
peakfn,
template_EXPx[iMin:iMax + 1],
lst[iMin:iMax + 1],
p0 = [Peak, template_EXPx[iPeak], 5]
```

```
)

      #option to break if too wide
      #if EXPFitSolution[0][2] > 40:
      # break
      #/option

      sanit_ap = EXPFitSolution[0][0]
      sanit_mu = EXPFitSolution[0][1]
      sanit_sd = EXPFitSolution[0][2]
      lst_FITdist.append(EXPFitSolution[0])

      y_part = [peakfn(
item,
sanit_ap,
sanit_mu,
sanit_sd
) for item in template_EXPx]
      lst_vib_ExSani = list(map(add, lst_vib_ExSani, y_part))

      lst = list(map(sub, lst, y_part))
      num_iter += 1
    except KeyboardInterrupt:
      raise
    except:
      break
  return lst_vib_ExSani
#
def drape(peakfn, data, dra_Ap, dra_mu, dra_SD):
  base = []
  for i in data:
    soln = peakfn(i, dra_Ap, dra_mu, dra_SD)
    base.append(soln)
  return base
#
Phy_Enrg_F = 2625.49962     # Hartree to kJ/mol
```

```
Phy_Bolt_k = 1.380648813e-23    # Boltzmann constant
Phy_Avog_L = 6.0221412927e23    # Avogadro constant
Phy_Room_T = 298.15     # Room Temperature
#
def get_energy_kjmol(entry):
  lst_hart_instance = []
  for line in open(entry + '.out'):
    if (re.search('energy: ', line)
and re.search('hartrees', line)):
      lst_hart_instance.append(
float(line.split(' hartrees')[0].split( )[-1])
)
  return Phy_Enrg_F * float(lst_hart_instance[-1])
#
def boltz_calc(confs, horz_scale_factor,
wav_horz_shift, template_EXPx):
  build_ydata = [0 for i in template_EXPx]
  lst_enrgs_kjmol = []

  #get Boltzmann weights from confs
  for conf in confs:
    lst_enrgs_kjmol.append(get_energy_kjmol(conf))
  ground_st_enrg = min(lst_enrgs_kjmol)
  lst_rel_enrgs = [item - ground_st_enrg for item in lst_enrgs_kjmol]
  lst_rel_boltz = [math.exp(
-1000 * item / Phy_Bolt_k / Phy_Avog_L / Phy_Room_T
) for item in lst_rel_enrgs]
  lst_abs_boltz = [item / sum(lst_rel_boltz) for item in lst_rel_boltz]

  for i in range(len(confs)):
    for item in get_calpeak(confs[i], wav_min = 975, wav_max = 2052):
      calc_wav = horz_scale_factor * float(item[1]) + wav_horz_shift
      calc_amp = lst_abs_boltz[i] * float(item[2])
      lst_add = drape(loren, aEXPxData, calc_amp, calc_wav, 5)
      build_ydata = list(map(add, build_ydata, lst_add))
  return build_ydata
```

```
#
def dft_ground(confs, horz_scale_factor,
wav_horz_shift, template_EXPx):
  lst_enrgs_kjmol = []

  #get Boltzmann weights from confs
  for conf in confs:
    lst_enrgs_kjmol.append(get_energy_kjmol(conf))
  ground_st_enrg = min(lst_enrgs_kjmol)
  ind_min = lst_enrgs_kjmol.index(ground_st_enrg)
  ground_state = get_calpeak(confs[ind_min],
wav_min = 975, wav_max = 2052)
  build_ydata = [0 for i in template_EXPx]
  for item in ground_state:
    calc_wav = horz_scale_factor * float(item[1]) + wav_horz_shift
    calc_amp = float(item[2])
    lst_add = drape(loren, aEXPxData, calc_amp, calc_wav, 5)
    build_ydata = list(map(add, build_ydata, lst_add))
  return build_ydata
#
def equal_portion(confs, horz_scale_factor,
wav_horz_shift, template_EXPx):
  build_ydata = [0 for i in template_EXPx]
  for conf in confs:
    for item in get_calpeak(conf,
wav_min = 975, wav_max = 2052):
      calc_wav = horz_scale_factor * float(item[1]) + wav_horz_shift
      calc_amp = float(item[2])
      lst_add = drape(loren, aEXPxData, calc_amp, calc_wav, 5)
      build_ydata = list(map(add, build_ydata, lst_add))
  return build_ydata
#
def ground_state_calc(confs, horz_scale_factor,
wav_horz_shift, template_EXPx):
  ground_state = get_calpeak(confs[0],
wav_min = 975, wav_max = 2052)
```

```
  build_ydata = [0 for i in template_EXPx]
  for item in ground_state:
    calc_wav = horz_scale_factor * float(item[1]) + wav_horz_shift
    calc_amp = float(item[2])
    lst_add = drape(loren, aEXPxData, calc_amp, calc_wav, 5)
    build_ydata = list(map(add, build_ydata, lst_add))
  return build_ydata
#
def score(getcalc, nam_exp_ydata, nam_cal_enant):
  lst_conf = get_confs(nam_cal_enant)

  #choose which method of ybuild to use
  CALCyData = boltz_calc(lst_conf, getcalc[0], getcalc[1], getcalc[2])

  check_equal_length(nam_exp_ydata, CALCyData)

  nam_exp_ydata = sq_scale(nam_exp_ydata, CALCyData)

  multiplied = []
  for i in range(len(CALCyData)):
    multiplied.append(nam_exp_ydata[i] * CALCyData[i])
  return sum(multiplied) / sum([n**2 for n in CALCyData])

#main
opt_subtrac = input('To apply the SubtractiveCorrection Method, ' +
'please type 1: ') == str(1)
opt_setcurv = input('To apply the SetCurve Baseline Correction, ' +
'please type 2: ') == str(2)
opt_setcuSC = input('To apply the ScaledSC Baseline Correction, ' +
'please type 3: ') == str(3)
opt_polyfit = input('To use the Polynomial Baseline Correction, ' +
'please type 4: ') == str(4)
opt_constan = input('To apply the Constant Baseline Correction, ' +
'please type 5: ') == str(5)
opt_sanitis = input('To apply the Experimental Data Sanitiser, ' +
'please type 6: ') == str(6)
```

```
horiz_scale = float(input('Please type the value of the desired SF: '
os.chdir('Compounds')
for fol_compnd in sorted(os.listdir('.')):
  #fol_compnd = 'Limonen'
  os.chdir(fol_compnd)
  print('Analysing ' + fol_compnd)

  #compile list of instruments
  lst_fol_instrm = []
  if 'Biotls' in os.listdir('.'):
    lst_fol_instrm.append('Biotls')
  if 'Bruker' in os.listdir('.'):
    lst_fol_instrm.append('Bruker')

  #go to instrument directory
  for fol_instrm in lst_fol_instrm:
    os.chdir(fol_instrm)
    lst_BIOspec = list(glob.glob('*_VCD.PRN'))
    lst_BRUspec = list(glob.glob('*vcd*.dpt'))

    #print('BIO' + str(lst_BIOspec))
    #print('BRU' + str(lst_BRUspec))

    check_empty(lst_BIOspec, lst_BRUspec)
    lst_nam_allspec = sorted(lst_BIOspec + lst_BRUspec)
    check_two(lst_nam_allspec)

    if fol_instrm == 'Biotls':
      aName = lst_nam_allspec[0]
      aEXPxData = get_biot_xdata(aName,
wav_min = 1000, wav_max = 2000)
      aEXPyData = get_biot_ydata(aName,
wav_min = 1000, wav_max = 2000)
      bName = lst_nam_allspec[1]
      bEXPxData = get_biot_xdata(bName,
wav_min = 1000, wav_max = 2000)
```

```
      bEXPyData = get_biot_ydata(bName,
wav_min = 1000, wav_max = 2000)
      get_base_val = get_biot_val
      BASmean = BIOmean
      FITmin, FITmax = 0.5, 2.0
    elif fol_instrm == 'Bruker':
      aName = lst_nam_allspec[0]
      aEXPxData = get_bruk_xdata(aName,
wav_min = 1000, wav_max = 2000)
      aEXPyData = get_bruk_ydata(aName,
wav_min = 1000, wav_max = 2000)
      bName = lst_nam_allspec[1]
      bEXPxData = get_bruk_xdata(bName,
wav_min = 1000, wav_max = 2000)
      bEXPyData = get_bruk_ydata(bName,
wav_min = 1000, wav_max = 2000)
      get_base_val = get_bruk_val
      BASmean = BRUmean
      FITmin, FITmax = 0.6, 1.4

    #baseline correction
    if opt_subtrac: #subtractive
      aEXPyData = list(map(sub, aEXPyData, bEXPyData))
      bEXPyData = [-i for i in aEXPyData]
      opt_setcurv = opt_setcuSC = opt_polyfit = opt_constan = 0

    if opt_setcurv:   #set curve
      setc_scale = 1.0

      if opt_setcuSC:   #scaling required for set curve
        aFITscale = np.mean(
[abs(item) for item in aEXPyData]) / BASmean
        bFITscale = np.mean(
[abs(item) for item in bEXPyData]) / BASmean

          #only if both confs need scale down,
```

```
        if aFITscale < FITmin and bFITscale < FITmin:
          #reduce setc_scale by factor of 20
          setc_scale *= 0.05

        elif ((aFITscale < FITmin and bFITscale > FITmin) or
        (aFITscale > FITmin and bFITscale < FITmin)):
          print(
'ERROR: Setcurve fitting values differ significantly')
          exit()

      def get_setc_val(item):   #multiply by scale factor
        return setc_scale * get_base_val(item)
      aEXPyData = list(map(
sub, aEXPyData, [get_setc_val(item) for item in aEXPxData]))
      bEXPyData = list(map(
sub, bEXPyData, [get_setc_val(item) for item in bEXPxData]))

    if opt_polyfit:   # fits a 4th order polynomial
      polybas = np.polyfit(aEXPxData, aEXPyData, 4)
      pBasData = [fitpoly(polybas, i) for i in aEXPxData]
      aEXPyData = list(map(sub, aEXPyData, pBasData))

      polybas = np.polyfit(bEXPxData, bEXPyData, 4)
      pBasData = [fitpoly(polybas, i) for i in bEXPxData]
      bEXPyData = list(map(sub, bEXPyData, pBasData))

    if opt_constan:   # subtracts a constant
      constmean = np.mean(aEXPyData)
      aEXPyData = list(map(
sub, aEXPyData, [constmean for item in aEXPyData]))

      constmean = np.mean(bEXPyData)
      bEXPyData = list(map(
sub, bEXPyData, [constmean for item in bEXPyData]))

    #Sanitise
```

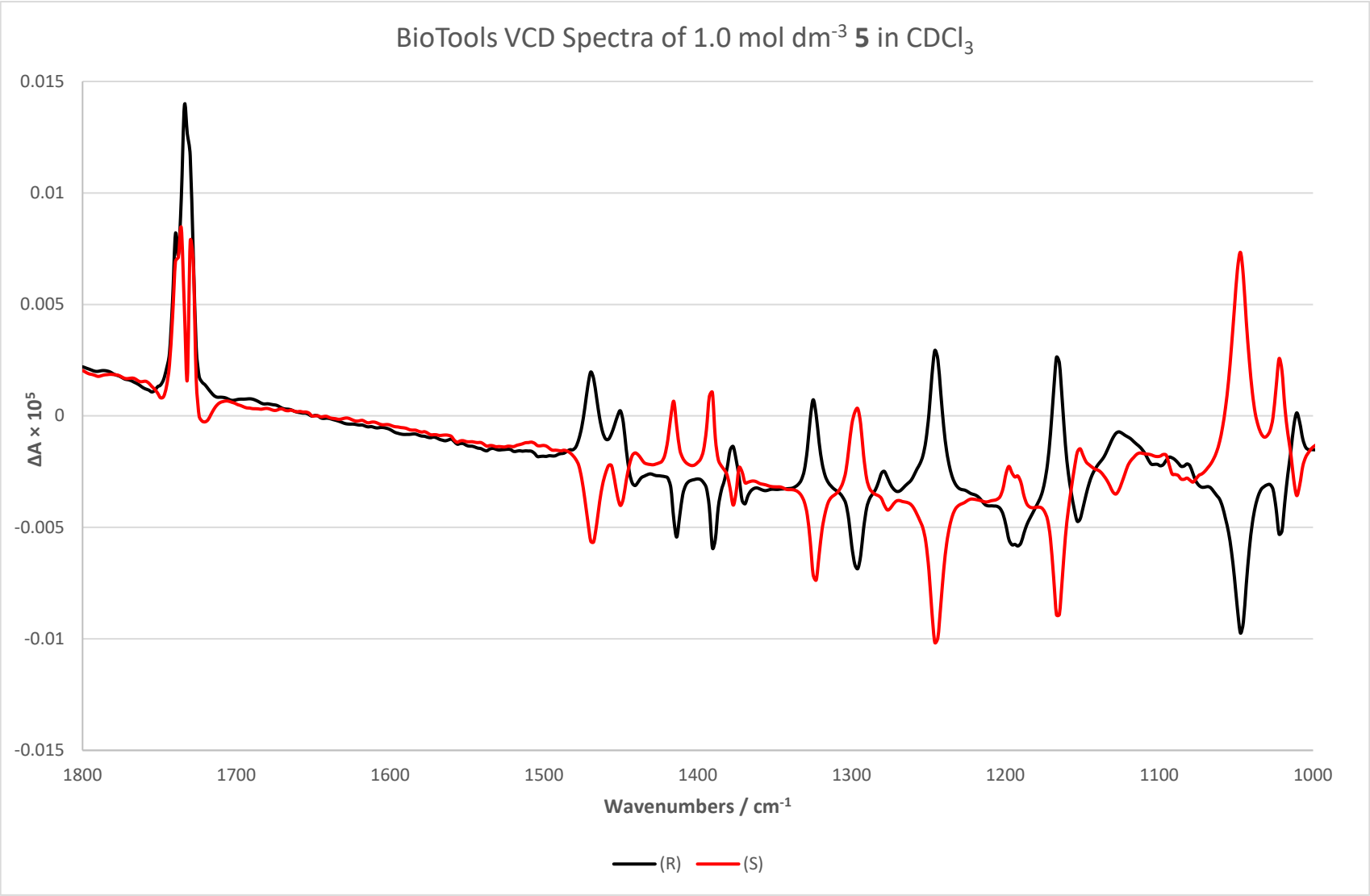
```
    if opt_sanitis:
      #bp()
      aEXPyData = sanit(aEXPyData, loren, aEXPxData)
      #bp()
      bEXPyData = sanit(bEXPyData, loren, bEXPxData)

    os.chdir('..')
    check_same(aEXPxData, bEXPxData)

    #get calculation types
    lst_fol_CalTyp = glob.glob('B3LYP*')
    for fol_CalTyp in lst_fol_CalTyp:

      os.chdir(fol_CalTyp)
      if not '-' in glob.glob('*.out')[0]:
        ind_namchar = 7
      else:
        ind_namchar = 22

      lst_enant = get_enants(ind_namchar)
      check_two(lst_enant)

      def func_to_min_a(num):
        score_parameters = [horiz_scale, num, aEXPxData]  #0.975
        final_score = score(score_parameters,
aEXPyData, lst_enant[0])
        return -final_score**2

      sft_a = optimize.minimize(func_to_min_a, 0).x
      score_parameters = [horiz_scale, sft_a, aEXPxData]  #0.975
      a_score = score(score_parameters, aEXPyData, lst_enant[0])
  #   print(fol_compnd, fol_CalTyp, a_score[0])

      b_score = score(score_parameters, bEXPyData, lst_enant[1])
  #   print(fol_compnd, fol_CalTyp, b_score[0])
```

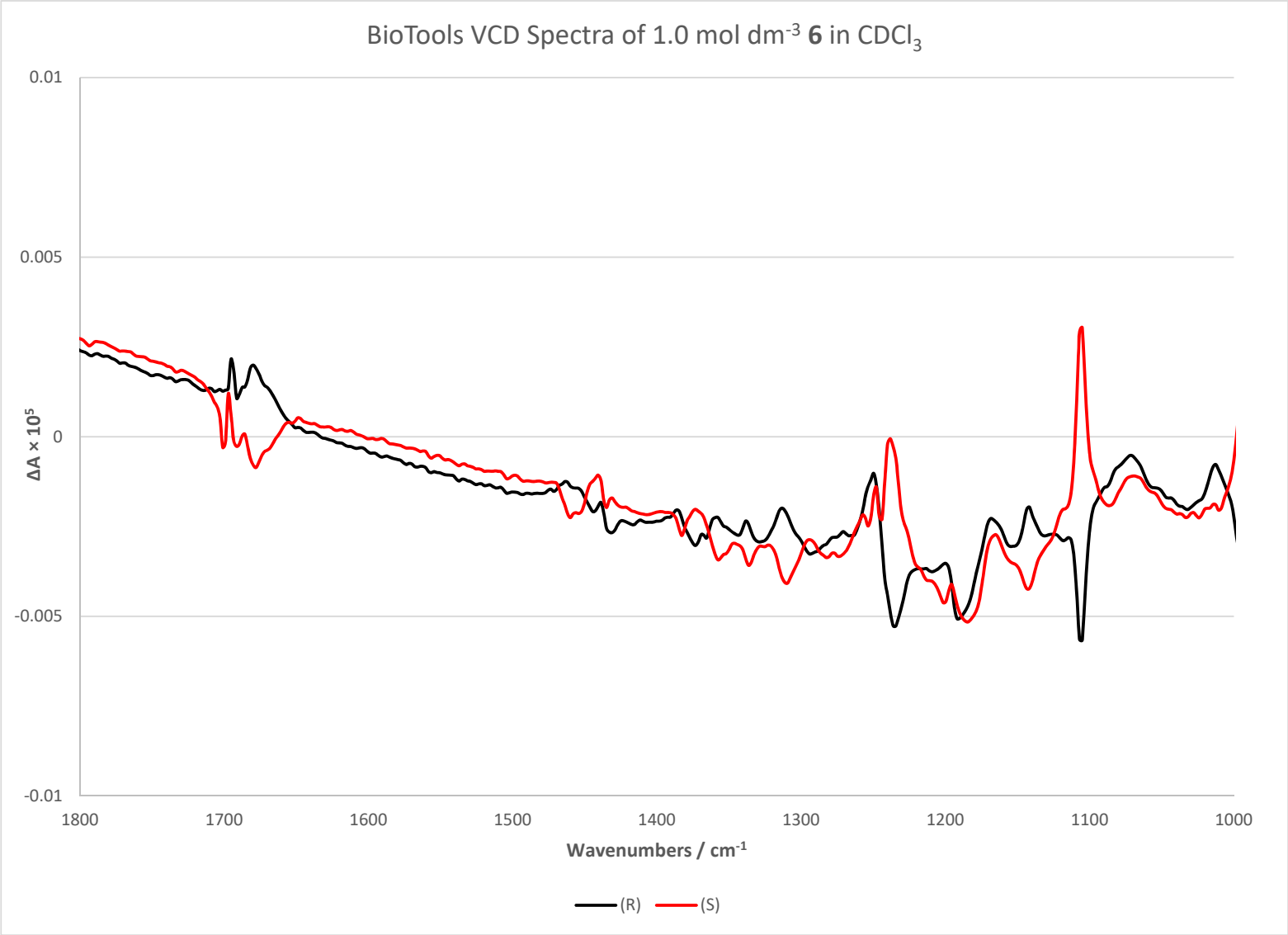
```
        c_score = -score(score_parameters, aEXPyData, lst_enant[1])
  #     print(fol_compnd, fol_CalTyp, c_score[0])

        d_score = -score(score_parameters, bEXPyData, lst_enant[0])
  #     print(fol_compnd, fol_CalTyp, d_score[0])
        print(fol_compnd, fol_CalTyp, np.mean(
[a_score[0],
b_score[0],
c_score[0],
d_score[0]]))

        os.chdir('..')
  os.chdir('..')
```

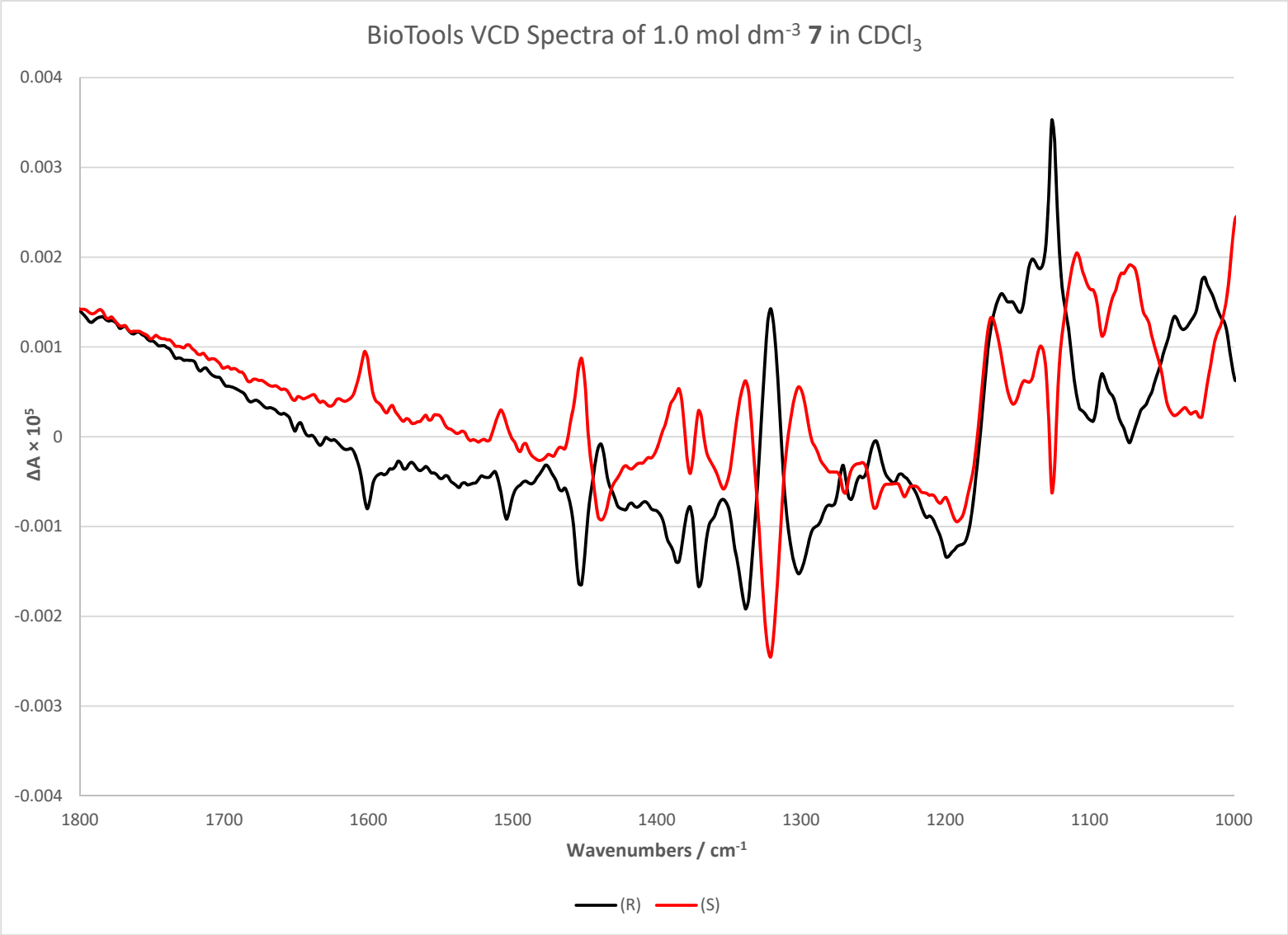
Appendix C

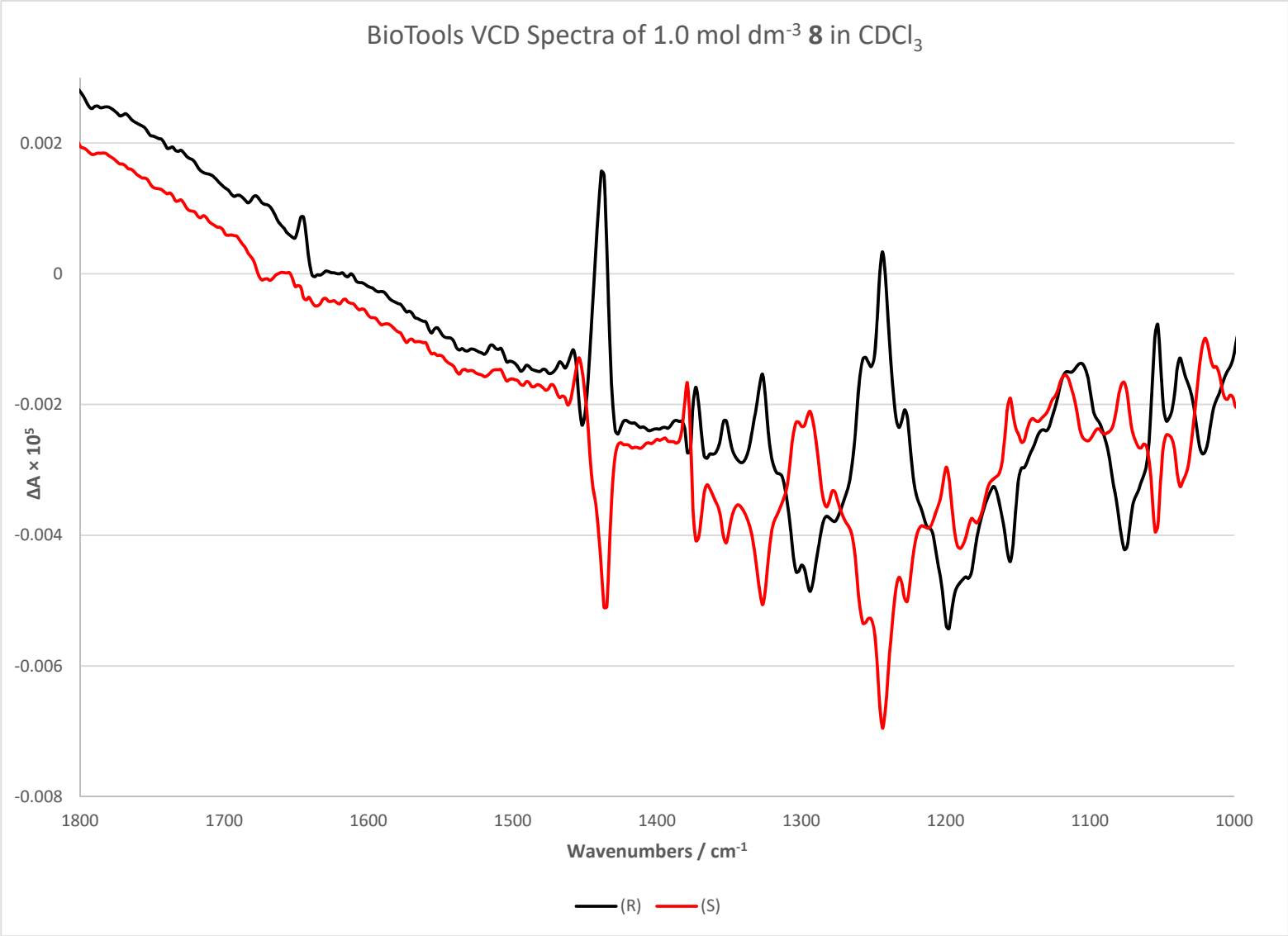
List of experimental VCD spectra

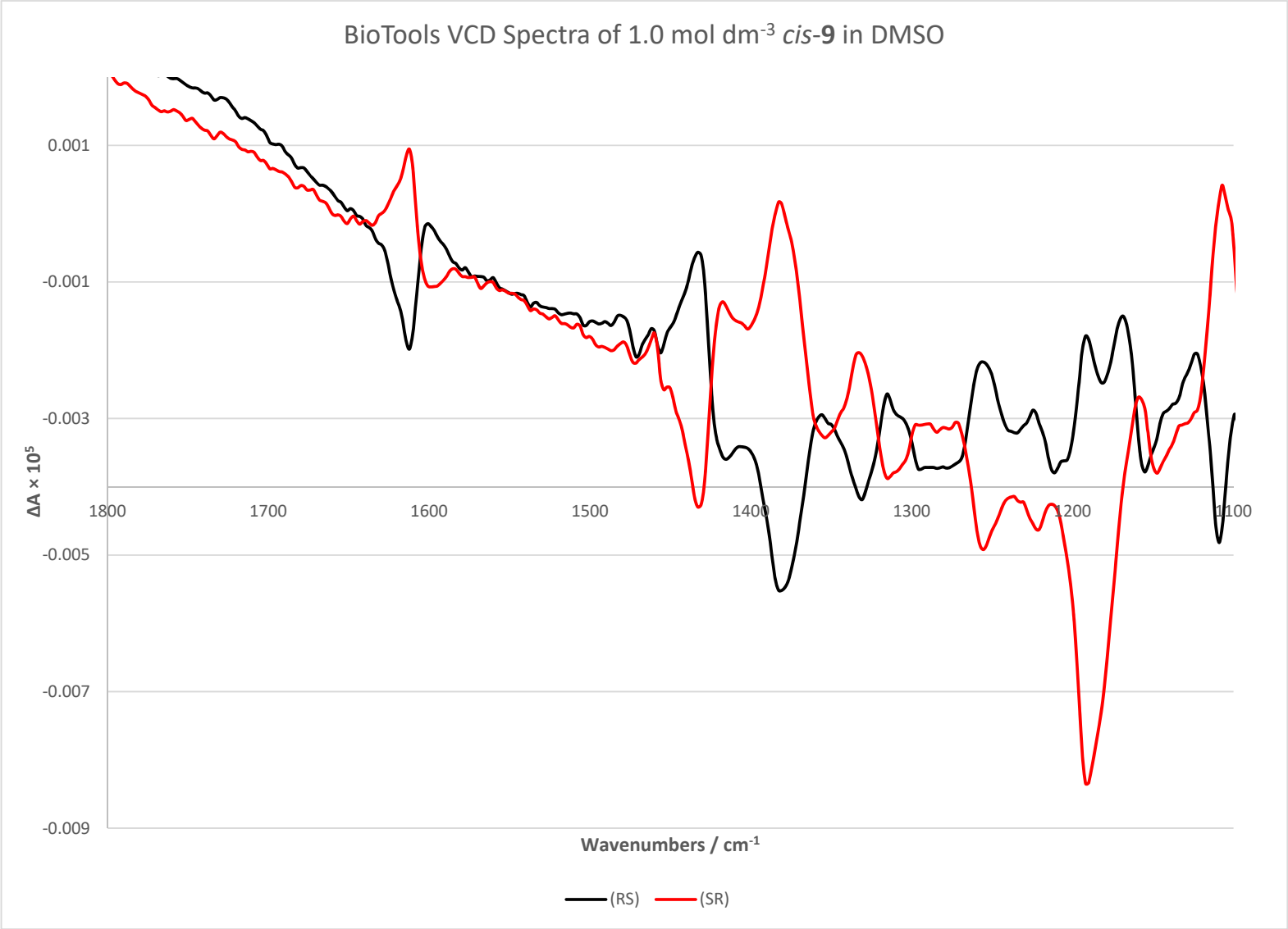
The spectra recorded over the course of this work are plotted here. Experimental details can be found in Chapter 3, Table 3.1.

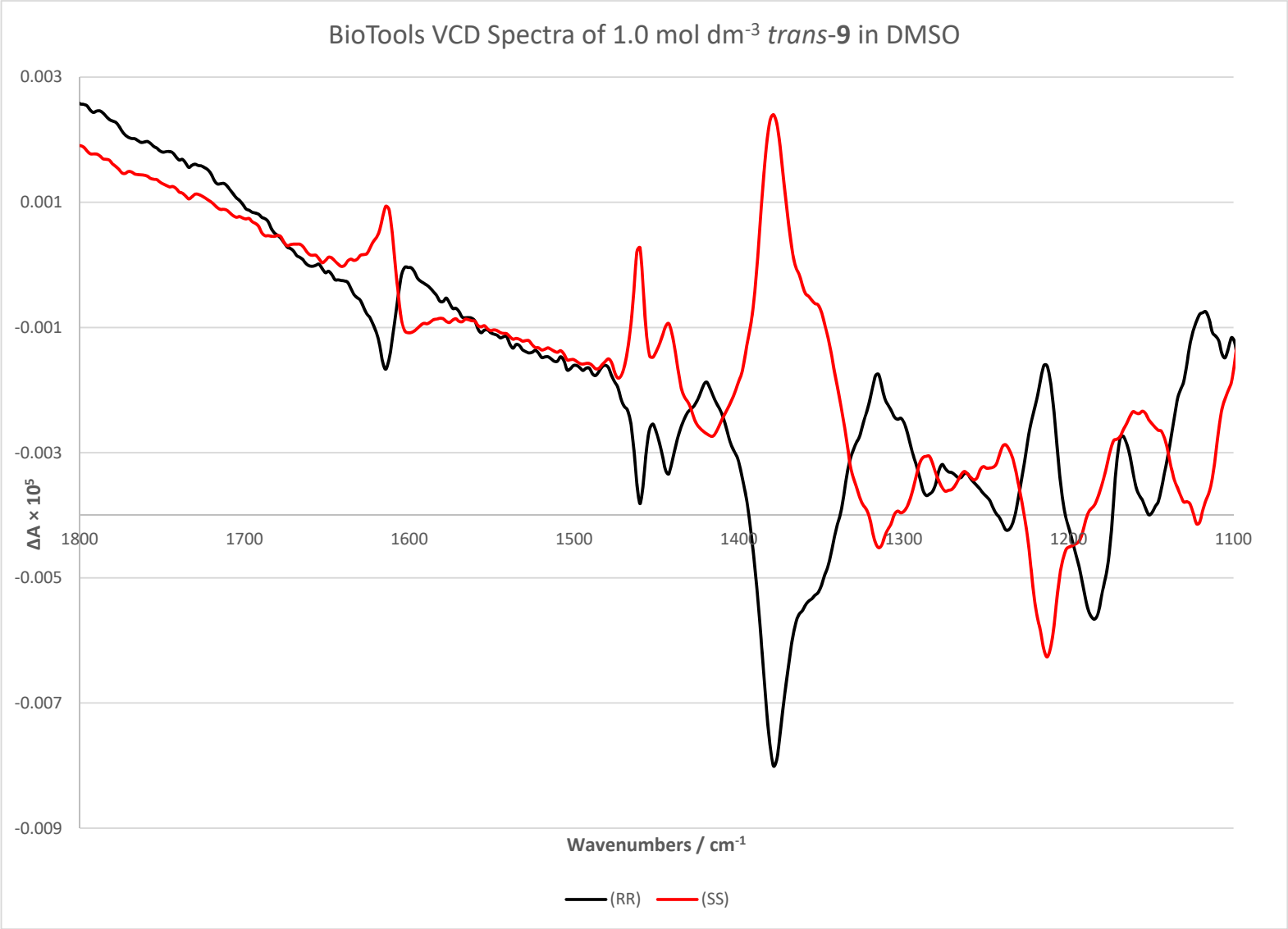

BioTools VCD Spectra of 1.0 mol dm-3 5 in CDCl3
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

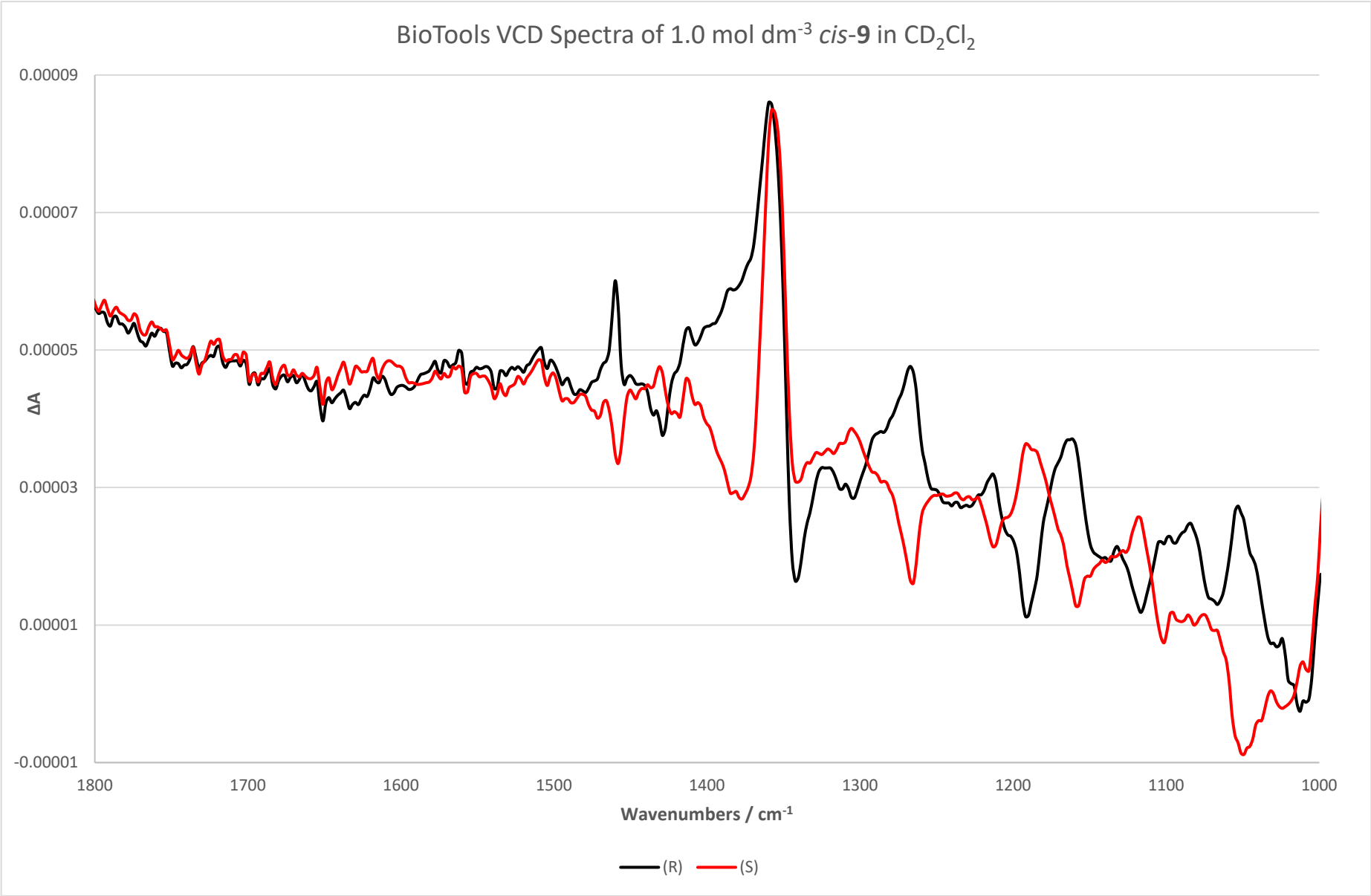

BioTools VCD Spectra of 1.0 mol dm-3 6 in CDCl3
0.01
0.005
0
-0.005
-0.01
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

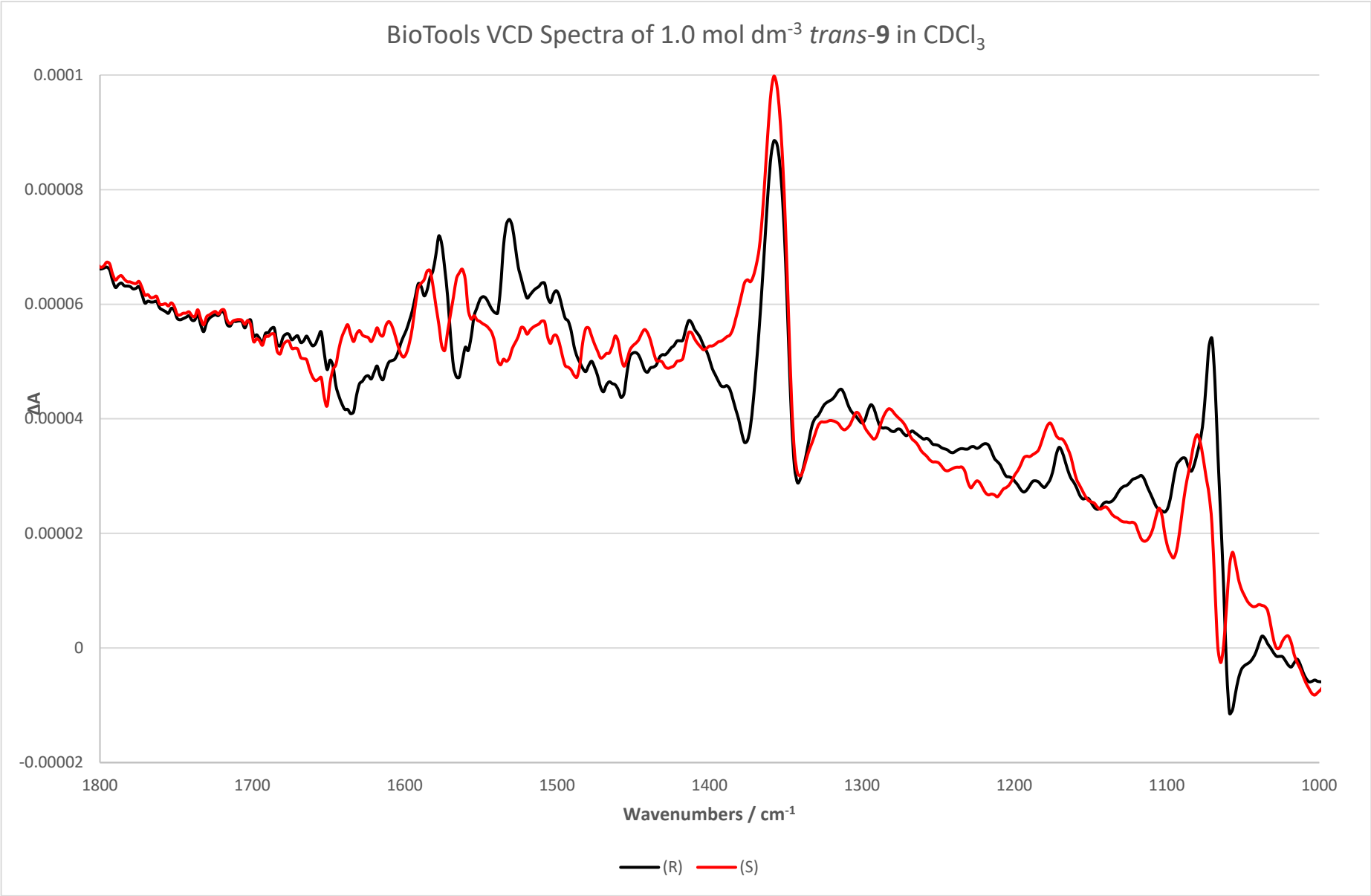

BioTools VCD Spectra of 1.0 mol dm-3 7 in CDCl3
0.004
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
-0.004
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

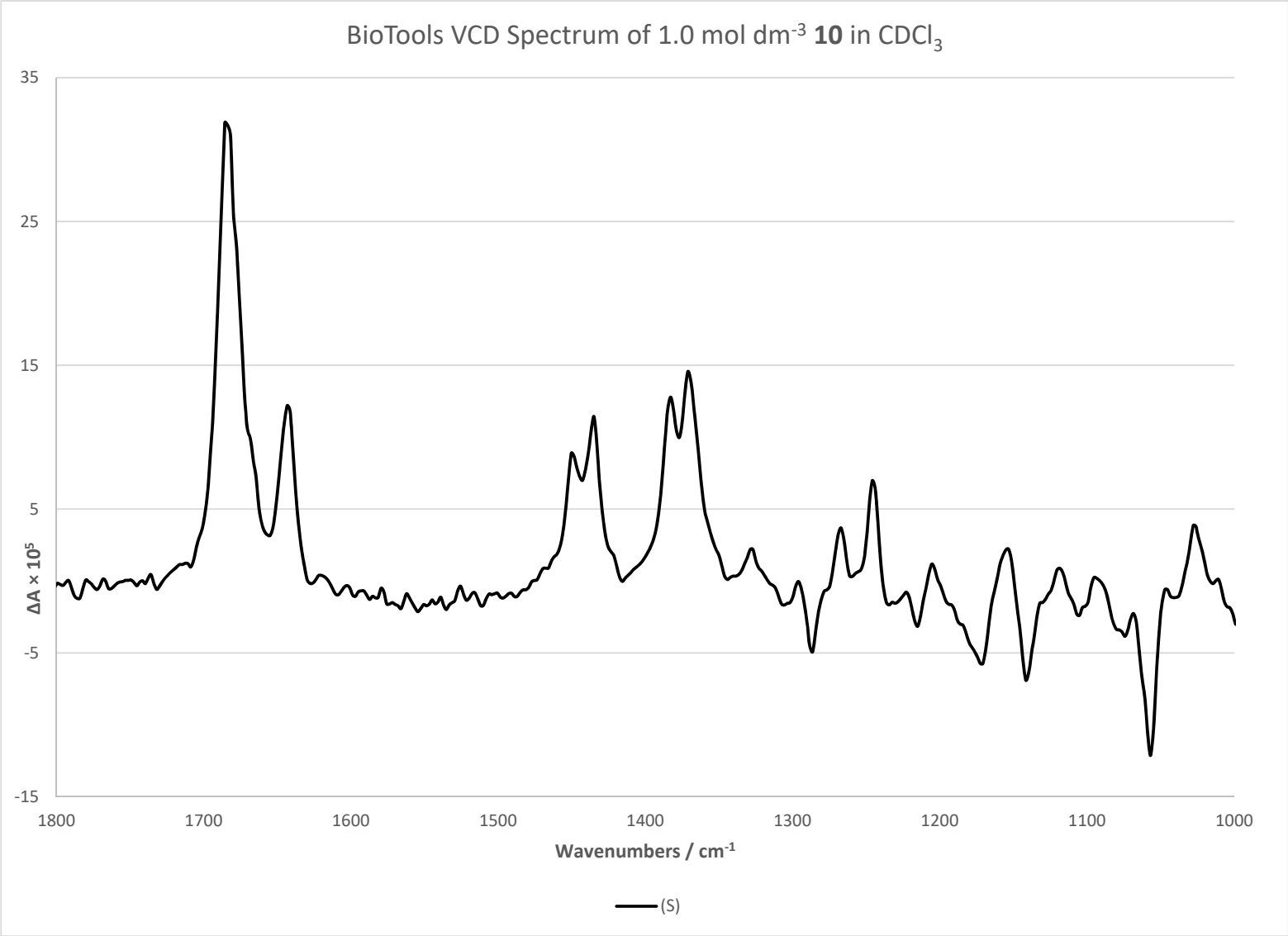

BioTools VCD Spectra of 1.0 mol dm-3 8 in CDCl3
0.002
0
-0.002
-0.004
-0.006
-0.008
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

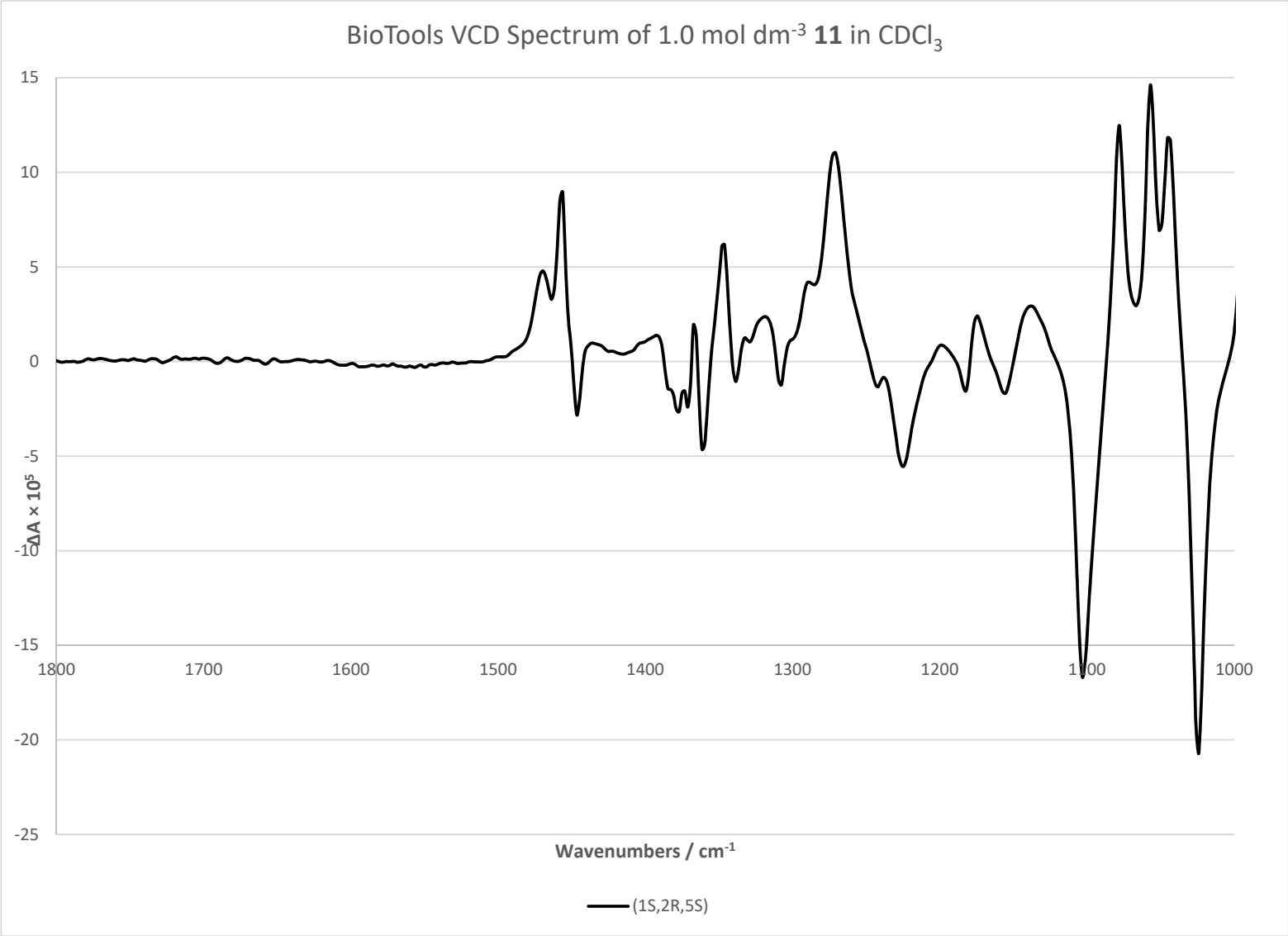

BioTools VCD Spectra of 1.0 mol dm-3 cis-9 in DMSO
0.001
-0.001
-0.003
-0.005
-0.007
-0.009
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
Wavenumbers / cm-1
(RS)
(SR)

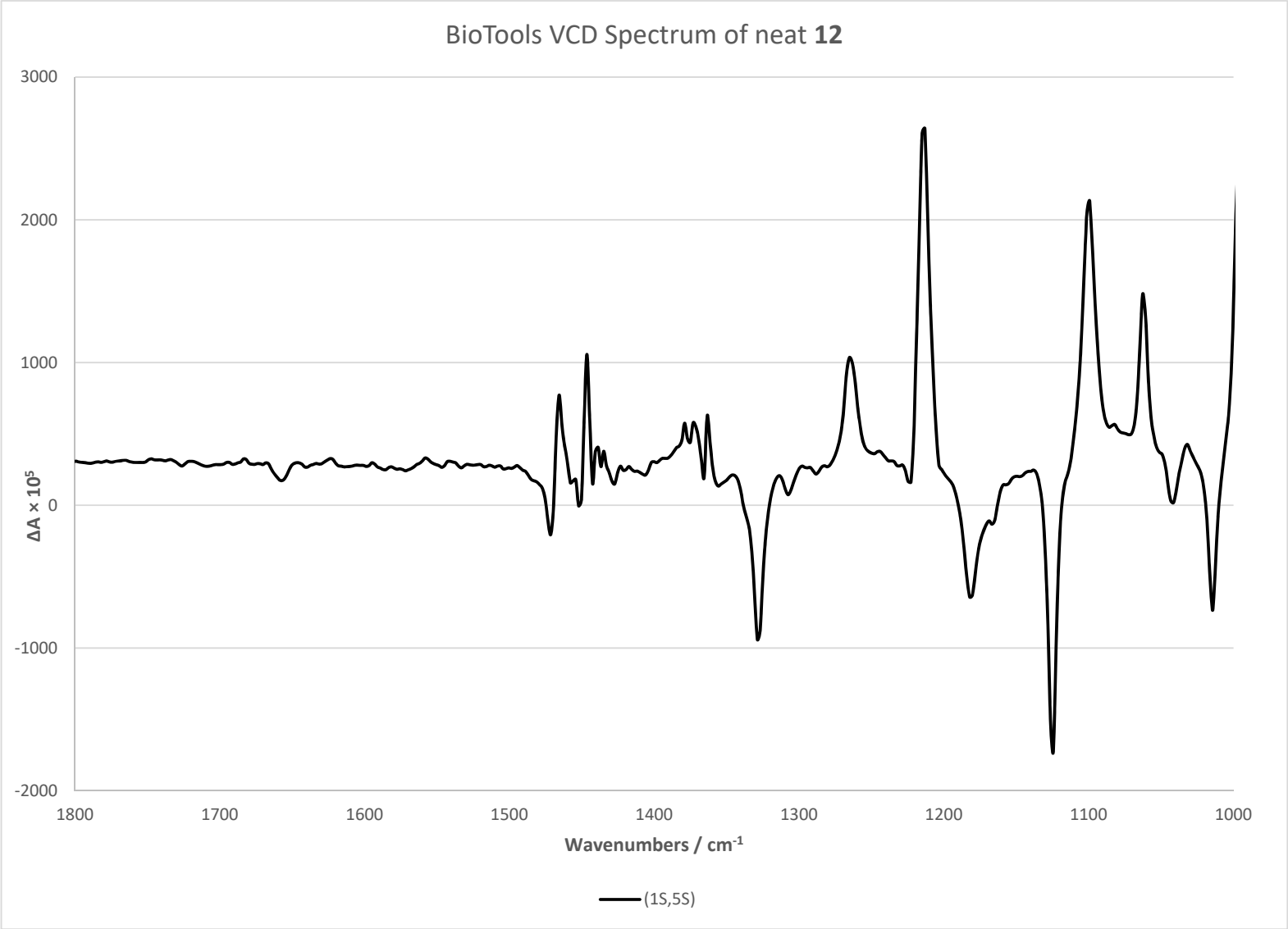

BioTools VCD Spectra of 1.0 mol dm-3 trans-9 in DMSO
0.003
0.001
-0.001
-0.003
-0.005
-0.007
-0.009
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
Wavenumbers / cm-1
(RR)
(SS)

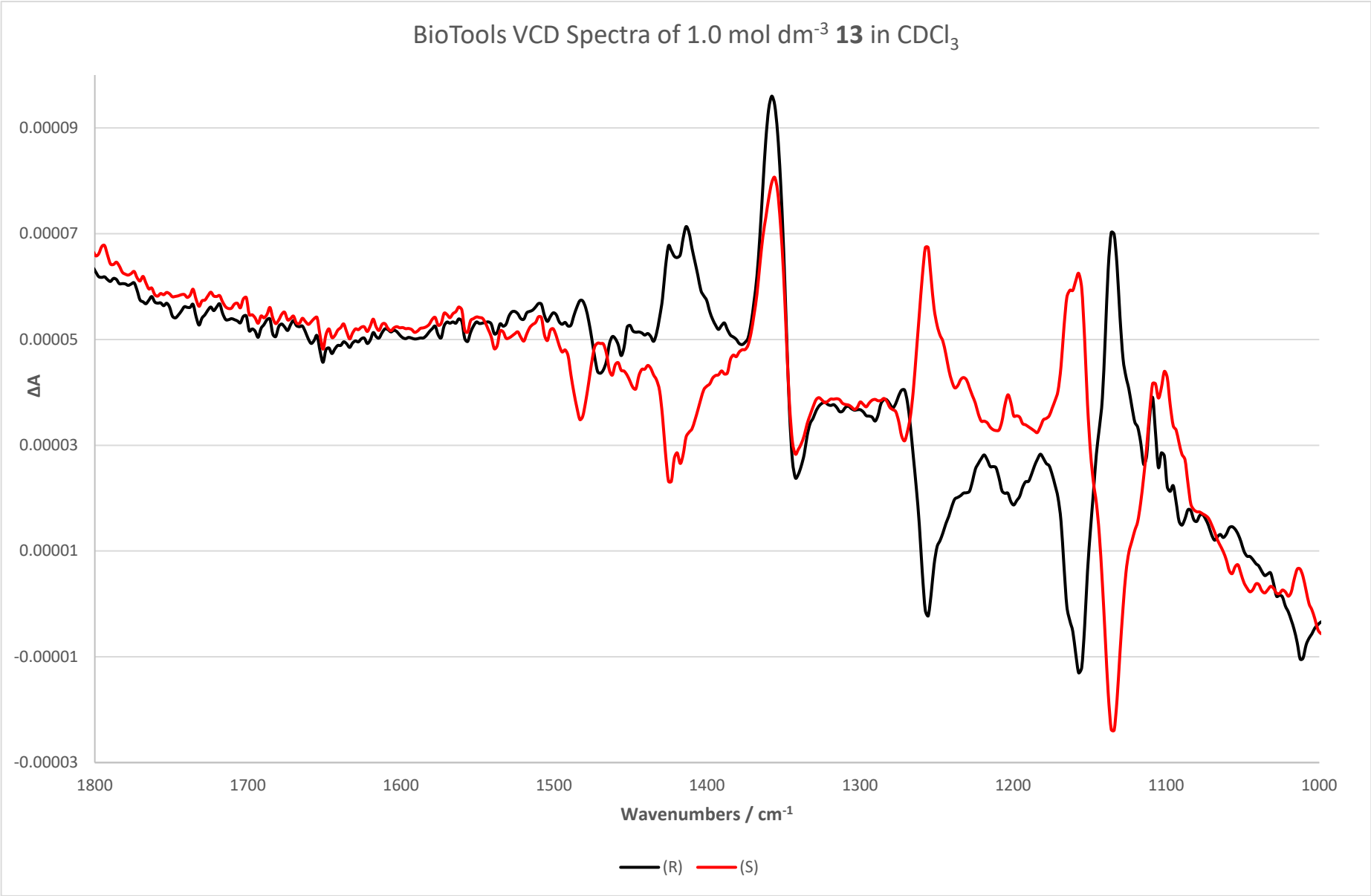

BioTools VCD Spectra of 1.0 mol dm-3 cis-9 in CD2Cl2
0.00009
0.00007
0.00005
0.00003
0.00001
-0.00001
ΔA
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

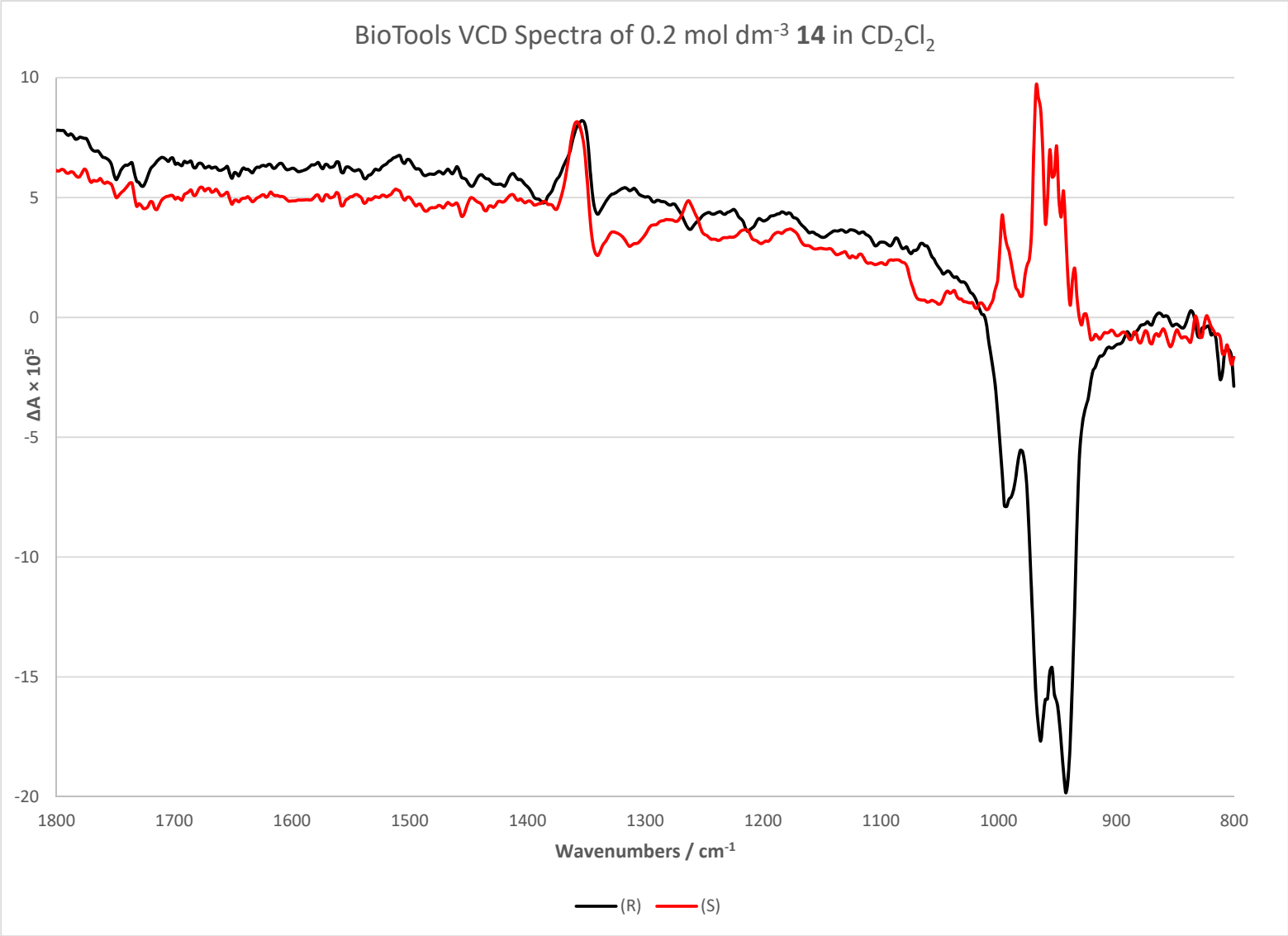

BioTools VCD Spectra of 1.0 mol dm-3 trans-9 in CDCl3
0.0001
0.00008
0.00006
0.00004
0.00002
0
-0.00002
ΔA
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

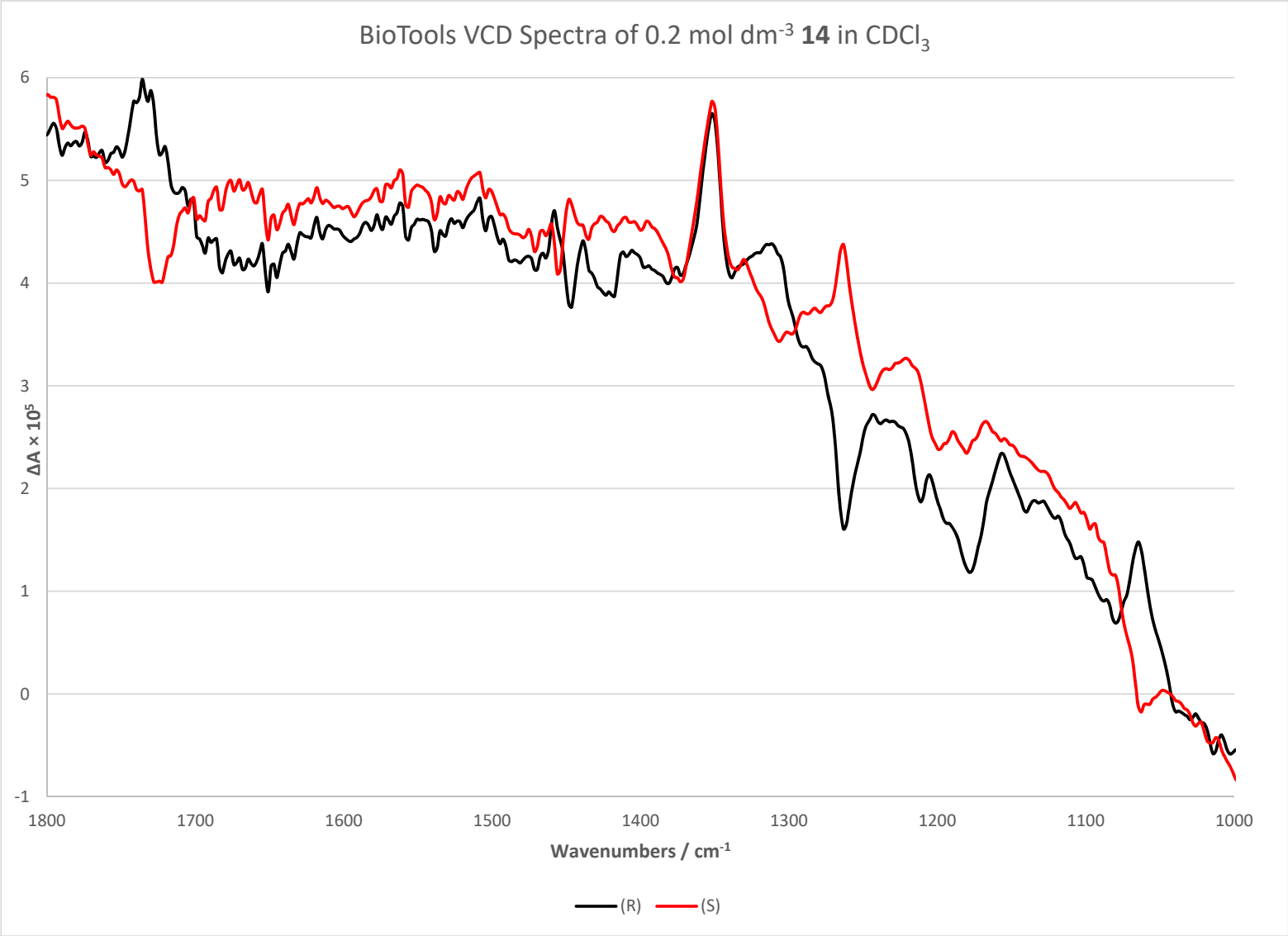

BioTools VCD Spectrum of 1.0 mol dm-3 10 in CDCl3
35
25
15
5
-5
-15
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(S)

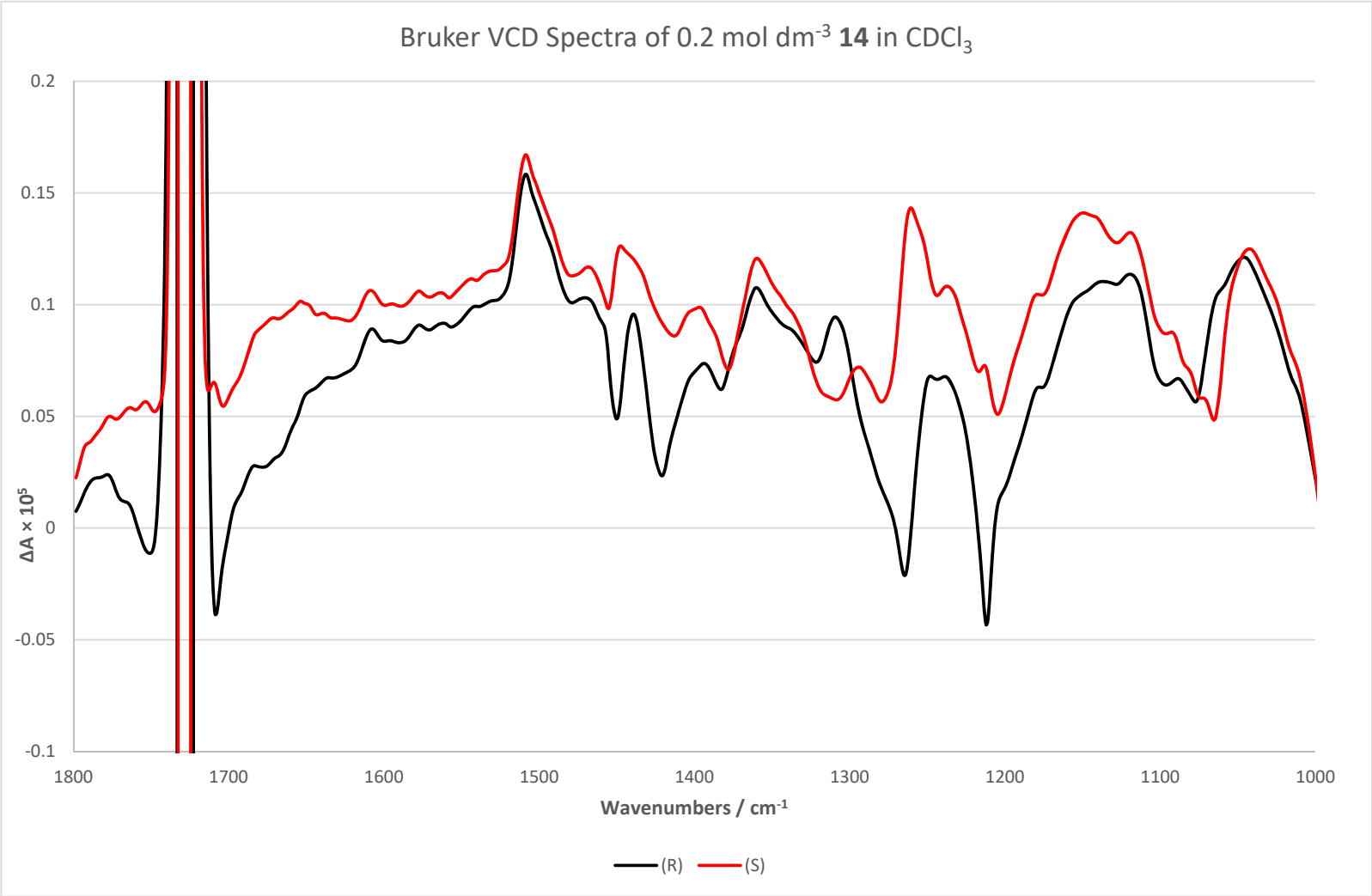

BioTools VCD Spectrum of 1.0 mol dm-3 11 in CDCl3
15
10
5
0
-5
-10
-15
-20
-25
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(1S,2R,5S)

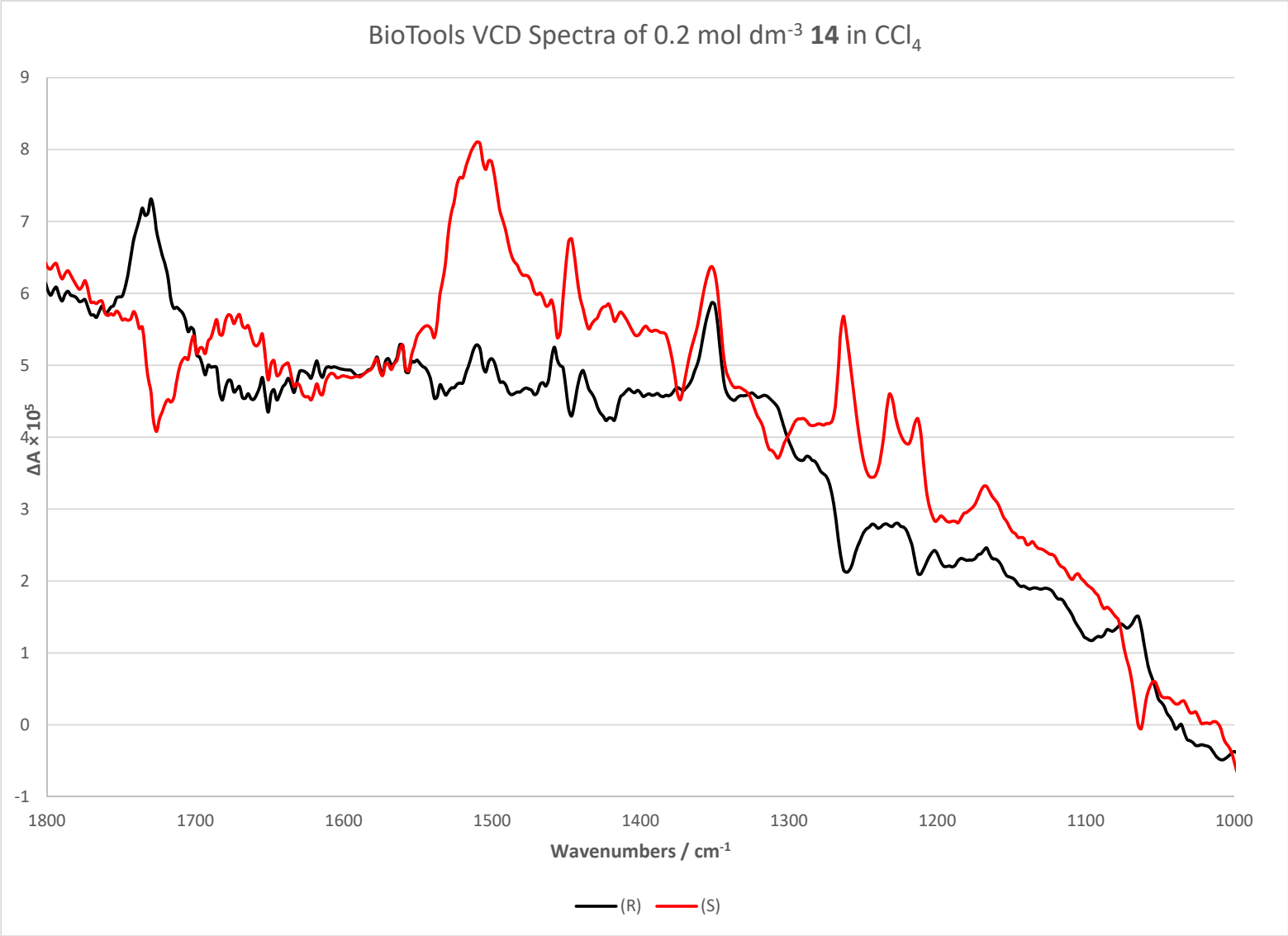

BioTools VCD Spectrum of neat **12**
3000
2000
1000
0
-1000
-2000
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm⁻¹
(1S,5S)

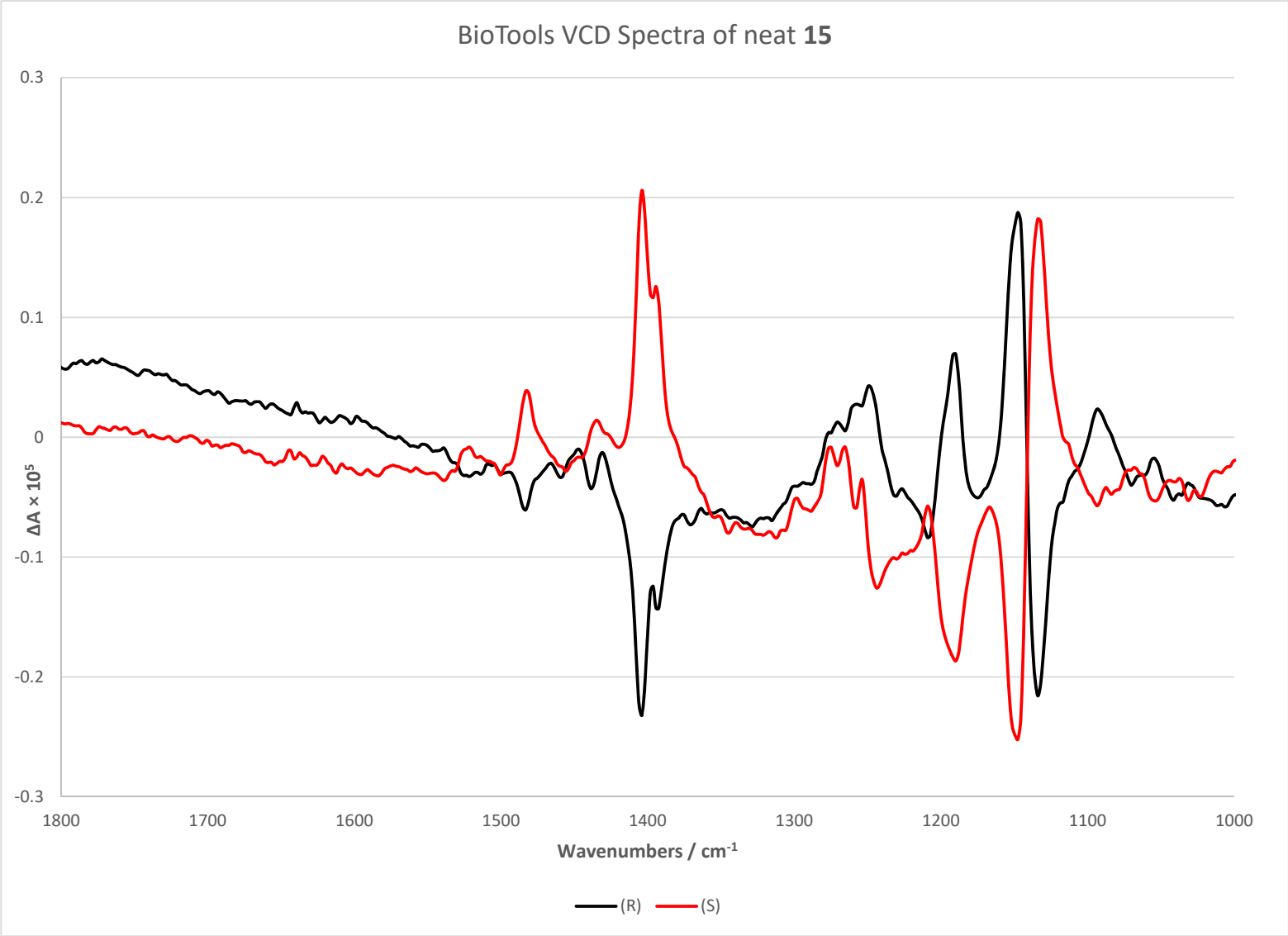

BioTools VCD Spectra of 1.0 mol dm-3 13 in CDCl3
0.00009
0.00007
0.00005
0.00003
0.00001
-0.00001
-0.00003
ΔA
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

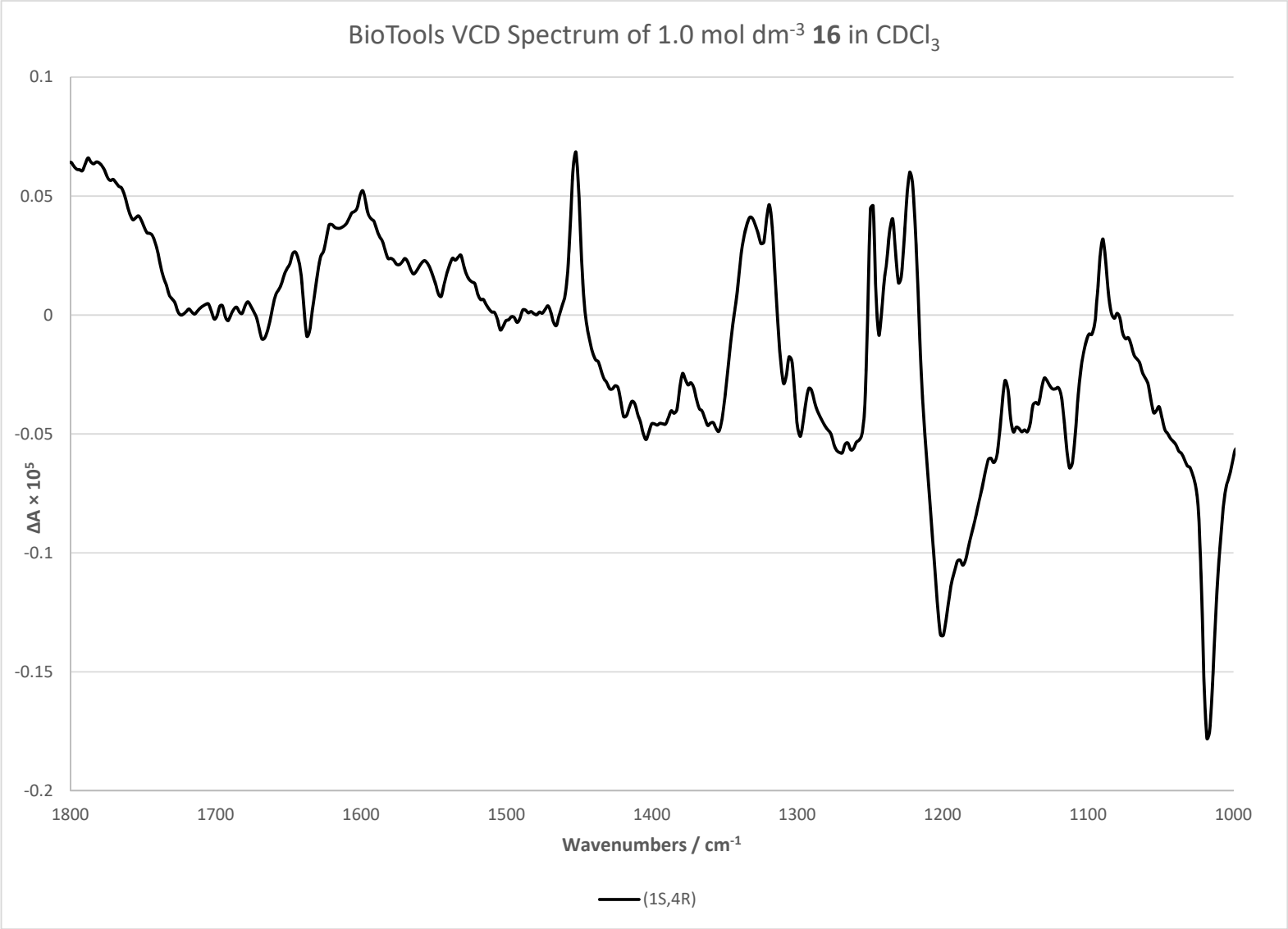

BioTools VCD Spectra of 0.2 mol dm-3 14 in CD2Cl2
10
5
0
-5
-10
-15
-20
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
Wavenumbers / cm-1
(R)
(S)

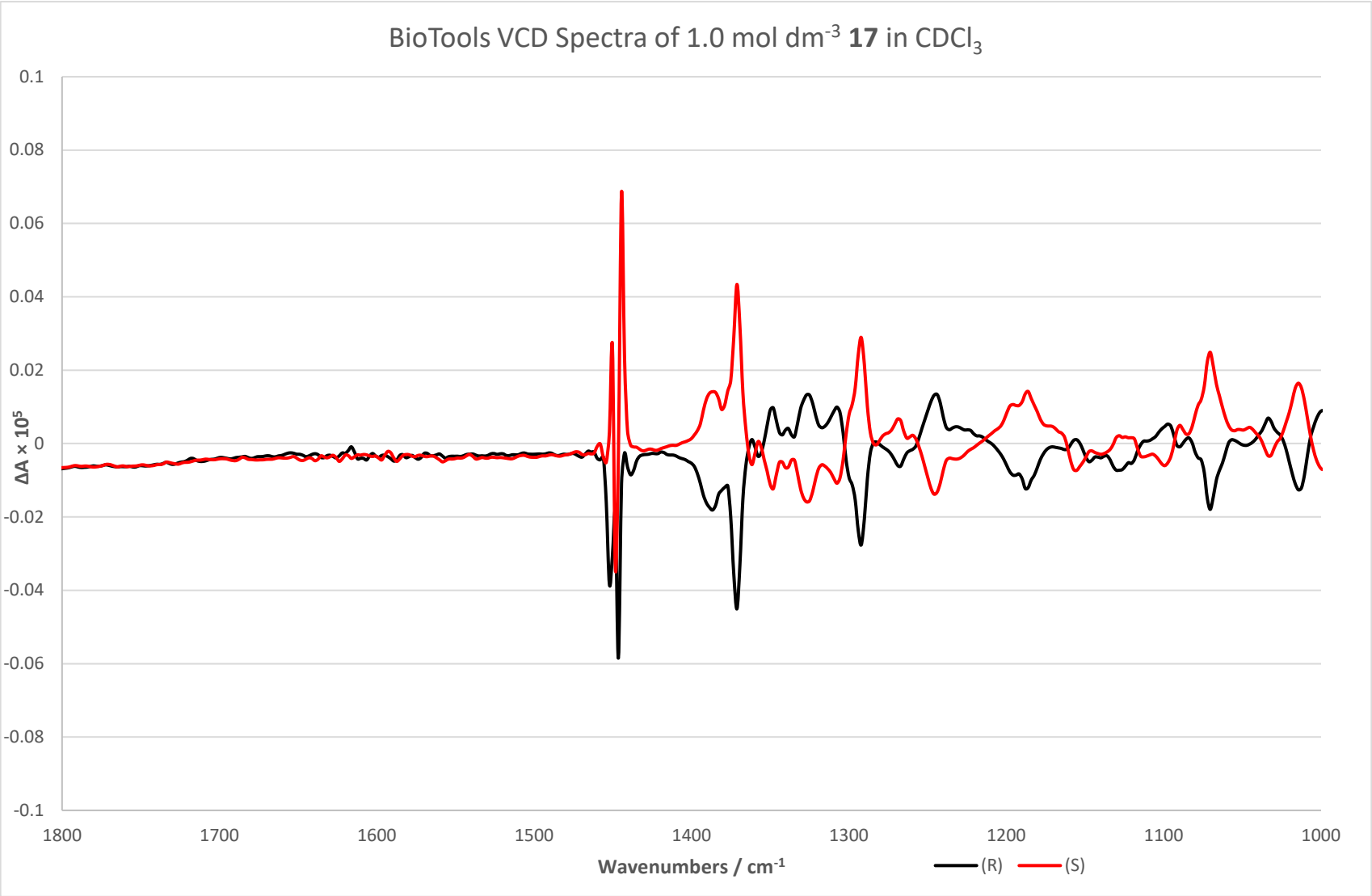

BioTools VCD Spectra of 0.2 mol dm-3 14 in CDCl3
6
5
4
3
2
1
0
-1
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

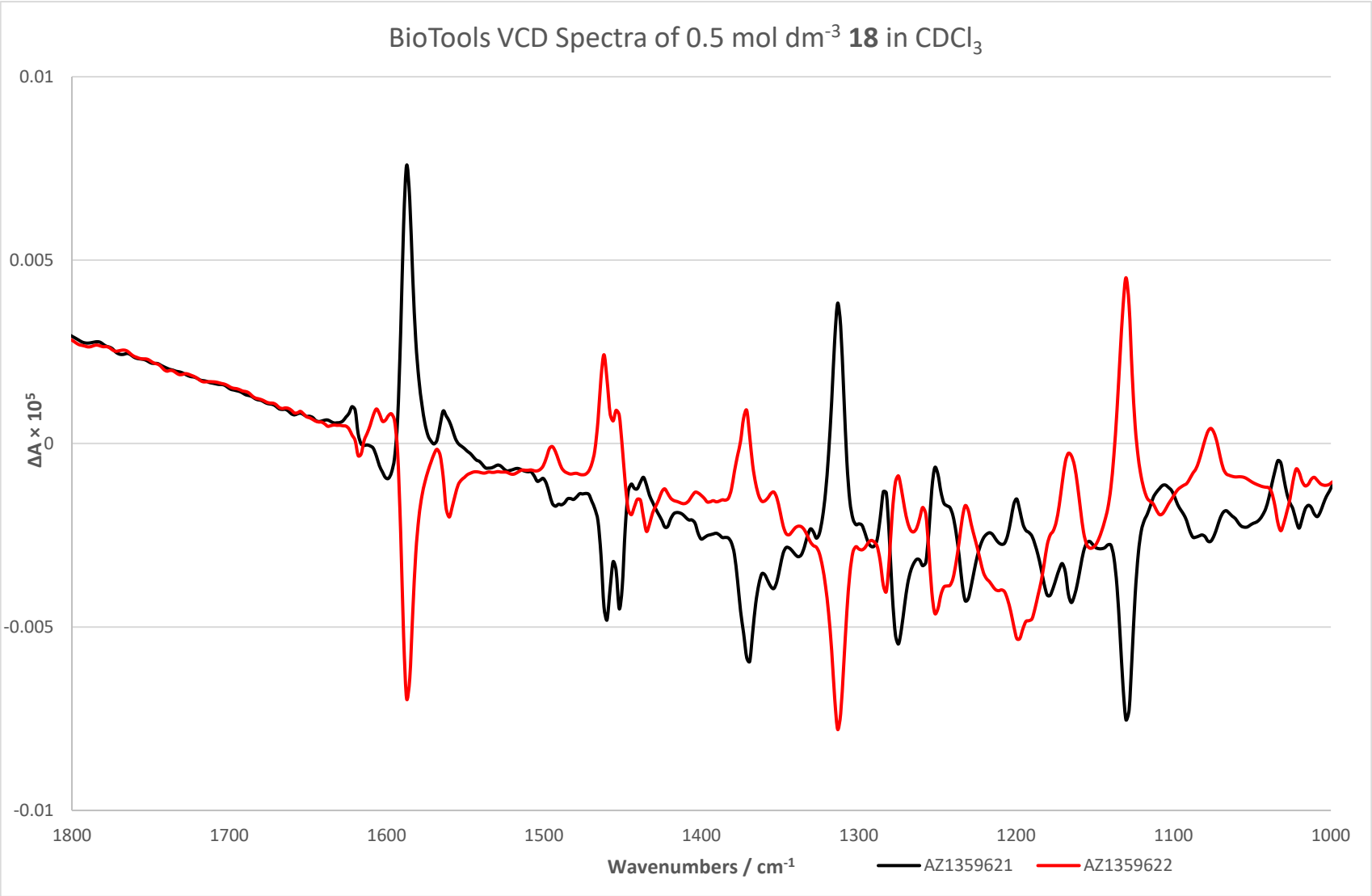

Bruker VCD Spectra of 0.2 mol dm-3 14 in CDCl3
0.2
0.15
0.1
0.05
0
-0.05
-0.1
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

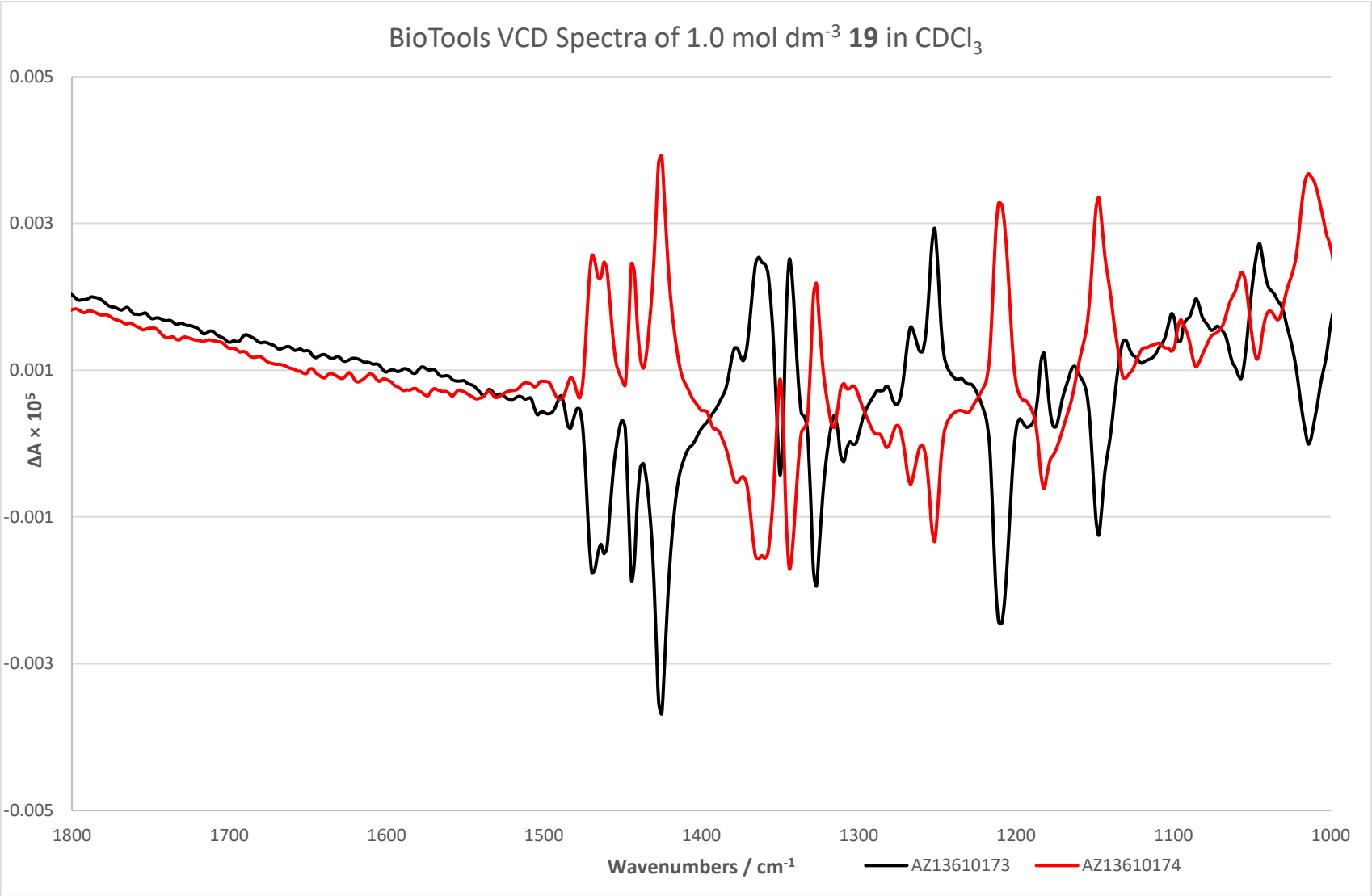

BioTools VCD Spectra of 0.2 mol dm-3 14 in CCl4
9
8
7
6
5
4
3
2
1
0
-1
ΔA x 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

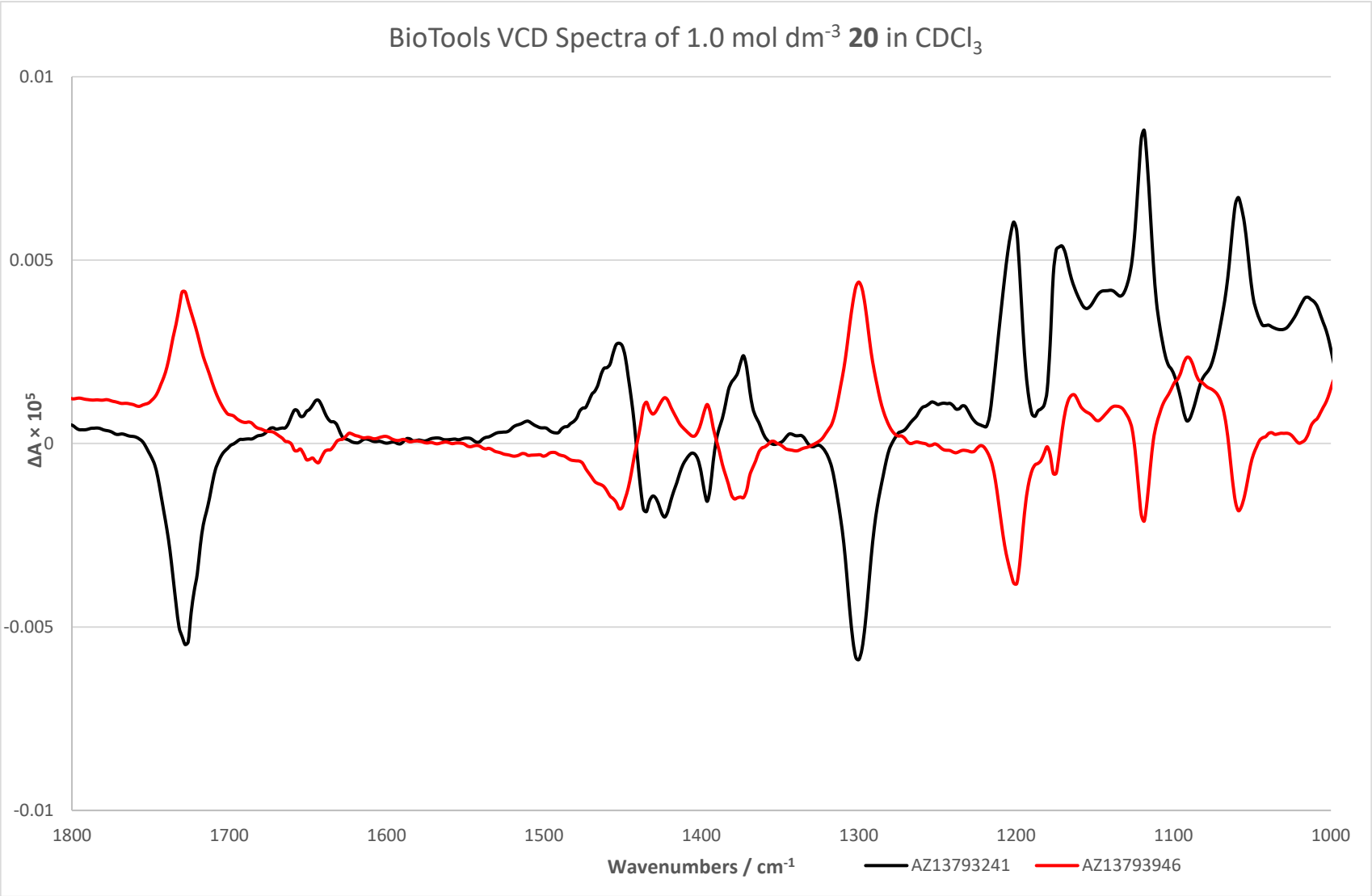

BioTools VCD Spectra of neat **15**
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm⁻¹
(R)
(S)

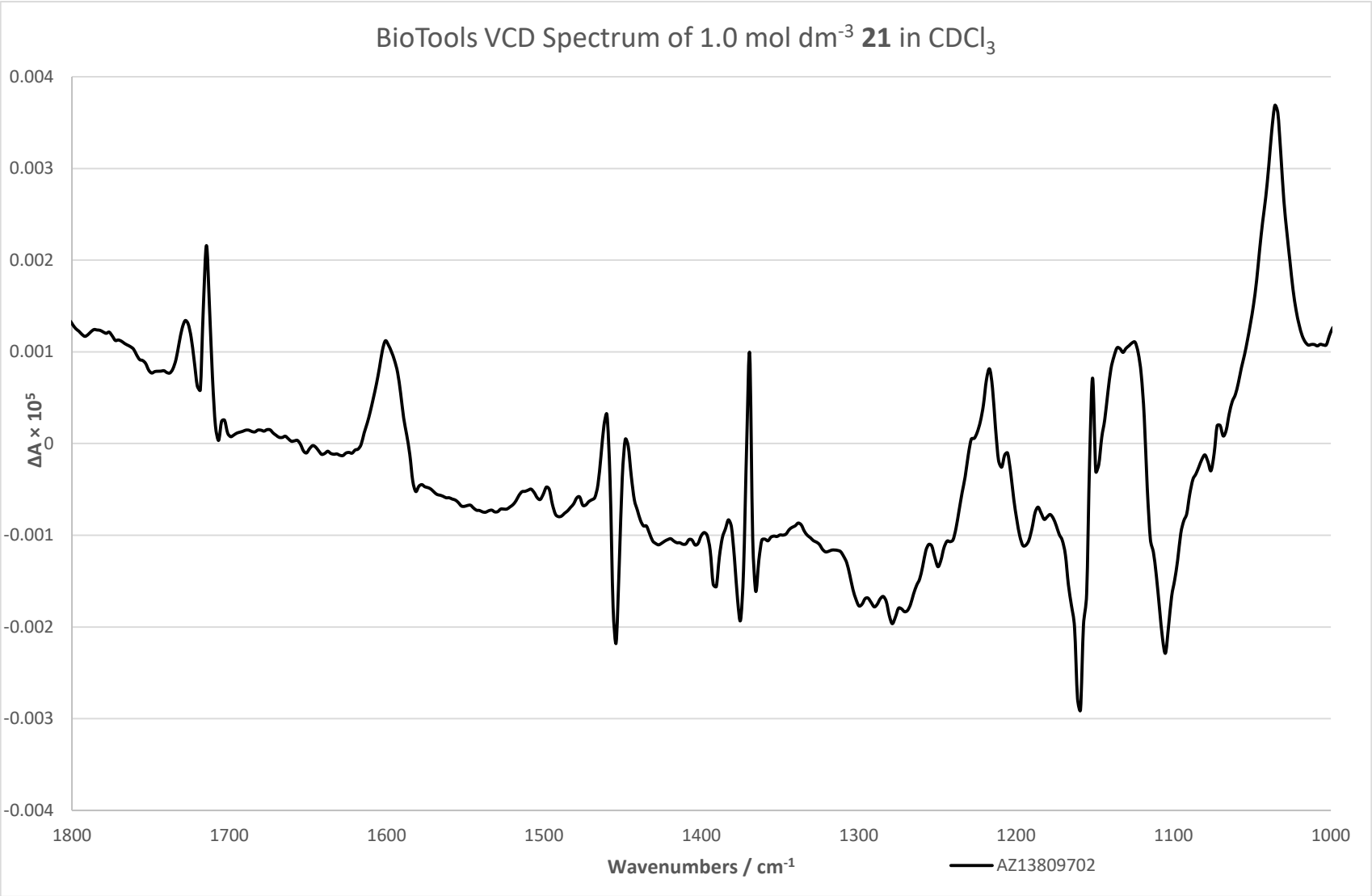

BioTools VCD Spectrum of 1.0 mol dm-3 16 in CDCl3
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(1S,4R)

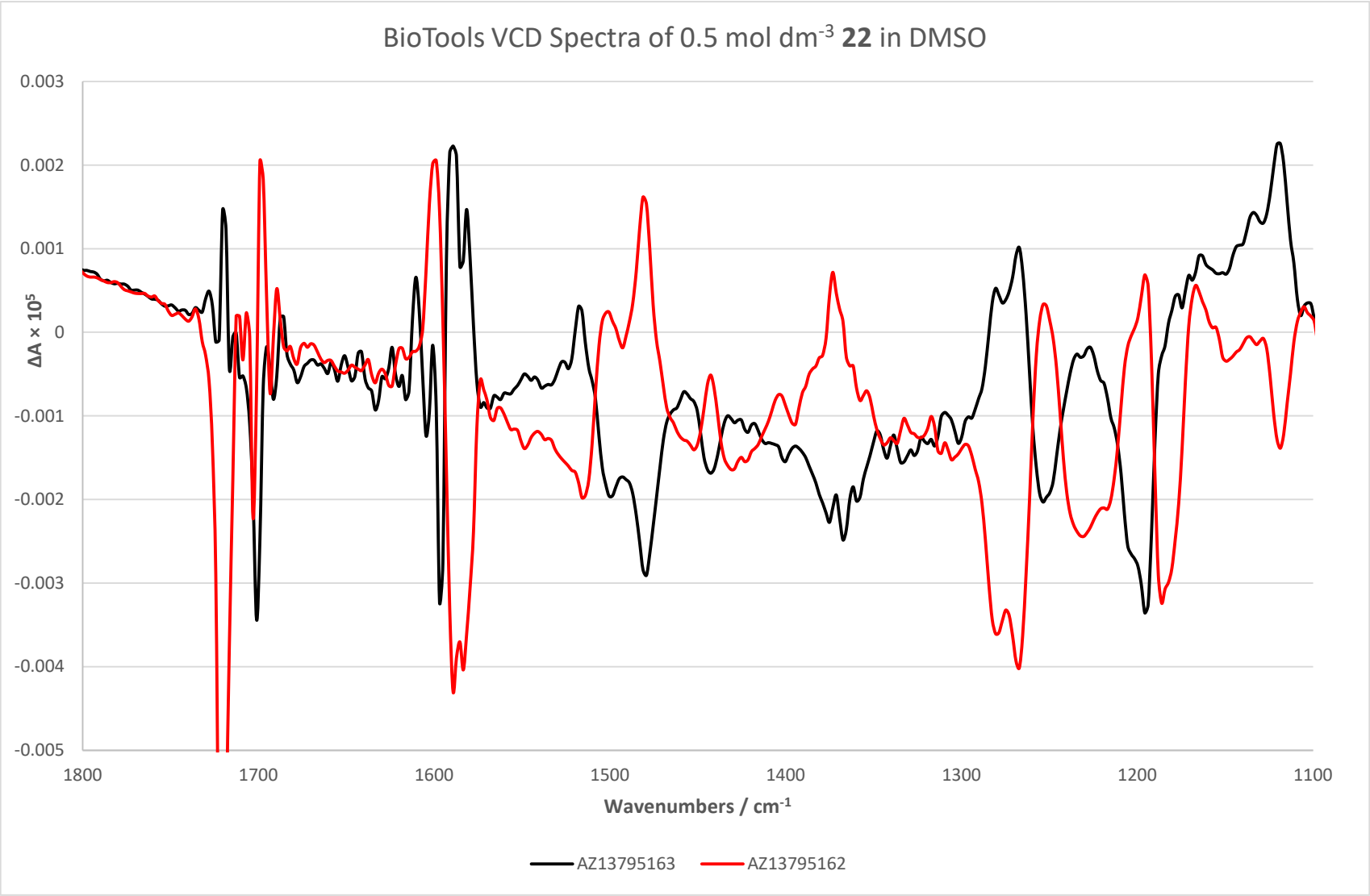

BioTools VCD Spectra of 1.0 mol dm-3 17 in CDCl3
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

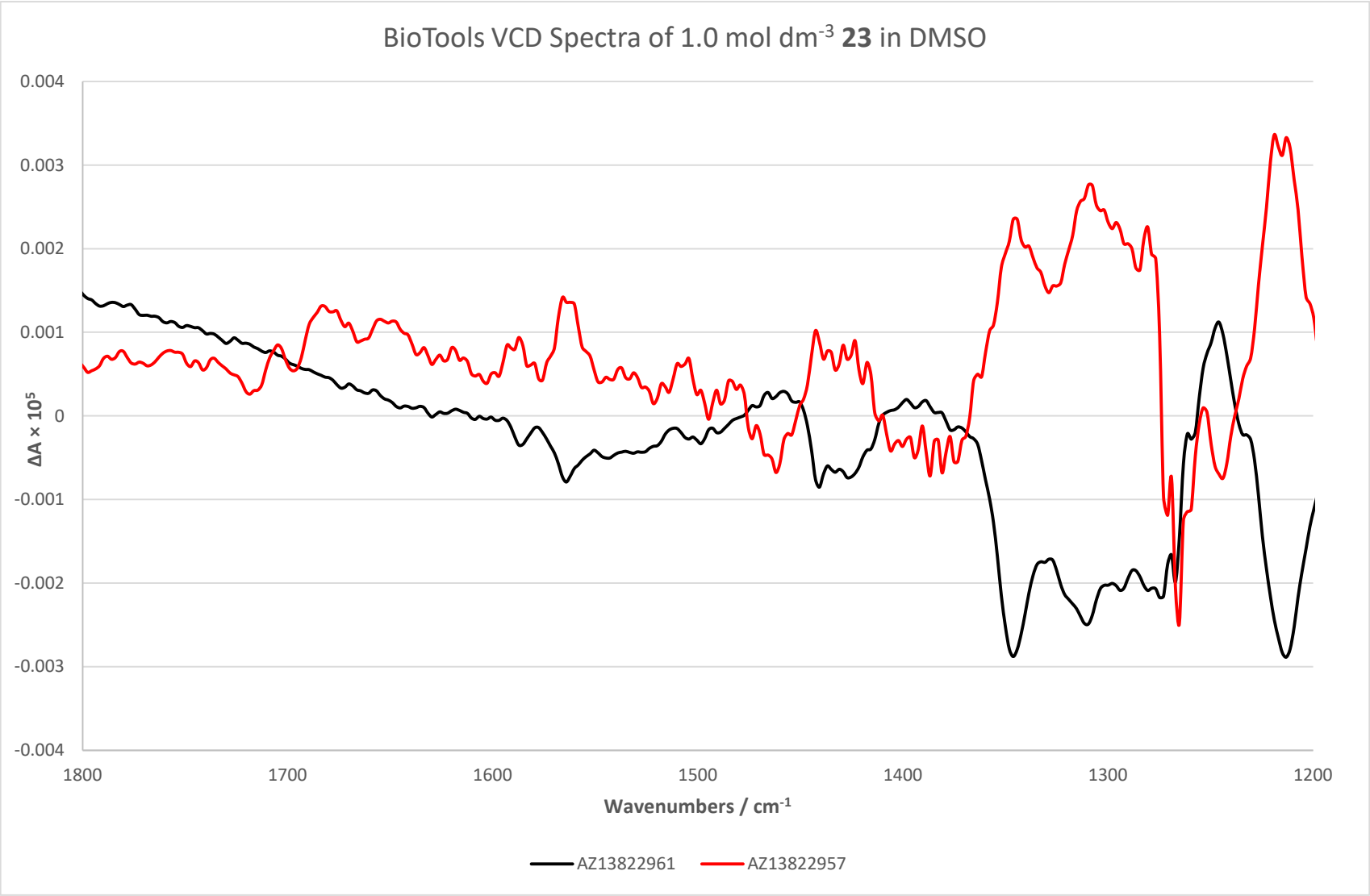

BioTools VCD Spectra of 0.5 mol dm-3 18 in CDCl3
0.01
0.005
0
-0.005
-0.01
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ1359621
AZ1359622

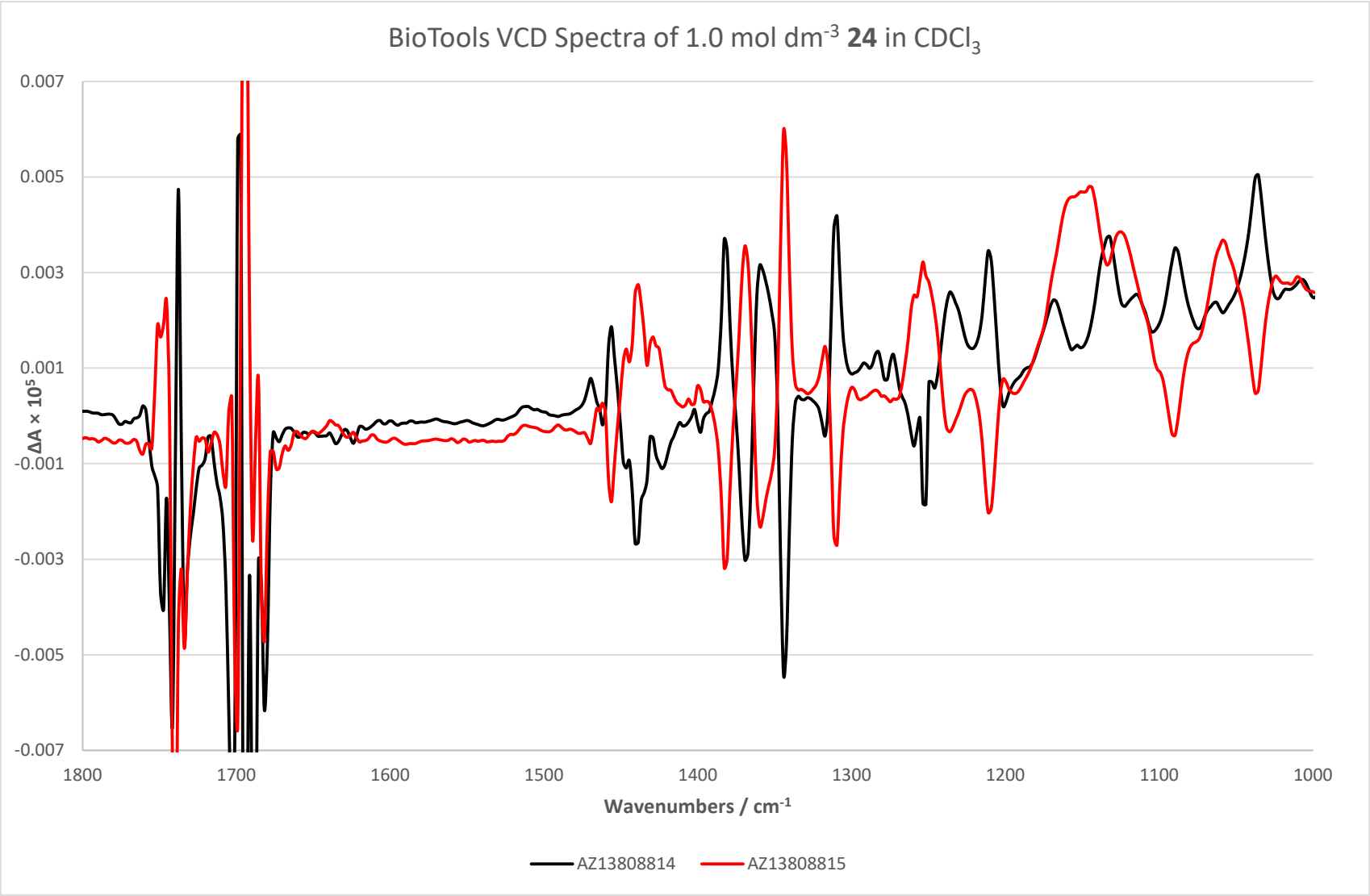

BioTools VCD Spectra of 1.0 mol dm-3 19 in CDCl3
0.005
0.003
0.001
-0.001
-0.003
-0.005
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ13610173
AZ13610174

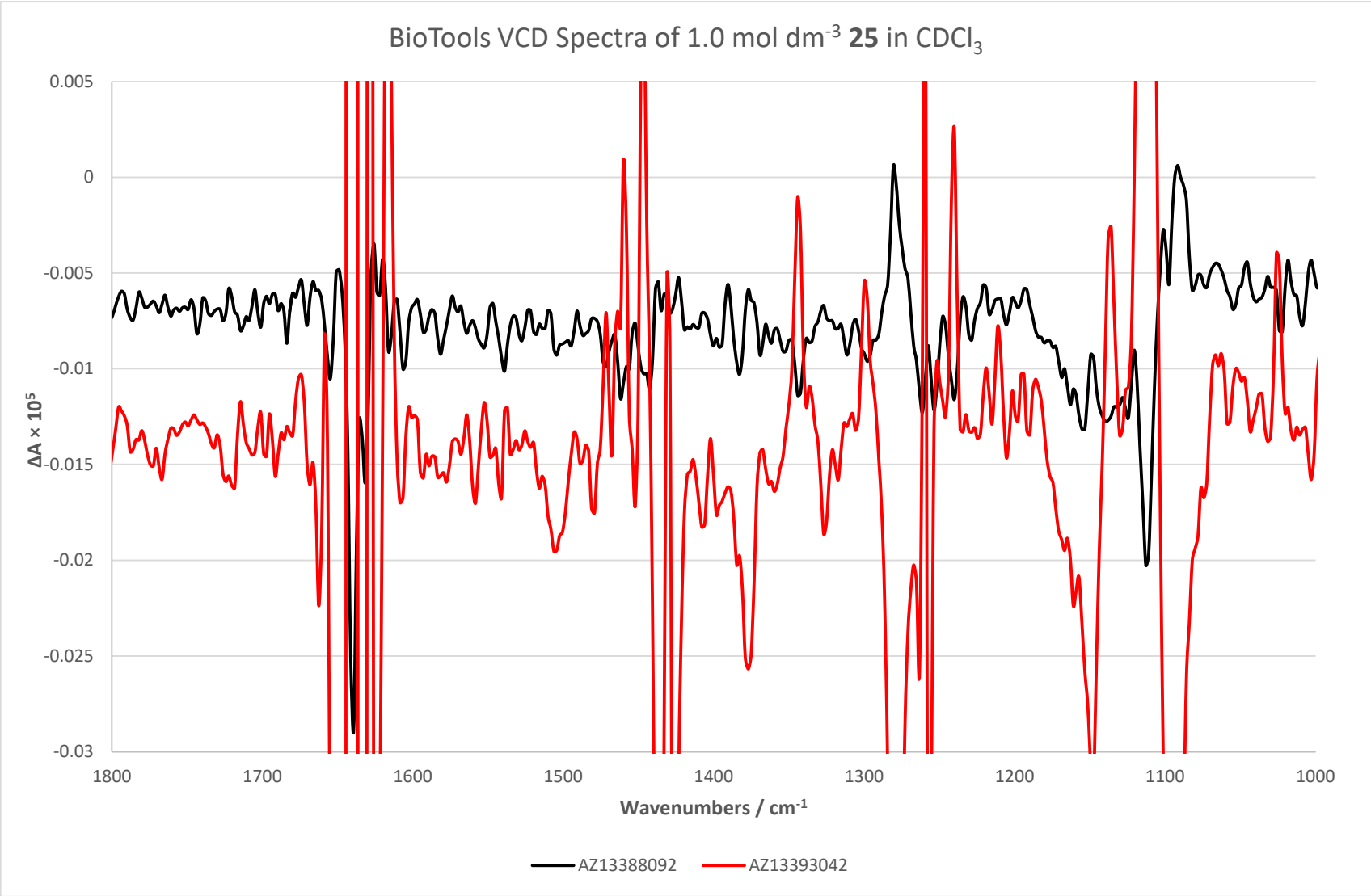

BioTools VCD Spectra of 1.0 mol dm-3 20 in CDCl3
0.01
0.005
0
-0.005
-0.01
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ13793241
AZ13793946

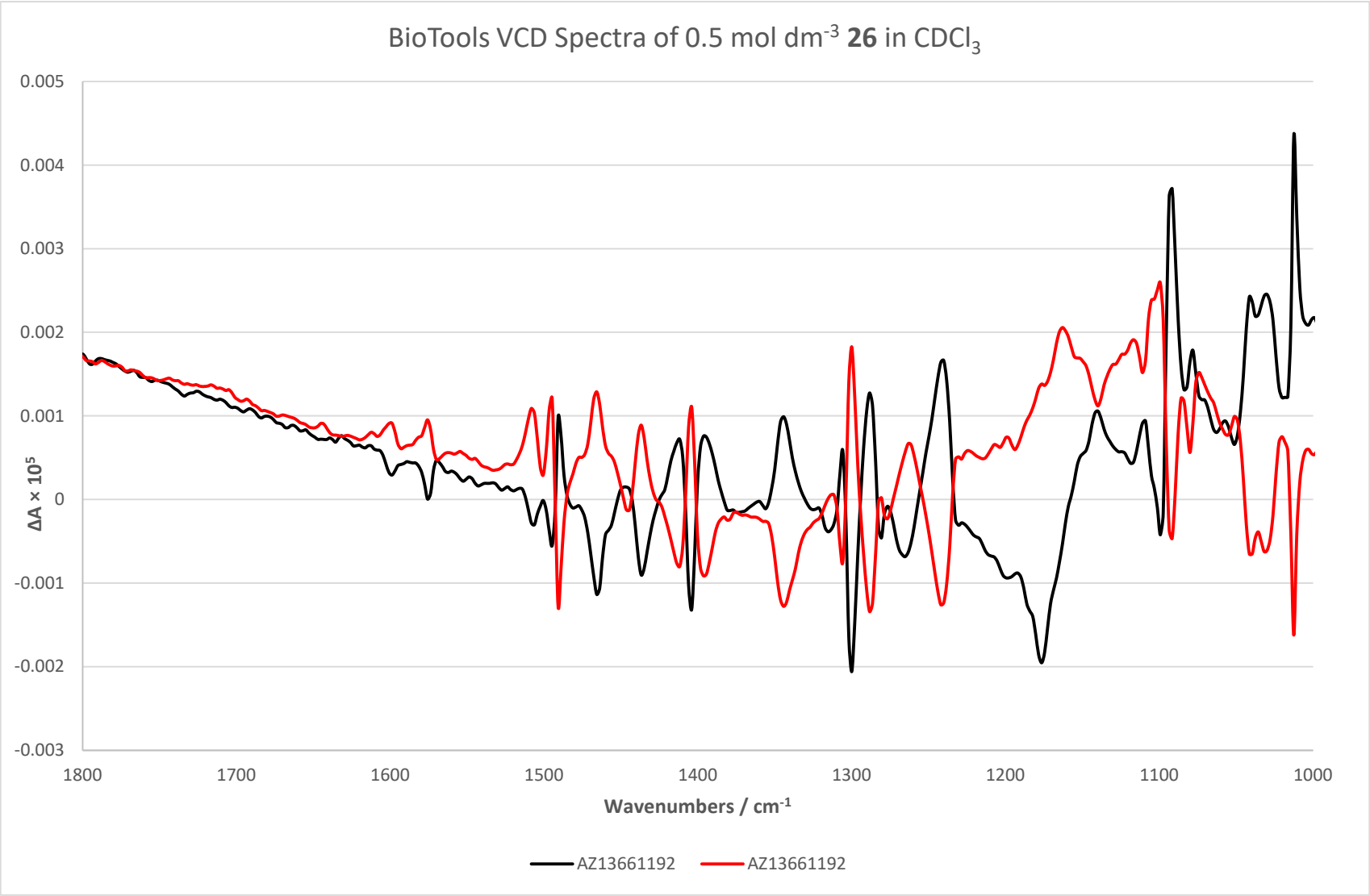

BioTools VCD Spectrum of 1.0 mol dm-3 21 in CDCl3
0.004
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
-0.004
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ13809702

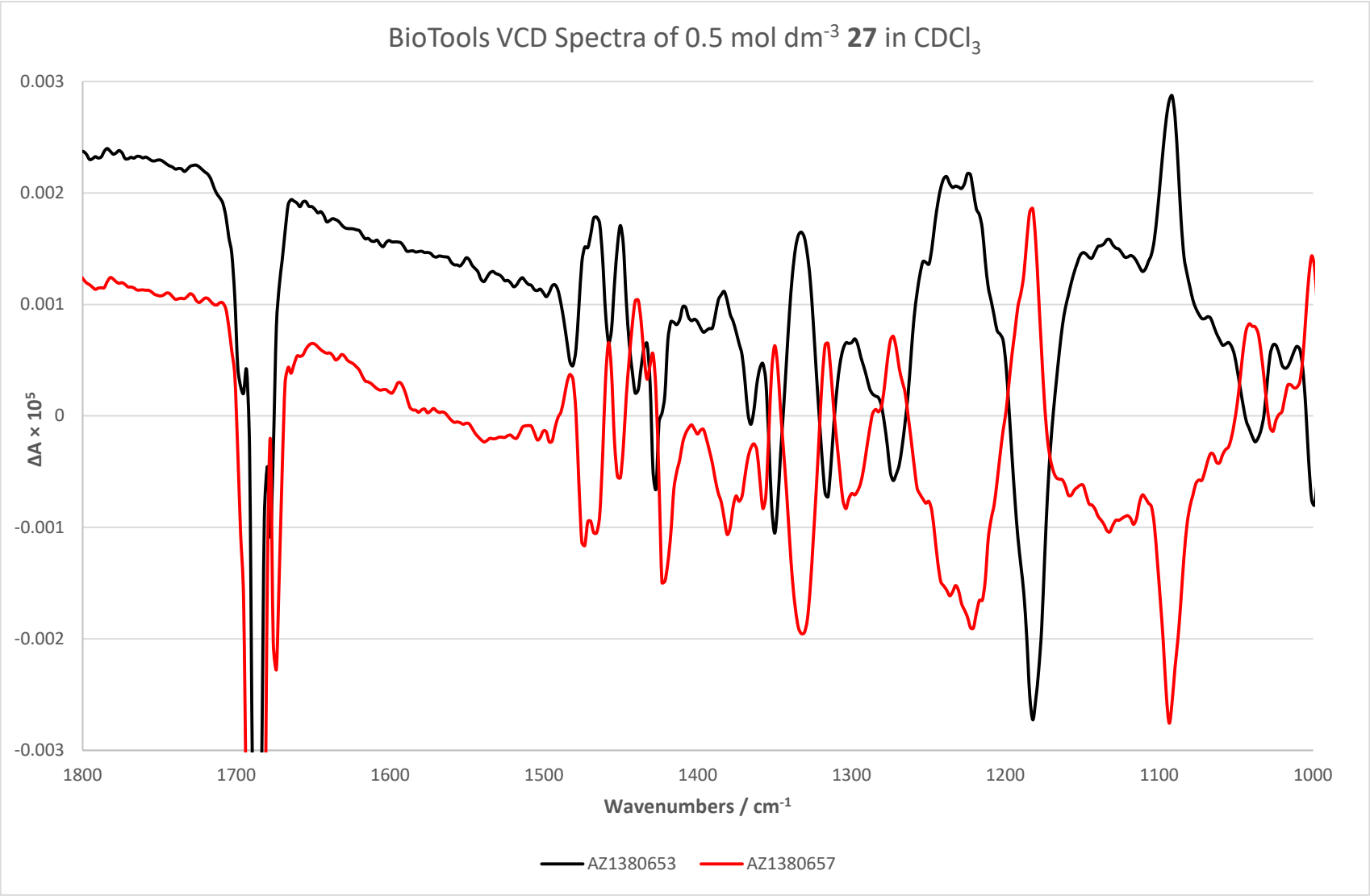

BioTools VCD Spectra of 0.5 mol dm-3 22 in DMSO
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
-0.004
-0.005
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
Wavenumbers / cm-1
AZ13795163
AZ13795162

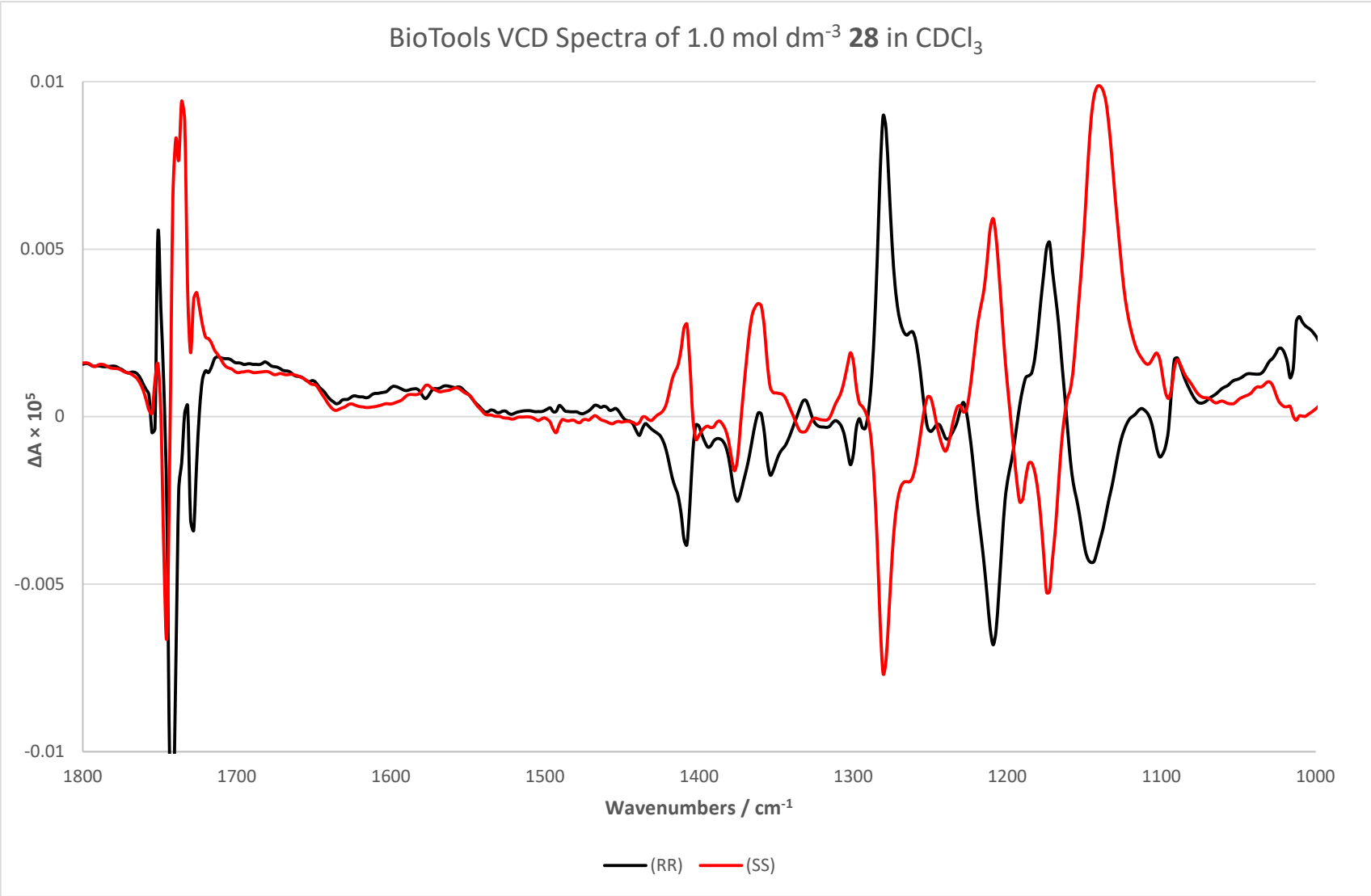

BioTools VCD Spectra of 1.0 mol dm-3 23 in DMSO
0.004
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
-0.004
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
Wavenumbers / cm-1
AZ13822961
AZ13822957

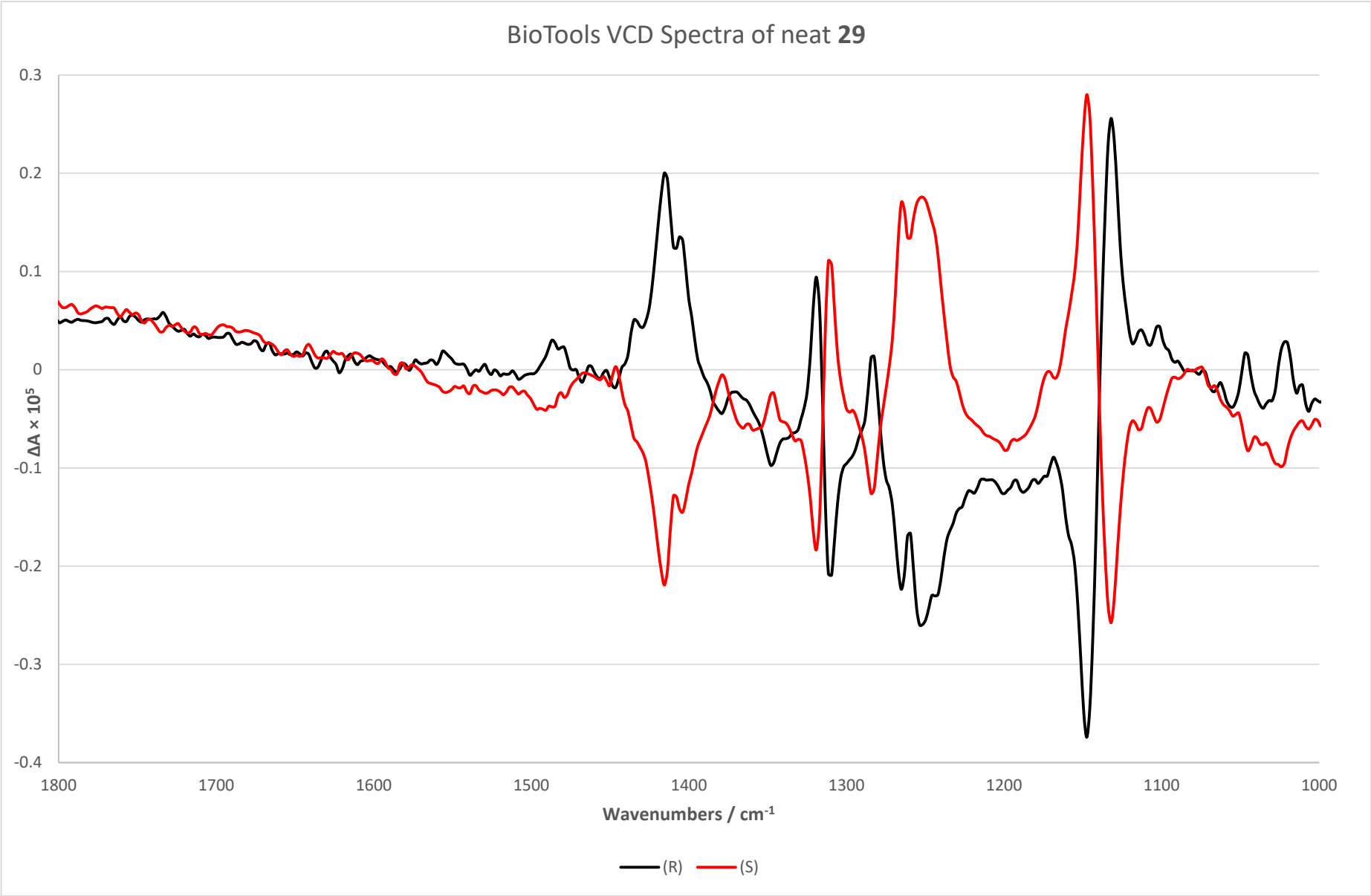

BioTools VCD Spectra of 1.0 mol dm-3 24 in CDCl3
0.007
0.005
0.003
0.001
-0.001
-0.003
-0.005
-0.007
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ13808814
AZ13808815

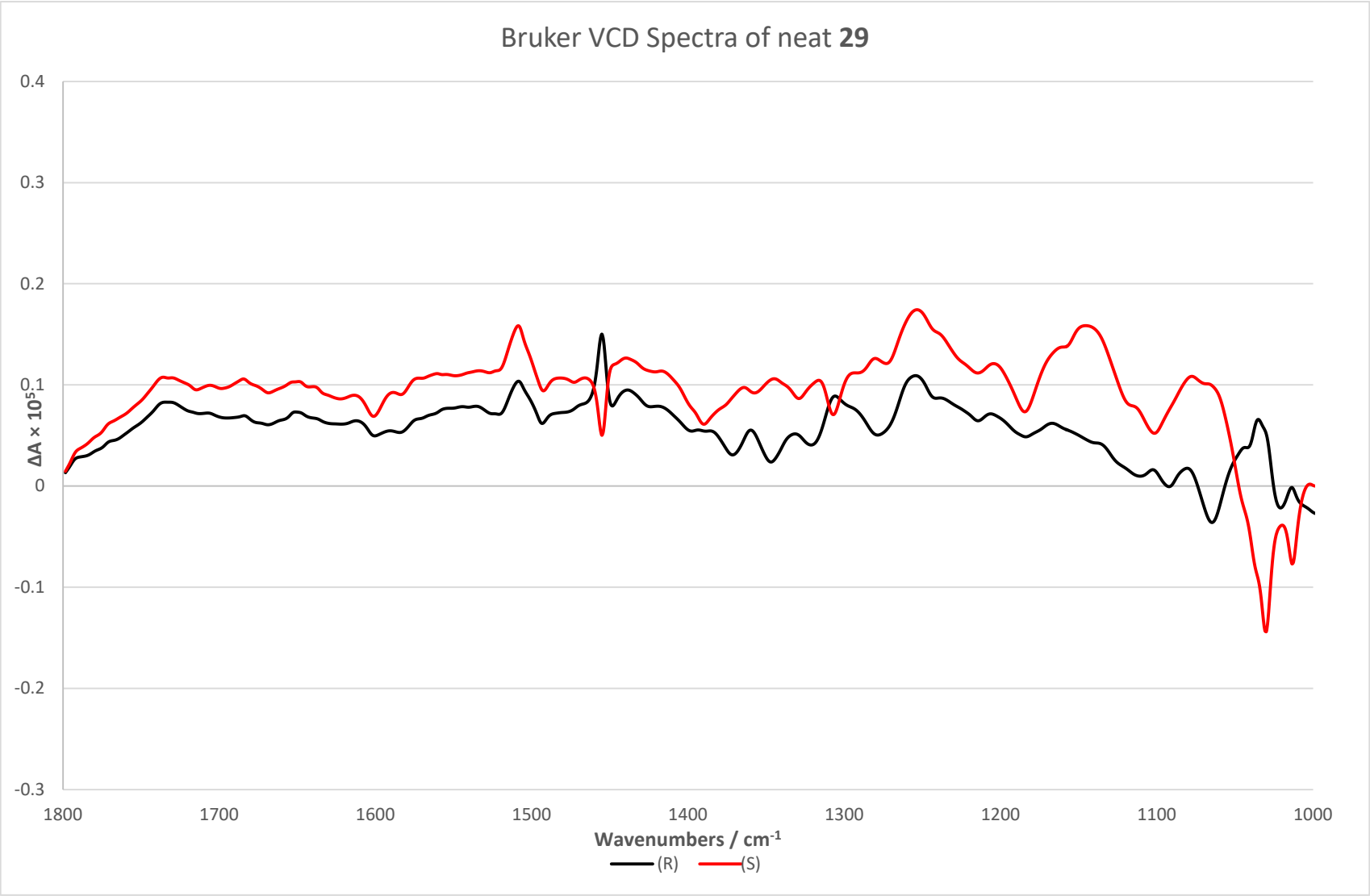

BioTools VCD Spectra of 1.0 mol dm-3 25 in CDCl3
0.005
0
-0.005
-0.01
-0.015
-0.02
-0.025
-0.03
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ13388092
AZ13393042

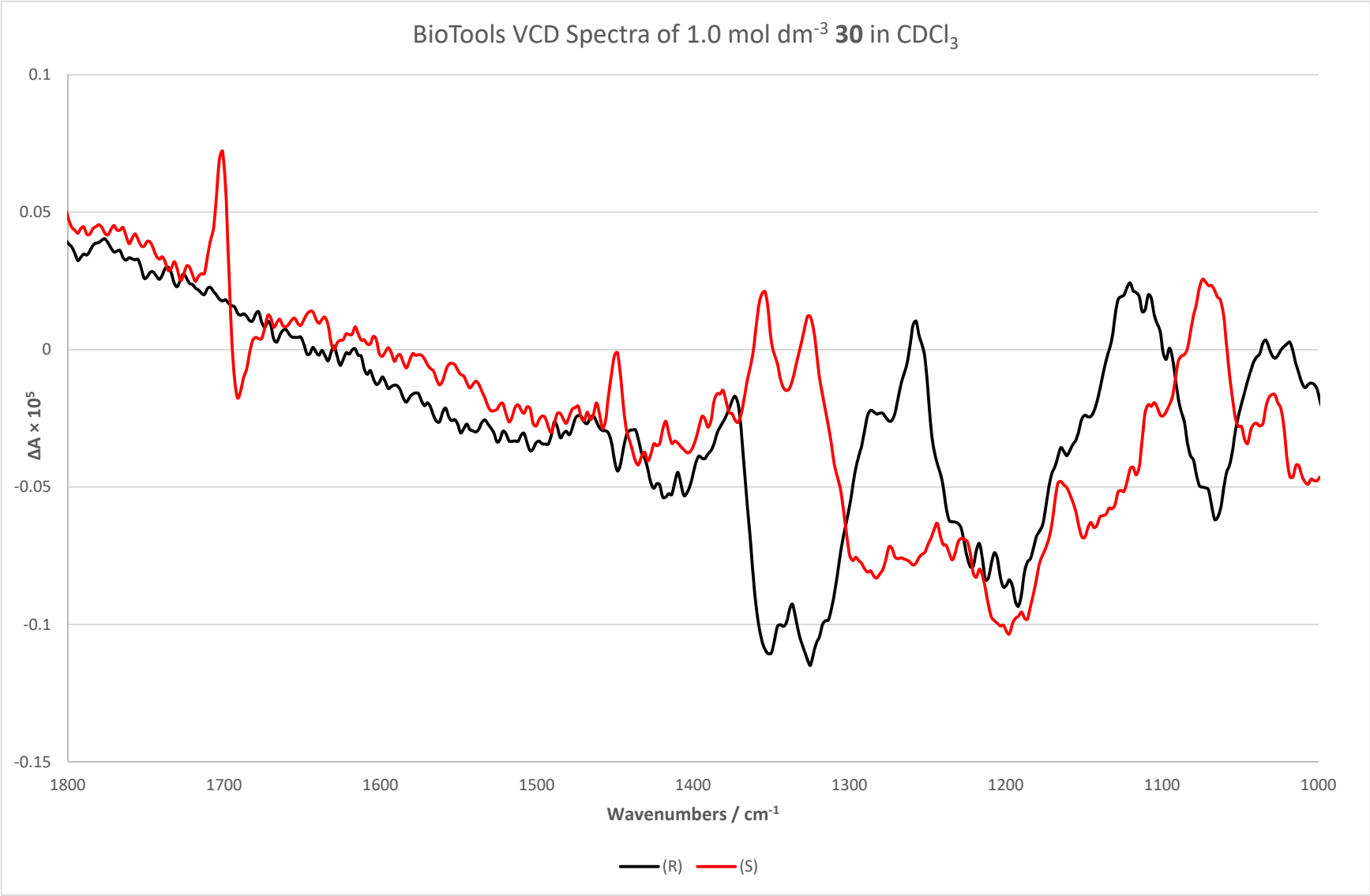

BioTools VCD Spectra of 0.5 mol dm-3 26 in CDCl3
0.005
0.004
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ13661192
AZ13661192

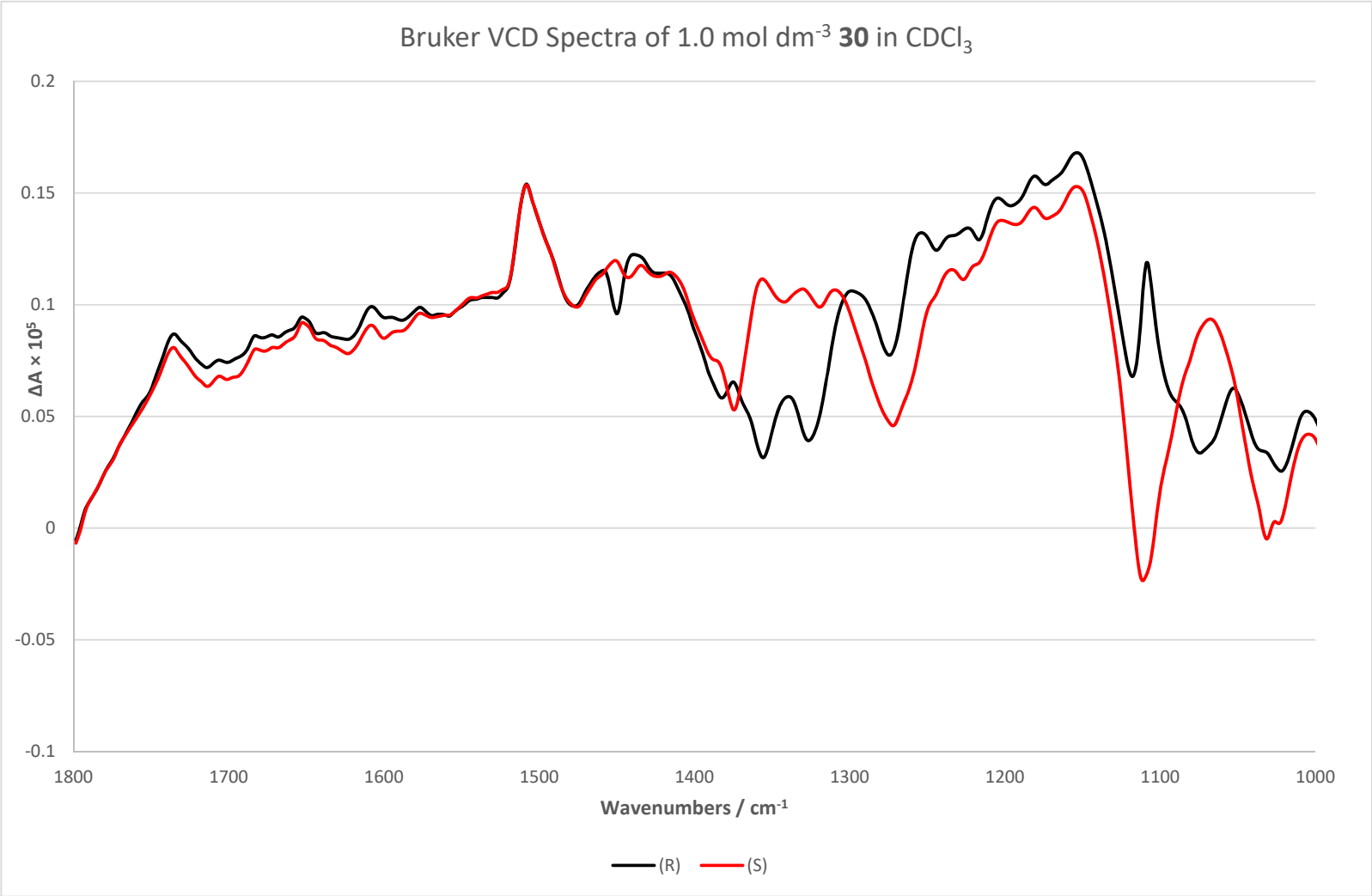

BioTools VCD Spectra of 0.5 mol dm-3 27 in CDCl3
0.003
0.002
0.001
0
-0.001
-0.002
-0.003
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
AZ1380653
AZ1380657

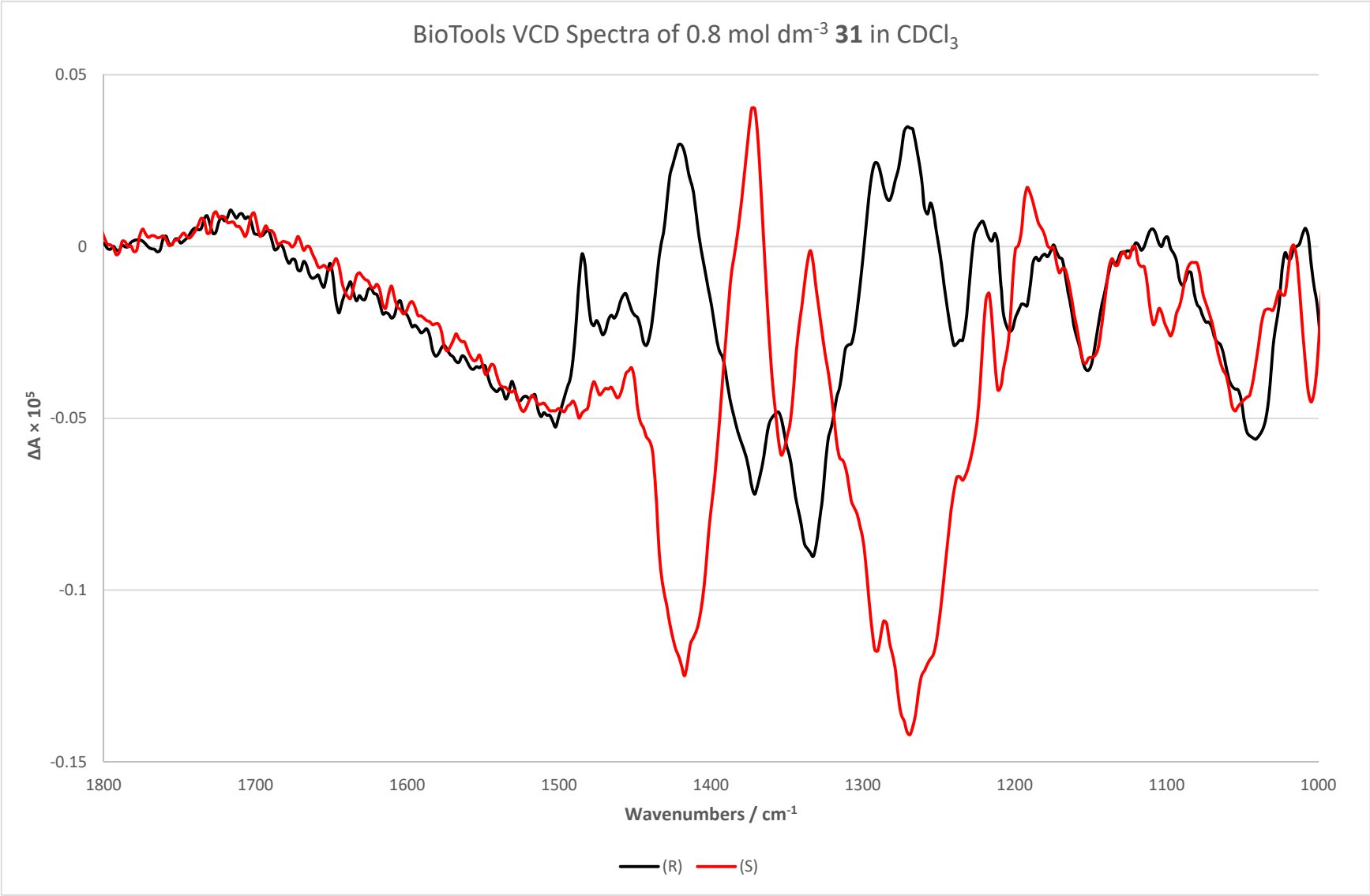

BioTools VCD Spectra of 1.0 mol dm-3 28 in CDCl3
0.01
0.005
0
-0.005
-0.01
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(RR)
(SS)

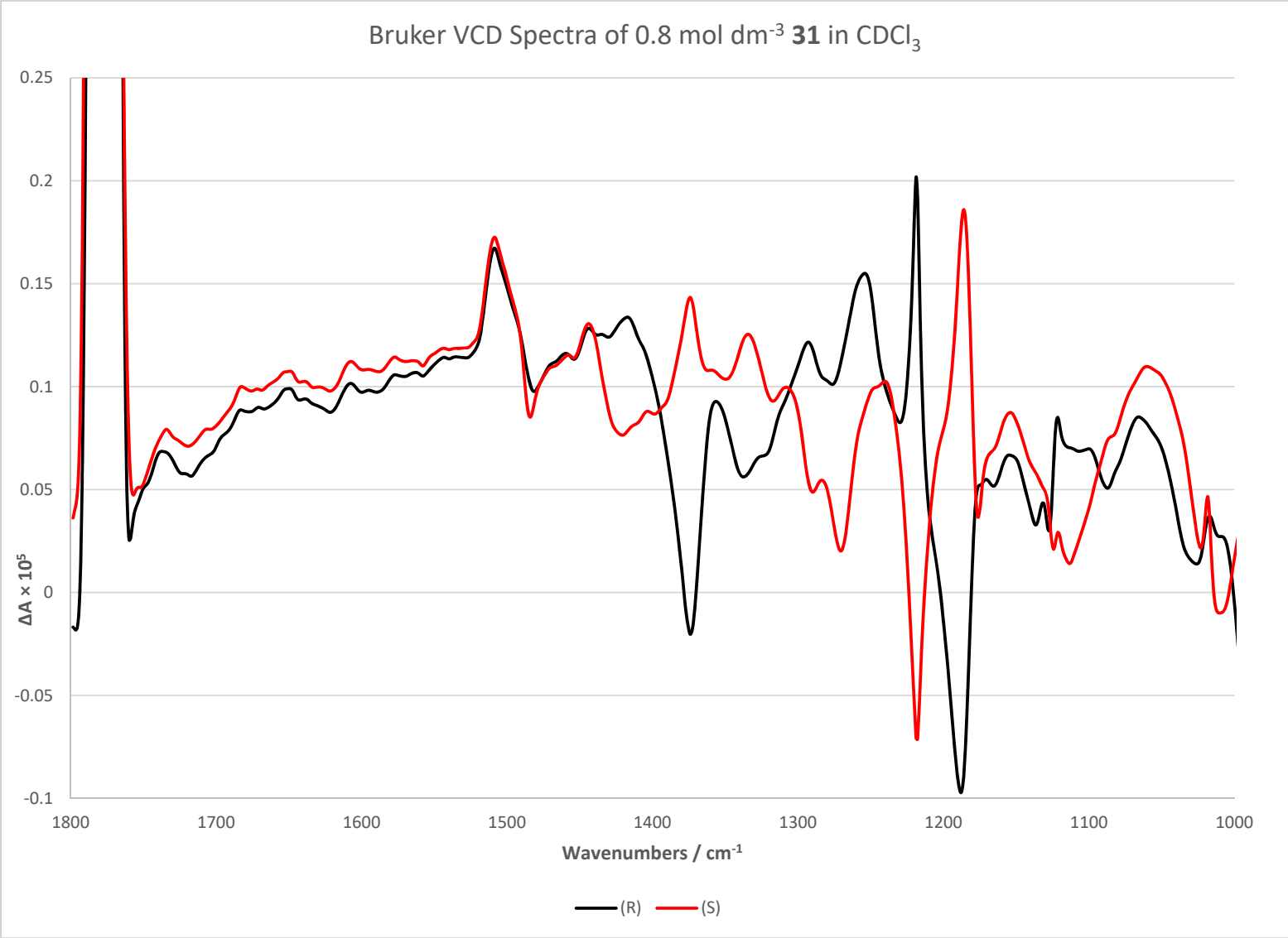

BioTools VCD Spectra of neat 29
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

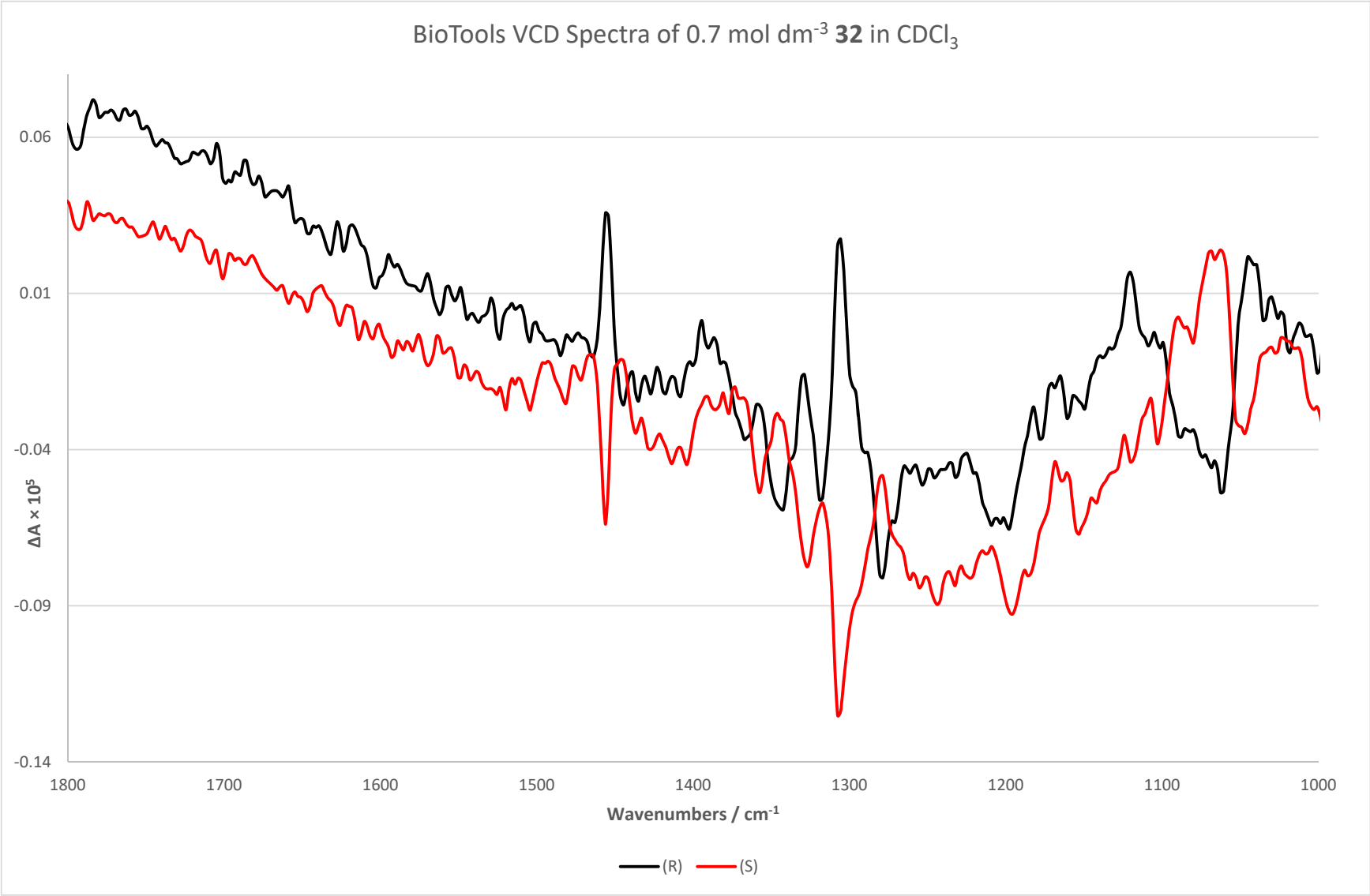

Bruker VCD Spectra of neat 29
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm⁻¹
(R)
(S)

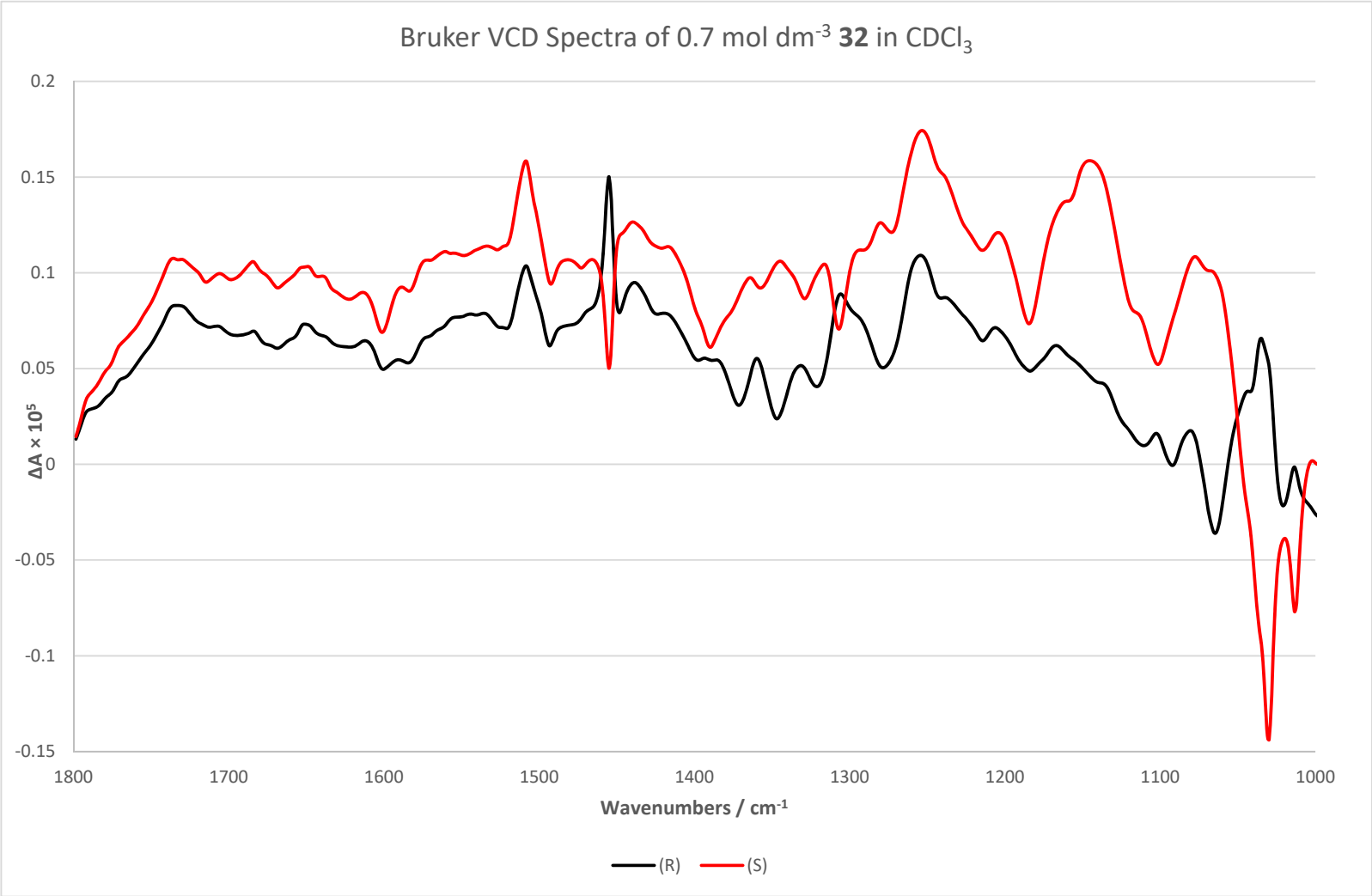

BioTools VCD Spectra of 1.0 mol dm-3 30 in CDCl3
0.1
0.05
0
-0.05
-0.1
-0.15
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

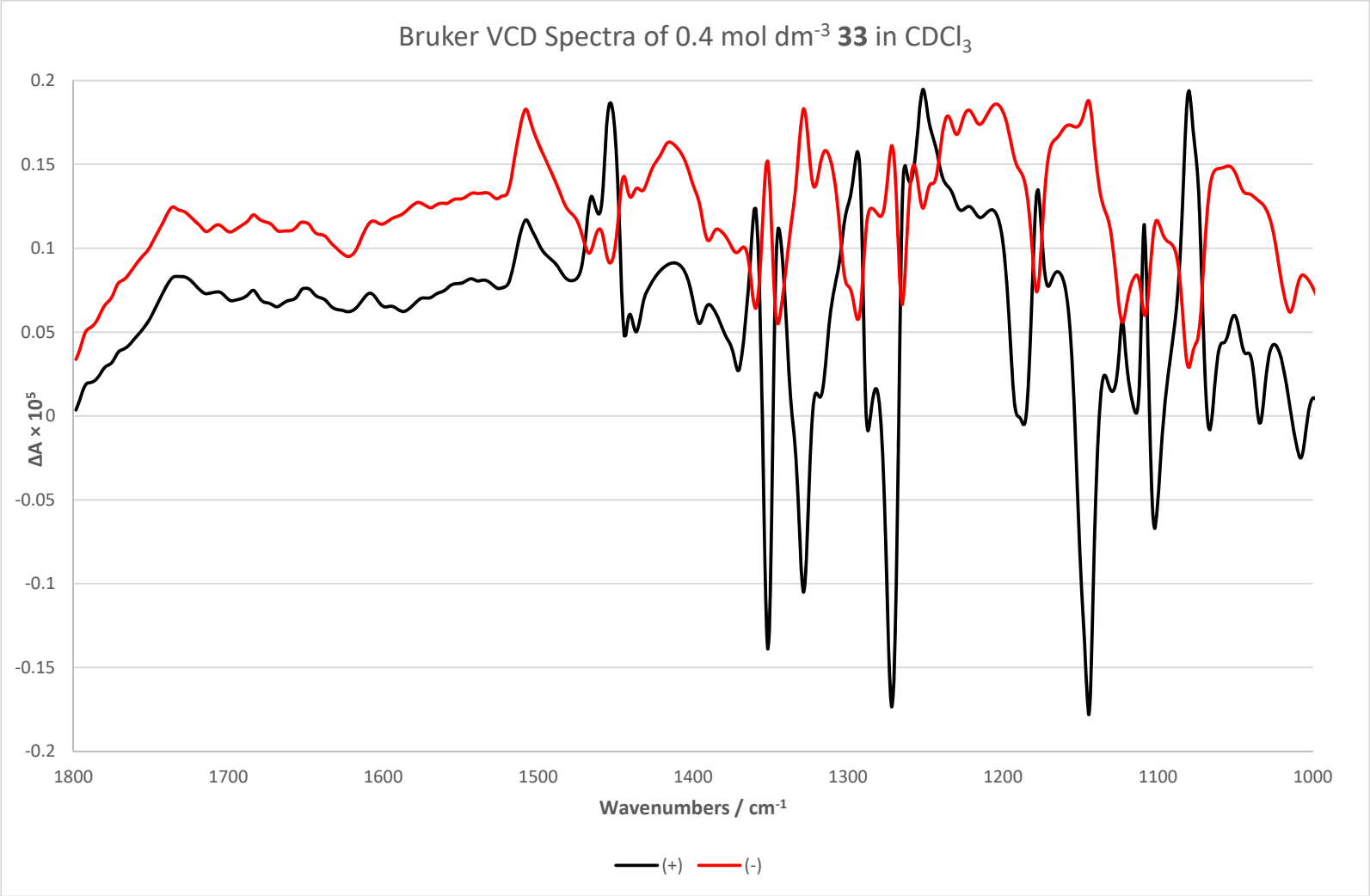

Bruker VCD Spectra of 1.0 mol dm-3 30 in CDCl3
0.2
0.15
0.1
0.05
0
-0.05
-0.1
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

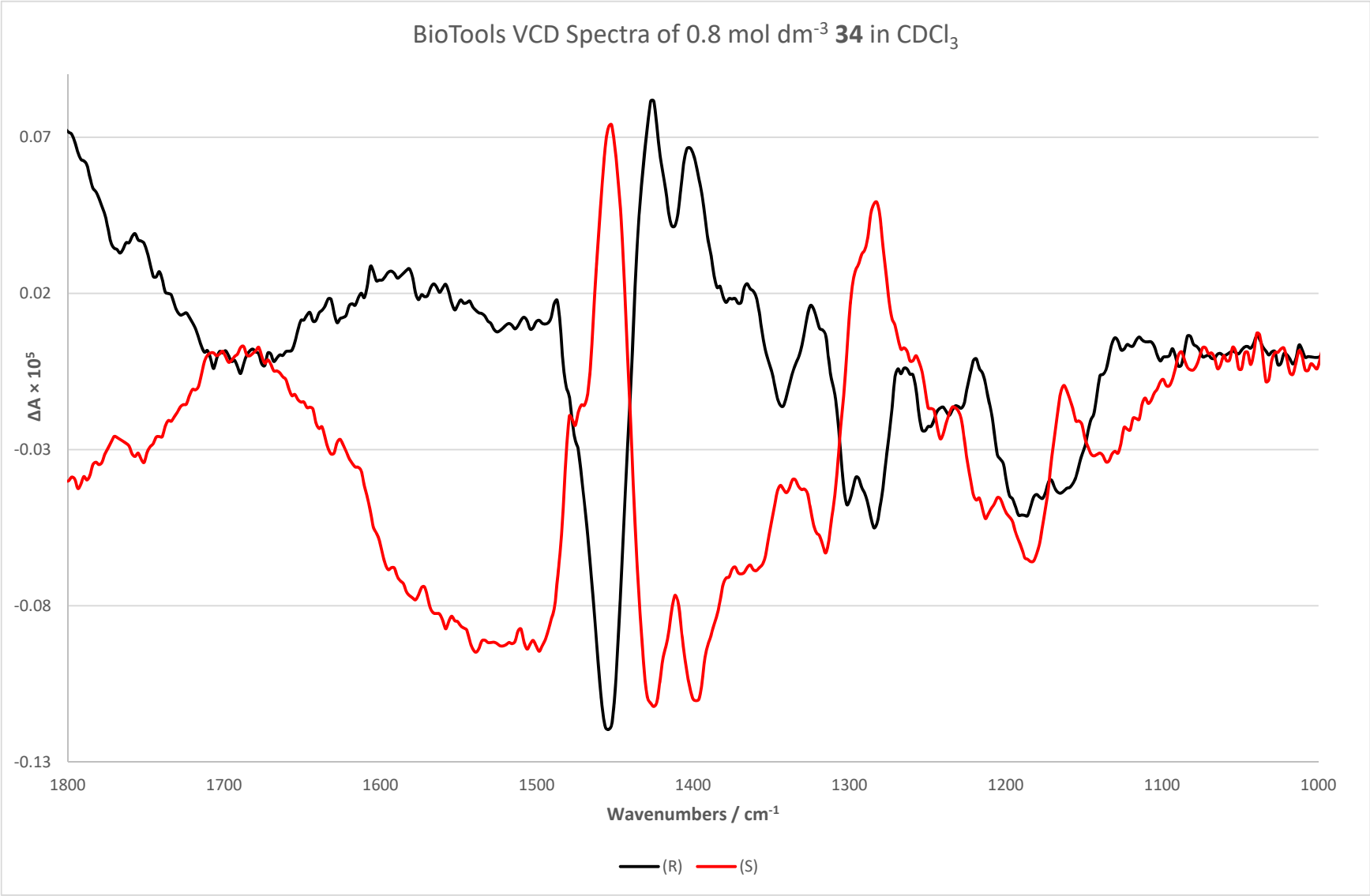

BioTools VCD Spectra of 0.8 mol dm-3 31 in CDCl3
0.05
0
-0.05
-0.1
-0.15
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

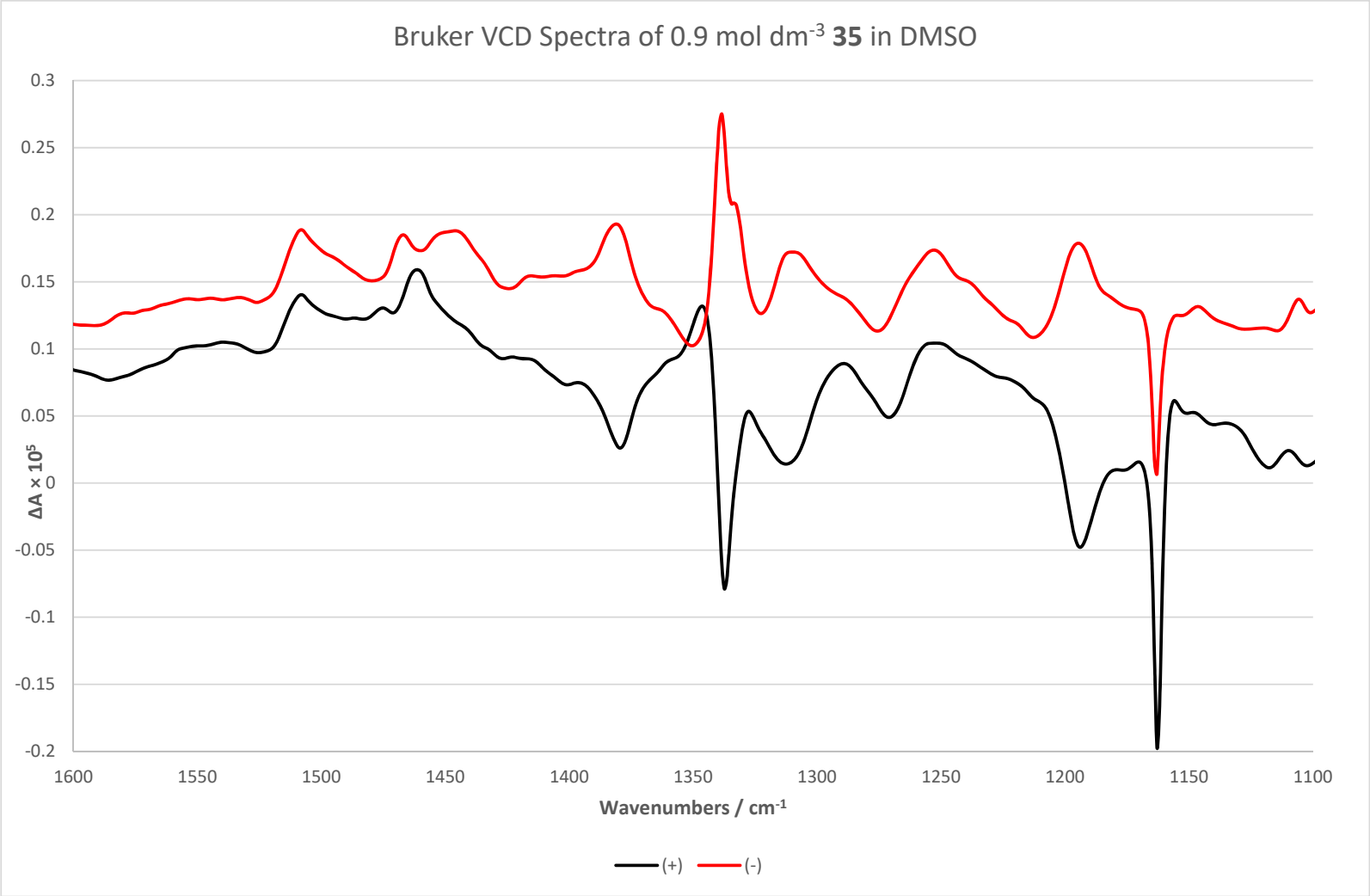

Bruker VCD Spectra of 0.8 mol dm-3 31 in CDCl3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

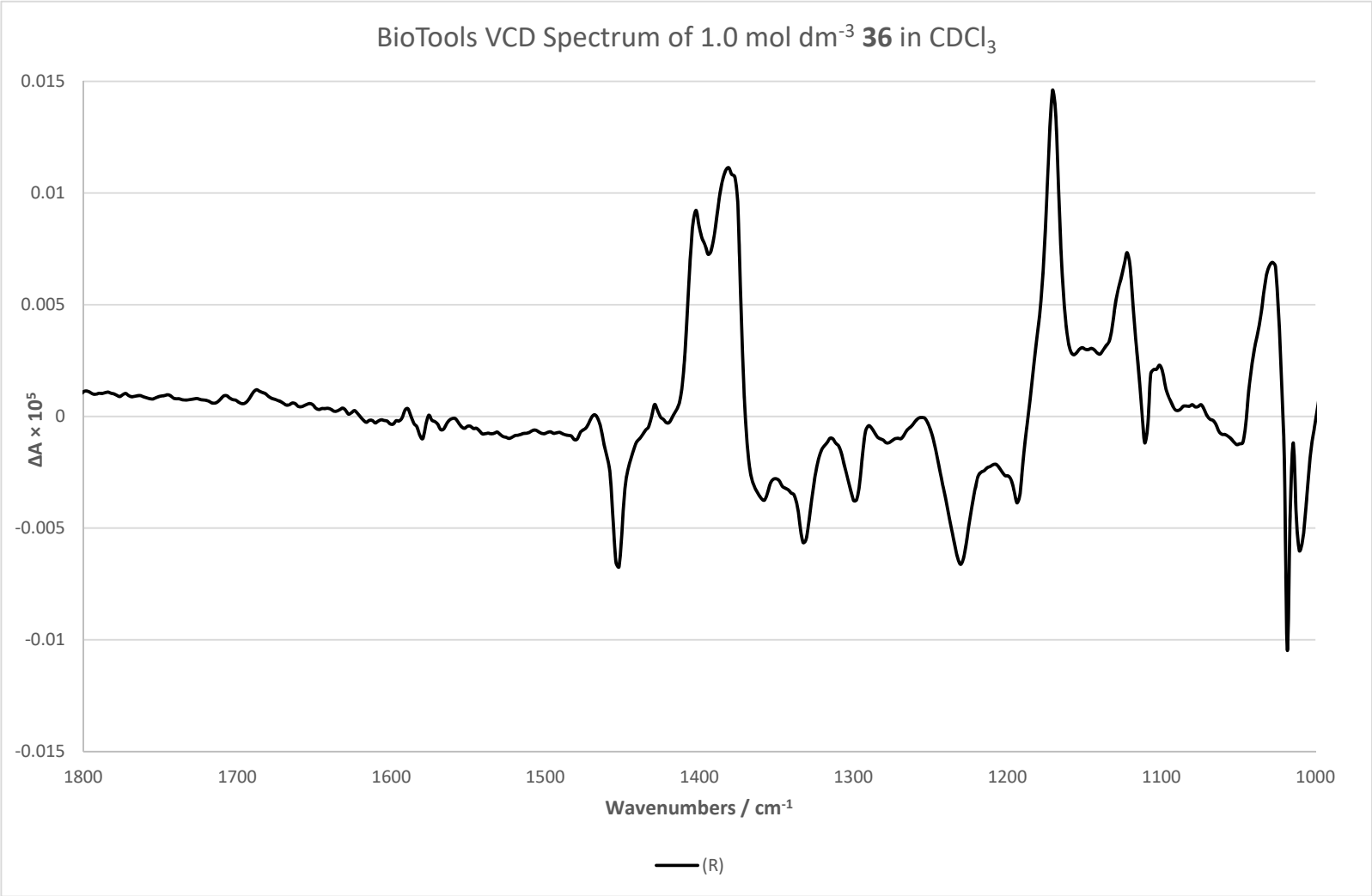

BioTools VCD Spectra of 0.7 mol dm-3 32 in CDCl3
0.06
0.01
-0.04
-0.09
-0.14
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

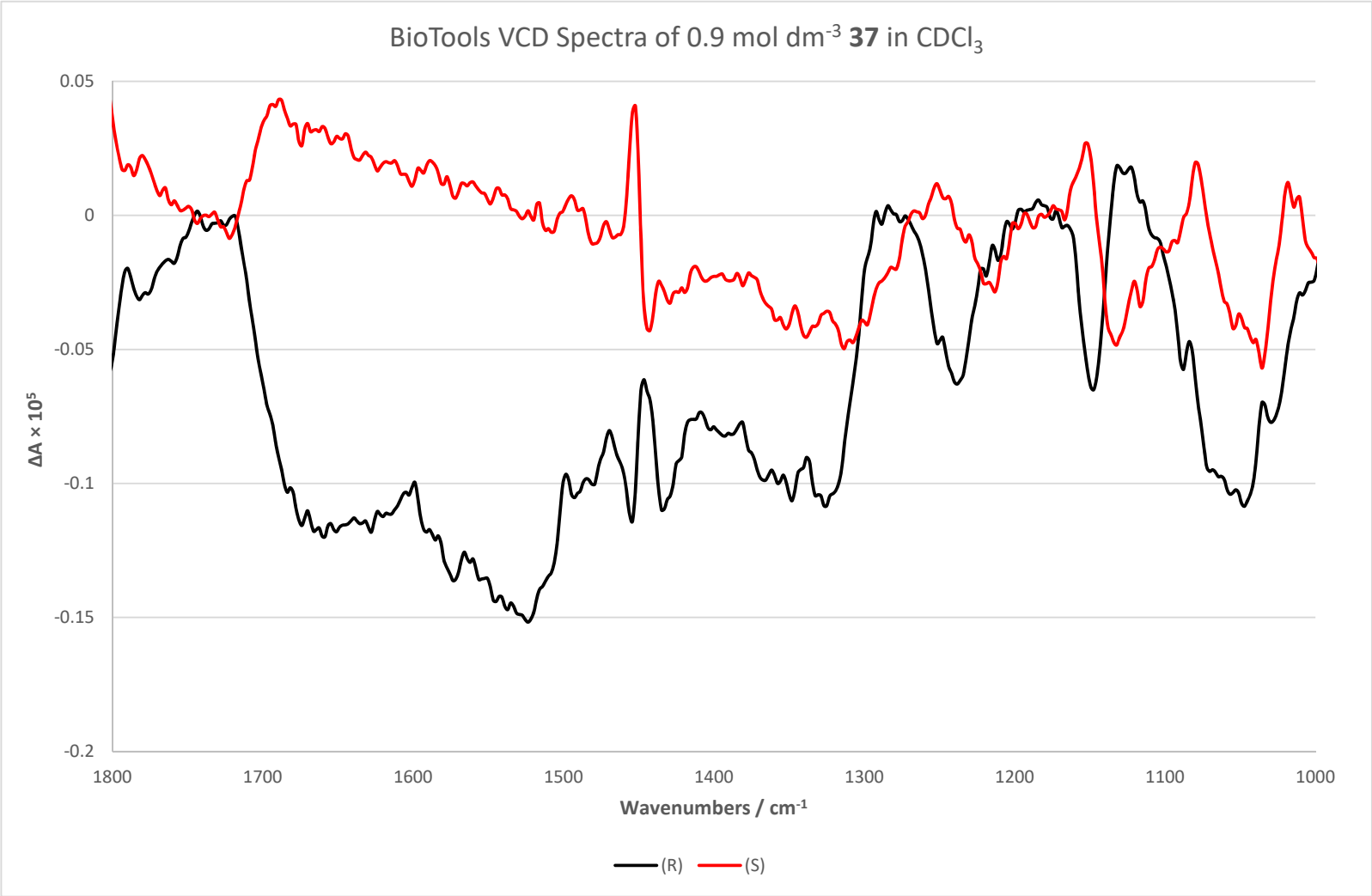

Bruker VCD Spectra of 0.7 mol dm-3 32 in CDCl3
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

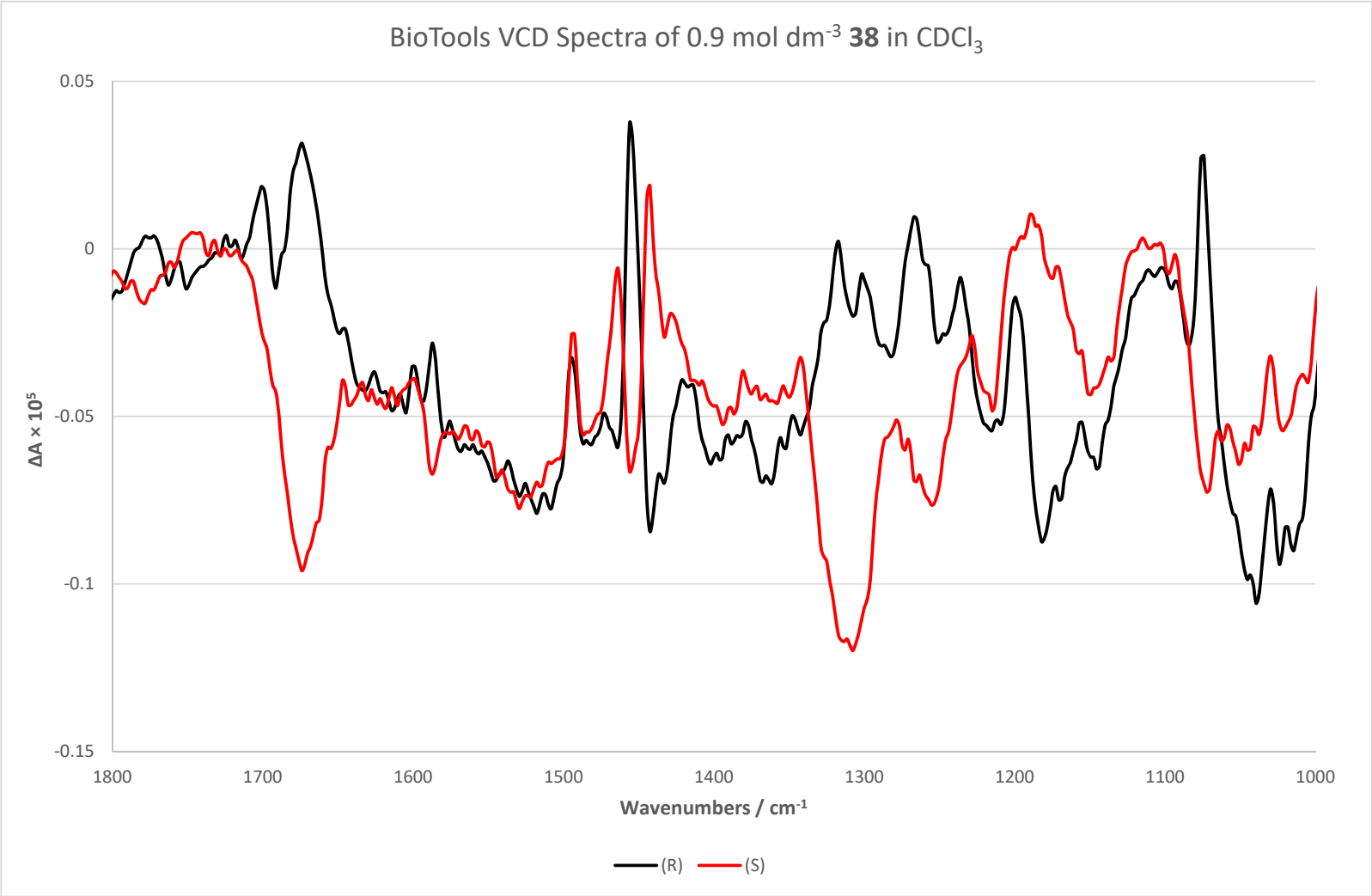

Bruker VCD Spectra of 0.4 mol dm-3 33 in CDCl3
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(+)
(-)

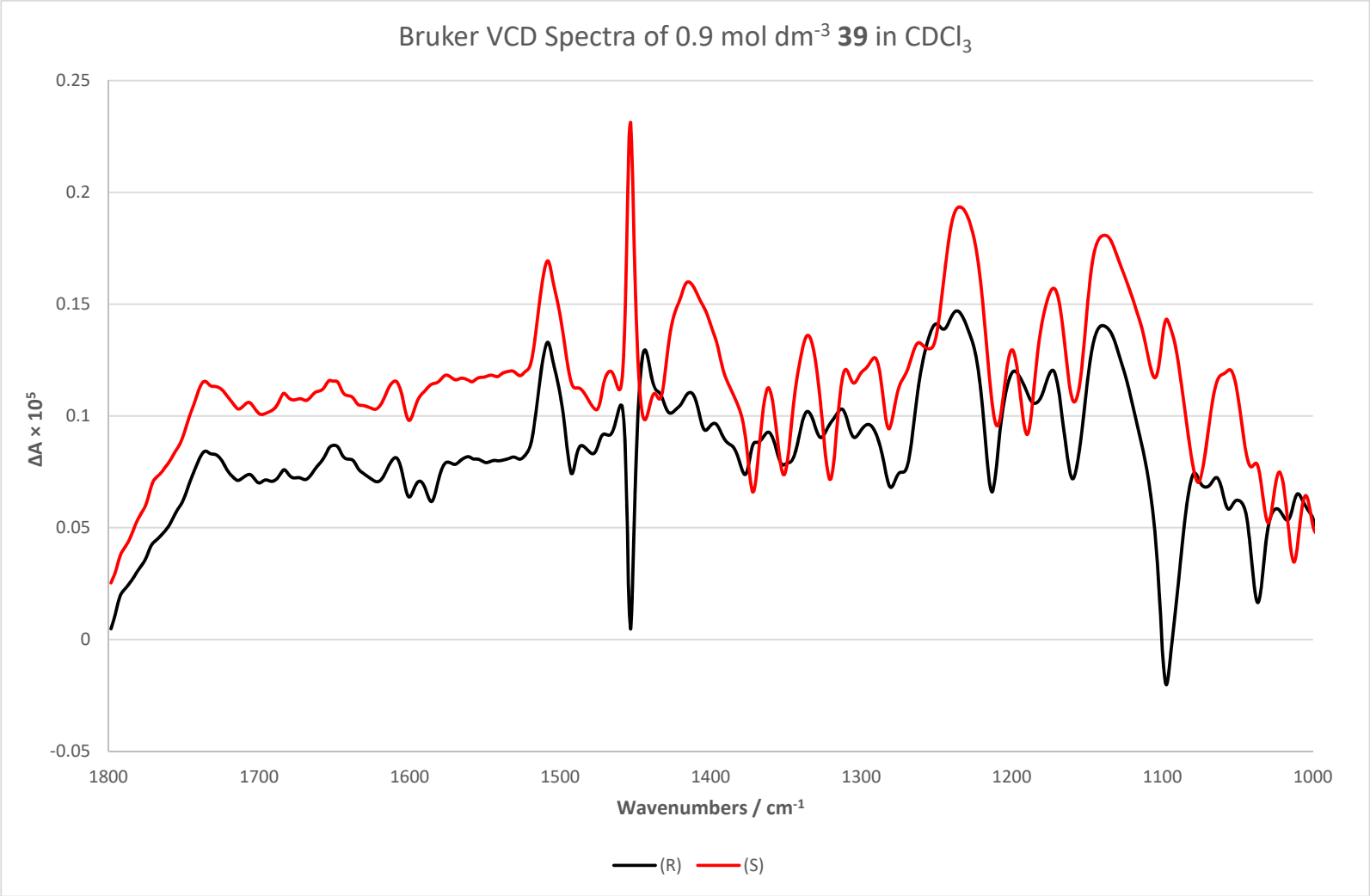

BioTools VCD Spectra of 0.8 mol dm-3 34 in CDCl3
0.07
0.02
-0.03
-0.08
-0.13
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

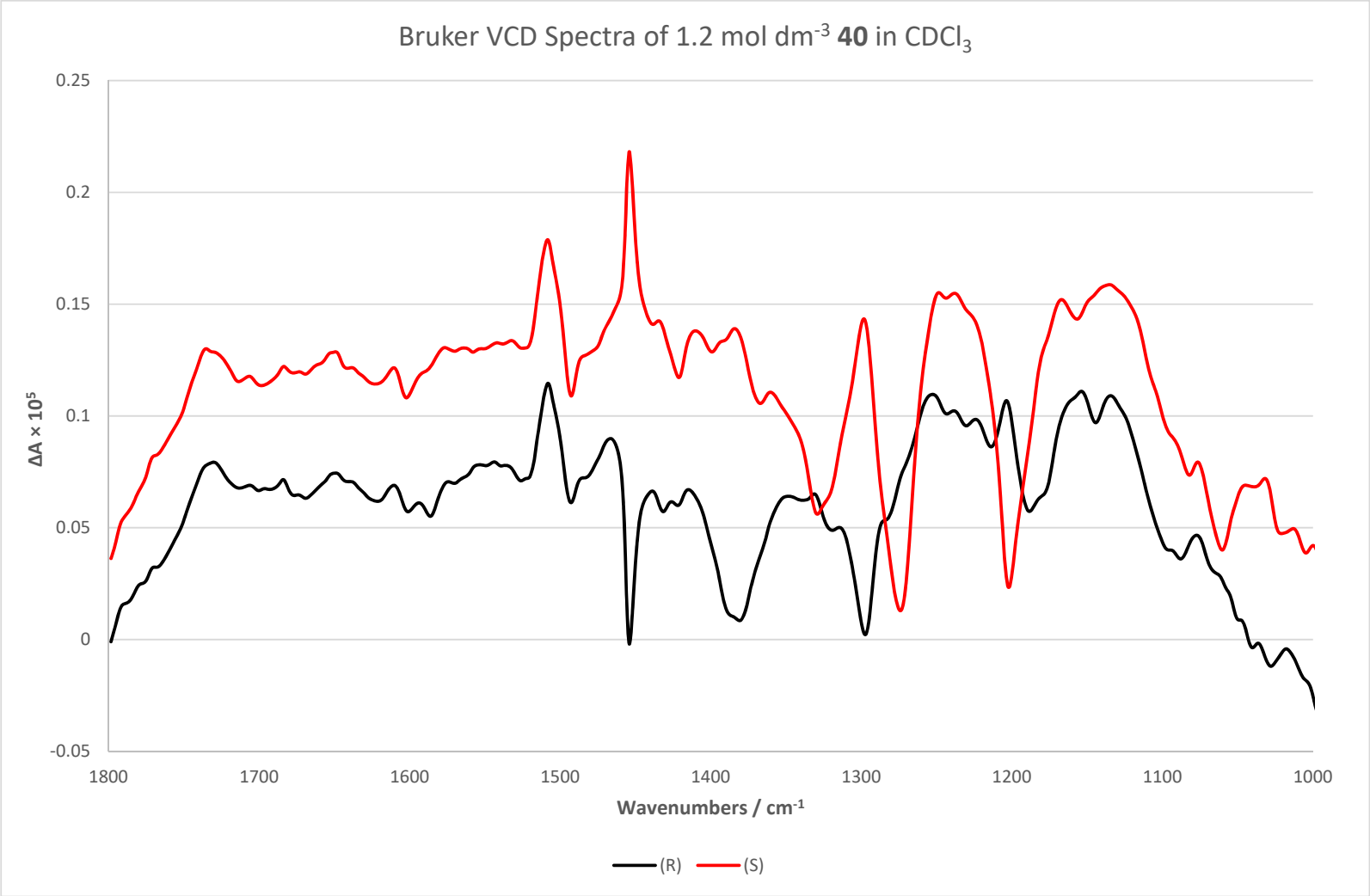

Bruker VCD Spectra of 0.9 mol dm-3 35 in DMSO
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
ΔA × 10^5
1600
1550
1500
1450
1400
1350
1300
1250
1200
1150
1100
Wavenumbers / cm-1
(+)
(-)

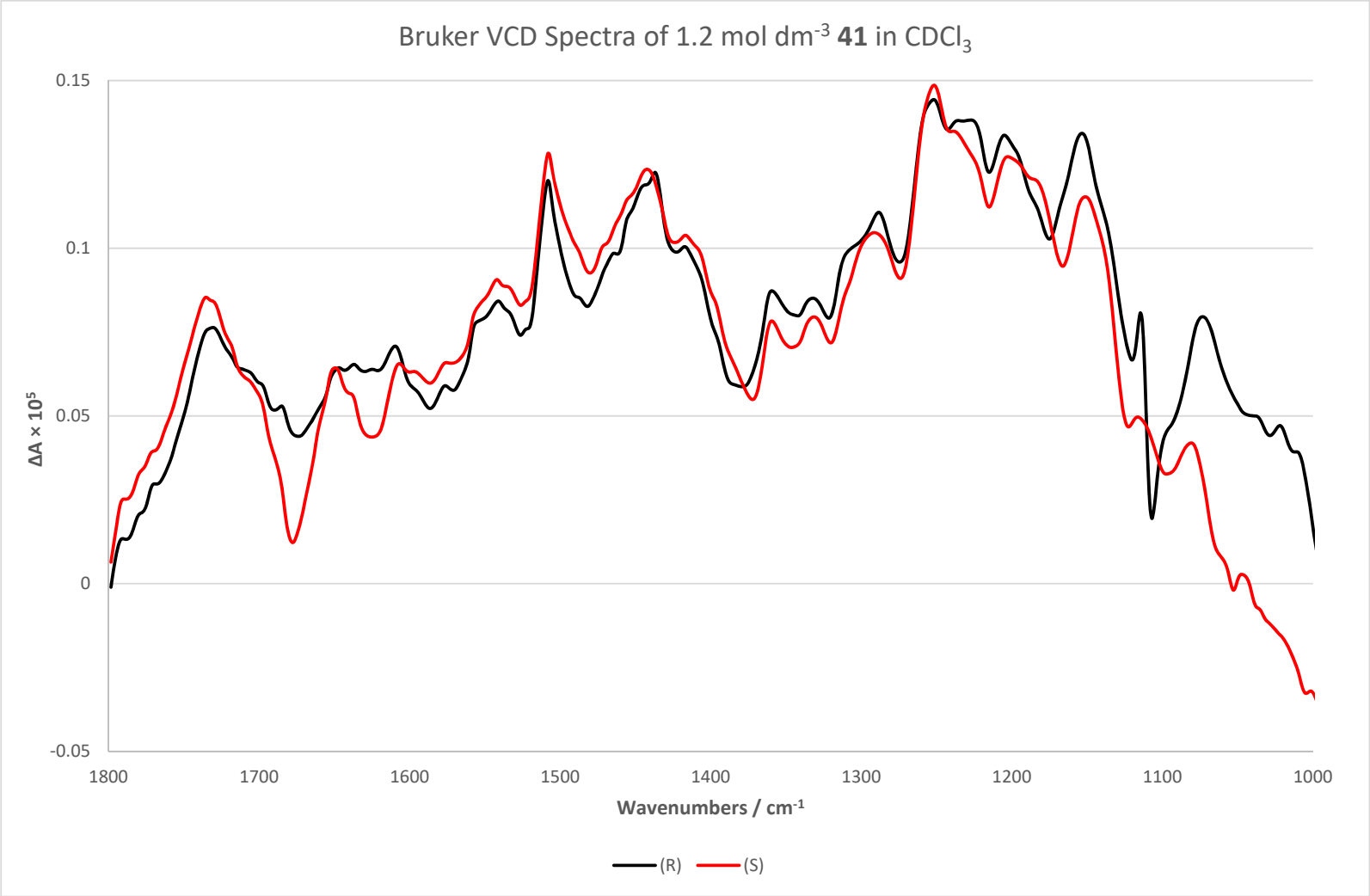

BioTools VCD Spectrum of 1.0 mol dm-3 36 in CDCl3
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)

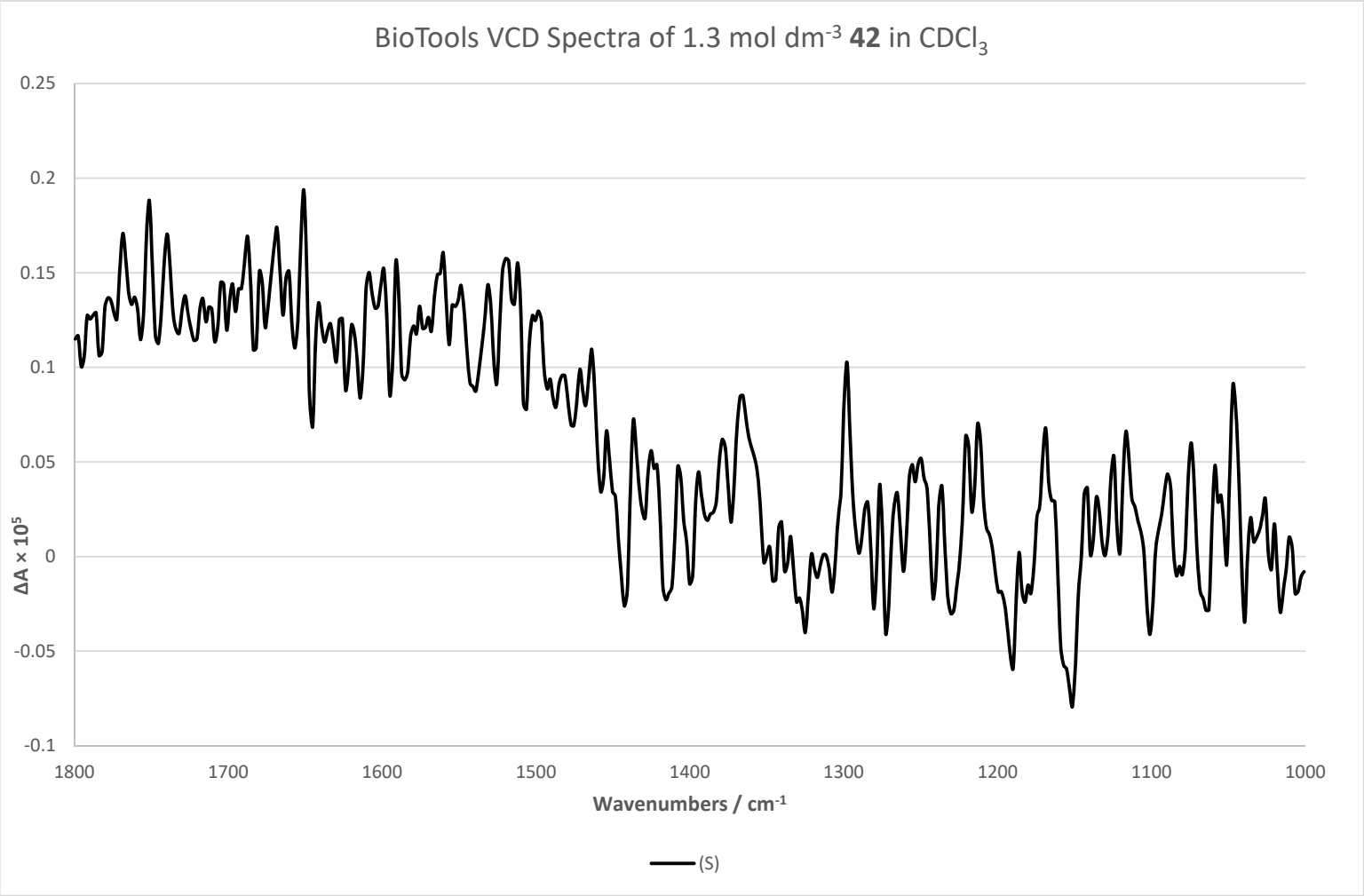

BioTools VCD Spectra of 0.9 mol dm-3 37 in CDCl3
0.05
0
-0.05
-0.1
-0.15
-0.2
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

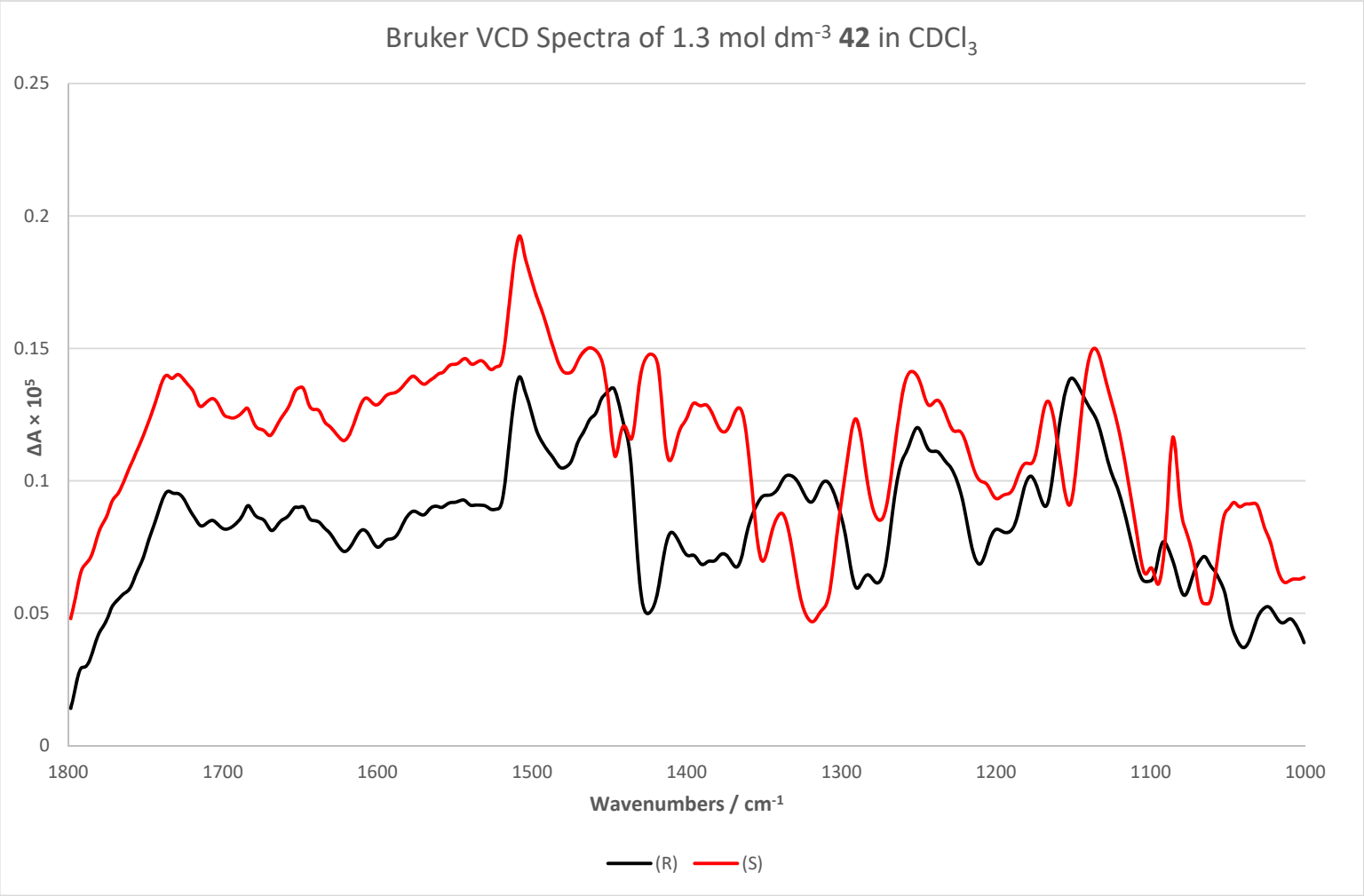

BioTools VCD Spectra of 0.9 mol dm-3 38 in CDCl3
0.05
0
-0.05
-0.1
-0.15
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

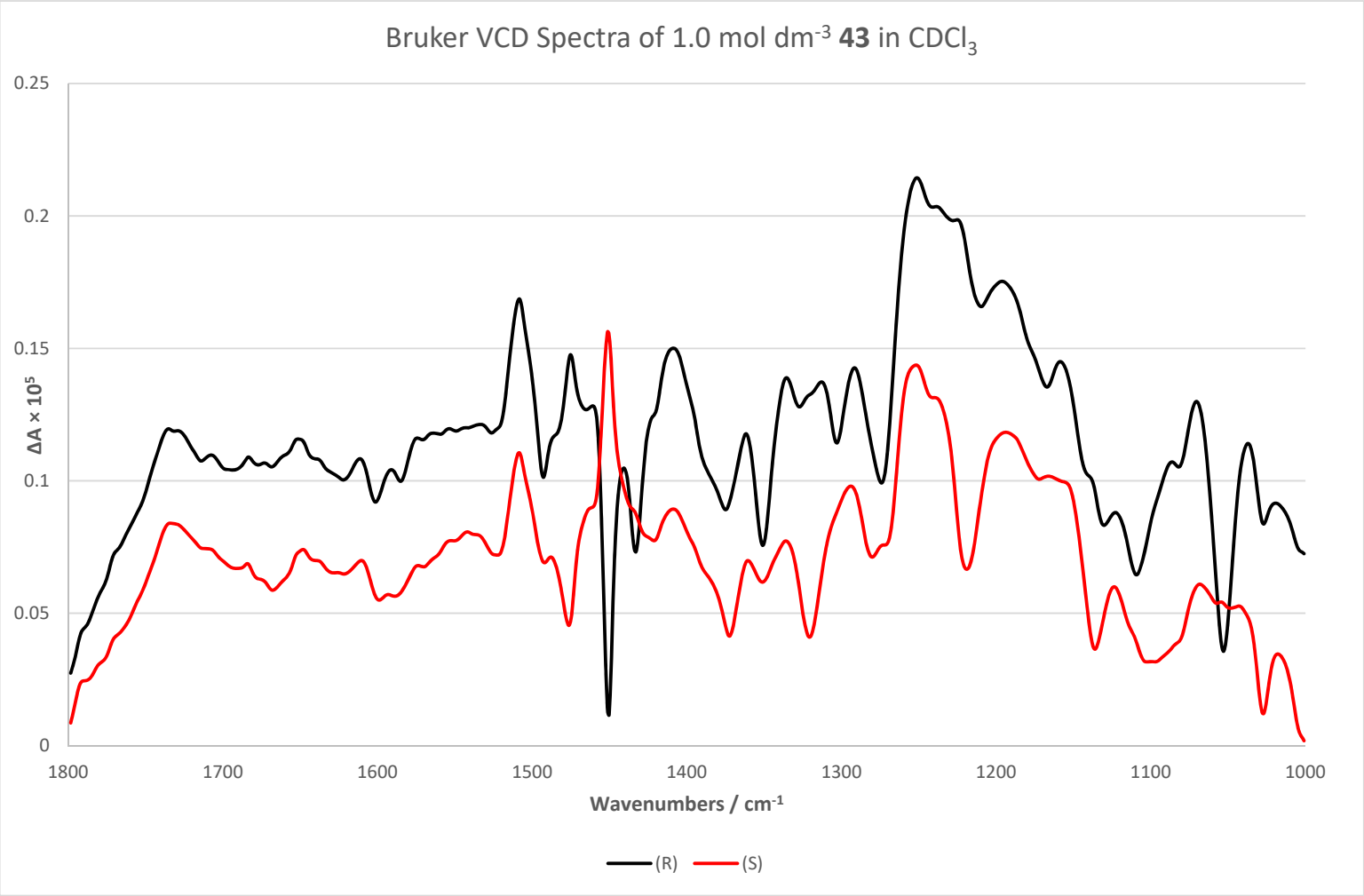

Bruker VCD Spectra of 0.9 mol dm-3 39 in CDCl3
0.25
0.2
0.15
0.1
0.05
0
-0.05
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)

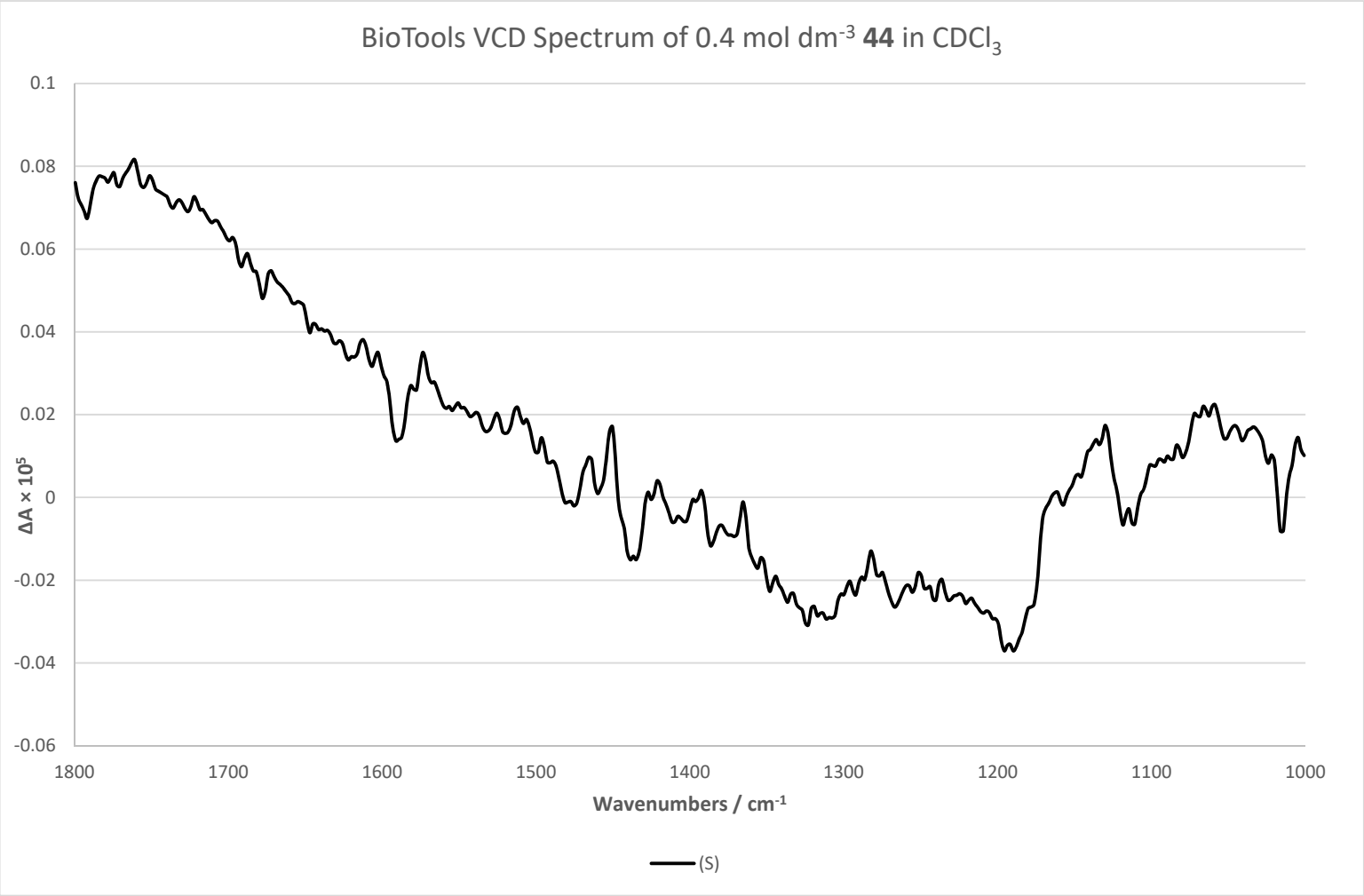

Bruker VCD Spectra of 1.2 mol dm-3 40 in CDCl3
0.25
0.2
0.15
0.1
0.05
0
-0.05
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)


Bruker VCD Spectra of 1.2 mol dm-3 41 in CDCl3
0.15
0.1
0.05
0
-0.05
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)


BioTools VCD Spectra of 1.3 mol dm-3 42 in CDCl3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(S)


Bruker VCD Spectra of 1.3 mol dm-3 42 in CDCl3
0.25
0.2
0.15
0.1
0.05
0
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)


Bruker VCD Spectra of 1.0 mol dm-3 43 in CDCl3
0.25
0.2
0.15
0.1
0.05
0
ΔA × 10⁵
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(R)
(S)


BioTools VCD Spectrum of 0.4 mol dm-3 44 in CDCl3
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
ΔA × 10^5
1800
1700
1600
1500
1400
1300
1200
1100
1000
Wavenumbers / cm-1
(S)