Machine learning for vibrational circular dichroism

Constructing novel and accelerating established applications

Proefschrif voorgelegd tot het behalen van de graad van doctor in de Wetenschappen: Chemie te verdedigen door

Tom Vermeyen

Universiteit Antwerpen

UNIVERSITEIT GENT

Promotor: Prof. Dr. Wouter Herrebout, University of Antwerp
Promotor: Prof. Dr. Patrick Bultinck, Ghent University
Faculteit Wetenschappen, Departement Chemie
Antwerpen 2023

fwo Fonds Wetenschappelijk Onderzoek
Vlaanderen
Opening new horizons

VLAAMS
SUPERCOMPUTER
CENTRUM
Vlaanderen
is supercomputing

This work was supported by the Research Foundation - Flanders (FWO).
Computational resources were provided by the Flemish Supercomputer Center (VSC).

Tom Vermeyen

Machine Learning for vibrational circular dichroism
Constructing novel and accelerating established applications

Thesis submitted in fulfilment of the requirements for the degree of Doctor in Science: Chemistry at the University of Antwerp and Ghent University, 2023.

Reviewers: Prof. Dr. Wout Bittremieux, Prof. Dr. Ir. An Ghysels,
Prof. Dr. Ir. Pieter Geiregat, Dr. Ir. Ann Vos,
Prof. Dr. Valentin Paul Nicu
Comittee chair: Prof. Dr. Filip Lemière
Promotors: Prof. Dr. Wouter Herrebout and Prof. Dr. Patrick Bultinck

University of Antwerp
TSM^2 Group
Department of Chemistry
Groenenborgerlaan 171
2020 Antwerpen

Ghent University
GQCG
Department of Chemistry
Krijgslaan 281
9000 Gent

Acknowledgements

Aan het einde van dit vijfjarige traject zou ik graag mijn dank willen betuigen aan verschillende personen die, zij het direct of indirect, bijgedragen hebben tot dit eindresultaat. In eerste instantie zou ik graag mijn promotor **Wouter** enorm willen bedanken. Tijdens mijn bachelor- en masterjaren heb je een stevige hand (de hand van de meester genoemd) gehad in mijn wetenschappelijke ontwikkeling. Ook de afgelopen jaren heb je mij steeds opportuniteiten gegeven en ervoor gezorgd dat ik mijn wetenschappelijke vaardigheden kon ontplooien en aanscherpen. Bovendien waren de fietssuggesties altijd verhelderend, ondanks je verwondering over hoe frequent deze reparaties nodig bleken. Met dit werk afgerond, zijn we gekomen aan de laatste 'we zullen wel zien'. Ik wil ook graag mijn promotor **Patrick** bedanken voor de steun de afgelopen jaren. Ik heb de vrijheid gekregen om mijn eigen ideeën te verkennen en de hulp gekregen om deze verder te verfijnen. Bovendien is mijn schrijfstijl er enorm op vooruitgegaan, dankzij je oog voor heldere communicatie en ingeslopen dutchisms. Dank jullie wel voor het vertrouwen om mij de sprong te laten wagen in de onbekende wereld van 'Machine Learning'.

I would like to express my gratitude towards the researchers with whom I worked on fruitful collaborations. My first thanks go to **João Batista Jr** for developing the idea for the monoterpene project, working on meticulously assigning the spectral patterns and providing the experimental monoterpene dataset. Secondly, I thank **Jure Brence** for his guiding hand when I was first exploring the capabilities of Machine Learning, helping me to become more comfortable and knowledgeable on the various techniques. Also, I want to express my gratitude towards **Christian Merten** for hosting me in Bochum during my master studies and helping to get me my first paper (Prost!). **Ana Cunha**, thank you for the assistance with the details of Machine Learning and computations. Finally, I would like to express my gratitude to (now Dr.!) **Hien Le**, along with **Emmy Tuenter** and **Luc Pieters**, for their seemingly never-ending list of plant samples to perform chiral analysis on.

Aan de bende Molspeccers die mij voorgingen en achternakwamen ben ik zeker ook dank verschuldigd. De Molspec groep (sindskort TSM^2) is een warm nest waarin ik in me sinds mijn bachelorjaren altijd welkom heb gevoeld. Ik heb deze traditie zo ver mogelijk verder gezet, hetzij op mijn volkomen eigen manier. Beste **Christian**, **Isabelle**, **Linny**, **Yannick**, **Carl**, **Evelien**, **Filip**, **Roberta**, **Jo**, **Robin**, **Jente**, **Dimitri**, **Sjobbe**, **Dyma** en **Kiana**. Een dikke merci voor de gezellige momenten, grappige voorvallen en interessante gesprekken (gesponserd door de koffiemachine). Ik wil specifiek ook **Jo** bedanken voor de hulp bij zowat alles waar ik hem mee lastig viel. Jouw feedback heeft alvast mijn zicht verbreed en je hebt mij structuur aangereikt op de meest nodige momenten. Tenslotte wil ik ook mijn partner-in-crime/science **Roy** uitdrukkelijk bedanken voor de kameraadschap en de vele inzichten de afgelopen tien jaren. Het heeft net iets langer geduurd voor me, maar we staan nu allebei aan de eindmeet van onze universitaire studies.

Ook buiten de academische sfeer hebben verschillende personen mij op het pad gehouden naar dit einddoel. Hierbij wil ik iedereen bedanken met wie ik een podium heb gedeeld en mij geholpen hebben om dit te verkennen. Ik zou specifiek **Edith** willen bedanken voor de sterke vriendschap en de verschillende muzikale escapades. Ook wil ik graag **Amber** en **Nick** bedanken, samen met de mensen van **Klub Kultuur**, **Al Dente** en **Jong-KVCV**.

Natuurlijk wil ik ook mijn familie bedanken. **Make**, **Pake**, **Pieter** en **Helena**. Sinds deze jonge knaap op 15-jarige leeftijd plots besliste om chemist te worden, stonden jullie steeds paraat en gaven jullie steun en liefde. Jullie hebben dit ganse traject van dichtbij gevolgd en mij stelselmatig geholpen om de opgedoken barrieres te overwinnen. Ik ben niet zeker of jullie mijn fascinatie met dit vakgebied ooit hebben begrepen, maar ik hoop dat ik jullie trots heb gemaakt.

Evelyne, **Luc**, **Maxim**. Het hielp enorm dat ik wist dat er bij jullie altijd een plaatsje vrij was, zelfs wanneer covid om de hoek loerde, waardoor het bij jullie al snel aanvoelde als een tweede thuis. Ik wil jullie, samen met de rest van de familie **De Vocht** en **Van Nueten**, enorm bedanken voor de warme verwelkoming.

Tenslotte wil ik **Margot** uitermate bedanken om mijn luidste supporter te zijn en voor haar schijnbaar onuitputtelijke geduld tijdens de lange schrijfsessies. De onverlate steun en onvoorwaardelijke liefde die je mij gaf, hielpen me doorheen de meest uitdagende beklimmingen van dit parcours en brachten me steeds terug naar het juiste pad richting de eindmeet. Het ziet ernaar uit dat we samen aan de finish zijn geraakt.

List of publications

The following papers were published as a result of the work performed by the author of this thesis.

i. T. Vermeyen and C. Merten, Solvation and the secondary structure of a proline-containing dipeptide: Insights from VCD spectroscopy, *Phys. Chem. Chem. Phys.*, 2020, **22**, 15640-15648. (IF 3.9)

ii. K. Knippen, B. Bredenkötter, L. Kanschat, M. Kraft, T. Vermeyen, W. Herrebout, K. Sugimoto, P. Bultinck and D. Volkmer, CFA-18: A homochiral metal-organic framework (MOF) constructed from rigid enantiopure bistriazolate linker molecules, *Dalton Trans.*, 2020, **49**, 15758-15768. (IF 4.6)

iii. M. Baldé, E. Tuenter, M. Troaré, L. Peeters, A. Matheeussen, P. Cos, G. Caljon, T. Vermeyen, W. Herrebout, A. Balde, K. Foubert and L. Pieters, Antiplasmodial Oleanane Triterpenoids from Terminalia albida Root Bark, *J. Nat. Prod.*, 2021, **84**, 666-675. (IF 4.8)

iv. J. Bogaerts, R. Aerts, T. Vermeyen, C. Johannessen, W. Herrebout and J. Batista, Tackling Stereochemistry in Drug Molecules with Vibrational Optical Activity, *Pharmaceuticals*, 2021, **14**, 877. (IF 5.2)

v. T. Vermeyen, J. Brence, R. Van Echelpoel, R. Aerts, G. Acke, P. Bultinck and W. Herrebout, Exploring machine learning methods for absolute configuration determination with vibrational circular dichroism, *Phys. Chem. Chem. Phys.*, 2021, **23**, 19781-19789. (IF 3.9) This paper is reproduced in chapter 4 of this thesis.

vi. H.T.N. Le, T. Vermeyen, R. Aerts, W. Herrebout, L. Pieters and E. Tuenter, Epimeric mixture analysis and absolute configuration determination using an integrated spectroscopic and computational approach - A case study on two epimers of 6-hydroxyhippeastidine, *Molecules*, 2023, **28**, 214. (IF 4.9)

vii. H.T.N. Le, S. De Jonghe, K. Erven, T. Vermeyen, W. Herrebout, J. Neyts, C. Pannecouque, A. Baldé , L. Pieters and E. Tuenter, Anti-SARS-CoV-2 activity and cytotoxicity of Amaryllidaceae alkaloids from Hymenocallis littoralis. *Molecules*, 2023, **28**, 3222. (IF 4.9)

viii. T. Vermeyen, A. Batista, A. Valvedere, W. Herrebout, J. Batista. Pushing the boundaries of VCD spectroscopy in natural product chemistry. *Phys. Chem. Chem. Phys.*, 2023, **25**, 13825-13832. (IF 3.9) This paper is reproduced in chapter 6 of this thesis.

ix. T. Vermeyen, A. Cunha, P. Bultinck, W. Herrebout, Impact of conformation and intramolecular interactions on vibrational circular dichroism spectra identified with machine learning. *Commun. Chem.*, 2023, **6**, 148. (IF 7.2) The results of this paper are reproduced in chapter 5 of this thesis.

x. H.T.N. Le, S. De Jonghe, K. Erven, J. Neyts, C. Pannecouque, T. Vermeyen, W. Herrebout, L. Pieters, E. Tuenter, Comprehensive study of alkaloids from Scadoxus multiflorus by HPLC-PDA-SPE-NMR and evaluation of their anti-SARS-CoV-2 activity. *Phytochem. Lett.*, 2023, **57**, 156-162. (IF 1.7)

xi. H.T.N. Le, S. De Jonghe, K. Erven, J. Neyts, C. Pannecouque, T. Vermeyen, W. Herrebout, L. Pieters, E. Tuenter, A new alkaloid from Pancratium maritimum - Structure elucidation using computer-assisted structure elucidation (CASE) and evaluation of anti-SARS-CoV-2 activity. *Phytochem. Lett.*, 2023, **58**, 1-7. (IF 1.7)

Data management plan and resources

For Machine Learning (ML) applications, a lot of attention is given to the exact models and their performance for different tasks. Nonetheless, the most important part for the success of ML models arguably lies in the data behind the model. Following the current guidelines of FWO and UGent, the research data generated in this work is made publicly available on the zenodo repository.

- Exploring machine learning methods for absolute configuration determination with vibrational circular dichroism (Chapter 4): the VCD spectral dataset for the different combinations of functionals, basis sets and FWHM are available at `doi.org/10.5281/zenodo.8142819`. The PCA & t-SNE transformed data, linear coefficients and feature importances for the RF models can be found at `https://github.com/TomVermeyen/AC_Determination.git`.

- Impact of conformation and intramolecular interactions on vibrational circular dichroism spectra identified with machine learning (Chapter 5): The conformers spectra for all compounds, and the stereoisomers considered, are available at `doi.org/10.5281/zenodo.8009874`.

- Pushing the boundaries of VCD spectroscopy in natural product chemistry (Chapter 6): The experimental VCD and IR terpenes, mixtures and natural oils dataset can be found at `doi.org/10.5281/zenodo.7875469`

It is my opinion that the future of ML applications within the field will stand or fall with availability of a VCD spectral database. The creation of VCD spectral databases has been discussed frequently within the community. I have heard that currently active steps are being undertaken to build such a database. I hope that by making this data publicly available I can contribute to the future maturity of the subject.

To the PhD student who has been given this work as a starting guide for ML within VCD, I would like to make the following comment. At the beginning of this work, my background in ML was limited to standard linear models and some basic knowledge on QSAR. Knowledge of these subjects is easiest to gain by playing with data and reproducing results. In this sense, I would recommend to use the data from Chapter 4 and fit models to them, all models were run on my home computer within minutes. Create a python notebook (Jupyter/Google Colab), open a webpage at `scikit-learn.org` and follow your curiosity.

Contents

List of abbreviations

2D	two-Dimensional
3D	three-Dimensional
AC	Absolute Configuration
AI	Artificial Intelligence
CA	Classification Accuracy
CD	Circular Dichroism
CPL	Circularly Polarised Light
CIP	Cahn-Ingold-Prelog
DFT	Density Functional Theory
ECD	Electronic Circular Dichroism
EDTM	Electric Dipole Transition Moment
ELU	Exponential Linear Unit
FNN	Feedforward Neural Network
FT	Fourier-Transform
FWHM	Full Width at Half Maximum
GC	Gas Chromatography
GC-MS	Gas Chromatography - Mass Spectrometry
GUI	Graphical User Interface
IR	InfraRed
kNN	k-Nearest Neighbours
Lasso	Least absolute shrinkage and selection operator
LCPL	Left handed Circularly Polarised Light
LogReg	Logistic Regression
MCT	MercuryCadmiumTelluride
MDTM	Magnetic Dipole Transition Moment
ML	Machine Learning
MSE	Mean Squared Error
NB	Naive Bayes
NMR	Nuclear Magnetic Resonance
OD	Optical Density
OR	Optical Rotation
PC	Principal Component

PCA	Principal Component Analysis
PCM	Polarisable Continuum Model
PEM	PhotoElastic Modulator
RCPL	Right handed Circularly Polarised Light
ReLU	Rectified Linear Unit
ROA	Raman Optical Activity
RF	Random Forest
SeLU	Scaled exponential Linear Unit
SI	Sampling Interval
SVM	Support Vector Machine
t-SNE	t-Stochastic Neighbour Embedding
XRD	X-Ray Diffraction
UV-VIS	UltraViolet-VISible
VCD	Vibrational Circular Dichroism

Chapter 1

General introduction

1.1 Chirality

Chirality is the geometric property of a rigid object describing that the object cannot be superimposed on its mirror image.[1] In essence, the mirror images of a chiral object possess some dissimilar characteristics and - if they do not freely interconvert - are regarded as distinct objects. This concept of chirality was introduced by Lord Kelvin in the late 19th century using the following definition[2]:

> I call any geometrical figure, or group of points, chiral, and say that it has chirality if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself.

The word chiral is derived from the Greek word $\chi\epsilon\iota\rho$ or hand, indicating that the human hand itself is chiral.[3] Indeed, a right hand can only be superimposed on another right hand. While left and right hands are often seen as identical, they interact differently with other chiral objects: Two right hands can properly shake hands, whereas a left and right hand cannot. Likewise, a glove is a perfect fit to either a left or right hand, but not to both. So, mirror images of a chiral object can be distinguished within a chiral environment e.g. when interacting with another chiral object like a glove. Outside of a chiral environment, the properties of these mirror images objects are practically identical (without considering the chemically negligible effects of parity violation). Regardless of handedness, both hands can equally well hold a glass of water, throw a ball and carry a grocery bag.

Chirality also plays a role in much smaller objects like molecules. Louis Pasteur observed this in tartaric acid in the 19th century while continuing the research of Arago and Biot on optical rotation.[4] Crystals of sodium ammonium tartrate were found to exist in two mirror-image forms, which after dissolving in water rotated polarised light to equal but opposite extent. Later, Pasteur discovered that the source of this rotation lies within the chiral arrangement of the atoms in tartaric acid.

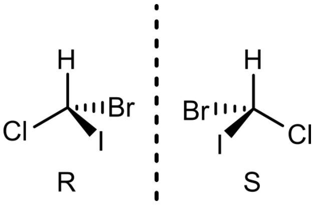


Figure 1.1: Enantiomers of CHClBrI, a molecule with a single chiral center. Wedged and hashed bonds indicate that the atom is oriented towards and away from the reader respectively.

In small molecules, chirality is often the result of a stereogenic center such as a tetrahedral carbon with four different substituents attached, as shown in Figure 1.1. These mirror image structures are referred to as enantiomers and different naming conventions have been established to distinguish them. The IUPAC convention recommends the use of the R/S description, which is most universal in use.[1] The chiral center is given the R or S label according the handedness, dictated by the Cahn-Ingold-Prelog priority rules[5]. The spatial arrangement of these substituents and the appointed label is typically referred to as the Absolute Configuration (AC). For amino acids and sugars, the D/L convention remains a commonly used alternative based on the orientation of substituents within the Fischer projection. Interestingly, most naturally occurring sugars occur as the D-enantiomer and amino acids are typically found as the L-enantiomer.[6,7] Both the R/S and D/L naming conventions label the molecular chirality using the geometric arrangement of the substituents. The +/- convention instead relies directly on an experimental property dependent on the molecular chirality. Enantiomers typically rotate the plane of incoming polarised light in an equal but opposite manner.[8] So, enantiomers can be distinguished according to the sense of this rotation, denoting an enantiomer as either (+) or (-). This +/- naming convention does not provide any direct insight on the spatial arrangement of the nuclei, it merely provides a way to distinguish an enantiomer rotating light clockwise and one rotating it anti-clockwise. Therefore, the R/S convention is typically preferred if the link between the optical rotation and its AC is unknown.

When multiple stereogenic centers are present in a molecule, enantiomers are obtained upon inverting the configuration of each chiral element. If between a set of stereoisomers not all elements are inverted, they are called diastereomers instead. Diastereomers that differ in the configuration of a single stereogenic center are called epimers. The differences between the types of stereoisomers are illustrated in Figure 1.2 for menthol. As enantiomers are perfect mirror images, they only obtain divergent properties within a chiral environment. Diastereomers and epimers are not perfect mirror images and, as a result, have different properties -even- in an achiral environment. For this reason, they can be distinguished with common techniques like melting point determination, thin-layer chromatography and nuclear magnetic resonance spectroscopy.[9,10]

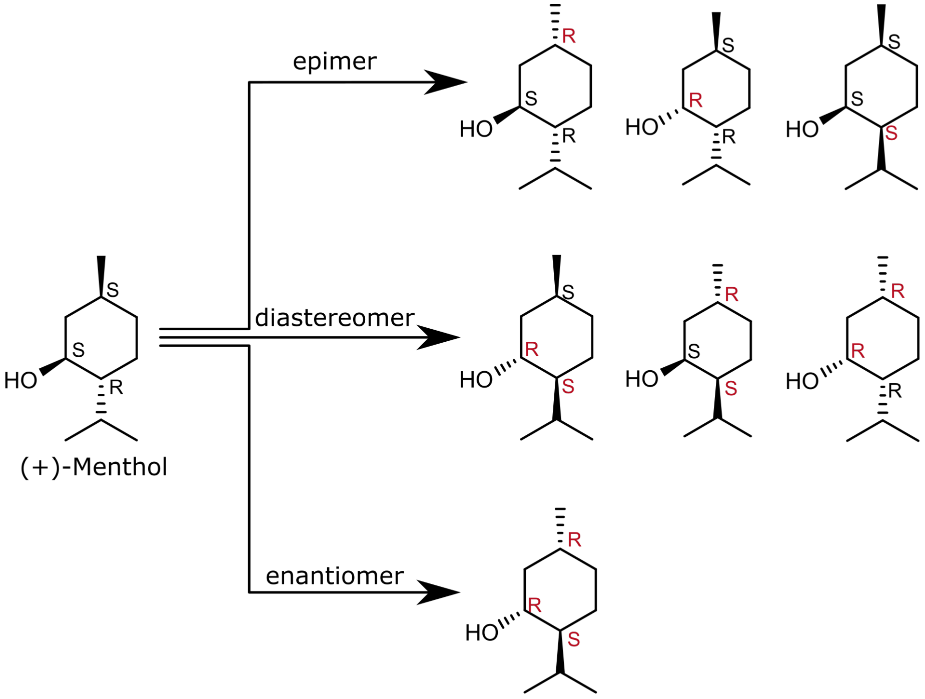


Figure 1.2: Overview of the stereoisomers (epimers, diastereomers and enantiomer) of (+)-menthol. The R/S labels of inverted chiral centres are highlighted in red.

It is important to note that while many chiral compounds contain one or more asymmetric atom(s), it is neither a required nor a sufficient criterion for chirality. A well-known counterexample is the meso (*R*,*S*) stereoisomer of tartaric acid, which contains two asymmetric carbon atoms (Figure 1.3). The (*R*,*R*) stereoisomer is chiral as it cannot be superimposed onto the (*S*,*S*) stereoisomer. However, the (*R*,*S*) stereoisomer is achiral due to the mirror plane found in its structure. If a compound contains an inversion center or mirror plane, the mirror image structures of the compound are superimposable. Therefore, the (*R*,*S*) and (*S*,*R*) stereoisomers of tartaric acid are indistinguishable.

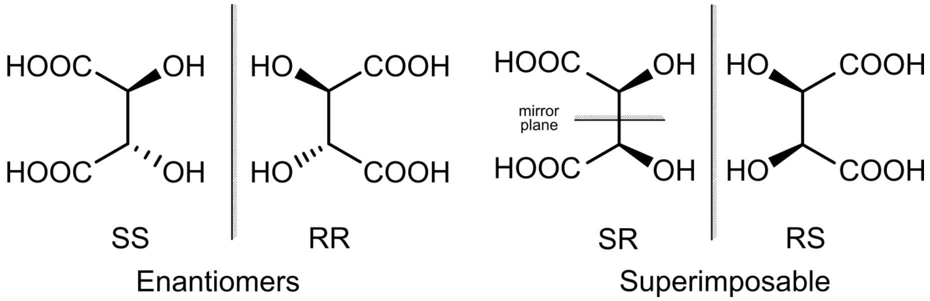


Figure 1.3: Stereoisomers of tartaric acid.

Aside from a stereogenic center, chirality of a compound can also arise from a stereogenic axis, a chiral plane or a helical structure. Examples of compounds containing such chiral elements are shown in Figure 1.4.[5]

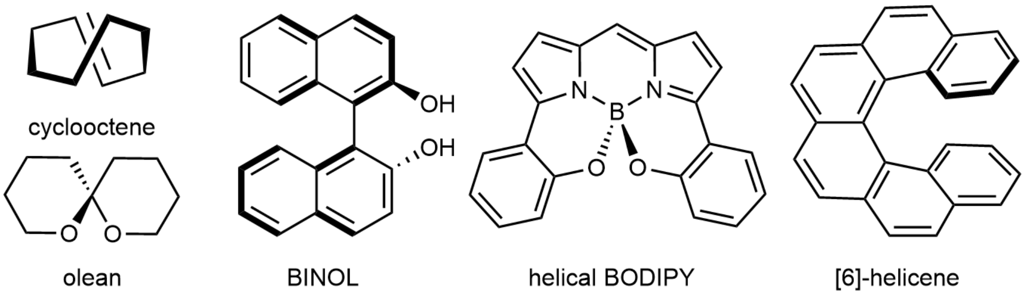


Figure 1.4: Chiral compounds containing a chiral axis (olean & BINOL), chiral plane (cyclooctene) and helical structure ([6]-helicene & a boron-chelated BODIPY[11]).

The enantiospecific interaction between chiral compounds has important implications, even for life itself. Life is built from homochiral (the exclusive presence of a single enantiomer) building blocks such as amino acids and carbohydrates, creating a chiral environment.[6,7] The link between life and chiral molecules is so prevalent that homochiral compounds have been used as a biosignature for ancient and extraterrestrial life.[12,13] So, it is not surprising that enantiomers can provoke different responses within our own body, influencing how we perceive their smell and taste. Depending on the absolute configuration, limonene smells either like lemons ((*R*)-enantiomer) or oranges ((*S*)-enantiomer) and carvone like spearmint ((*R*)-enantiomer) or caraway ((*S*)-enantiomer).[14–16] The perceived taste of amino acids, in turn, is influenced by the chirality of the α carbon.[17,18] The origin for the chiral recognition is often explained on the basis of a three-point interaction between the chiral compound and a chiral receptor as illustrated in Figure 1.5.[19]

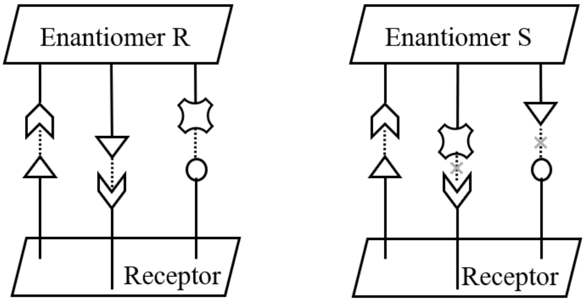


Figure 1.5: Different interactions between receptor and mirror image compounds.

Enantiomers of chiral drugs often provoke different pharmacological activity.[20–22] The (*S*)-enantiomer of atenolol, a well-known β-blocker, is 100 times more potent than the (*R*)-enantiomer.[23] Similarly, the (*S*)-enantiomer of ibuprofen is 100 times more potent than its mirror image and the enantiomers have different metabolic profiles.[20,24,25] Differences in toxicity are also found in enantiomers as illustrated by the infamous thalidomide crisis. Thalidomide was commercially distributed as a racemate i.e. a mixture of both enantiomers.[26] Whereas the (*R*)-enantiomer of thalidomide provided the desired activity, the (*S*)-enantiomer was in fact teratogenic. Thalidomide had been administered to many pregnant women in the late 1950s to early 1960s to treat morning sickness, resulting in many stillborn and physically disabled infants. This tragedy signified the importance of studying the properties of drug enantiomers and developing methods that could distinguish them. In the aftermath of the thalidomide crisis, the Food and Drug Administration introduced more strict requirements in their drug evaluations.[27] Eventually in the early 1990s, they included a provision that both enantiomers need toxicological screening if a drug is administered as a racemate and the absolute configuration of enantiopure ingredients has to be identified.[27,28] Two years later the European Medicines Agency introduced a similar policy.[29]

Interest in using enantiopure drug compounds has increased in the past two decades, with racemic drugs being replaced by an enantiopure alternative (a so-called ‘chiral switch’).[20,30–32] Nevertheless, some drug compounds -such as Ibuprofen- are still manufactured as racemates because of the cost of extra chiral separation and asymmetric synthesis. Moreover, (R)-Ibuprofen is partially enzymatically converted within the body to (S)-Ibuprofen whereas (S)-Ibuprofen does not undergo chiral inversion, limiting the added value of supplying the enantiopure drug.[22,25] New agrochemicals are more and more often chiral and enantiomers often exhibit different properties.[33,34] While most chiral agrochemicals are still manufactured as racemates to keep manufacturing costs low, enantioselective analytical methods remain in demand for registration and monitoring purpose.

1.2 Determination of molecular chirality

Distinguishing enantiomers is not a straightforward task as most of their chemical and physical properties are identical. Many analytical structure determination methods cannot distinguish enantiomers as they rely on properties that are independent of AC. One way to circumvent this problem is to convert a set of enantiomers to a set of diastereomers instead, which behave differently in an achiral environment. Doing so, conventional analytical methods can be used without directly modifying the instruments involved. Other techniques invoke chiral sensitivity using an internal chiral reference or by polarising radiation such that it obtains a specific handedness. This section provides an overview of the most commonly used chiral determination methods.

X-Ray Diffraction (XRD) has seen wide-spread use in determining molecular chirality, either in an indirect or direct manner. For the indirect method, other enantiopure compounds are incorporated in the crystal via co-crystallisation or reaction with the analyte (forming diastereomers).[20,35,36] The resulting crystals will then scatter the incoming X-rays differently based on the chirality of the original analyte. The direct method instead relies on the phenomenon of anomalous dispersion. Here, the resonant scattering of X-rays introduces a phase shift, resulting in small deviations for the scattering patterns of mirror image crystals that can be linked back to the molecular chirality.[37,38] The resonant scattering relies on selective excitation of core electrons in heavy atoms, making the direct approach less accessible to organic compounds containing only first and second row atoms. XRD applications also require a sufficiently large single crystal of the analyte. Therefore, alternative tools need to be used when the analyte is obtained as an oil/liquid or contains impurities preventing crystal growth within a reasonable time frame.

Nuclear Magnetic Resonance (NMR) is arguably the preferred spectroscopic method of many organic chemists to study molecular structures. NMR is transparent to molecular chirality and therefore can only distinguish diastereomers. These diastereomers are typically formed upon reaction of the chiral analyte with a chiral reagent of known chirality. A well-known chiral reagent is the Mosher's acid used for secondary alcohols.[39] The process of converting the chiral analyte is labour-intensive and the success of the method depends heavily on the availability of a suitable chiral reagent. Alternatively, chiral solvents and chiral additives are used to create a chiral environment. This approach introduces peak shifts in the NMR spectrum whose sign depends on the chirality of the analyte.

Chiroptical techniques do not require chiral molecules as internal references or derivation of the analyte. These techniques instead rely on the interaction of chiral compounds with circularly polarised light (CPL). CPL is inherently chiral, coming in either a left handed (LCPL) or right handed (RCPL) form. Enantiomers interact differently with these chiral forms, allowing to distinguish enantiomers. An example of a chiroptical technique is optical rotation, which is known as the first analytical tool capable of detecting molecular chirality. As discussed in section 1.1, enantiomers rotate the polarisation plane of linearly polarised light in opposite directions. The origin of the rotation lies in a slight variation in the refractive index for the LCPL and RCPL components of the light beam, resulting in circular birefringence. The chiral compound present in the analysed fluid is then labelled according to the sign of the optical rotation. When the optical rotation is recorded for different wavelengths, this technique is referred to as Optical Rotary Dispersion (ORD). The experimental setup is fairly simple, making it a cheap tool to monitor the chirality and enantiomeric purity of the compound. The method sees less use nowadays in establishing the chirality of newly discovered or synthesised compounds. Optical rotation experiments do not reveal direct information on the spatial arrangement of the nuclei, merely on the rotation sense. A common practice is to compare the rotation to that of other compounds containing similar structural elements, and when these values match sufficiently, equate their AC. This practice makes it prone to error accumulation and less reliable for AC determination. Though, once the link between the optical rotation and the molecular geometry has been established for the compound, optical rotation provides a fast way to check the chirality and enantiomeric purity for other samples. Recent advances show that this link can be established with quantum chemical computations, improving the reliability of the assignments.[40] Nonetheless, ORD experiments only provide a limited amount of information to establish the molecular chirality and more informative chiroptical techniques are typically favoured over it nowadays.

Another chiroptical technique is circular dichroism, where the different absorption of LCPL and RCPL is recorded. Chiral compounds capable of absorbing light at a specific wavelength, demonstrate a subtle preference to absorb either LCPL or RCPL. In the case of Electronic Circular Dichroism (ECD), this preference is measured for wavelengths in the UV-VIS region. With the requirement of absorption in the UV-VIS region, chiral compounds lacking chromophores cannot be studied with ECD. Introducing chromophores into such compounds requires additional synthetic steps, making the procedure very labour-intensive. A solution to this problem is to record the CD phenomenon in the infrared (IR) region instead. This chiroptical technique is known as Vibrational Circular Dichroism (VCD) and a general description of its applications is given in section 1.3. Further details on the method are provided in Chapter 2. A chiroptical technique mentioned alongside VCD is Raman Optical Activity (ROA). ROA is the chiroptical version of Raman spectroscopy and has been used to determine the molecular chirality of pharmaceuticals alongside VCD.[41–45]

A final technique worth mentioning is rotational spectroscopy (or molecular rotational resonance). Recent efforts have posed rotational spectroscopy as a new tool for chiral analysis, along with commercialisation of their experimental setup. The method relies on the complexation of the chiral analyte with another enantiopure compound (e.g. propylene oxide or 3-butyn-2-ol), transforming the enantiomeric forms of the analyte into diastereomeric complexes.[46–49]. As these diastereomeric complexes have very distinct rotational constants, the chirality of the analyte can be retrieved from the rotational spectrum.

1.3 Vibrational circular dichroism

VCD is a chiroptical technique which measures the preference to absorb LCPL or RCPL in the IR region. As the name implies, the absorbance is the result of vibrational transitions, circumpassing the need for UV-VIS chromophores. VCD is a weak phenomenon that contains a wealth of chiral information due to the large number of accessible vibrational transitions. As with any circular dichroism technique, the spectra of enantiomers are perfect mirror images. Therefore, enantiomers can be distinguished fairly easily with VCD, as illustrated in Figure 1.6. However, the information present in the spectrum cannot be directly linked back to the AC. Some empirical rules linking spectrum and the chirality have been established, but they remain limited to small groups of compounds and thus are not generally applicable.[50–55] Instead, the link is established with the aid of quantum chemical predictions. For fairly rigid compounds of moderate size, this approach is fairly straightforward. When highly flexible compounds are involved, the approach becomes more cumbersome and requires substantial computational resources.[56–59]

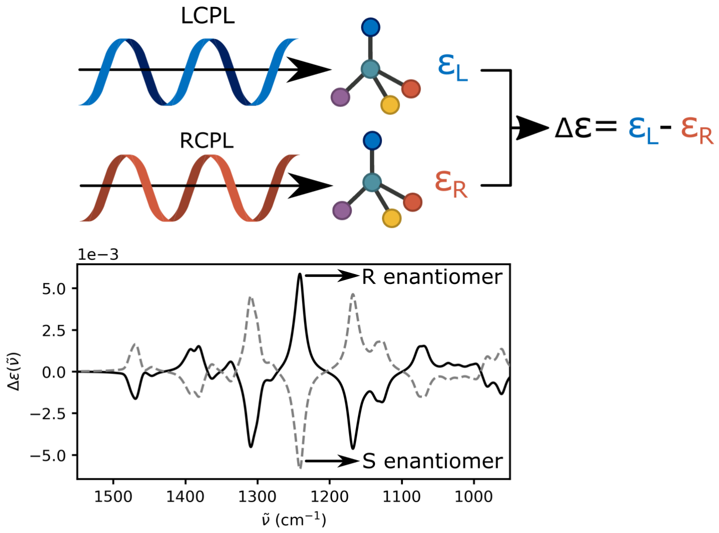


Figure 1.6: Illustration of the VCD phenomenon and the mirror image spectra of enantiomers.

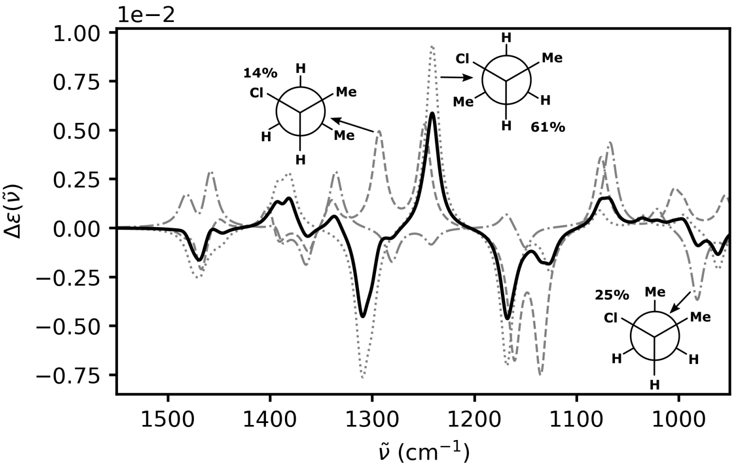


Figure 1.7: VCD spectrum of (R)-2-chlorobutane (black) and the contributions of its three conformers to the molecular spectrum (grey).[a]

Flexible compounds continuously rotate internally from one 3D orientation to another without breaking any bonds. These different orientations are called conformers and each of them contributes to the overall VCD spectrum of the compound. As illustrated in Figure 1.7, the contributions of these conformers differ substantially, leading to the high sensitivity of the molecular spectrum. To arrive at an accurate computed spectrum, one has to properly account for these individual contributions. A flexible compound favours adopting conformers of lower energy and their relative populations are shaped by the environment (e.g. solvent). Therefore, most VCD applications require performing the following steps:

- Measurement of experimental VCD spectrum.
- Identify all conformers the compound can adopt.
- Determine how abundant each conformer is.
- Compute the VCD spectrum for each conformer.
- Combine the individual contributions of each conformer into a molecular spectrum.

[a] technical details: conformer spectra calculated at B3LYP/aug-cc-pvdz with the CCl_4 solvent included as a polarisable continuum. Molecular spectrum obtained as Boltzmann average weighted according to the $\Delta H^0_{298.15}$ values of each conformer.

- Compare the obtained spectrum with the experimental one.

For flexible compounds, identifying all relevant conformers can prove challenging and requires expertise. Furthermore, the need to compute the VCD spectrum of each conformer increases the computational cost significantly. Nonetheless, this approach has been successfully used for both rigid and flexible compounds.[45,58,60–62]

VCD has become a well-established method to identify the chirality of organic compounds like pharmaceuticals.[63–100] VCD finds common use in pharmaceutical companies and institutions such as the U. S. Pharmacopeia, U.S. Food and Drug Administration and European Medicines Agency, that recognise it as a standard method for distinguishing enantiomers. Aside from AC determination, VCD is also employed to study the structure of flexible chiral compounds. The high sensitivity of VCD allows to study the conformational population of compounds in solution. As a result, the conformational properties of many pharmaceuticals have been studied with VCD, often in tandem with the chiral analysis.[84–109] The high sensitivity of VCD to the 3D structure is also leveraged to study the structural characteristics of substantially larger systems, including polypeptides/proteins[110–114], polynucleotides[111,115–119], polymers[120–123], crystals[119,124–130], ionic liquids[131,132] and gels[133].

1.4 Scope

VCD spectra contain a lot of information on the chirality of organic compounds, making it a reliable technique in the chiral analysis toolkit. The link between the patterns within these spectra and the handedness of the compound remains rather opaque. Traditionally, DFT calculations have been employed to establish this link. While this methodology has unlocked the potential of VCD for many applications, these calculations are expensive and require expertise. This work explores the possibility of establishing this link with Machine Learning (ML) instead and creating new VCD applications powered by ML. ML has been successfully applied in many chemical applications to link experimentally measurable properties to molecular structures or speed up expensive computer models. This has resulted in an increasing number of spectroscopic applications powered by ML based analysis and allows to predict spectral patterns for much larger systems than previously thought possible. At the beginning of my research, supervised ML methods had not yet been applied to VCD spectroscopy and it was unknown whether they were applicable to VCD workflows. Therefore, many approaches to include ML within VCD workflows were tested in this thesis and the results for these approaches could range from very promising to failing immediately. In the end we chose to initiate three different projects, each covering a different approach. Each project was designed to yield deeper insight into VCD or the future of ML within the field if the obtained results were unilaterally negative. The first project focuses on the main application of VCD, being AC determination. We generated a dataset of substituted α-pinene spectra and tasked different ML methods with extracting the chirality from these spectra. Here, we were especially interested in the balance between how accurate and interpretable each ML method was. The second project covers the link between conformers and their contributions to the molecular spectrum. Deep neural networks were tasked with predicting conformer spectra from the geometry of these conformers for 6 different compounds. These compounds were chosen such that differences in the performance of the model could be traced back to chemical influences such as intramolecular interactions or functional groups. The results were then interpreted within the scope of VCD applications e.g. the speed-up for the computational VCD workflow and the transferability of the model to other stereoisomers. Finally, for the third project we pushed the boundaries of VCD applications by performing ML-aided terpene mixture analysis. The project was instigated by prof. João Batista Junior from Federal University of Saõ Paulo, who was interested in detecting individual terpenes within terpene mixtures using the added discriminatory power of VCD. A significant number of VCD spectra of pure terpenes, terpene mixtures and natural oils were recorded and a visual analysis of the mixtures was performed by the co-authors. Using this dataset, an interpretable ML model was obtained and tested on different terpene mixtures and natural oils.

To contextualise the results obtained in these projects, an overview on VCD and ML is provided in Chapters 2 and 3. Chapter 2 covers the main concepts behind VCD, its theoretical foundation and the typical workflow employed for AC determination. In Chapter 3 the fundamentals of ML and the models used throughout Chapters 4-6 are explained. The chapter offers the necessary insight into different ML applications and provides an understanding of the strengths or weaknesses inherent to different ML models. Next, the design and results for the three projects are presented in Chapters 4-6. Finally, Chapter 7 provides concluding remarks on the results of this work.

References

[1] H. A. Favre and W. H. Powell, *Nomenclature of Organic Chemistry*, The Royal Society of Chemistry, 2013, ch. 9, pp. 1156 – 1292.

[2] W. T. B. Kelvin, *Baltimore Lectures on Molecular Dynamics and the Wave Theory of Light*, C.J. Clay and Sons, Cambridge, United Kingdom, 1904.

[3] Oxford English Dictionary: 'chiral, adj.', `https://www.oed.com/view/Entry/31848?redirectedFrom=chiral`, Accessed: 2023-06-28.

[4] J. Gal, *Nat. Chem.*, 2017, **9**, 604–605.

[5] R. Cahn, C. Ingold and V. Prelog, *Angew. Chem. Int. Ed.*, 1966, **5**, 385–415.

[6] R. Breslow, *Tetrahedron Lett.*, 2011, **52**, 4228–4232.

[7] J. E. Hein and D. G. Blackmond, *Acc. Chem. Res.*, 2012, **45**, 2045–2054.

[8] P. L. Polavarapu, *Chirality*, 2002, **14**, 768–781.

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

[10] A. C. Allen, D. A. Cooper, W. O. Kiser and R. C. Cottreli, *J. Forensic Sci.*, 1981, **26**, 11325J.

[11] R. Clarke, K. L. Ho, A. A. Alsimaree, O. J. Woodford, P. G. Waddell, J. Bogaerts, W. Herrebout, J. G. Knight, R. Pal, T. J. Penfold and M. J. Hall, *ChemPhotoChem*, 2017, **1**, 513–517.

[12] D. P. Glavin, A. S. Burton, J. E. Elsila, J. C. Aponte and J. P. Dworkin, *Chem. Rev.*, 2020, **120**, 4660–4689.

[13] D. Avnir, *New Astron. Rev.*, 2021, **92**, 101596.

[14] G. F. Russell and J. I. Hills, *Science*, 1971, **172**, 1043–1044.

[15] J. C. Brookes, A. P. Horsfield and A. M. Stoneham, *J. R. Soc. Interface*, 2009, **6**, 75–86.

[16] J. Gal, *Chirality*, 2012, **24**, 959–976.

[17] M. Kawai, Y. Sekine-Hayakawa, A. Okiyama and Y. Ninomiya, *Amino Acids*, 2012, **43**, 2349–2358.

[18] S. S. Schiffman, T. B. Clark and J. Gagnon, *Physiol. Behav.*, 1982, **28**, 457–465.

[19] A. Berthod, *Anal. Chem.*, 2006, **78**, 2093–2099.

[20] A. Calcaterra and I. D'Acquarica, *J. Pharm. Biomed. Anal.*, 2018, **147**, 323–340.

[21] E. Sanganyado, Z. Lu, Q. Fu, D. Schlenk and J. Gan, *Water Res.*, 2017, **124**, 527–542.

[22] L. A. Nguyen, H. He and C. Pham-Huy, *Int. J. Biomed. Sci.*, 2006, **2**, 85–100.

[23] A. M. Barrett and V. A. Cullum, *Br. J. Pharmacol.*, 1968, **34**, 43–55.

[24] A. M. Evans, *Clin. Rheumatol.*, 2001, **20**, 9–14.

[25] A. Gliszczyńska and E. Sánchez-López, *Pharmaceutics*, 2021, **13**, 414.

[26] W. Rehman, L. M. Arfons and H. M. Lazarus, *Ther. Adv. Hematol.*, 2011, **2**, 291–308.

[27] H. Brooks, C. Guida and G. Daniel, *Curr. Top. Med. Chem.*, 2011, **11**, 760–770.

[28] Administration F.a.D., *Development of new stereoisomeric drugs*, `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/development-new-stereoisomeric-drugs`, Accessed: 2023-06-28.

[29] European Medicines Agency, *Investigation Of Chiral Active Substances*, `https://www.ema.europa.eu/en/documents/scientific-guideline/investigation-chiral-active-substances_en.pdf`, Accessed: 2023-08-28.

[30] G. Hancu and A. Modroiu, *Pharmaceuticals*, 2022, **15**, 240.

[31] I. D'Acquarica and I. Agranat, *ACS Pharmacol. Transl. Sci.*, 2023, **6**, 201–219.

[32] I. Agranat, H. Caner and J. Caldwell, *Nat. Rev. Drug Discov.*, 2002, **1**, 753–768.

[33] P. Jeschke, *Pest Manag. Sci.*, 2018, **74**, 2389–2404.

[34] E. M. Ulrich, C. N. Morrison, M. R. Goldsmith and W. T. Foreman, in *Chiral Pesticides: Identification, Description, and Environmental Implications*, ed. D. M. Whitacre, Springer US, Boston, MA, 2012, pp. 1–74.

[35] H. D. Flack and G. Bernardinelli, *Chirality*, 2008, **20**, 681–690.

[36] A. L. Thompson and D. J. Watkin, *Tetrahedron: Asymmetry*, 2009, **20**, 712–717.

[37] J. M. Bijvoet, A. F. Peerdeman and A. J. van Bommel, *Nature*, 1951, **168**, 271–272.

[38] S. Parsons, *Tetrahedron: Asymmetry*, 2017, **28**, 1304–1313.

[39] J. A. Dale and H. S. Mosher, *J. Am. Chem. Soc.*, 1973, **95**, 512–519.

[40] F. Bohle, J. Seibert and S. Grimme, *J. Org. Chem.*, 2021, **86**, 15522–15531.

[41] J. Bogaerts, F. Desmet, R. Aerts, P. Bultinck, W. Herrebout and C. Johannessen, *Phys. Chem. Chem. Phys.*, 2020, **22**, 18014–18024.

[42] E. De Gussem, K. A. Tehrani, P. Herrebout and C. Johannessen, *ACS Omega*, 2019, **4**,

14133–14139.

[43] G.-S. Yu, D. Che, T. B. Freedman and L. A. Nafie, *Tetrahedron Asymmetry*, 1993, **4**, 511–516.

[44] V. Profant, A. Jegorov, P. Bouř and V. Baumruk, *J. Phys. Chem. B*, 2017, **121**, 1544–1551.

[45] Y. He, B. Wang, R. K. Dukor and L. A. Nafie, *Appl. Spectrosc.*, 2011, **65**, 699–723.

[46] B. H. Pate, L. Evangelisti, W. Caminati, Y. Xu, J. Thomas, D. Patterson, C. Perez and M. Schnell, *Chiral Analysis (Second Edition)*, Elsevier, Second Edition edn, 2018, ch. 17, pp. 679–729.

[47] R. E. Sonstrom, J. L. Neill, A. V. Mikhonin, R. Doetzer and B. H. Pate, *Chirality*, 2022, **34**, 114–125.

[48] S. R. Domingos, C. Pérez, M. D. Marshall, H. O. Leung and M. Schnell, *Chem. Sci.*, 2020, **11**, 10863–10870.

[49] J. L. Neill, L. Evangelisti and B. H. Pate, *Anal. Sci. Adv.*, 2023, **n/a**, 1–16.

[50] T. Taniguchi and K. Monde, *J. Am. Chem. Soc.*, 2012, **134**, 3695–3698.

[51] H. Izumi, S. Yamagami, S. Futamura, L. A. Nafie and R. K. Dukor, *J. Am. Chem. Soc.*, 2004, **126**, 194–198.

[52] F. J. Devlin, P. J. Stephens and P. Besse, *J. Org. Chem.*, 2005, **70**, 2980–2993.

[53] F. Passareli, A. N. L. Batista, A. J. Cavalheiro, W. A. Herrebout and J. M. Batista Junior, *Phys. Chem. Chem. Phys.*, 2016, **18**, 30903–30906.

[54] M. Z. M. Zubir, N. F. Maulida, Y. Abe, Y. Nakamura, M. Abdelrasoul, T. Taniguchi and K. Monde, *Org. Biomol. Chem.*, 2022, **20**, 1067–1072.

[55] F. M. dos Santos Jr., K. U. Bicalho, I. H. Calisto, G. S. Scatena, J. B. Fernandes, Q. B. Cass and J. M. Batista Jr., *Org. Biomol. Chem.*, 2018, **16**, 4509–4516.

[56] M. A. J. Koenis, Y. Xia, S. R. Domingos, L. Visscher, W. J. Buma and V. P. Nicu, *Chem. Sci.*, 2019, **10**, 7680–7689.

[57] L. Böselt, D. Sidler, T. Kittelmann, J. Stohner, D. Zindel, T. Wagner and S. Riniker, *J. Chem. Inf. Model.*, 2019, **59**, 1826–1838.

[58] J. Bogaerts, R. Aerts, T. Vermeyen, C. Johannessen, W. Herrebout and J. M. Batista, *Pharmaceuticals*, 2021, **14**, 877.

[59] K. D. R. Eikås, M. T. P. Beerepoot and K. Ruud, *J. Phys. Chem. A*, 2022, **126**, 5458–5471.

[60] C. Merten, T. P. Golub and N. M. Kreienborg, *J. Org. Chem.*, 2019, **84**, 8797–8814.

[61] P. L. Polavarapu and E. Santoro, *Nat. Prod. Rep.*, 2020, **37**, 1661–1699.

[62] C. Merten, *Phys. Chem. Chem. Phys.*, 2017, **19**, 18803–18812.

[63] Y. Zhang, M. R. Poopari, X. Cai, A. Savin, Z. Dezhahang, J. Cheramy and Y. Xu, *J. Nat. Prod.*, 2016, **79**, 1012–1023.

[64] E. C. Sherer, C. H. Lee, J. Shpungin, J. F. Cuff, C. Da, R. Ball, R. Bach, A. Crespo, X. Gong and C. J. Welch, *J. Med. Chem.*, 2014, **57**, 477–494.

[65] S. Qiu, E. De Gussem, K. Abbaspour Tehrani, S. Sergeyev, P. Bultinck and W. Herrebout, *J. Med. Chem.*, 2013, **56**, 8903–8914.

[66] S. S. Wesolowski and D. E. Pivonka, *Bioorganic Med. Chem. Lett.*, 2013, **23**, 4019–4025.

[67] P. J. Stephens, J. J. Pan, F. J. Devlin, K. Krohn and T. Kurtán, *J. Org. Chem.*, 2007, **72**, 3521–3536.

[68] E. Vass, M. Hollósi, E. Forró and F. Fülöp, *Chirality*, 2006, **18**, 733–740.

[69] E. Carosati, R. Budriesi, P. Ioan, G. Cruciani, F. Fusi, M. Frosini, S. Saponara, F. Gasparrini, A. Ciogli, C. Villani, P. J. Stephens, F. J. Devlin, D. Spinelli and A. Chiarini, *J. Med. Chem.*, 2009, **52**, 6637–6648.

[70] B. Nieto-Ortega, J. Casado, E. W. Blanch, J. T. López Navarrete, A. R. Quesada and F. J. Ramírez, *J. Phys. Chem. A*, 2011, **115**, 2752–2755.

[71] N. Jiang, R. X. Tan and J. Ma, *J. Phys. Chem. B*, 2011, **115**, 2801–2813.

[72] Z. Ma, D. C.-H. Lin, R. Sharma, J. Liu, L. Zhu, A.-R. Li, T. Kohn, Y. Wang, J. J. Liu, M. D. Bartberger, J. C. Medina, R. Zhuang, F. Li, J. Zhang, J. Luo, S. Wong, G. R. Tonn and J. B. Houze, *Bioorganic Med. Chem. Lett.*, 2016, **26**, 15–20.

[73] J. J. Chen, W. Qian, K. Biswas, C. Yuan, A. Amegadzie, Q. Liu, T. Nixey, J. Zhu, M. Ncube, R. M. Rzasa, F. Chavez, N. Chen, F. DeMorin, S. Rumfelt, C. M. Tegley, J. R. Allen, S. Hitchcock, R. Hungate, M. D. Bartberger, L. Zalameda, Y. Liu, J. D. McCarter, J. Zhang, L. Zhu, S. Babu-Khan, Y. Luo, J. Bradley, P. H. Wen, D. L. Reid, F. Koegler, C. Dean, D. Hickman, T. L. Correll, T. Williamson and S. Wood, *Bioorganic Med. Chem. Lett.*, 2013, **23**, 6447–6454.

[74] F. Gonzalez-Lopez de Turiso, D. Sun, Y. Rew, M. D. Bartberger, H. P. Beck, J. Canon, A. Chen, D. Chow, T. L. Correll, X. Huang, L. D. Julian, F. Kayser, M.-C. Lo, A. M.

Long, D. McMinn, J. D. Oliner, T. Osgood, J. P. Powers, A. Y. Saiki, S. Schneider, P. Shaffer, S.-H. Xiao, P. Yakowec, X. Yan, Q. Ye, D. Yu, X. Zhao, J. Zhou, J. C. Medina and S. H. Olson, *J. Med. Chem.*, 2013, **56**, 4053–4070.

[75] N. Nishimura, M. H. Norman, L. Liu, K. C. Yang, K. S. Ashton, M. D. Bartberger, S. Chmait, J. Chen, R. Cupples, C. Fotsch, J. Helmering, S. R. Jordan, R. K. Kunz, L. D. Pennington, S. F. Poon, A. Siegmund, G. Sivits, D. J. Lloyd, C. Hale and D. J. J. St. Jean, *J. Med. Chem.*, 2014, **57**, 3094–3116.

[76] J. G. Allen, M. P. Bourbeau, G. E. Wohlhieter, M. D. Bartberger, K. Michelsen, R. Hungate, R. C. Gadwood, R. D. Gaston, B. Evans, L. W. Mann, M. E. Matison, S. Schneider, X. Huang, D. Yu, P. S. Andrews, A. Reichelt, A. M. Long, P. Yakowec, E. Y. Yang, T. A. Lee and J. D. Oliner, *J. Med. Chem.*, 2009, **52**, 7044–7053.

[77] D. J. Minick, R. C. Copley, J. R. Szewczyk, R. D. Rutkowske and L. A. Miller, *Chirality*, 2007, **19**, 731–740.

[78] O. McConnell, A. Bach II, C. Balibar, N. Byrne, Y. Cai, G. Carter, M. Chlenov, L. Di, K. Fan, I. Goljer, Y. He, D. Herold, M. Kagan, E. Kerns, F. Koehn, C. Kraml, V. Marathias, B. Marquez, L. McDonald, L. Nogle, C. Petucci, G. Schlingmann, G. Tawa, M. Tischler, R. T. Williamson, A. Sutherland, W. Watts, M. Young, M.-Y. Zhang, Y. Zhang, D. Zhou and D. Ho, *Chirality*, 2007, **19**, 658–682.

[79] E. Santoro, G. Mazzeo, A. G. Petrovic, A. Cimmino, J. Koshoubu, A. Evidente, N. Berova and S. Superchi, *Phytochemistry*, 2015, **116**, 359–366.

[80] O. R. Thiel, M. Achmatowicz, C. Bernard, P. Wheeler, C. Savarin, T. L. Correll, A. Kasparian, A. Allgeier, M. D. Bartberger, H. Tan and R. D. Larsen, *Org. Process Res. Dev.*, 2009, **13**, 230–241.

[81] N. A. Tamayo, Y. Bo, V. Gore, V. Ma, N. Nishimura, P. Tang, H. Deng, L. Klionsky, S. G. Lehto, W. Wang, B. Youngblood, J. Chen, T. L. Correll, M. D. Bartberger, N. R. Gavva and M. H. Norman, *J. Med. Chem.*, 2012, **55**, 1593–1611.

[82] T. C. Lourenço, J. ao M. Batista Jr, M. Furlan, Y. He, L. A. Nafie, C. C. Santana and Q. B. Cass, *J. Chromatogr. A*, 2012, **1230**, 61–65.

[83] J. Kong, L. A. Joyce, J. Liu, T. M. Jarrell, J. C. Culberson and E. C. Sherer, *Chirality*, 2017, **29**, 854–864.

[84] S. Abbate, G. Longhi, F. Lebon and M. Tommasini, *Chem. Phys.*, 2012, **405**, 197–205.

[85] M. Górecki, *Org. Biomol. Chem.*, 2015, **13**, 2999–3010.

[86] D. E. Pivonka and S. S. Wesolowski, *Appl. Spectrosc.*, 2013, **67**, 365–370.

[87] D. Dunmire, T. B. Freedman, L. A. Nafie, C. Aeschlimann, J. G. Gerber and J. Gal, *Chirality*, 2005, **17**, S101–S108.

[88] P. J. Stephens, F. J. Devlin, F. Gasparrini, A. Ciogli, D. Spinelli and B. Cosimelli, *J. Org. Chem.*, 2007, **72**, 4707–4715.

[89] J. Bogaerts, F. Desmet, R. Aerts, P. Bultinck, W. Herrebout and C. Johannessen, *Phys. Chem. Chem. Phys.*, 2020, **22**, 18014–18024.

[90] D. Rossi, R. Nasti, S. Collina, G. Mazzeo, S. Ghidinelli, G. Longhi, M. Memo and S. Abbate, *J. Pharm. Biomed. Anal.*, 2017, **144**, 41–51.

[91] E. De Gussem, K. A. Tehrani, W. A. Herrebout, P. Bultinck and C. Johannessen, *ACS Omega*, 2019, **4**, 14133–14139.

[92] P. L. Polavarapu, C. Zhao, A. L. Cholli and G. G. Vernice, *J. Phys. Chem. B*, 1999, **103**, 6127–6132.

[93] C. Zhao, P. Polavarapu, H. Grosenick and V. Schurig, *J. Mol. Struct.*, 2000, **550-551**, 105–115.

[94] P. L. Polavarapu, A. L. Cholli and G. Vernice, *J. Am. Chem. Soc.*, 1992, **114**, 10953–10955.

[95] T. B. Freedman, N. Ragunathan and S. Alexander, *Faraday Discuss.*, 1994, **99**, 131–149.

[96] C. Ashvar, P. Stephens, T. Eggimann and H. Wieser, *Tetrahedron: Asymmetry*, 1998, **9**, 1107–1110.

[97] B. E. Maryanoff, D. F. McComsey, R. K. Dukor, L. A. Nafie, T. B. Freedman, X. Cao and V. W. Day, *Bioorg. Med. Chem.*, 2003, **11**, 2463–2470.

[98] T. B. Freedman, R. K. Dukor, P. J. C. M. van Hoof, E. R. Kellenbach and L. A. Nafie, *Helv. Chim. Acta*, 2002, **85**, 1160–1165.

[99] J. Shen, J. Yang, W. Heyse, H. Schweitzer, N. Nagel, D. Andert, C. Zhu, V. Morrison, G. A. Nemeth, T.-M. Chen, Z. Zhao, T. A. Ayers and Y.-M. Choi, *J. Pharm. Anal.*, 2014, **4**, 197–204.

[100] M. Górecki and J. Frelek, *International Journal of Molecular Sciences*, 2022, **23**, 273.

[101] M. R. Poopari, Z. Dezhahang and Y. Xu, *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, 2015, **136**, 131–140.

[102] P. Fagan, L. Kocourková, M. Tatarkovič, F. Králík, M. Kuchař, V. Setnička and P. Bouř, *ChemPhysChem*, 2017, **18**, 2258–2265.

[103] F. Králík, P. Fagan, M. Kuchař and V. Setnička, *Chirality*, 2020, **32**, 854–865.

[104] A. Sathya, T. Prabhu and S. Ramalingam, *Heliyon*, 2020, **6**, e03433.

[105] B. Nieto-Ortega, J. Casado, E. W. Blanch, J. T. López Navarrete, A. R. Quesada and F. J. Ramírez, *J. Phys. Chem. A*, 2011, **115**, 2752–2755.

[106] H. Izumi, A. Ogata, L. A. Nafie and R. K. Dukor, *J. Org. Chem.*, 2009, **74**, 1231–1236.

[107] H. Izumi, A. Ogata, L. A. Nafie and R. K. Dukor, *J. Org. Chem.*, 2008, **73**, 2367–2372.

[108] H. Izumi, S. Futamura, N. Tokita and Y. Hamada, *J. Org. Chem.*, 2007, **72**, 277–279.

[109] F. Wang, C. Zhao and P. L. Polavarapu, *Biopolymers*, 2004, **75**, 85–93.

[110] T. A. Keiderling, *Chem. Rev.*, 2015, **120**, 3381–3419.

[111] A. Polyanichko, V. Andrushchenko, P. Bouř and H. Wieser, Circular dichroism: theory and spectroscopy, 2012, pp. 67–126.

[112] M. Pazderková, T. Pazderka, M. Shanmugasundaram, R. K. Dukor, I. K. Lednev and L. A. Nafie, *Chirality*, 2017, **29**, 469–475.

[113] S. Ma, X. Cao, M. Mak, A. Sadik, C. Walkner, T. B. Freedman, I. K. Lednev, R. K. Dukor and L. A. Nafie, *J. Am. Chem. Soc.*, 2007, **129**, 12364–12365.

[114] D. Kurouski, R. A. Lombardi, R. K. Dukor, I. K. Lednev and L. A. Nafie, *Chem. Commun.*, 2010, **46**, 7154–7156.

[115] P. Bouř, V. Andrushchenko, M. Kabeláč, V. Maharaj and H. Wieser, *J. Phys. Chem. B*, 2005, **109**, 20579–20587.

[116] V. Andrushchenko, H. Wieser and P. Bouř, *J. Phys. Chem. A*, 2007, **111**, 9714–9723.

[117] V. Andrushchenko and P. Bouř, *Chirality*, 2010, **22**, E96–E114.

[118] V. Andrushchenko, D. Tsankov, M. Krasteva, H. Wieser and P. Bouř, *J. Am. Chem. Soc.*, 2011, **133**, 15055–15064.

[119] M. Krupová, P. Leszczenko, E. Sierka, S. Emma Hamplová, R. Pelc and V. Andrushchenko, *Chem. Eur. J*, 2022, **28**, e202201922.

[120] M. A. J. Koenis, A. Osypenko, G. Fuks, N. Giuseppone, V. P. Nicu, L. Visscher and W. J. Buma, *J. Am. Chem. Soc.*, 2020, **142**, 1020–1028.

[121] A. Osypenko, E. Moulin, O. Gavat, G. Fuks, M. Maaloum, M. A. J. Koenis, W. J. Buma and N. Giuseppone, *Chem. Eur. J*, 2019, **25**, 13008–13016.

[122] C. Merten, T. Kowalik and A. Hartwig, *Appl. Spectrosc.*, 2008, **62**, 901–905.

[123] C. Merten and A. Hartwig, *Macromolecules*, 2010, **43**, 8373–8378.

[124] D. Kurouski, J. D. Handen, R. K. Dukor, L. A. Nafie and I. K. Lednev, *Chem. Commun.*, 2015, **51**, 89–92.

[125] J. Frelek, M. Górecki, M. łaszcz, A. Suszczyńska, E. Vass and W. J. Szczepek, *Chem. Commun.*, 2012, **48**, 5295–5297.

[126] I. Kawamura and H. Sato, *Anal. Biochem.*, 2019, **580**, 14–20.

[127] V. Declerck, A. Pérez-Mellor, R. Guillot, D. J. Aitken, M. Mons and A. Zehnacker, *Chirality*, 2019, **31**, 547–560.

[128] J. J. L. González, F. P. Ureña, J. R. A. Moreno, I. Mata, E. Molins, R. M. Claramunt, C. López, I. Alkorta and J. Elguero, *New J. Chem.*, 2012, **36**, 749–758.

[129] S. Jähnigen, A. Scherrer, R. Vuilleumier and D. Sebastiani, *Angew. Chem. Int. Ed.*, 2018, **57**, 13344–13348.

[130] S. Jähnigen, K. Le Barbu-Debus, R. Guillot, R. Vuilleumier and A. Zehnacker, *Angew. Chem. Int. Ed.*, 2023, **62**, e202215599.

[131] P. Oulevey, S. Luber, B. Varnholt and T. Bürgi, *Angew. Chem. Int. Ed.*, 2016, **55**, 11787–11790.

[132] J. Blasius, R. Elfgen, O. Hollóczki and B. Kirchner, *Phys. Chem. Chem. Phys.*, 2020, **22**, 10726–10737.

[133] H. Sato, T. Yajima and A. Yamagishi, *Chirality*, 2015, **27**, 659–666.

Chapter 2

Vibrational circular dichroism

To follow the results presented in Chapters 4-6, proper understanding of the physics behind VCD is required. Section 2.1 focuses on the basic theoretical foundation of VCD. The main references for this section are the books written by Stephens *et al.*[1] and Nafie[2], along with the review written by Magyarfalvi *et al.*[3]. Section 2.3 covers the workflow commonly used for AC determination along with practical implications.

2.1 Theoretical background

2.1.1 VCD & IR intensities

At the basis of Circular Dichroism (CD) lies the difference in interaction of a chiral compound with Left Circularly Polarised Light (LCPL) and Right Circularly Polarised Light (RCPL). In Electronic Circular Dichroism (ECD), the most well-known form of CD, the differential absorption for the circularly polarised light in electronic transitions is measured. For Vibrational Circular Dichroism (VCD), the CD phenomenon is instead measured for the vibrational transitions of a chiral molecule. These vibrational transitions typically occur in the IR region and, therefore, VCD is often seen as the chiral version of IR spectroscopy. The absorption $A(\nu)$ describes the decrease in intensity of a light beam passing through a sample:

$$A(\nu) = -log_{10}\left(\frac{I(\nu)}{I_0(\nu)}\right) \tag{2.1}$$

where $I_0(\nu)$ is the incident intensity of the light beam at frequency ν and $I(\nu)$ the transmitted intensity. For an achiral compound, the absorption is identical for both polarisation types. For chiral compounds, this symmetry is broken and there is a small difference between these absorbances $\Delta A(\nu)$:

$$\Delta A(\nu) = A_L(\nu) - A_R(\nu) \tag{2.2}$$

where $A_L(\nu)$ and $A_R(\nu)$ are the absorbance of LCPL and RCPL respectively. The absorption by the chiral compound is taken as the mean of $A_L(\nu)$ and $A_R(\nu)$:

$$A(\nu) = \frac{1}{2}\left(A_L(\nu) + A_R(\nu)\right) \tag{2.3}$$

Circular dichroism is a notably weak phenomenon, especially so when vibrational transitions are involved. The ratio between $\Delta A(\nu)$ and $A(\nu)$, known as the anisotropic ratio, is typically 10^{-4} or lower. Additionally, the absorption of light due to vibrational transitions is rather weak compared to electronic transitions. To extract these weak signals from the noise, VCD experiments require larger amount of sample and longer measurement times compared to other spectroscopic techniques.

Within the limit of the Lambert-Beer law, $A(\nu)$ can be transformed into a molar absorbance $\varepsilon(\nu)$:

$$\varepsilon(\nu) = \frac{A(\nu)}{lC} \tag{2.4}$$

where l is the cel length and C the molar concentration of the absorbing compound. By combining Eq. 2.2 and 2.4, a molar quantity $\Delta\varepsilon(\nu)$ can be introduced for VCD, known as the molar absorbance difference:

$$\Delta\varepsilon(\nu) = \frac{\Delta A(\nu)}{lC} = \frac{A_L(\nu) - A_R(\nu)}{lC} \tag{2.5}$$

The molar absorbances $\varepsilon(\nu)$ and $\Delta\varepsilon(\nu)$ are less dependent on the exact experimental set-up, compared to $A(\nu)$ and $\Delta A(\nu)$. Therefore, these molar quantities are the preferred quantities for analysis.

2.1.2 Dipole and rotational strength

To compute the IR and VCD spectrum, the molar absorbance and molar absorbance difference of the relevant vibrational transitions have to be determined. Here, the key properties to determine are the so-called dipole strength and rotational strength. The molar absorbance ε_{if} of the vibrational transition from state $|i\rangle$ to $|f\rangle$, associated with normal mode a, is directly proportional to the dipole strength D^a_{if}:[1]

$$\varepsilon_{if} = \frac{8\pi^3 N_A \nu_{if}}{3.10^3 hc \ln 10} D^a_{if} \tag{2.6}$$

where h is the Planck's constant, c the speed of light in vacuum, N_A Avogadro's number, ν_{if} the frequency and D^a_{if} the dipole strength of the vibrational transition. The dipole strength D^a_{if} results from the electric dipole transition moment $\langle i|\hat{\boldsymbol{\mu}}|f\rangle_a$:

$$D^a_{if} = |\langle i|\hat{\boldsymbol{\mu}}|f\rangle_a|^2 \tag{2.7}$$

An expression similar to Eq. 2.6 can be found for the molar absorbance difference $\Delta\varepsilon_{if}$:

$$\Delta\varepsilon_{if} = \frac{32\pi^3 N_A \nu_{if}}{3.10^3 hc \ln 10} R^a_{if} \tag{2.8}$$

Here, we have introduced the rotational strength R^a_{if} of the vibrational transition. The rotational strength itself depends on both the electric dipole transition moment and the magnetic dipole transition moment $\langle f|\hat{\boldsymbol{m}}|i\rangle_a$:

$$R^a_{if} = \mathfrak{Im}[\langle i|\hat{\boldsymbol{\mu}}|f\rangle_a \cdot \langle f|\hat{\boldsymbol{m}}|i\rangle_a] \tag{2.9}$$

The molar absorbance ε_{if} depends solely on the size of the electric dipole transition moment. The molar absorbance difference $\Delta\varepsilon_{if}$, in turn, depends on the size of the electric and magnetic dipole transition moment, along with the angle ξ^a_{if} between these moments. The rotational strength can be both positive or negative, whereas the dipole strength is strictly positive. The scalar product of the transition moments changes sign upon reflection or inversion, resulting in mirror image results for enantiomeric pairs.

To summarise, the IR and VCD intensity for a vibrational transition are directly proportional to the dipole and rotational strength respectively. This entails evaluating the electric and magnetic dipole transition moment for said vibrational transition, along with the corresponding vibrational frequency.

2.1.3 Evaluation of the dipole transition moments

The dipole transition moments are evaluated with a truncated Taylor expansion of the dipole moments around the equilibrium geometry $\boldsymbol{R}^0$. Within the Born-Oppenheimer approximation and the harmonic oscillator approximation, these dipole transition moments can be formulated in terms of the so-called Atomic Polar Tensors (APTs) and Atomic Axial Tensors (AATs). Here, greek letters (α, β, γ) are used to denote Cartesian components of vectors and matrices. The β'th component of electronic dipole transition moment is equal to:

$$\langle i|\hat{\mu}_\beta|f\rangle_a = \sqrt{\frac{\hbar}{4\pi\nu_a}} \sum_J^N \sum_\alpha^{x,y,z} S_{J\alpha,a} P^J_{\alpha\beta} \tag{2.10}$$

where $P^J_{\alpha\beta}$ is the APT of nucleus J, N is the number of nuclei in the compound, ν_a the vibrational frequency of normal mode a and $S_{J\alpha,a}$ describes the nuclear displacements along the normal mode coordinate Q_a:

$$S_{J\alpha,a} = \left(\frac{\partial R_{J\alpha}}{\partial Q_a}\right)_{\boldsymbol{R}=\boldsymbol{R}^0} \tag{2.11}$$

where $R_{J\alpha}$ is the α'th Cartesian coordinate of nucleus J. Here, the notation $\boldsymbol{R} = \boldsymbol{R}^0$ indicates that the derivative is evaluated at the equilibrium geometry $\boldsymbol{R}^0$. $P^J_{\alpha\beta}$ contains both an electronic contribution $E^J_{\alpha\beta}$ and a nuclear contribution $N^J_{\alpha\beta}$:

$$P^J_{\alpha\beta} = E^J_{\alpha\beta} + N^J_{\alpha\beta} \tag{2.12}$$

The nuclear contribution $N^J_{\alpha\beta}$ to the APT is fairly straightforward, requiring only the charges Z_J of the nuclei:

$$N^J_{\alpha\beta} = (Z_J e)\delta_{\alpha\beta} \tag{2.13}$$

where the elements of Kronecker delta $\delta_{\alpha\beta}$ are equal to 1 for identical Cartesian directions (i.e. $\alpha = \beta$) and 0 if the directions are different. The electronic contribution to the APT is:

$$E^J_{\alpha\beta} = \left(\frac{\partial \langle \psi_i | \hat{\mu}_\beta | \psi_i \rangle}{\partial R_{J\alpha}}\right)_{\boldsymbol{R}=\boldsymbol{R}^0} \tag{2.14}$$

where $|\psi_i\rangle$ is the electronic wavefunction.

To evaluate the rotational strength, we also require the magnetic dipole transition moment. The magnetic dipole transition moment is determined with the AATs $M^J_{\alpha\beta}$:[4]

$$\langle i|\hat{m}_\beta|f\rangle_a = \sqrt{4\pi\hbar^3\nu_a}\sum_J^N \sum_\alpha^{x,y,z} S_{J\alpha,a} M^J_{\alpha\beta} \tag{2.15}$$

Similar to the APTs, $M^J_{\alpha\beta}$ consists of an electronic and nuclear contribution:

$$M^J_{\alpha\beta} = I^J_{\alpha\beta} + J^J_{\alpha\beta} \tag{2.16}$$

The nuclear contribution $J^J_{\alpha\beta}$ is fairly simple, requiring only the nuclear charges and their Cartesian coordinates in the equilibrium geometry $R^0_{J\gamma}$:

$$J^J_{\alpha\beta} = \frac{i}{4\hbar c}\sum_\gamma^{x,y,z} (Z_J e) R^0_{J\gamma} \epsilon_{\alpha\beta\gamma} \tag{2.17}$$

where the elements of the Levi-Civita tensor $\epsilon_{\alpha\beta\gamma}$ are +1 for an even number of permutations of (x, y, z), -1 for an odd number and 0 otherwise. The electronic contribution of $I^J_{\alpha\beta}$ is less straightforward to obtain, as it does not exist within the Born-Oppenheimer approximation. The vibronic coupling theory[5] shows that the electronic contribution is non-zero when the wavefunctions of excited electronic states are mixed with the wavefunction of the electronic ground state. Sadly, the resulting formulation involves a summation over all excited electronic states, preventing its practical implementation. However, this problem can be circumvented with the magnetic field perturbation formalism[4,6], which effectively removes the sum over states. Here, the electronic contribution can be rewritten in terms of the perturbation of the electronic wavefunction by an external magnetic field $\boldsymbol{H}$:

$$I^J_{\alpha\beta} = \left\langle \left(\frac{\partial \tilde{\psi}_i}{\partial R_{J\alpha}}\right)_{\boldsymbol{R}=\boldsymbol{R}^0} \middle| \left(\frac{\partial \tilde{\psi}_i}{\partial H_\beta}\right)_{\boldsymbol{R}=\boldsymbol{R}^0} \right\rangle \tag{2.18}$$

where $\tilde{\psi}_i$ is the perturbed electronic wavefunction. While there is no external magnetic field present in the VCD experiment, the computational trick offers a convenient alternative to the sum over states formalism. A last hurdle to evaluating the magnetic dipole transition moment lies in its origin-dependence. This problem is addressed by switching to gauge-including atomic orbitals[7,8], which yield gauge-invariant magnetic dipole transition moments.

So, calculating the intensities ε_{if} and $\Delta\varepsilon_{if}$ requires the APTs and AATs, along with the nuclear displacements and frequency for each normal mode. To obtain these properties, the equilibrium geometry $\boldsymbol{R}^0$ of the molecule has to be known beforehand or through geometry optimisation. At this geometry, the $3N$-6 largest eigenvalues λ_a of the mass-weighted Hessian yield the vibrational frequencies:

$$\nu_a = \sqrt{\frac{\lambda_a}{4\pi^2 c^2}} \tag{2.19}$$

and the eigenvectors define the transformation matrix $S_{J\alpha,a}$ from Cartesian displacement to the normal coordinates.

2.1.4 Molecular VCD spectrum

With the dipole transition moment formulations given in section 2.1.3, the dipole strength and rotational strength can be determined with Eqs. 2.7 and 2.9. In turn, this allows to evaluate the values of ε_{if} and $\Delta\varepsilon_{if}$ according to Eqs. 2.6 and 2.8. The full IR spectrum $\varepsilon(\nu)$ and the VCD spectrum $\Delta\varepsilon(\nu)$ consists of contributions from all accessible vibrational excitations with frequency ν_{if}:

$$\varepsilon(\nu) = \sum_{i,f} \varepsilon_{if}\ f_{if}(\nu_{if}, \nu) \tag{2.20}$$

$$\Delta\varepsilon(\nu) = \sum_{i,f} \Delta\varepsilon_{if}\ f_{if}(\nu_{if}, \nu) \tag{2.21}$$

where $f_{if}(\nu_{if}, \nu)$ is a function describing the band broadening for the $i \to f$ transition. This function is approximated as a Lorentzian with a half-width at half-max γ:

$$f_{if}(\nu_{if}, \nu) = \frac{1}{\pi} \frac{\gamma}{(\nu_{if} - \nu)^2 + \gamma^2} \tag{2.22}$$

The value of γ is set to an empirical value, typically ranging from 5 to 7.5 cm^{-1}. Up until now, the implicit assumption was made that the molecule only adopted a single geometry $\boldsymbol{R}^0$. This assumption no longer holds for flexible compounds as they are able to adopt multiple conformers. Each conformer has its own VCD spectrum and these conformer spectra are distinct as demonstrated previously in Fig. 1.7. The molecular IR and VCD spectrum consists of contributions $\Delta\varepsilon_c^{conf}(\nu)$ and $\varepsilon_c^{conf}(\nu)$ from each relevant conformer c:

$$\varepsilon^{mol}(\nu) = \sum_c^C w_c \varepsilon_c^{conf}(\nu) \tag{2.23}$$

$$\Delta\varepsilon^{mol}(\nu) = \sum_c^C w_c \Delta\varepsilon_c^{conf}(\nu) \tag{2.24}$$

where the weight w_c describes the relative contribution of conformer c to the molecular IR spectrum $\varepsilon^{mol}(\nu)$ and VCD spectrum $\Delta\varepsilon^{mol}(\nu)$ and C is the number of conformers for the compound. These weights are taken as the relative population of the conformers within the Boltzmann distribution:

$$w_c = \frac{\mathrm{e}^{-\Delta H^0_{298.15,c}/RT}}{\sum_c^C \mathrm{e}^{-\Delta H^0_{298.15,c}/RT}} \tag{2.25}$$

where R is the universal gas constant and T the temperature. Note that here conformer enthalpies $\Delta H^0_{298.15,c}$ are used instead of the Gibbs free energies. The entropic contributions to the Gibbs free energies are not straightforward to calculate. These entropic contributions are roughly identical for conformers of the same compound.[9] For this reason, the Gibbs free energy relative to the lowest-energy conformer can be approximated with the enthalpy.

2.1.5 Solvent effects

In deriving the formulations of ε_{if} and $\Delta\varepsilon_{if}$, any interaction between the chiral molecule and its environment is neglected. In most applications, the VCD spectrum is recorded for solutions or neat liquids of a chiral compound. Here, the solute is perturbed by the neighbouring solvent molecules. The strength of this perturbation depends on the chosen solvent and the properties of the chiral compound. The influence of the solvent can have the following effects[10–12]: change the conformer population, shift vibrational frequencies, enhance/diminish signal intensity or flip its sign, introduce non-zero rotational strength for vibrational modes of the solvent.

For apolar solvents, the perturbation of the VCD spectrum is relatively limited. The explicit solvent-solute interactions can be safely ignored for most cases, with only a handful of exceptions[13,14]. Implicit solvation models, such as the Polarisable Continuum Model (PCM)[15,16] or COnductor-like Screening MOdel (COSMO)[17], are typically used to account for the electrostatic interaction between the molecule and the solvent continuum. The lack of strong solvent-solute interactions can promote the aggregation of the solute. Carboxylic acids are well-known to form stable dimers within apolar environments and calculations need to explicitly account for these aggregates.[18–20] Recently, the additive 7-Azaindole was shown to prevent this aggregation and simplify the computational workflow .[21]

For more polar solvents, the solvent-solute interactions influence the VCD spectrum more significantly. Implicit solvation models may not capture all of these interactions and an atomistic description of these interactions is required instead. One such approach is to create so-called microclusters and evaluate their energies, vibrational frequencies and VCD intensities.[11,18–20,22–27] These microclusters are complexes consisting of the solute surrounded by a small number (typically 1-3) of solvent molecules. This approach has been shown to improve upon the solvent description when implicit solvent models reproduce the experimental spectrum poorly. However, the need to manually place explicit solvent molecules makes this approach labour-intensive and prone to user bias, especially if the compound is flexible and contains multiple moieties favouring hydrogen bonding. An alternative approach is to simulate the solution using Molecular Dynamics (MD) and perform VCD calculations on the extracted snapshots, with or without explicit solvent molecules.[28–34] These MD based methods include more complex solvation configurations, though the required computational power is significantly larger and might not be accessible to many VCD users. Explicit solvation models add complexity and computational cost to the workflow. Therefore, they are chiefly only employed if the AC cannot be determined with an implicit model.

2.2 VCD instrumentation

VCD instruments are typically built using the setup shown in Figure 2.1, upon which modifications are added to improve reliability and noise level. Similar to IR spectrometers, VCD spectrometers operate in a Fourier-Transform (FT) configuration. Here, the beam generated by the mid-IR source is modulated using an interferometer with a characteristic Fourier frequency. The modulated IR beam emerging from the interferometer is linearly polarised under an angle of 45 degrees (respective to the optical axis) and passed to a photoelastic modulator (PEM). In the PEM, the IR light passes through a ZnSe crystal which is stretched and compressed in a sinusoidal manner with a characteristic frequency (typically between 35-60 kHz). The stress on the crystal induces a difference in the velocity of the linear components, parallel and perpendicular to the PEM's optical axis, of the IR light. As a result, the linearly polarised IR light is transformed such that it continuously alternates between LCPL and RCPL. Next, the IR beam passes through the sample and the detector records the transmitted IR intensity. The sample cell should be constructed from material transparent to the IR range of interest, typically BaF_2 or CaF_2. Additionally, a detector with an adequate response time is needed to handle the high-frequency PEM modulation, with the HgCdTe (CMT) detector being a popular choice.

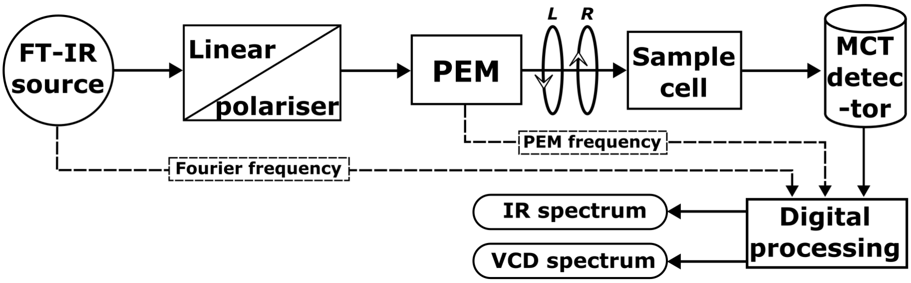


Figure 2.1: Schematic overview of an FT-VCD spectrometer, inspired by the block diagram presented by Bogaerts *et al.*[35].

The signal recorded by the detector is doubly modulated at both the low Fourier frequency and the high PEM frequency. This signal is separated into a low frequency path and a high frequency path by passing it through a low-pass and high-pass filter respectively. The IR transmission spectrum is obtained by applying Fourier transformation to the interferogram of the low frequency path, along with the IR absorbance spectrum using equation 2.1. The high frequency path contains the differential signal spectrum, which is needed to produce the VCD spectrum. A lock-in amplifier is used to demodulate the PEM modulation, which provides the interferogram of the high frequency path. The Fourier transform of this interferogram yields the differential signal spectrum. The VCD spectrum can then be obtained from the ratio of the differential signal to the mean signal from the low frequency path. The Chiral*IR*-2x spectrometer from Biotools Inc. used in chapter 6 uses a second PEM, placed between the sample and the MCT detector, to improve the stability of the VCD baseline. Additional details on the instrumentation can be found in the book written by Nafie[2] and the review of Keiderling[36].

2.3 VCD workflow for AC determination

This section covers the workflow used to determine the AC of a compound. The main steps of the workflow are illustrated in Fig. 2.2. We assume that all aspects of the 2D molecular structure have been resolved without any information on the 3D structure. Parts of the workflow can be simplified if the details such as the relative configuration or conformer populations have already been identified with other analytical tools.

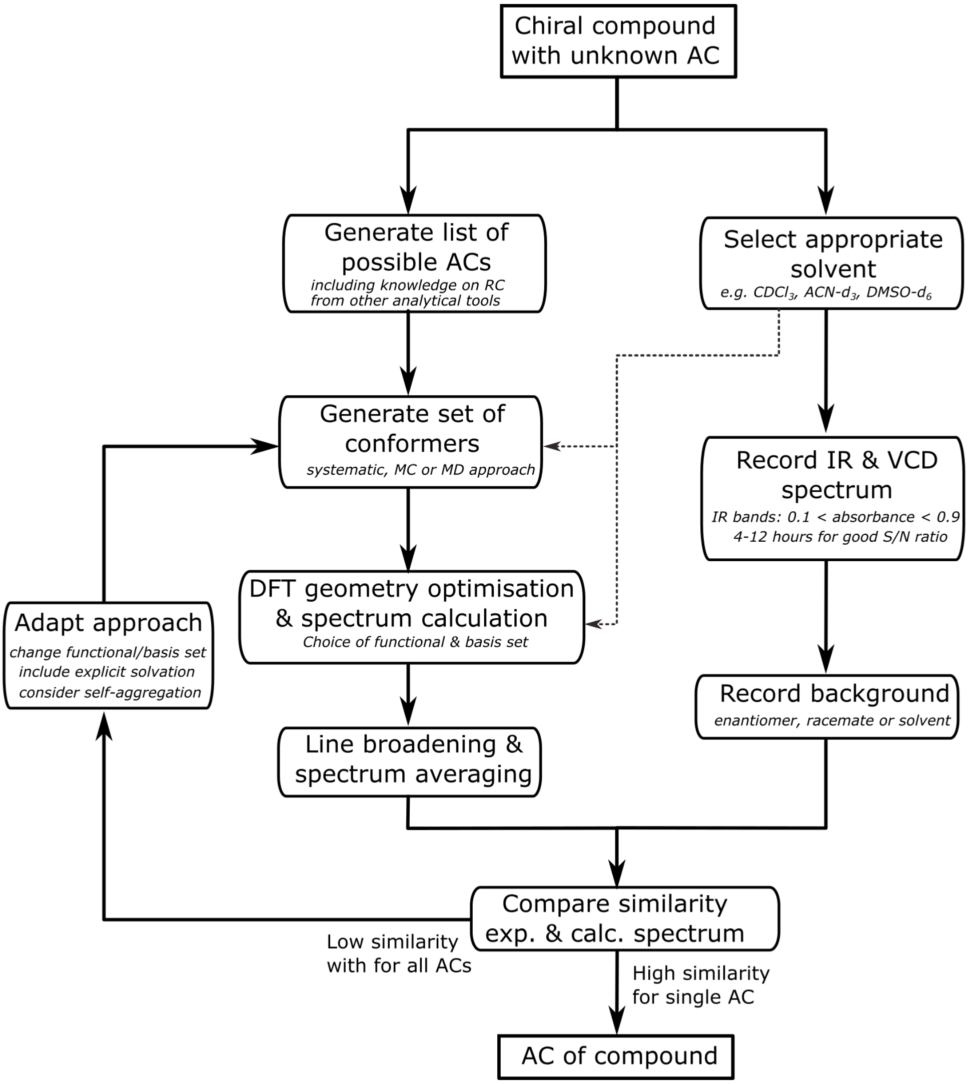

Chiral compound with unknown AC
Generate list of possible ACs
including knowledge on RC from other analytical tools
Select appropriate solvent
e.g. CDCl$_3$, ACN-d$_3$, DMSO-d$_6$
Generate set of conformers
systematic, MC or MD approach
Record IR & VCD spectrum
IR bands: 0.1 < absorbance < 0.9
4-12 hours for good S/N ratio
DFT geometry optimisation & spectrum calculation
Choice of functional & basis set
Adapt approach
change functional/basis set
include explicit solvation
consider self-aggregation
Record background
enantiomer, racemate or solvent
Line broadening & spectrum averaging
Compare similarity exp. & calc. spectrum
Low similarity with for all ACs
High similarity for single AC
AC of compound

Figure 2.2: Typical workflow for AC determination with VCD. The figure is inspired by the workflow presented by Bogaerts *et al.*[35].

2.3.1 Experimental workflow

To obtain a high-quality experimental VCD spectrum, an appropriate selection of solvent, sample concentration and cell length is necessary. Ideally, intensities of the IR bands remain within the range of 0.1-0.9 absorption units. Achieving this involves optimising the sample concentration and cell length. Also, the chosen solvent must adequately dissolve the sample to attain the desired concentration. The optimal concentration and cell length can differ among frequency regions and, as a result, these regions may require separate measurements. The solvent has to be transparent in the region of interest to maximise the spectral window. Therefore, deuterated solvents lacking any C-C bonds (e.g. DMSO-d_6, ACN-d_3 or $CDCl_3$) are preferred for measurements in the fingerprint region. The solvent choice also impacts the computational workflow as discussed in section 2.1.5. For polar solvents, VCD calculations may involve explicit treatment of the solute-solvent interactions. For this reason, apolar solvents are favoured when the compound is easily solvated in them and aggregation is not an issue. To extract the weak VCD signals, longer measurement times are needed to reduce the noise level. The collection time of a VCD spectrum generally ranges between 4 and 12 hours, though solutions with notably small VCD intensities require even longer measurement times. Finally, VCD spectra contain a background dependent on the instrumentation and experimental setting. To correct for this background, a baseline spectrum is subtracted from each VCD spectrum. The baseline should contain the VCD spectrum measured under identical conditions, but without the chiral response of the solute, and substracted from the spectrum of the chiral sample. The VCD spectrum of the racemate typically provides a reliable baseline. If the racemate is unavailable, the solvent spectrum can be used as an approximate baseline. The most accurate method for removing the backgroud is to record the VCD spectrum of the enantiomer and use the half-sum spectrum of both enantiomers as the baseline.

2.3.2 Computational workflow

In most cases, extracting the AC directly from the experimental VCD spectrum is not possible. Instead, quantum chemical predictions of the VCD spectrum are performed for each possible AC. Upon comparing these predictions with the experimental spectrum, the AC of the sample can be identified. The methodology behind these quantum chemical predictions is discussed here. Details on the comparison of the predictions with the experiment(s) are provided in section 2.3.3

The VCD patterns of enantiomers are identical expect for their inverted signs. Therefore, it suffices to compute just the spectrum of one enantiomer and invert the VCD intensities for the other enantiomer. Furthermore, knowledge of the relative configuration obtained with other analytical tools (e.g. NMR, XRD) further reduces the number of stereoisomers to consider. First, a 3D geometry is generated for each potential stereoisomer or, if available, retrieved from an external dataset or the results of another analytical procedure. The following procedure is then repeated for each stereoisomer considered.

A conformational search is performed, using the 3D geometry as starting point, to identify the different conformers of the compound. This step is non-trivial for more flexible and complex compounds and many algorithms have been designed to tackle this issue. For such systems, the individual shortcomings of a single method can be mitigated by pooling the conformers found by different algorithms. In most applications, this step is performed through either a systematic search, Monte Carlo approach or MD simulation with a force field. For small compounds with limited flexibility, it may be worthwhile to perform the conformational search using more accurate methods than a force field.

The geometry of each conformer is then further optimised using Density Functional Theory (DFT), followed by frequency calculations for each optimised geometry. The absence of imaginary frequencies identifies whether the optimised geometry is in fact a conformer and not a transition state. Then, the dipole strengths and rotational strengths are determined for the normal modes of each conformer. A general recommendation by the community is to combine the B3LYP or B3PW91 hybrid functional with 6-31G(d) as minimal basis set. The actual basis set is chosen based on the size of the system and the computational resources available.

Next, the conformer spectra are broadened using a Lorentzian function with a FWHM between 10 and 15 cm^{-1}. The conformer spectra are then weighted according to their relative population and combined linearly. The resulting molecular spectrum is inverted, yielding the spectrum of the other enantiomer. The relative population of the conformers is determined according to the Boltzmann distribution using the enthalpies (see section 2.1.4) or, alternatively, the zero-point corrected energies or Gibbs free energies. Alternatively, the conformer weights can be taken from extensive MD simulations.

2.3.3 Comparison of the experimental and calculated spectrum

To establish the chirality of the compound, the predicted VCD spectra of each AC are compared with the experimental one. The calculated vibrational frequencies cannot perfectly match the experiment as a result of the harmonic approximation and the limited basis set. Therefore, the wavenumbers of the calculated spectra are scaled with either a single scaling factor or one per frequency region. Their value is either taken from literature or optimised using the IR spectrum, with values ranging between 0.95-1.05.

The similarity between the calculated and experimental spectrum can be determined either through visual inspection or with quantitative methods. During visual inspection, the user identifies the main features in each spectrum and compares the computed spectra with the experimental one based on the features they share. This method is susceptible to user bias and requires an expert eye to assign the AC in a reliable manner. The quantitative methods rely on a similarity metric to compare computed spectra to the experimental one without introducing user bias. The cosine similarity (or overlap integral)[37,38] quantifies the normalised overlap between a computed and experimental spectrum, yielding values between -1 and 1 for VCD spectra due to their signed intensities. A value of 1 indicates a perfect match, while -1 signifies perfect mirror images. The stereoisomer that results in the largest similarity is assigned to the experiment. For a robust assignment, the similarity values obtained for the other stereoisomers should be significantly smaller. The Tanimoto similarity[39,40] is another frequently used metric, whose values also range between -1 and 1. Recently, new methods have been proposed that use sequence alignment procedures[41,42] or introduce leniancy on the exact conformer weights[43,44] for the similarity assessment. While important, the exact quantitative methods chosen should not impact the AC assigned to the compound.

The reliability of the AC determination directly depends on the accuracy of the DFT predicted spectra. The DFT predictions need to accurately represent the general spectral pattern for each AC. This is especially important when multiple enantiomer pairs are considered. If none of the predicted spectra match the experiment, the computational workflow has to be adjusted. Here, the goal becomes to improve the accuracy of the computational method. To achieve this, one may increase the size of the basis set, change the functional or add conformers generated with another algorithm. At an increased cost, a model-averaged VCD spectrum can be used instead to account for spectral variations due to the details of the DFT method used.[45] Additionally, the methods discussed in section 2.1.5 can be used to correct for self-aggregation or explicit interaction with the solvent. However, their inclusion is labour intensive and challenging for non-experts.

References

[1] P. J. Stephens, F. J. Devlin and J. R. Cheeseman, *VCD spectroscopy for organic chemists*, CRC Press, Boca Raton, FL, 2012.

[2] L. A. Nafie, *Vibrational Optical Activity: Principles and Applications*, Wiley, Hoboken, NJ, 2011.

[3] G. Magyarfalvi, G. Tarczay and E. Vass, *WIREs Computational Molecular Science*, 2011, **1**, 403–425.

[4] P. J. Stephens, *J. Phys. Chem.*, 1985, **89**, 748–752.

[5] L. A. Nafie and T. B. Freedman, *J. Chem. Phys.*, 1983, **78**, 7108–7116.

[6] A. Buckingham, P. Fowler and P. Galwas, *Chem. Phys.*, 1987, **112**, 1–14.

[7] F. London, *J. Phys. Radium*, 1937, **8**, 397–409.

[8] R. Ditchfield, *Mol. Phys.*, 1974, **27**, 789–807.

[9] D. A. McQuarrie, *Statistical Thermodynamics*, Harper and Row, New York, NY, 1973.

[10] C. Merten, *Phys. Chem. Chem. Phys.*, 2017, **19**, 18803–18812.

[11] V. P. Nicu, E. J. Baerends and P. L. Polavarapu, *J. Phys. Chem. A*, 2012, **116**, 8366–8373.

[12] V. P. Nicu, J. N. and E. J. Baerends, *J. Phys. Chem. A*, 2008, **112**, 6978–6991.

[13] E. Debie, L. Jaspers, P. Bultinck, W. Herrebout and B. V. D. Veken, *Chem. Phys. Lett.*, 2008, **450**, 426–430.

[14] V. P. Nicu, E. Debie, W. Herrebout, B. Van der Veken, P. Bultinck and E. J. Baerends, *Chirality*, 2009, **21**, 287–297.

[15] S. Miertuš, E. Scrocco and J. Tomasi, *Chem. Phys.*, 1981, **55**, 117–129.

[16] J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999–3094.

[17] A. Klamt and G. Schüürmann, *J. Chem. Soc., Perkin Trans. 2*, 1993, 799–805.

[18] K. Bünnemann and C. Merten, *Phys. Chem. Chem. Phys.*, 2017, **19**, 18948–18956.

[19] K. Bünnemann, C. H. Pollok and C. Merten, *J. Phys. Chem. B*, 2018, **122**, 8056–8064.

[20] M. Losada, H. Tran and Y. Xu, *J. Chem. Phys.*, 2008, **128**, 014508.

[21] C. Grassin, E. Santoro and C. Merten, *Chem. Commun.*, 2022, **58**, 11527–11530.

[22] J. Kubelka, R. Huang and T. A. Keiderling, *J. Phys. Chem. B*, 2005, **109**, 8231–8243.

[23] A. S. Perera, J. Thomas, M. R. Poopari and Y. Xu, *Front. Chem.*, 2016, **4**, 9.

[24] M. R. Poopari, Z. Dezhahang and Y. Xu, *Phys. Chem. Chem. Phys.*, 2013, **15**, 1655–1665.

[25] L. Weirich and C. Merten, *Phys. Chem. Chem. Phys.*, 2019, **21**, 13494–13503.

[26] L. Weirich, K. Blanke and C. Merten, *Phys. Chem. Chem. Phys.*, 2020, **22**, 12515–12523.

[27] L. Weirich, J. Magalhães de Oliveira and C. Merten, *Phys. Chem. Chem. Phys.*, 2020, **22**, 1525–1533.

[28] A. Scherrer, R. Vuilleumier and D. Sebastiani, *J. Chem. Phys.*, 2016, **145**, 084101.

[29] T. Giovannini, M. Olszòwka and C. Cappelli, *J. Chem. Theory Comput.*, 2016, **12**, 5483–5492.

[30] M. Thomas and B. Kirchner, *J. Phys. Chem. Lett.*, 2016, **7**, 509–513.

[31] K. Le Barbu-Debus, J. Bowles, S. Jähnigen, C. Clavaguéra, F. Calvo, R. Vuilleumier and A. Zehnacker, *Phys. Chem. Chem. Phys.*, 2020, **22**, 26047–26068.

[32] S. Ghidinelli, S. Abbate, J. Koshoubu, Y. Araki, T. Wada and G. Longhi, *J. Phys. Chem. B*, 2020, **124**, 4512–4526.

[33] S. Jähnigen, D. Sebastiani and R. Vuilleumier, *Phys. Chem. Chem. Phys.*, 2021, **23**, 17232–17241.

[34] D. R. Galimberti, *J. Chem. Theory Comput.*, 2022, **18**, 6217–6230.

[35] J. Bogaerts, R. Aerts, T. Vermeyen, C. Johannessen, W. Herrebout and J. M. Batista, *Pharmaceuticals*, 2021, **14**, 877.

[36] T. A. Keiderling, *Molecules*, 2018, **23**, 2404.

[37] E. Debie, E. De Gussem, R. K. Dukor, W. Herrebout, L. A. Nafie and P. Bultinck, *ChemPhysChem*, 2011, **12**, 1542–1549.

[38] J. Vandenbussche, P. Bultinck, A. K. Przybył and W. A. Herrebout, *J. Chem. Theory Comput.*, 2013, **9**, 5504–5512.

[39] J. Shen, C. Zhu, S. Reiling and R. Vaz, *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, 2010, **76**, 418–422.

[40] P. L. Polavarapu and C. L. Covington, *Chirality*, 2014, **26**, 539–552.

[41] L. Böselt, D. Sidler, T. Kittelmann, J. Stohner, D. Zindel, T. Wagner and S. Riniker, *J. Chem. Inf. Model.*, 2019, **59**, 1826–1838.

[42] L. Böselt, R. Aerts, W. Herrebout and S. Riniker, *Phys. Chem. Chem. Phys.*, 2023, **25**, 2063–2074.

[43] G. Marton, M. A. J. Koenis, H.-B. Liu, C. A. Bewley, W. J. Buma and V. P. Nicu, *Angew. Chem. Int. Ed.*, 2023, e202307053.

[44] M. A. J. Koenis, Y. Xia, S. R. Domingos, L. Visscher, W. J. Buma and V. P. Nicu, *Chem. Sci.*, 2019, **10**, 7680–7689.

[45] G. Monaco, F. Aquino, R. Zanasi, W. Herrebout, P. Bultinck and A. Massa, *Phys. Chem. Chem. Phys.*, 2017, **19**, 28028–28036.

Chapter 3

Machine learning

In the past years, the use of ML in chemistry has grown immensely, tackling a vast range of chemical problems. The absence of ML applications for VCD indicates that many researchers within the VCD community are likely less familiar with the core concepts of ML algorithms. As the focus of this thesis lies heavily on ML applications, the main part of this chapter will introduce the basis of the techniques and concepts used in ML.[1–7] The wine dataset[8] is used throughout this chapter to improve understanding of the concepts introduced. This toy dataset consists of 13 attributes of wines coming from three Portuguese cultivators. These attributes are related to the chemical composition and interaction with light, such as magnesium content, alcohol percentage and UV-VIS absorbance. The aim of this chapter is to provide a general understanding of the ML concepts at the basis of the VCD applications covered in the following chapters. For those interested in deepening their knowledge beyond this, I suggest Bishop[1] for a deeper dive into the ML concepts, Géron[2] as starting point for implementation in Python and Cartwright[4] or Schütt *et al.*[7] for an overview of chemical applications.

3.1 Fundamental concepts of ML

3.1.1 What is machine learning?

Many people are familiar with the term Artificial Intelligence (AI) either through its ubiquity in media, fiction or business applications. Recent examples of popular AI applications are large language models (GPT-4[9], GPT-3[10]) and generative art models (Dall-E[11], Midjourney[12], Stable Diffusion[13]). Despite the popularity of the concept, no general definition of AI is widely accepted.[14] The field of AI arose from the 1956 Dartmouth AI conference[15] where the concept was first proposed. Here, a machine capable of performing tasks which are seen as intelligent by humans, is referred to as AI. This working definition was largely shaped to combine research lines under a single banner. As such, the perception of AI is usually guided by the milestones and techniques that arise from it.[14] Many researchers will likely not contest its use to describe programs such as AlphaGo[16] which outperform human professional players in the complex game of Go. However, the boundaries of when a program can still be seen as intelligent are murky and can change over time.[17,18] AI research is more often categorised according to the different subdomains of AI, defined by the problems they tackle or the theoretical foundations/models used. Machine Learning (ML) is one such subdomain.

ML is primarily concerned with extracting patterns from data using mathematical methods.[6] The patterns are directly inferred from these data without explicit instruction on the nature of these patterns. The increasing wealth and availability of data makes ML a valuable tool for scientists to plough through the information in datasets and to catch interesting patterns. The resulting patterns are then used to gain a deeper understanding of connections between data points and features or to replace previously existing workflows. The scope of the patterns is influenced either directly by the ML method chosen or indirectly by the application area covered by the data itself. Thus, the success of any ML application will depend on the following issues: is the ML method suitable for the task and is the data appropriate for the envisioned application. As a result, a diverse collection of ML methods have been developed covering a large range of applications. Deep learning is a subdomain of ML where so-called deep neural networks (see section 3.4) are used. Deep learning has been extensively studied for its ability to infer very complex patterns from large amounts of data by combining different levels of data representations. As a result, it has become an integral part of many groundbreaking ML applications such as computer vision[19], natural language processing[10], art generation[11] and prediction of protein folding[20].

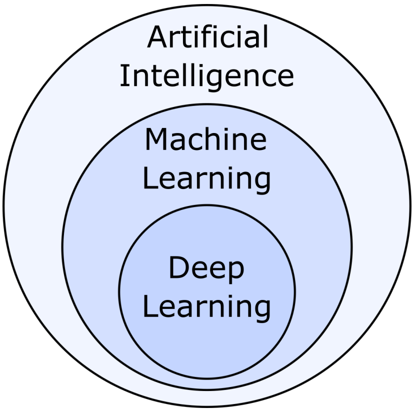


Figure 3.1: Venn diagram of artificial intelligence, machine learning and deep learning. Figure was inspired by Alzubaidi *et al.*[21]

3.1.2 Types of ML

ML methods can be divided into broad categories based on how they learn patterns from data instances and which problem is being solved. Here, the term 'data instance' refers to a single observation (i.e. data point or sample) in a dataset, like the analytical results for a wine bottle. Two major categories are supervised and unsupervised learning. For supervised ML methods the goal is to learn the connection between a set of features describing a data instance and the label(s) given to that instance. The set of data instances used by the ML method to establish this connection is referred to as the training set. So, input data is fed to the ML model and its internal parameters are adjusted to produce the label as output. The patterns extracted from the input features are tailored to the label and are used to predict the labels for new unlabelled data instances.

Classification is a form of supervised learning where the labels are categorical and the goal is to assign categorical values based on the input. An example of a classification problem for the wine dataset would be to identify the wine cultivator of a wine bottle based on the physicochemical properties of the wine. When these classes are mutually exclusive, such as identifying the cultivator of a wine bottle out of three possible cultivators, this problem is known as multi-class classification. For regression tasks, the label is typically a continuous or numeric quantity (e.g. predicting the price of the wine bottle). ML methods are specifically designed for either classification or regression problems and different metrics are used to evaluate their performance. These limitations can be circumvented by transforming the labels, allowing classification models to be used for regression tasks and vice versa. A classification task can be performed with a regression model by predicting probabilities for a categorical label, instead of predicting the label value directly. By converting continuous values into categorical values such as ‘high’, ‘medium’ or ‘low’, a classification model can be applied to a regression problem.

Supervised learning can also be performed on data instances with multiple labels, instead of a single label, which is known as multi-task or multi-output supervised learning. Here, for each label a separate regression and/or classification problem is defined. A common example of multi-task supervised learning is encountered for classification with labels that are not mutually exclusive, known as multi-label classification, such as identifying whether a wine bottle is expired and/or is of Portuguese origin. Only a limited selection of ML methods (e.g. decision trees, neural networks) can be trained directly on multi-task problems. For methods that do not support multi-task learning, a separate model needs to be trained for each single problem individually.

In unsupervised learning, ML methods are tasked with extracting the underlying structure present in unlabelled data. Only input data is provided to the ML method and the goal is to discover the connections between data instances or features. In the absence of an output, the ML method is not given a ‘correct‘ solution for this task and the nature of the patterns discovered will depend on the ML method chosen. Clustering and dimensionality reduction are typical applications of unsupervised learning. Clustering ML methods group data points together based on their relative similarity. Doing so, distinct clusters and patterns in the dataset can be identified. In dimensionality reduction, an ML method is used to project the data into a lower-dimensional representation, reducing the number of features in the dataset, while preserving most of the information present in the data. Another application of unsupervised learning is found in so-called generative methods. Here, an ML model learns the underlying distribution of the data and generates new data instances from this distribution. These generative methods have been used to, among others, generate new molecules in drug design[22–27] and pictures or 3D meshes of human faces[28,29].

Aside from supervision, ML methods are categorised on how interpretable the extracted patterns and decision making are to a human. White box ML methods are transparent and the patterns they extract are more easily understood. Black box ML methods, on the other hand, can extract much more complex patterns but are more difficult to interpret. White box ML methods are favoured for applications that require transparent decision-making such as high-stakes scenarios in healthcare or the justice system.[30–39] Linear ML methods (see section 3.3) and decision trees (see section 3.2.1) are examples of such white box methods. Concerns around the use of black box ML models in automated decision-making processes have brought the European union to implement a 'right to explanation' provision, requiring organisations to provide explanations for decision making involving a person.[40] With the increasing demand to make ML methods more interpretable, researchers have devoted significant effort to developing model-agnostic analysis methods and modifying black box ML methods to enhance their interpretability.[41–48] However, current white box models are not able to tackle all ML tasks. For complex problems, black box models like neural networks remain highly used. Black box models can extract more complex patterns from the data, resulting in improved accuracy for these problems. The choice for either white or black box ML methods depends on the required accuracy and level of explainability.

3.1.3 Model training and hyperparameter tuning

To reiterate, ML methods directly infer patterns from a dataset without explicit instruction on their exact nature. In supervised learning, the data is labelled and the ML model learns to reproduce these labels by adjusting its internal parameters. The process of adjusting the internal parameters of an ML model is known as ML model training. The dataset used for the model training is known as the training set. It is important to keep in mind that any ML method can only be taught to reproduce the patterns already present within the dataset provided. So, if the training set does not cover the future application area of the model, the accuracy of its predictions will likely be lacking. An ML model trained on Portuguese wines will provide more accurate predictions for other Portuguese wines than for e.g. German wines. If the training set contains significant unintended biases, the ML model will be biased likewise. A highly debated issue regarding model bias is the presence of racial bias in the algorithms used in applications

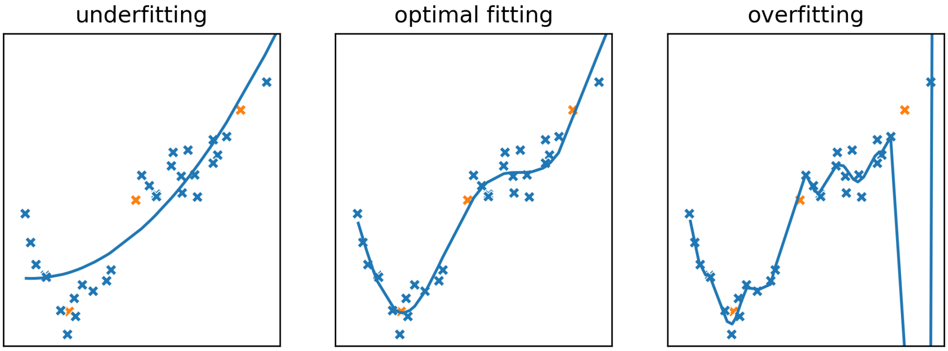


Figure 3.2: Illustration of under- and overfitting for a regression problem with a single feature. The training set (blue) and test set (orange) data instances are generated using a third degree polynomial and Gaussian noise.

such as health systems and computer vision.[49–56] Even if the dataset is unbiased and properly reflects the application area, the patterns extracted from this dataset might not necessarily result in accurate predictions.

Underfitting occurs when the ML method fails to replicate the data labels in the training set. The underfitting results from a mismatch between the complexity of the problem and either the size of the dataset, the flexibility of the ML method or procedure for the model training. The ML model can also fit the training set too closely to the point that it has memorised the entire dataset including noise, resulting in the overfitting of an ML model.[57,58] The patterns extracted by the model perform very well on the training set but are hardly transferable to new data instances. Overfitting can be caused by an excess of flexibility in the ML method, training the model on too many features or letting model training continue for too long. As underfitting results in low accuracy of the model it is easier to spot than overfitting, which requires predictions on new unseen data. An optimal balance between under- and overfitting is necessary to obtain an accurate ML model, as illustrated in Figure 3.2. To detect overfitting and evaluate generalisability, a subset of the dataset is kept aside and is not involved in the training process. This subset is known as a holdout set or test set, and is only used after all aspects of the model are finetuned. Doing so, the test set provides a final evaluation of the ML model's capabilities.

Many ML models have so-called hyperparameters that control the details of the model's structure and the training process. In contrast to the internal parameters learned during training, the values of these hyperparameters must be set prior to the model training. The hyperparameters can significantly impact the accuracy and generalisability of a model, along with the cost of the model training step. The process of finding the optimal hyperparameters for an ML model, also known as ML model optimization, is not a trivial task. During optimisation, a large range of values have to be considered for each hyperparameter. Typically, there is no one-size-fits-all approach to determining the optimal values for the hyperparameters.[59] Instead, the ML model is trained multiple times with different hyperparameter values, and the values that produce the most accurate model are ultimately chosen. The hyperparameter optimisation can be performed in one of the following ways: manual tuning based on trial and error, grid search, random search or Bayesian optimisation. In Bayesian optimisation, probabilistic models are used to map the values for each hyperparameter to the accuracy of the model.[60] As a result, a distribution of the accuracy in the hyperparameter space is obtained. In a next step, the learned distribution is used to suggest a new set of hyperparameter values. For models with many hyperparameters that can adopt a large range of values, random search and Bayesian optimisation are typically preferred.

To detect whether a set of hyperparameters results in overfitting of the model, the accuracy of the model during optimisation needs to be evaluated on 'unseen' data. One common method is to split another subset from the dataset, the validation set, which is exclusively used to evaluate the accuracy of each trained model during optimisation. So, the training set is used to establish the internal parameters, the validation set to optimise the hyperparameters and the test set as a final evaluation tool. Alternatively, the performance can be established with cross-validation. Here, the dataset is split into k subsets, referred to as 'folds'. One of the folds is used as the validation set while the remaining folds make up the training set. This procedure is then repeated using another fold as validation set, up until each fold has been used as validation set exactly once. The accuracy of the model is then averaged across all folds. Cross-validation helps reduce the influence of random variations between a training and validation set and provides an estimate for the uncertainty upon the model's accuracy. One of the main drawbacks of cross-validation lies in its increased cost for training and optimising the model. The training step is repeated k times for every set of hyperparameters, increasing the cost by a factor k. For ML applications involving large datasets or computationally expensive models, the use of a single validation set is therefore be preferred.

3.2 Decision trees and ensemble methods

3.2.1 Decision tree

Decision trees are white-box models that are suited for simple classification tasks. Their inner workings resemble how humans make decisions and intuitively identify patterns, making these models transparent and easy to interpret. For this reason, it sees common use in decision making and workflows in companies or governments. The model uses a tree-like structure to isolate the dataset into smaller populations that share the same label.[61] A decision tree is constructed with three types of nodes: a root node, branch nodes and leaf nodes. The root node is supplied with the entire dataset and splits the data into two subsets. The node finds the decision rule that isolates data instances with a different label as best as possible. The exact nature of these decision rules and isolation metric will be discussed later. A branch node is then generated for each of the two split subsets. The branch node is then tasked with splitting these subsets further with new decision rules. This branching of decision rules results in a tree-like structure with an exponentially increasing number of decision rules along the depth of the tree. When the tree reaches a prespecified depth or a certain criterion is met for a node, it will not split the data any further and the nodes are turned into leaf nodes. The samples found in each leaf node are given the same predicted label. Figure 3.3 shows a decision tree trained on a binary classification task on the wine dataset.

The success of a branch node in isolating data instances with M different labels in node a is expressed with the gini index G_a:[62]

$$G_a = 1 - \sum_{k=1}^{M} P_{a,k}^2 \tag{3.1}$$

where $P_{a,k}$ is the relative population of data instances with label k in node a. The

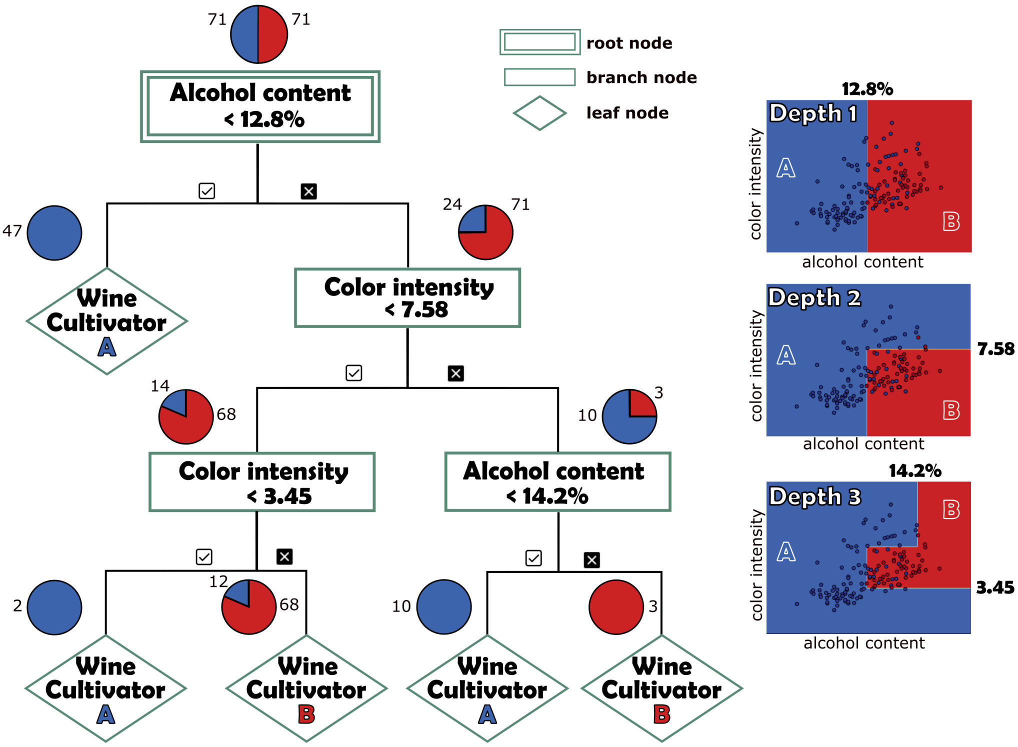


Figure 3.3: A decision tree of depth 3 trained to distinguish the wines of two cultivators A and B from the alcohol content and color intensity. For each node, the relative population of wines coming from cultivator A (blue) and B (red) is illustrated with a pie chart. The decision border in feature space is shown for each new layer of branch nodes, with the areas coloured according to the predicted cultivator (blue=A, red=B).

decision rule is of the form ‘if the value of feature X is above/below a threshold value Y’. This means that a branch node will always split the dataset along one of feature axes in feature space. Such decision rules see frequent use in chemistry, such as ‘if pH > 7 the solution is basic’ or ‘if the IR spectrum has a intense band around 1700 cm^{-1} suspect a carbonyl moiety’ However, the decision rules can be too limiting for certain ML tasks. For instance, if the actual decision border is a linear combination of 2 features, decision trees have limited success and, at best, form a rough stair-like decision border. Also, decision trees are very susceptible to overfitting. Each new layer of branch nodes increases the number of trainable parameters exponentially. To limit overfitting, methods are used to regularise the growth of new branches. This is achieved by transforming nodes into leaf nodes, known as tree trimming, when prespecified criteria are not met. Examples of such regularisation methods are imposing a minimum number of samples in a branch node or leaf node.

3.2.2 Random forest

To tackle the shortcomings of overfitting, multiple decision trees are combined into a single ensemble model. Ensemble models combine multiple ML models and the predictions of these ensemble models are more powerful. In bagging algorithms, the predicted labels of each model are used as votes.[63] These votes are then combined into a single label using a voting procedure. In hard-voting algorithms, the label predicted by the ensemble model is the majority vote. This is similar to voting systems in democratic processes, where each vote is given the same importance. In soft-voting, the uncertainty of each vote is taking into account into the vote averaging, where more importance is given to votes with high certainty. For decision trees, the uncertainty is based on the gini index of the leaf node.

The predictions coming from the ensemble model are more accurate than the individual models it contains. This makes ensemble models a great choice to compensate for the overfitting tendencies of decision trees. A bagging model constructed with only decision trees is referred to as a Random Forest[64] (RF). If the same training procedure is performed for each model in the ensemble, we simply obtain an ensemble of identical copies. To prevent this, the RF algorithm introduces randomness in the training of the decision trees. Each decision tree is trained on a random subset of the data instances and a random selection of features. Doing so, the individual trees are less accurate, but the predictions of the ensemble model become more reliable and generalisable. However, the increase in accuracy is accompanied by a loss in interpretability. These ensemble models are constructed with hundred or thousands of decision trees. Therefore, identifying the exact pattern used by the model becomes too time consuming for most applications. However, the relative importance of each feature in its decision making is relatively easy to extract. The importance of a feature is determined as the average increase in gini index when said feature is used for the decision rule.

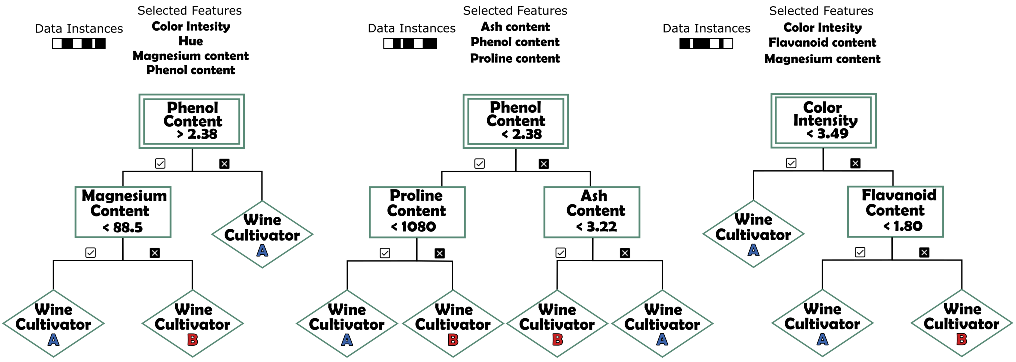


Figure 3.4: A random forest constructed with 3 independent decision trees. Each decision tree is trained separately on 3-4 randomly selected features of the wine dataset and a random subset of the samples in the dataset. The label predicted by the random forest is the majority vote of the 3 decision trees.

3.3 Linear methods and extensions

3.3.1 Linear regression

Linear regression is one of the simplest ML models used for regression tasks. In linear regression, the label of a sample i, denoted y_i, is assumed to be the result of a linear combination of the values of the feature vector $\boldsymbol{x}_i$ of said sample:

$$\begin{aligned} y_i^{pred} &= \boldsymbol{w}^\top \boldsymbol{x}_i + b \\ &= b + \sum_{j=1}^{D} w_j \cdot x_{ij} \end{aligned} \tag{3.2}$$

where D is the total number of features, j is the feature index, $\boldsymbol{w}$ is the weight vector of length D containing the weights for each feature, b is the intercept or bias, $\boldsymbol{x}_i$ is a vector of length D containing the features' values for sample i and y_i^{pred} is the label predicted by the regression model. An overview of the notation used in this section is provided in Table 3.1.

The linear model receives a feature vector $\boldsymbol{x}_i$ as input, and produces a predicted label y_i^{pred} as output for a given $\boldsymbol{w}$ and b. For the model to be accurate, y_i^{pred} needs to approximate the real label y_i as close as possible. During training, $\boldsymbol{w}$ and b are tuned to minimise the so-called loss $\mathcal{L}$ using a set of N samples. For a linear regression model, the loss is typically determined as the mean squared

symbol	description
j	feature index
D	number of features
i	sample index
N	number of samples
$\boldsymbol{x}_i$	vector of length D containing the features' values for sample i
$\boldsymbol{w}$	vector of length D containing the weights for all D features
b	intercept value
y_i	label value for sample i
y_i^{pred}	predicted label value for sample i

Table 3.1: Overview of the variables involved with a linear regression model.

error on the training set.

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i^{pred} - y_i \right)^2 \tag{3.3}$$

Linear models fail in situations where the link between the features and the label is non-linear or where features interact with each other. Here, it can help to transform the features or the label in a non-linear manner e.g. replacing a feature by the square or logarithm of its value. For instance, linear models in drug discovery often include the logarithm of the n-octanol:water partition coefficient as a feature.

3.3.2 Logistic regression

Linear models can also be used for binary classification tasks with some modification. In binary classification, a model learns to predict whether a data instance i belongs to a certain class (y_i=1) or not (y_i=0). In logistic regression, a weighted average of the input features $\boldsymbol{w}^\top \boldsymbol{x}_i + b$ is passed to a sigmoid function $\sigma(z)$:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.4}$$

The output from the sigmoid function is interpreted as the probability of the data instance belonging to the class, given the values of the features and a given set of weights:

$$p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}, b) = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \tag{3.5}$$

Once the probability for a data instance is predicted, a decision border for the probability has to be chosen. By default, a probability of 50% is used as a threshold:

$$y_i^{pred} = \begin{cases} 1 & \text{if } p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}, b) \geq .5 \\ 0 & \text{if } p(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}, b) < .5 \end{cases} \tag{3.6}$$

This threshold value can be changed to balance the relative importance of false positives and false negatives. Increasing the probability threshold results in less false positives at the cost of more false negatives. The primary limitation of logistic regression is the assumption that the data is linearly separable. Nonetheless, logistic regression can be sufficient for simple classification tasks. Furthermore, the decision making of the model is transparent and, therefore, easy to interpret.

3.3.3 Regularisation

Linear models are susceptible to fitting the training set too strongly, especially when the number of features is significant or the features are only partially independent. Models that are overfitted typically have large weights for many features. The large weights make the model overly sensitive to low information features and less stable. Regularisation limits overfitting by introducing a penalty term to the loss function. During training, the model balances the accuracy of the prediction and the size of feature weights. As the model is punished for having large weights, the obtained solution after training is often simpler, i.e. the predictions of the model are mainly based on a handful of features. For this reason, regularisation is also used as a tool for feature selection. Aside from linear models, regularisation is also used to limit overfitting in other ML methods such as neural networks (see section 3.4.3). There are two commonly used types of regularisation, namely L2- and L1-regularisation. Both regularisation types are frequently used for linear and logistic regression models.

L2-regularisation

In L2-regularisation, the penalty term is the L2-norm of the model's weight vector. The penalty term is then added to the loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i)^2 \ + \alpha_{L2}\sum_{j=1}^{D} w_j^2 \tag{3.7}$$

Here, the hyperparameter α_{L2} governs the importance of the penalty term and is called the L2 regularisation strength. Choosing a proper α_{L2} value is not trivial and often different values need to be considered. To illustrate the optimisation of α_{L2}, let us consider two cases where a very large and a very small α_{L2} value are used. A large α_{L2} will focus the model training on minimising only the weights, resulting in a large difference between the predicted and real labels and thus an underfitted model. For a small α_{L2} the resulting model is indistinguishable from a non-regularised linear regression model. If the absence of the regularisation results in overfitting, there is typically an optimum α_{L2} which balances underfitting and overfitting. Finding this optimum requires training the model with different values for the hyperparameter and evaluate the resulting models on the validation set.

L1-regularisation

In L1-regularisation, the sum of the absolute values of the weights is chosen as penalty term:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i)^2 \ + \alpha_{L1}\sum_{j=1}^{D} |w_j| \tag{3.8}$$

The L1-regularisation strength α_{L1} controls the importance of the regularisation. The value of α_{L1} is tuned as a hyperparameter of the model during optimisation. L1-regularisation favours solutions where many weights are equal to zero, inducing more sparsity in the ML model upon increasing α_{L1}. Sparse models predict labels using only the most important features. Therefore, L1-regularised models have feature selection included as a part of the training process. L2-regularisation also reduces the size of the weights, but does not set them to zero.[65]

Combining L1- and L2-regularisation

If the intended regularisation lies in between L1- and L2-regularisation, both terms can be added to the loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i)^2 \ + \alpha_{L1}\sum_{j=1}^{D}|w_j| + \alpha_{L2}\sum_{j=1}^{D}w_j^2 \tag{3.9}$$

Here, changing α_{L1} and α_{L2} allows to tune and balance the importance of each regularisation type. A downside of including both is the increased complexity of the optimisation with the additional hyperparameter.

Influence of scaling

For both L1- and L2-regularisation, the penalty terms are a non-weighted sum. The implicit assumption is made that the weights, and thus the features, are provided in the same unit of measurement. In many applications, the features are provided in different units. The features of the wine dataset are expressed in different units and cover different ranges as shown in Table 3.2. If these features are not standardised, regularisation will result in a model that focuses on the proline content and ignores the non-flavanoid phenols. Even if these features are expressed using the same physical units, a difference of 0.5 mg/mL in the concentration of non-flavanoid phenols or proline are not directly comparable. Therefore, it is common practice to standardise each feature by subtracting the mean and scaling to unit variance. Doing so, the features are expressed in similar units.

feature	range of values
proline (mg/L)	280-1680
alcohol (%v/v)	11.0-14.8
OD_{280}/OD_{315}	1.3-4.0
non-flavanoid phenols (mg/L)	0.13-0.66

Table 3.2: Different scales of some features in the wine dataset.

3.4 Feedforward neural networks

Neural networks are a family of ML algorithms that draw inspiration from how the brain operates. In the brain, pattern recognition results from the connections between neurons. Signals are passed between connected neurons and these information flows are combined in a non-linear manner. Neural networks mimic this architecture, with interconnected artificial neurons that transform incoming signals using a non-linear function. The number of neurons in neural networks can be immense, such as the 100 billion neurons in ChatGPT, to extract very complex patterns. This structure enables neural networks to extract complex patterns from data and makes them the most flexible models in the ML toolbox. As a result, neural networks are effectively seen as universal function approximators.[66] Neural networks come in many forms, each optimised for tackling different tasks. This section will mainly tackle Feedforward Neural Networks (FNNs) with fully-connected layers. The emphasis will lie on introducing the concepts needed to interpret the results presented in Chapters 4-5.

3.4.1 Basic architecture

The simplest unit in the FNN's architecture is the perceptron, which takes in information from multiple input neurons and generates a non-linear response. This is done by taking a weighted average of the input $\boldsymbol{x}_i$ containing the feature values. The weighted average is then transformed with a non-linear function $g(\cdot)$ into a scalar output y_i^{pt}:

$$y_i^{pt} = g(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \tag{3.10}$$

where the vector $\boldsymbol{w}$ contains the weights and b is the bias. Training a perceptron essentially involves finding the optimal values of the weights and the bias. These weights represent the different connections between the input neurons containing the values of each feature and the output neuron, as shown in Figure 3.5. The non-linear function $g(\cdot)$ is referred to as the activation function. The exact nature of the activation function will be discussed later. Interestingly, the logistic regression covered in section 3.3.2 can be interpreted as a perceptron with a sigmoid activation function.

The predictive power of a single perceptron is limited and the model can only

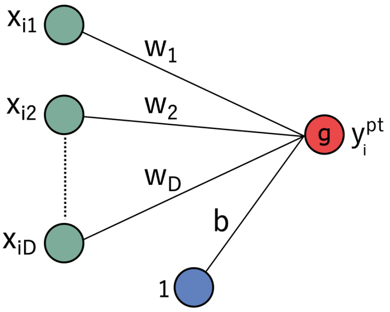


Figure 3.5: Structure of a single perceptron. The model takes in the feature vector $\boldsymbol{x}_i$ (with D features) of a sample i and generates a non-linear response y_i^{pt}.

produce a scalar output. By introducing multiple output nodes and connecting them to the input neurons, the model outputs a vector instead (see Figure 3.6). This model is a multi-output perceptron, where each input neuron is connected to each output neuron. These connections are determined for each output neuron separately with different weights and values of b. For a multi-output perceptron with D input neurons (i.e. features) and M output neurons, a weight matrix $\boldsymbol{W}$ of dimensions $D \times M$ and bias vector $\boldsymbol{b}$ of length M are defined. The output vector $\boldsymbol{y}_i^{sl}$ for a given input vector $\boldsymbol{x}_i$ then becomes:

$$\boldsymbol{y}_i^{sl} = g(\boldsymbol{W}^\top \boldsymbol{x}_i + \boldsymbol{b}) \tag{3.11}$$

The model can be seen as consisting of two layers, namely an input layer and output layer. The multi-output perceptron is still limited in which patterns it can learn, as it consists of M distinct perceptrons.

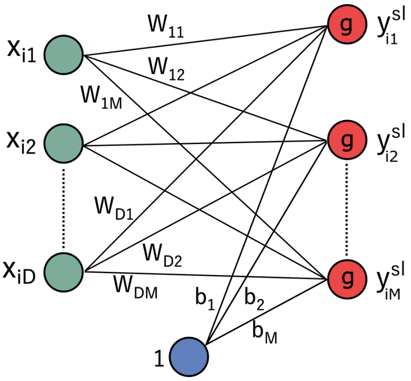


Figure 3.6: Structure of a multi-output perceptron. The model takes in the feature vector $\boldsymbol{x}_i$ (with D features) of a sample i and generates a vector $\boldsymbol{y}_i^{sl}$ of length M.

To improve the predictive power of the model, one or multiple layers of fully-connected neurons are inserted between the input and output layer. These layers are referred to as hidden layers and enable the model to learn more complex patterns. A shallow FNN contains a single hidden layer, whereas a deep FNN involves multiple hidden layers. By stacking the hidden layers, each additional hidden layer extracts more general patterns from the previous layer. The structure of an FNN with two hidden layers is illustrated in Figure 3.7.

The number of hidden layers and the number of neurons in each layer are hyperparameters of the model. As the hidden layers are fully-connected, a weight matrix $\boldsymbol{W}^{(l)}$ and bias vector $\boldsymbol{b}^{(l)}$ is defined for each layer (l). The output of layer (l), denoted $\boldsymbol{h}_i^{(l)}$, results from the activation function $g^{(l)}(\cdot)$ and the output of the previous layer $\boldsymbol{h}_i^{(l-1)}$:

$$\boldsymbol{h}_i^{(l)} = g^{(l)}(\boldsymbol{W}^{(l)\top}\boldsymbol{h}_i^{(l-1)} + \boldsymbol{b}^{(l)}) \tag{3.12}$$

where $\boldsymbol{h}_i^{(0)} = \boldsymbol{x}_i$ and $\boldsymbol{y}_i = \boldsymbol{h}_i^{(L)}$ for an FNN consisting of L layers.

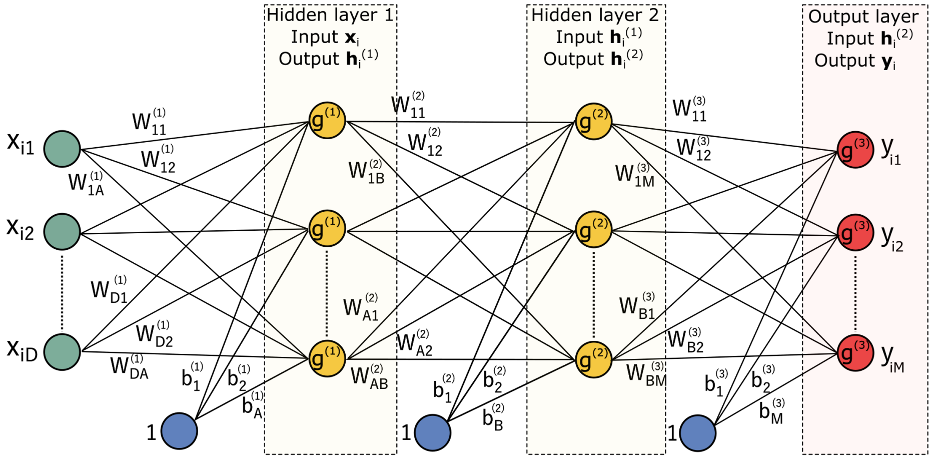


Figure 3.7: Structure of an FNN with two hidden layers. A and B are the number of neurons in hidden layer (1) and (2) respectively. The first hidden layer uses the matrix $\boldsymbol{W}^{(1)}$ and vector $\boldsymbol{b}^{(1)}$ to convert $\boldsymbol{x}_i$ into a vector $\boldsymbol{h}_i^{(1)}$ of length A. The second hidden layer converts $\boldsymbol{h}_i^{(1)}$ into a vector $\boldsymbol{h}_i^{(2)}$ of length B with the weight matrix $\boldsymbol{W}^{(2)}$ and bias vector $\boldsymbol{b}^{(1)}$. The output of the final layer $\boldsymbol{y}_i$, a vector of length M, is obtained with the weight matrix $\boldsymbol{W}^{(3)}$ and bias vector $\boldsymbol{b}^{(3)}$.

To illustrate this concept, consider the FNN with two hidden layers in Figure 3.7, which has A neurons in the first hidden layer and B neurons in the second hidden layer. For a given feature vector $\boldsymbol{x}_i$ of sample i the first hidden layer creates a vector $\boldsymbol{h}_i^{(1)}$ of length A:

$$\boldsymbol{h}_i^{(1)} = g^{(1)}(\boldsymbol{W}^{(1)\top}\boldsymbol{x}_i + \boldsymbol{b}^{(1)}) \tag{3.13}$$

For the first layer, the bias vector $\boldsymbol{b}^{(1)}$ has a length of A and $\boldsymbol{W}^{(1)}$ is a $(D \times A)$ matrix. The second layer uses the vector $\boldsymbol{h}_i^{(1)}$ as input, creating a new vector $\boldsymbol{h}_i^{(2)}$ of length B:

$$\boldsymbol{h}_i^{(2)} = g^{(2)}(\boldsymbol{W}^{(2)\top}\boldsymbol{h}_i^{(1)} + \boldsymbol{b}^{(2)}) \tag{3.14}$$

where the vector $\boldsymbol{b}^{(2)}$ has a length of B and $\boldsymbol{W}^{(2)}$ is a $(A \times B)$ matrix. The final layer of the FNN uses $\boldsymbol{h}_i^{(2)}$ to generate an output vector $\boldsymbol{y}_i$ of length M:

$$\boldsymbol{y}_i = g^{(3)}(\boldsymbol{W}^{(3)\top}\boldsymbol{h}_i^{(2)} + \boldsymbol{b}^{(3)}) \tag{3.15}$$

where the vector $\boldsymbol{b}^{(3)}$ has a length of M and $\boldsymbol{W}^{(3)}$ is a $(B \times M)$ matrix. The components of $\boldsymbol{W}^{(1)}$,$\boldsymbol{W}^{(2)}$,$\boldsymbol{W}^{(3)}$, $\boldsymbol{b}^{(1)}$,$\boldsymbol{b}^{(2)}$ and $\boldsymbol{b}^{(3)}$ are obtained through training. An example of the FNN applied to the wine dataset can be found in the appendix.

symbol	description
i	sample index
N	number of samples
D	number of features
A	number of neurons in layer (1)
B	number of neurons in layer (2)
M	number of output neurons
$\boldsymbol{x}_i$	the feature vector of length D for sample i, input for layer (1)
$\boldsymbol{W}^{(1)}$	$(D \times A)$ weight matrix of layer (1)
$\boldsymbol{b}^{(1)}$	bias vector of layer (1) of length A
$\boldsymbol{h}_i^{(1)}$	output vector of layer (1) of length A for sample i, input for layer (2)
$\boldsymbol{W}^{(2)}$	$(A \times B)$ weight matrix of layer (2)
$\boldsymbol{b}^{(2)}$	bias vector of layer (2) of length B
$\boldsymbol{h}_i^{(2)}$	output vector of layer (2) of length B for sample i, input for layer (3)
$\boldsymbol{W}^{(3)}$	$(B \times M)$ weight matrix of layer (3)
$\boldsymbol{b}^{(3)}$	bias vector of layer (2) of length M
$\boldsymbol{y}_i$	output vector of layer (3) of length M for sample i

Table 3.3: Overview of the variables involved with an FNN with two hidden layers.

3.4.2 Activation functions

At the heart of the FNN architecture lies the activation function. In the case of a linear activation function, the model would only create linear combinations of other linear combinations. By adding a non-linear activation function, the FNN can be taught complex patterns. An overview of the most popular activation functions is given in Figure 3.8. In the 1990s, the tanh and sigmoid function were typically chosen as activation functions. However, training deep neural networks using these activation functions proved problematic.[67] For most modern applications, the Rectified Linear Unit (ReLU) has become the default activation function.[68,69] Its success largely comes from its simplicity and its low-cost derivative. One problem encountered during FNN training is that neurons with ReLU activation can end up in a dead state. The dead neurons only generate '0' as output and their weights are no longer updated during training. To address this problem, alternative activation functions have been developed such as LeakyReLU[70], PReLU[71], ELU[72] and SELU[73]. While these activation functions can improve performance and model training, they are slower to compute or introduce an additional hyperparameter. Therefore, it can prove more beneficial to increase the number of neurons to compensate for the dead neuron problem. The output layer typically does not contain any activation function for a regression task. For classification tasks, the output layer uses an activation function such as the sigmoid function to limit the output values between 0 and 1.

3.4.3 Training & regularisation

During training, the weights and biases of the model are tuned to improve the accuracy of the predicted labels. After initialising the model weights[71], the model is trained by minimising the loss with an optimisation algorithm such as Adam[74]. The optimisation algorithm requires the gradient of the loss with regards to the FNN's parameters (weights and biases), which is obtained through backpropagation.[75] Optimisation algorithms typically contain 1-3 hyperparameters (e.g. the learning rate) to tune the learning process.

The large number of weights and biases make FNNs notably susceptible to overfitting. The flexibility of an FNN to fit to almost any pattern, allows it to

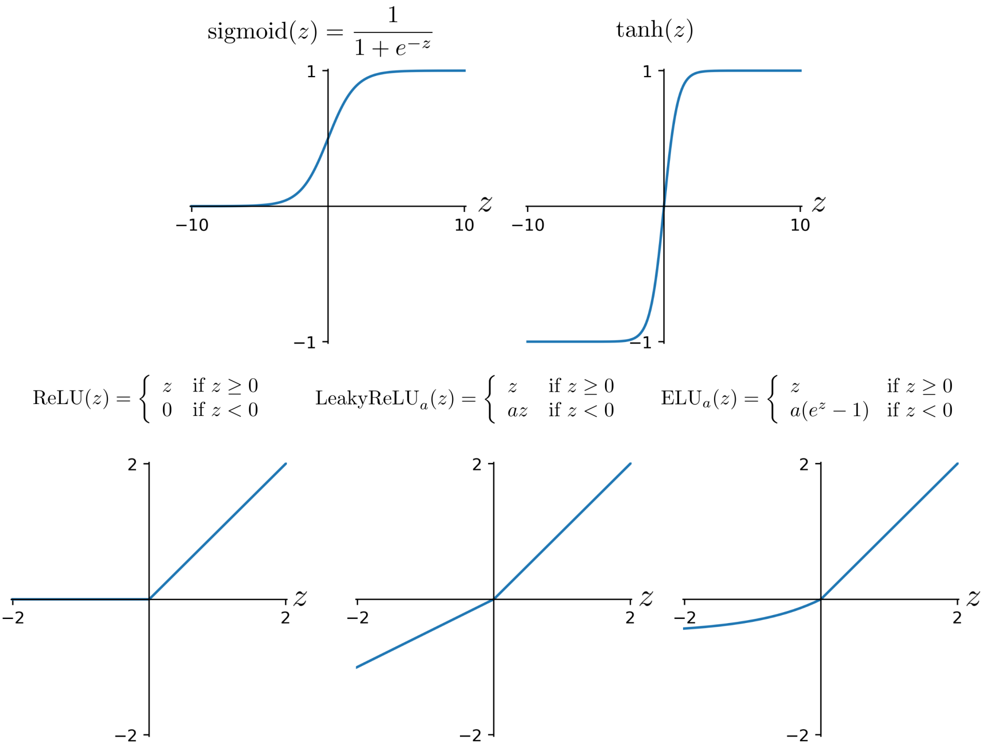


Figure 3.8: Overview of common activation functions used in FNNs. The hyperparameter a for LeakyReLU and ELU is set to 0.5.

learn spurious correlations, resulting in poor performance outside of the training set. When the network has too many parameters for the size of the dataset, the model starts learning the data instances by heart, instead of extracting a generalisable pattern.[76] To prevent this overfitting, the model has to be constrained during training. In most applications, the L1- or L2-norm of the weights is added to the loss as a penalty term, similar to the regularisation found in linear models (see section 3.3.3). The L1/L2-regularisation is combined with other regularisation methods, such as early stopping and dropout. In early stopping, the validation set is used to monitor overfitting. During each step of the training process, the performance of the model on the validation set is evaluated. When the validation loss has reached a minimum (i.e. does not improve), the model no longer learns generalisable patterns and the training process is stopped.

Another popular regularisation technique is Dropout[77,78]. Here, a randomly chosen subset of the neurons is temporarily made inactive, or ‘dropped out’, at every training step. The output of these neurons is set to zero and their weights are frozen, as illustrated in Figure 3.9. The neurons in the FNN learn to be less dependent on individual neurons and are less inclined to over-specialise. This results in a more robust FNN with improved performance on new data instances. The probability of a neuron to be dropped out is referred to as the dropout rate. Increasing this value makes the model less dependent on individual neurons, but also slows down the training process. Typical values for the hyperparameter range between 0.2 and 0.5.

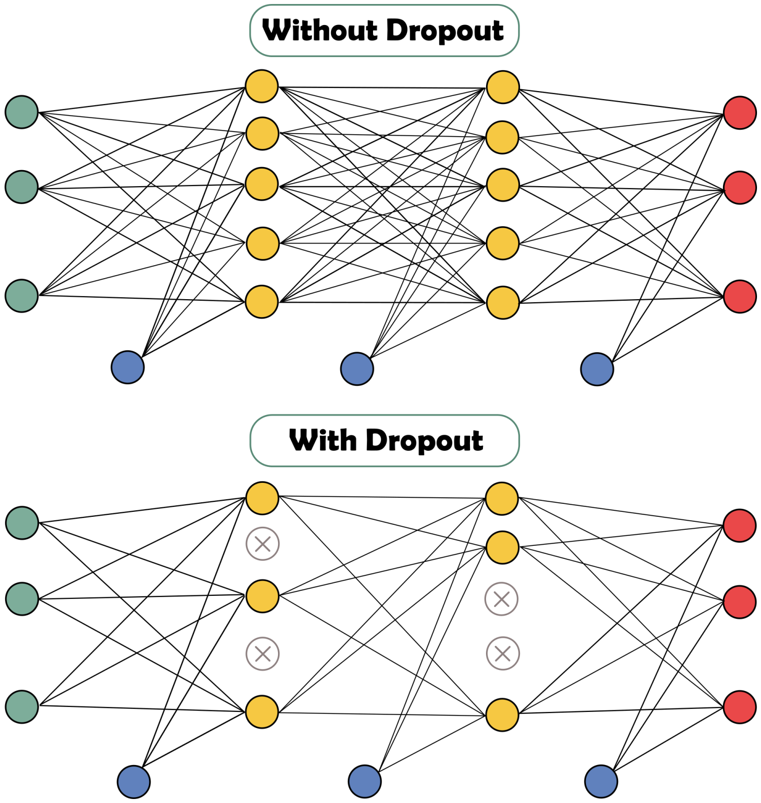


Figure 3.9: Regularisation of an FNN with two hidden layers with dropout.

3.5 ML applications for VCD

The research presented in this work explores the capacity of ML techniques to extract patterns in VCD spectra. In this context, the following three ML tasks are considered:

1. predicting the AC of a compound from their VCD spectrum

2. predicting the VCD spectrum of a conformer from its geometry

3. predicting the components present in mixtures from the VCD and/or IR spectrum.

In the first project, classification models are trained to predict the AC from a VCD spectrum. These models are trained on the VCD spectral dataset of decorated α-pinene structures. The application domain of these models is limited to the molecular motifs found in the dataset. Therefore, we are especially interested in the accuracy of these models achieved with smaller training set sizes. The cost of training the ML models for this project proves to be negligible compared to the cost of recording or computing a single VCD spectrum. As a result of the low cost, cross-validation can be incorporated in this project without any significant drawbacks. As a first step, we explore the dataset with unsupervised ML methods and simple white-box models, such as decision trees and linear models. This initial step allows us to gain a deeper understanding of the underlying VCD patterns in the data and assess the complexity of the ML task. Additionally, these white box-models enable us to identify the VCD patterns that are characteristic for the AC. As we lack prior knowledge on which classification models are most suitable for this task, a substantial number of ML methods are evaluated on the spectral dataset. The best performing models are selected for further investigation on smaller training sets and on lower-dimensional representations, obtained with unsupervised ML methods or RF feature ranking, of the VCD spectra. The resulting models are also tested for any bias towards certain molecular structures. Finally, the robustness of the proposed method is evaluated by assessing its performance -after retraining the models- on spectra obtained with different computational settings.

The second project involves a multi-output regression task: predicting the VCD intensities at multiple wavenumbers for conformers of the same compound. The molecular geometries are represented as a set of intramolecular coordinates describing the molecular flexibility and the intramolecular interactions in the compound. The application domain of a trained model is the set of conformers that the molecule can adapt. Given the complexity of predicting VCD spectra of conformers, multi-output FNNs are chosen as model for their ability to identify and learn complex patterns. Given the black-box nature of the FNN, extracting any chemically meaningful information from a model is challenging, if not impossible. To learn more about the shortcomings of the suggested approach, the FNN is trained on six different compounds. These model compounds are chosen such that lower performance of the model can be traced back to characteristics of the compound, such as intramolecular interactions. The speed-up achievable with the FNN depends strongly on the fraction of conformers present in the training set. Therefore, the accuracy of the FNN predicted spectra is evaluated for different training set sizes. The FNN's hyperparameters are optimised with Bayesian optimisation. Doing so, the model optimisation is automated and can be easily accounted for in the reported speed-up. Additionally, restarting the optimisation of each model from scratch prevents information leaking between different optimisation runs. To limit the cost of training and optimising the FNNs, cross-validation is not used in this project. The resulting FNNs are checked for any bias towards the presence/absence of intramolecular interactions in conformers or their relative energy. The performance of the approach is then evaluated within the context of the AC determination workflow. Here, the accuracy of the Boltzmann weighted molecular VCD spectrum obtained with the FNN is compared to the one consisting of only DFT computed spectra. Finally, the possibility of applying these FNN outside of their original application domain is explored by predicting spectra for different AC's.

In the final project, we test whether supervised ML models can help determine the composition of monoterpene mixtures. The IR/VCD spectrum of such a mixture can be seen as, approximately, a linear combination of the monoterpene spectra. Therefore, the use of linear ML models is an obvious choice for this ML task. Given the low cost of these models, cross-validation can be implemented in the workflow without any substantial drawbacks. Ideally, the Ml model is trained on an extensive set of mixture spectra. However, constructing such a dataset of sufficient size costs an enormous amount of effort. Therefore, the ML model is trained instead on a set of in-silico mixture spectra - noisy linear combinations of the monoterpene spectra. As a result, we can explore this ML application with only having to record the pure monoterpene spectra. After training, the performance of the model is evaluated on a small set of monoterpene mixtures of known composition. In a final stage, the resulting model is used to detect the monoterpenes in essential oils.

References

[1] C. M. Bishop, *Pattern recognition and machine learning*, Springer, Cambridge, United Kingdom, 2016.

[2] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow*, O'Reilly Media, Inc., Sebastopol, CA, 2nd edn, 2019.

[3] I. Goodfellow, Y. Bengio and C. Aaron, *Deep Learning*, MIT Press, Cambridge, MA, 2016.

[4] *Machine Learning in Chemistry*, ed. H. M. Cartwright, The Royal Society of Chemistry, Cambridge, United Kingdom, 2020.

[5] F. Chollet, *Deep learning with python*, Manning Publications, Shelter Island, NY, 2018.

[6] K. P. Murphy, *Machine Learning*, MIT Press, Cambridge, MA, 2012.

[7] *Machine Learning Meets Quantum Physics*, ed. K. Schütt, S. Chmiela, A. von Lilienfeld, A. Tkatchenko, K. Tsuda and K.-R. Müller, Springer, Cham, Switzerland, 2020.

[8] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, *Decis. Support Syst.*, 2009, **47**, 547–553.

[9] OpenAI, *GPT-4 Technical Report*, 2023, arXiv: 2303.08774.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, *Language Models are Few-Shot Learners*, 2020, arXiv: 2005.14165.

[11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, *Hierarchical Text-Conditional Image Generation with CLIP Latents*, 2022, arXiv: 2204.06125.

[12] *Midjourney*, `https://midjourney.com`, Accessed: 2023-05-02.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, 2022, arXiv: 2112.10752.

[14] P. D. König, T. D. Krafft, W. Schulz and K. A. Zweig, in *Essence of AI: What Is AI?*, Cambridge University Press, 2022, pp. 18–34.

[15] J. McCarthy, M. L. Minsky, N. Rochester and C. E. Shannon, *AI Mag.*, 2006, **27**, 12.

[16] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, *Nature*, 2016, **529**, 484–489.

[17] P. McCorduck, *Machines who think*, A K Peters, Natick, MA, 2nd edn, 2004.

[18] M. Haenlein and A. Kaplan, *Calif. Manage. Rev.*, 2019, **61**, 5–14.

[19] A. Krizhevsky, I. Sutskever and G. E. Hinton, Advances in Neural Information Processing Systems, 2012.

[20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

[21] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, *J. Big Data*, 2021, **8**, 53.

[22] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.

[23] W. P. Walters and R. Barzilay, *Acc. Chem. Res.*, 2021, **54**, 263–270.

[24] M. Xu, J. Cheng, Y. Liu and W. Huang, 2021 IEEE Symposium on Computers and Communications (ISCC), 2021, pp. 1–6.

[25] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, *Wiley Interdiscip. Rev. Comput. Mol.*, 2022, **12**, e1608.

[26] J. Polanski, *Int. J. Mol.*, 2022, **23**, 2797.

[27] A. E. Blanchard, C. Stanley and D. Bhowmik, *J. Cheminformatics*, 2021, **13**, 14.

[28] T. Karras, S. Laine and T. Aila, *A Style-Based Generator Architecture for Generative Adversarial Networks*, 2019, arXiv: 1812.04948.

[29] A. Ranjan, T. Bolkart, S. Sanyal and M. J. Black, Proceedings of the European conference on computer vision (ECCV), 2018, pp. 704–720.

[30] C. Rudin, *Nat. Mach. Intell.*, 2019, **1**, 206–215.

[31] C. Rudin and B. Ustun, *Interfaces*, 2018, **48**, 449–466.

[32] W. Samek, T. Wiegand and K.-R. Müller, *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 2017, **1**, 1–10.

[33] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai and t. P. consortium, *BMC Med. Inform. Decis. Mak.*, 2020, **20**, 310.

[34] A. B. Tosun, F. Pullara, M. J. Becich, D. L. Taylor, J. L. Fine and S. C. Chennubhotla, *Adv. Anat. Pathol.*, 2020, **27**, 241–250.

[35] H. R. Tizhoosh and L. Pantanowitz, *J. Pathol. Inform.*, 2018, **9**, 38.

[36] B. Acs, M. Rantalainen and J. Hartman, *J. Intern. Med.*, 2020, **288**, 62–81.

[37] S. Kundu, *Nat. Med.*, 2021, **27**, 1328–1328.

[38] A. Deeks, *Columbia Law Review*, 2019, **119**, 1829–1850.

[39] P. Weber, K. V. Carl and O. Hinz, *Manag. Rev. Q.*, 2023, N.A.

[40] B. Goodman and S. Flaxman, *AI Mag.*, 2017, **38**, 50–57.

[41] M. T. Ribeiro, S. Singh and C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 2016, arXiv: 1602.04938.

[42] H. Lakkaraju, E. Kamar, R. Caruana and J. Leskovec, *Interpretable & Explorable Approximations of Black Box Models*, 2017, arXiv: 1707.01154.

[43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, 2016, arXiv: 1502.03044.

[44] M. T. Ribeiro, S. Singh and C. Guestrin, *Model-Agnostic Interpretability of Machine Learning*, 2016, arXiv: 1606.05386.

[45] H. K. B. Babiker and R. Goebel, *An Introduction to Deep Visual Explanation*, 2018, arXiv: 1711.09482.

[46] G. Montavon, W. Samek and K.-R. Müller, *Digit. Signal Process.*, 2018, **73**, 1–15.

[47] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian and D. Dou, *Knowl. Inf. Syst.*, 2021, **64**, 3197 – 3234.

[48] G. Ras, N. Xie, M. van Gerven and D. Doran, *Explainable Deep Learning: A Field Guide for the Uninitiated*, 2021, arXiv: 2004.14545.

[49] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, *A Survey on Bias and Fairness in Machine Learning*, 2022, arXiv: 1908.09635.

[50] Z. Obermeyer, B. Powers, C. Vogeli and S. Mullainathan, *Science*, 2019, **366**, 447–453.

[51] J. Zou and L. Schiebinger, *Nature*, 2018, **559**, 324–326.

[52] R. Courtland, *Nature*, 2018, **558**, 357–360.

[53] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson and D. Sculley, *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World*, 2017, arXiv: 1711.08536.

[54] J. Buolamwini and T. Gebru, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018, pp. 77–91.

[55] B. Wilson, J. Hoffman and J. Morgenstern, *Predictive Inequity in Object Detection*, 2019, arXiv: 1902.11097.

[56] B. Seligman, S. Tuljapurkar and D. Rehkopf, *SSM - Popul. Health*, 2018, **4**, 95–99.

[57] X. Ying, *J. Phys.: Conf. Ser.*, 2019, **1168**, 022022.

[58] M. M. Bejani and M. Ghatee, *Artif. Intell. Rev.*, 2021, **54**, 6391–6438.

[59] T. F. Sterkenburg and P. D. Grünwald, *Synthese*, 2021, **199**, 9979–10015.

[60] J. Bergstra, D. Yamins and D. Cox, Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013, pp. 115–123.

[61] T. Hastie, R. Tibshirani and J. H. Friedman, *The elements of statistical learning*, Springer, New York, NY, 2nd edn, 2009.

[62] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification And Regression Trees*, Routledge, Boca Raton, FL, 1984.

[63] L. Breiman, *Mach. Learn.*, 1996, **24**, 123–140.

[64] L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.

[65] S.-J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, *IEEE J. Sel. Top. Signal Process.*, 2007, **1**, 606–617.

[66] G. Cybenko, *Math. Control Signals Syst.*, 1989, **2**, 303–314.

[67] X. Glorot and Y. Bengio, *J. Mach. Learn. Res.*, 2010, **9**, 249–256.

[68] B. Xu, N. Wang, T. Chen and M. Li, *Empirical Evaluation of Rectified Activations in Convolutional Network*, 2015, arXiv: 1505.00853.

[69] P. Ramachandran, B. Zoph and Q. V. Le, *Searching for Activation Functions*, 2017, arXiv: 1710.05941.

[70] A. Y. N. Andrew L. Maas, Awni Y. Hannun, Proceedings of the 30th International Conference on Machine Learning, 2013, p. 3.

[71] K. He, X. Zhang, S. Ren and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, 2015, arXiv: 1502.01852.

[72] D.-A. Clevert, T. Unterthiner and S. Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*, 2016, arXiv: 1511.07289.

[73] G. Klambauer, T. Unterthiner, A. Mayr and S. Hochreiter, *Self-Normalizing Neural Networks*, 2017, arXiv: 1706.02515.

[74] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2017, arXiv: 1412.6980.

[75] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533–536.

[76] I. Wallach and A. Heifets, *Journal of Chemical Information and Modeling*, 2018, **58**, 916–932.

[77] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, 2012, arXiv: 1207.0580.

[78] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *J. Mach. Learn. Res.*, 2014, **15**, 1929–1958.

Chapter 4

Exploring machine learning methods for absolute configuration determination with vibrational circular dichroism

4.1 Introduction

Plenty of natural chemical compounds are chiral and their stereoisomers tend to interact differently with other chiral compounds. This is of great importance in for instance medicinal chemistry, where stereoisomers produce different therapeutic effects when engaging their chiral biological target. As a consequence, methods capable of reliably identifying the absolute configuration (AC) of these compounds are of high interest.[1,2] Probably the best known method is X-ray diffraction. This method, however, requires single crystals which are not always easily available or require additional manipulations. NMR does not distinguish enantiomers and so its use requires derivatisation of the compounds.[3–5]

Stereoisomers do not only interact differently with other chiral compounds but with chiral fields in general. This difference in interaction is exploited in so-called Circular Dichroism (CD) methods. There the difference is measured between the interaction of a specific compound with left- and right-circularly polarised radiation.[6] Probably the best-known CD method is electronic circular dichroism (ECD). This is the chiral counterpart of UV-VIS spectroscopy and hence relies on transitions in electronic state and requires the presence of chromophores. Infrared spectroscopy also has a chiral counterpart, known as Vibrational Circular Dichroism (VCD). As there are many more and better resolved vibrational transitions than there are electronic transitions in VCD and ECD respectively, VCD spectra usually offer much richer information to extract the AC from experimental spectra.[7,8] Moreover, VCD has the important advantage that it does not require single crystals, elaborate derivatisation or the presence of chromophores.

CD methods encapsulate the difference between enantiomers in a very simple way: the CD spectra of enantiomers are each other's mirror image. If one enantiomer slightly prefers to absorb left circularly polarised light at a specific wavelength, the other enantiomer will show the same size preference for right circularly polarised light at that same specific wavelength. Unfortunately, there is no easy way to link a spectrum to an AC using e.g. tabulated characteristics or empirical rules.[9] Methods such as VCD therefore benefited greatly from the advent of efficient algorithms to quantum chemically reliably compute VCD spectra for a chosen AC of a compound.[10] If the computed spectrum matches to sufficiently large extent the experimental spectrum, a confident assignment can be made.[8] Experience shows that Density Functional Theory (DFT) calculations with a well-chosen functional and basis set often give satisfactory agreement between theory and experiment. Where needed, many extensions to these calculations, such as proper solvent handling or ways to concentrate on the essential parts of a molecule may help make calculations better or even simply affordable.[11–16]

As mentioned, empirical rules for AC assignment from an experimental spectrum remain unknown. The current alternative is to compute spectra which must be done for every molecule and even conformer thereof separately. This requires much extra expertise and is both time and resource consuming.

This paper therefore explores a third way. Our research hypothesis is that Machine Learning (ML) techniques can extract yet unknown spectral features from VCD spectra and in this way allow determining the AC of new compounds. As the main strength of VCD lies in its ability to identify enantiomers, the study focuses on distinguishing enantiomers. Machine Learning (ML) methods have already been applied successfully in different areas in chemistry, including spectroscopy,[17–35] but not VCD spectroscopy. What follows is, to the best of our knowledge, the first critical and elaborate investigation of the performance of ML methods to extract AC from VCD spectra.

4.2 Methodology

4.2.1 Database design

Our research methodology is based on the following observation: the AC of a compound is encapsulated in its VCD spectrum although in a rather opaque way. On the other hand, it is not unlikely that similar molecules with the same AC would also encapsulate this information on the AC in a similar way. We propose to use ML techniques to establish whether these techniques actually show that the AC is encoded in VCD spectra in a tractable way for ML techniques. Beyond establishing this, we wish to examine whether ML can learn enough from a sufficiently large dataset to allow determining the AC for new similar molecules. In the following sections, we present in detail the methodology on how we prove that our central hypothesis actually holds.

As first step, we compose a database of spectral data. This dataset should contain sufficient information to allow ML techniques to extract the necessary knowledge to be able to assign the AC. Ideally, one would have access to a wealth of experimental spectra and use these as input. However, there are some problems with this approach. On the one hand, there is simply not enough data available and measuring more spectra comes at too high a cost. Second, for each spectrum one needs rock solid proof that the AC is known. This requires cross checking this information with at least another method, such as another spectroscopic method, or through the synthesis pathway. Both reasons entail that working with experimental spectra is not an option.

Theoretically computed spectra do not suffer these problems. One has without any doubt certainty of the absolute configuration chosen. Therefore, we here use, instead of experimental spectra, DFT computed spectra for a set of rigid compounds where solvent effects are expected to play a minor role. By only considering rigid compounds, any accumulation of errors from the conformational VCD spectra, along with the corresponding Boltzmann weights, can be prevented. Such an accumulation may in an unpredictable fashion impact on the conclusions on the performance of ML methods. One would obviously also want to include all possible elements, functional groups, etc. However, we largely exclude functional groups that can interact strongly with their environment. Even though DFT calculations on molecules with such functional groups pose no problem and the spectra could technically be used, the chemical value of the spectra is limited so we chose not to use them. Obviously, once experimental spectra become available in sufficient numbers, the dataset could be extended to also include flexible molecules, molecules that interact strongly with the environment etc. albeit that then the challenge is to have absolute certainty on the AC of the experimental sample. As will be discussed in section 4.3.1, the potential lower chemical diversity introduced by using computed spectra does not impact the diversity of the spectra themselves. We stress that the only role played by DFT calculations here is to generate the database and it is in no way used in the spectral analysis, as only ML techniques are considered there. So, the DFT calculations are used as generators of data and not as analysers of data.

α-pinene is a well-known standard reference compound in the VCD and ROA community. Due to its rigidity and minor solvent dependence of its spectra, the VCD spectrum can be calculated reliably using DFT methods.[36–38] In this work, we have chosen to use the skeleton of α-pinene as a scaffold to generate a very large number of other compounds by introducing a wide diversity of side chains. These side chains, shown in Figure 4.1, were substituted on six different carbon atoms in the scaffold, generating all possible substitution pattern combinations.

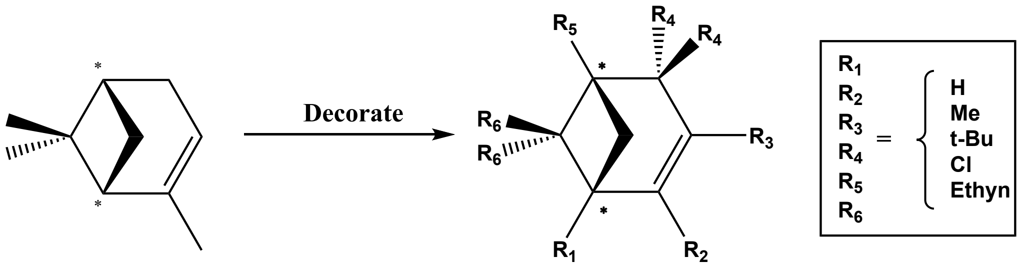


Figure 4.1: Decoration of the core structure of (−)-α-pinene. The carbon atoms involved are similarly to R_{1-6} defined as C_{1-6}.

Some restrictions were then applied. First, to avoid the creation of additional chiral centers, both C_4 and C_6 were always substituted twofold with the same substituent. Additionally, hydrogen was not used as substituent at C_6, to prevent rendering the compound achiral. Thirdly, structures with strong steric repulsion were excluded from the database. As such, structures that contained interatomic distances between their side chains smaller than 0.75 Å were omitted. This resulted in 3945 molecules sharing the (−)-α-pinene core, for which the VCD spectra were calculated. The spectra of the molecules sharing the (+)-α-pinene core were obtained by mirroring the calculated spectra of the corresponding enantiomers. The label used to identify the AC of the molecule was whether the molecule was based on the (−)- or the (+)-α-pinene core structure. CIP-rules were not used as the molecule contains two asymmetric carbons, labelled as (S,S) and (R,R) respectively for α-pinene, whose labelling can change for different decorations.

It should be noted that an imbalance in the dataset with respect to the relative presence of certain substituents has been introduced due to the abovementioned omission of certain structures based on steric clashes, as illustrated in Figure 4.2. This can leave certain structures underrepresented and more difficult to accurately classify with ML models. The relative presence of *t*-butyl is influenced the most, as it is the bulkiest substituent. Its complete absence at C_4 and C_6 will not impact the performance measure, as the model is not validated on structures decorated by *t*-butyl on these positions. However, its strong underrepresentation at C_1 might not provide enough samples in order for the ML model to process the influence that it can have on the VCD spectrum. An analysis of this is provided in the Supporting Information†.

4.2.2 Computational DFT settings

For the 3945 decorated (−)-α-pinene structures, geometry optimisation and subsequent gas phase VCD calculations were performed at B3PW91/6-31++G(d,p) level using Gaussian16[39]. Lorentz broadening was performed on the resulting line spectra, using a Full Width at Half Maximum (FWHM) of 10 cm^{-1}, ranging from 800 cm^{-1} to 1800 cm^{-1} with a sampling interval (SI) of 0.5 cm^{-1}.

4.2.3 ML methods

To fully gauge the capabilities of ML methods for VCD spectroscopy, multiple supervised and unsupervised methods were considered. These are introduced

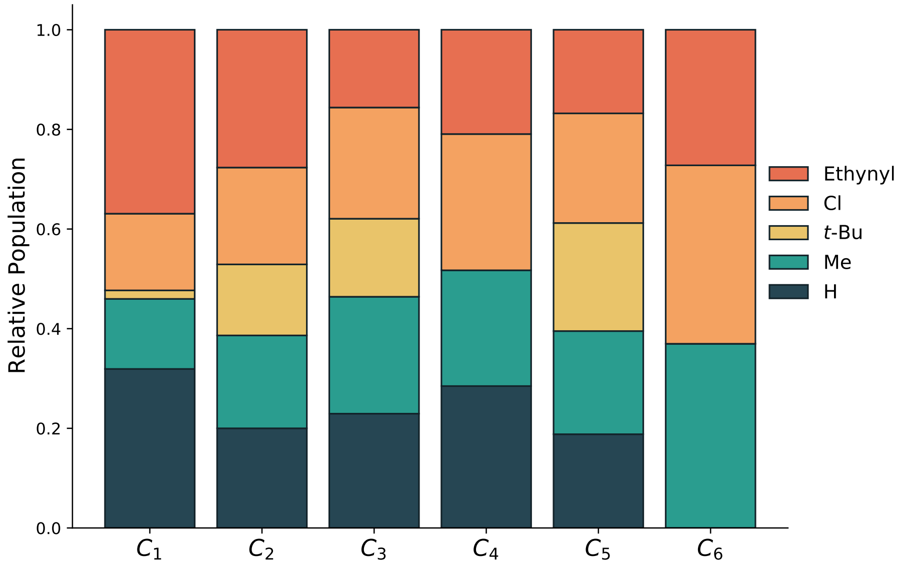


Figure 4.2: Relative representation of the substituents on the six different sites C_{1-6}.

succinctly below with their main features and, where applicable, the so-called hyperparameters that were optimised. For a more detailed description, we refer to the documentation of scikit-learn[40].

Principal component analysis (PCA)[41]

Principle: PCA is a linear method of dimensionality reduction that finds projections into lower-dimensional subspaces, such that the variance captured in these spaces is maximised. After this, dimensional reduction can be performed by only using the first n orthogonal components, which would capture the largest section of the variation of the data.
Hyperparameters: Not applicable.

t-Stochastic neighbour embedding (t-SNE)[42]

Principle: t-SNE is a method for visualisation of high-dimensional data that can model complex, non-linear dependencies. A distribution over pairs of samples is constructed both in the original and an embedding space. Divergence between the two distributions is minimised such that samples similar in the original space are placed close together in the embedding space with a high probability.

Hyperparameters: measure of perplexity, exaggeration.

Decision tree

Principle: A tree-structured model with class labels in leaves and descriptive features in branches. Trees are induced by recursively splitting the dataset in smaller subsets in each branch, such that the purity of the data (i.e. homogeneity of labels) in the leaves is maximised.
Hyperparameters: Tree depth.

Logistic regression (LogReg)

Principle: The method applies the techniques of linear regression to classification problems. A logistic function is fitted to rep-resent the probability of the sample belonging to a certain class. The predictive capabilities are typically improved by employing a regularisation method, such as lasso (l1)[43] and ridge (l2)[44] regularisation, to penalise large weights in regression.
Hyperparameters: Regularisation method and strength.

Naive Bayes (NB)[45]

Principle: A probabilistic method that uses Bayes' theorem to estimate the probability of a sample belonging to a certain class. The approach relies on a strong assumption that the attributes are conditionally independent.
Hyperparameters: Not applicable.

Support vector machines (SVM)[46]

Principle: A class of linear algorithms that finds a hyperplane separating two classes of data with as wide a margin as possible. Non-linear classification can be performed efficiently by mapping the inputs into high-dimensional feature spaces through invertible mathematical operations.
Hyperparameters: Kernel employed for mapping, cost, soft or hard margin.

k-Nearest neighbours (kNN)[47]

Principle: Each sample is classified to the class, most common among the k training points that are the closest to the sample according to a distance measure, such as the euclidean distance.
Hyperparameters: Number of neighbours, distance metric and weight.

Random forest (RF)[48]

Principle: An ensemble learning technique that constructs a large number of decision tree classifiers. Each tree is trained on a limited bootstrap sample from the original dataset. Furthermore, at each branch of the tree, only a restricted and random subset of features is considered. Each sample is classified according to a majority vote among the classifications of the individual trees. The relative importance of each feature for the model can be evaluated as the total increase in purity brought by that feature.
Hyperparameters: Number of trees, maximal tree depth.

Feedforward neural network (FNN)[49]

Principle: The data is classified by using a large network of interconnected artificial neurons, whose outputs are a non-linear function of the weighted sum of their inputs. The first layer of this network is the input layer, containing the input spectral data, and the final layer of this network is the output layer, giving the probability of belonging to a certain class. The inner layers, the so-called deep layers, construct complex features as every neuron combines the outputs of all the neurons in the previous layer in a non-linear manner.
Hyperparameters: Number of layers, number of neurons in each layer, optimiser, regularisation strength.

4.2.4 Model training

Each model was trained to classify the AC of the decorated molecules. As input the VCD intensity at every wavenumber is used and the output of the model is the AC label of the decorated molecule. The performance of the ML method is assessed based on the AC predicted versus the (known) true AC. For each model, the hyperparameters were optimised. To even out the probability that by chance a validation set would be used that is in any respect an outlier, 10 training and validation sets were used. In each case, 90% of the molecules were randomly included in the training set and the remaining 10% in the validation set. Equal representation of both enantiomers was imposed in each set. This will be referred to as a 9:1 training-validation split. The performance of each model was evaluated using the Classification Accuracy (CA) of the validation data. The CA is defined as the fraction of molecules with correctly determined AC. In the case of evaluation with multiple training-validation splits, the CA is taken as the mean accuracy on the validation set over the 10 iterations of the splits.

In case of RF and FNN, if an increase in the number of internal parameters of the model did not significantly increase its performance, the method with the lower number of internal parameters was retained. After optimisation of the hyperparameters for each ML model based solely on the B3PW91/6-31++G(d,p) spectra of sampling interval (SI) 0.5 cm^{-1}, the hyperparameters were frozen for the remainder of this study. These final hyperparameters are listed in Table 4.1†.

To investigate the number of spectra that need to be included to have decent classification accuracy, the procedure was repeated for various training-validation splits. In this study we considered the 9:1, 2:1, 1:1, 1:2, 1:4, 1:9, and 19:1 splits, which correspond to using 90%, 66%, 50%, 33%, 20%, 10%, and 5% of the total amount of data respectively as training set, and the remaining data as validation set. The models were built, trained, and validated using Orange3[50], a scikit-learn[40] based GUI. Using a desktop computer equipped with an Intel i5-8400 2.8 GHz processor and 32 GB of memory, all optimised models, except SVM, were trained in less than 1 minute on a single training set.

4.3 Results and discussion

Prior to deploying all of the aforementioned ML methods to conduct AC determination on the VCD spectral database, we checked whether no simple rules can be derived that would already allow a high CA. If such would be the case, the law of parsimony would already refute the use of ML methods. Due to the size of the database, finding characteristic bands or empirical patterns cannot be done by visual inspection.

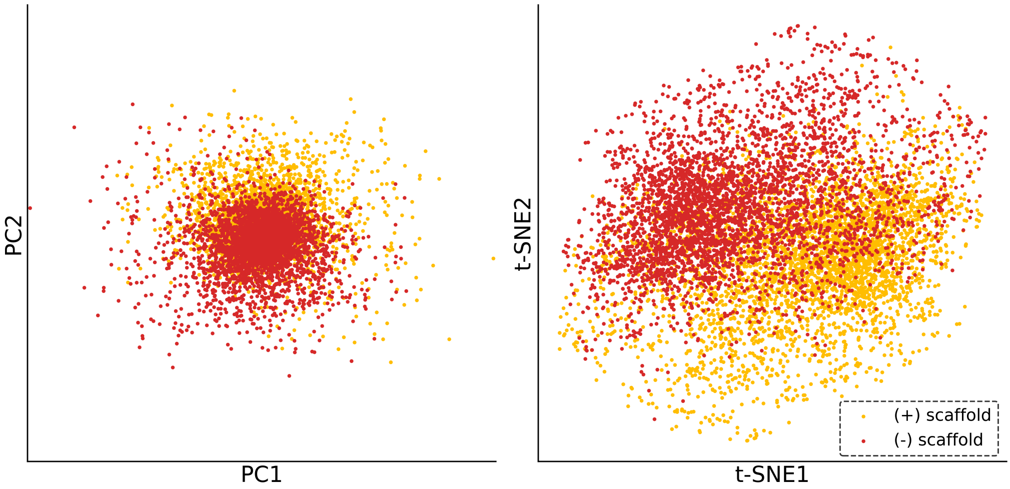


Figure 4.3: Visualisation of the spectra after dimensionality reduction with (left) 2D-PCA and (right) 2D-t-SNE, with yellow and red dots corresponding to the VCD spectra of decorated (+)- and (−)-α-pinene structures respectively.

To establish a baseline performance we rely on PCA and t-SNE, in combination with linear separation, and shallow decision trees, to possibly identify simple empirical patterns. The CA results of these methods are then used to gauge the performance of more advanced ML methods against.

4.3.1 Baseline performance with shallow decision trees, PCA and t-SNE

When a decision tree was trained on the entire dataset and using the entire spectra, a fraction of 0.766 was classified as the correct enantiomer for both tree depth 1 and 2. If instead of using the entire spectra, one uses the three most characteristic bands (provided they were separated by 8 cm^{-1}; 1184, 1424 and 1496 cm^{-1}), as identified by the decision tree, the decision tree classified a fraction of 0.785 properly, hence only a minor improvement.

For PCA, at least 62 components were needed to explain 95% of the total variance and >100 for 99%, which is indicative of the spectral complexity in the database. Furthermore, straightforward classification by linear separation using the first 2-3 principal components (logistic regression 9:1 split; CA 0.631-0.703) was not possible (see Figure 4.10†).

Finally, the use of t-SNE similarly showed that lower dimensional representations would not allow performant classification by linear separation (logistic regression 9:1 split on 2D-t-SNE; CA 0.791). The reason for the limited performance of linear separation on lower dimensional representations lies in the relatively large overlap of the (+)- and (−)-α-pinene populations, as illustrated in Figure 4.3 for both 2D-PCA and 2D-t-SNE, due to the absence of bands or patterns strongly characteristic for the AC. Keeping in mind that spectra of enantiomers are centrosymmetric in Figure 4.3 (see Supporting Information†), only a small part of the 2D-PCA plot remains characteristic for the (+)- and (−)-α-pinene based compounds. For 2D-t-SNE, the populations overlap to a lesser extent, creating larger regions dominated by a specific enantiomer. However, regions of strong overlap still occur which hamper proper discrimination of the ACs.

Altogether, some spectral patterns seem to be present in the data which can aid AC determination, but the resulting accuracy from these methods is far from convincing. One cannot conclude that there are empirical patterns or characteristic bands that allow a reliable AC determination. The information on the AC is buried within the VCD spectra in a complex manner. Therefore, more complex supervised ML methods are required.

4.3.2 Identification of best performing ML models

Each of the methods summarised in section 4.2.3 is trained and optimised to generate the AC label of the compound as output from the VCD spectrum input. The CA for the various ML methods is summarised in Figure 4.5 for all different train-validation splits. All methods were able to learn from the data and yielded better classification than obtained with shallow decision trees, with a CA of 0.766. NB was with an CA around 0.840 the least adequate for reliable AC determination. At first sight, LogReg showed promising accuracy. However, due to the very weak regularisation after optimisation it contained large coefficients for wavenumbers where only very faint intensities (tails from faraway bands) are present, as shown in Figure 4.14†. These coefficients would make the accuracy extremely unstable in the presence of any small deviations such as spectral noise (as expected in experimental spectra). When this overfitting was penalised with

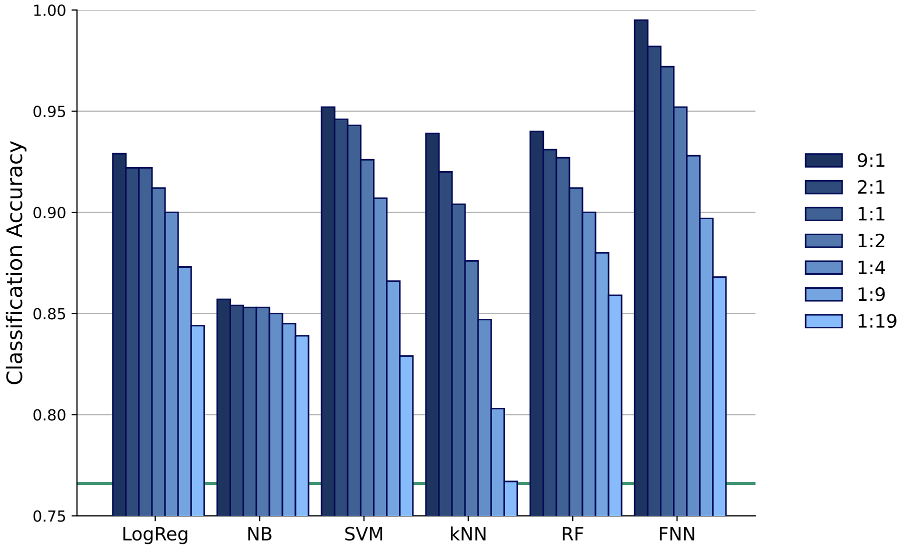


Figure 4.4: Classification accuracy of the spectra for several supervised ML models. The different train-validation split ratios are coloured as described in the legend.

stronger regularisation, the accuracy dropped significantly (see Figure 4.14†). Although SVM already showed promising improvement in performance, it remains the most computationally demanding method by far, requiring at least an order of magnitude more training time at the 9:1 split than the other methods. Moreover, its performance was noticeably dependent on the theoretical level used to perform the DFT calculations, making it less reliable in a general setting (see Figure 4.18†). kNN displayed a fairly high performance when using a large training set, but performed poorly in extracting the information connected to the AC when using a smaller training set.

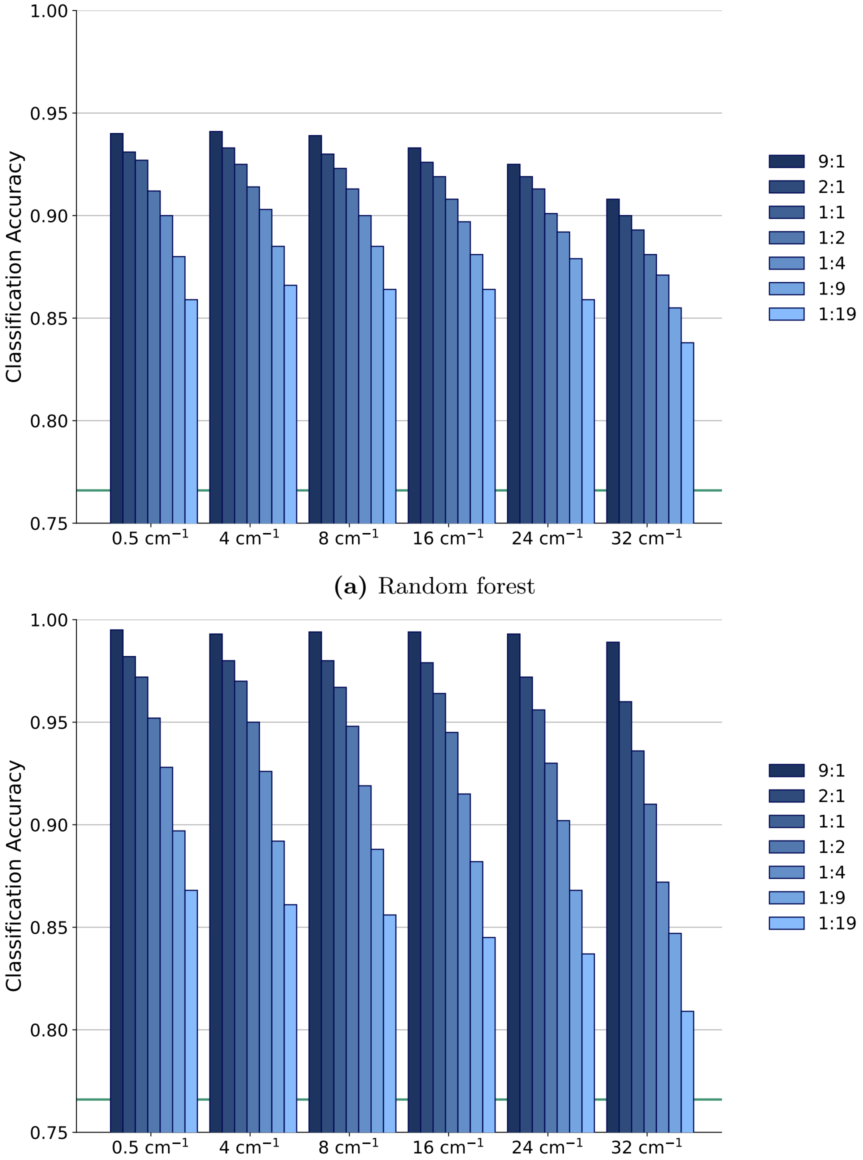
RF and FNN are overall the best performing models for identifying the ACs. In particular, FNN showed outstanding accuracy using larger training sets, with e.g. a CA of 0.995 for the 9:1 split, but still performed adequately when less training data was provided. RF did not outperform FNN, but still had fairly high accuracy across the various splits. The major advantage RF holds over FNN, is that the information extracted from the spectra and used in the algorithm to identify the AC is readily available, while this remains highly challenging to impossible for FNN and consequently limits it to remaining a black box model. As both methods clearly have their advantages, the remainder of this study focuses on RF and FNN.

4.3.3 Influence of spectral sampling interval

Thus far, all different models were trained on spectral data with a sampling interval (SI) of 0.5 cm^{-1}, providing them as much information as possible to train on in order to evaluate their maximal learning capabilities. However, considering that VCD spectrometers often record spectra at resolutions around 4-8 cm^{-1}, these models should additionally be evaluated at more representative SIs. Furthermore, models trained on data of larger SIs will more strongly repress possible overfitting tendencies, due to the lower dimensionality of the spectra. Therefore, the CA of both RF and FNN is evaluated for several SIs by subsampling the dimensions of the original spectral data.

Evaluating the differences between the SIs, shown in Figure 4.5, it becomes apparent that the performance of the models does not decrease significantly as long as the SI does not drop below 16 cm^{-1}. Changing the starting point of the spectra with an SI of 24 cm^{-1} influences the CA (see Figure 4.16†) but to a lesser extent than the SI itself. The absence of a specific wavenumber thus is not the main origin of the drop in the performance of the models. Instead, increasing the SI beyond 16 cm^{-1} causes loss of information in the VCD spectra, and prevents the models to identify the most AC representative patterns. Lowering the SI below 8 cm^{-1} does not improve model performance, which indicates that no new information is present in these representations. The strong correlation between adjacent wavenumbers for 0.5 cm^{-1} SI is reflected in only needing 62 PCs to explain 95% and more than 100 PCs to explain 99% of the total variance.

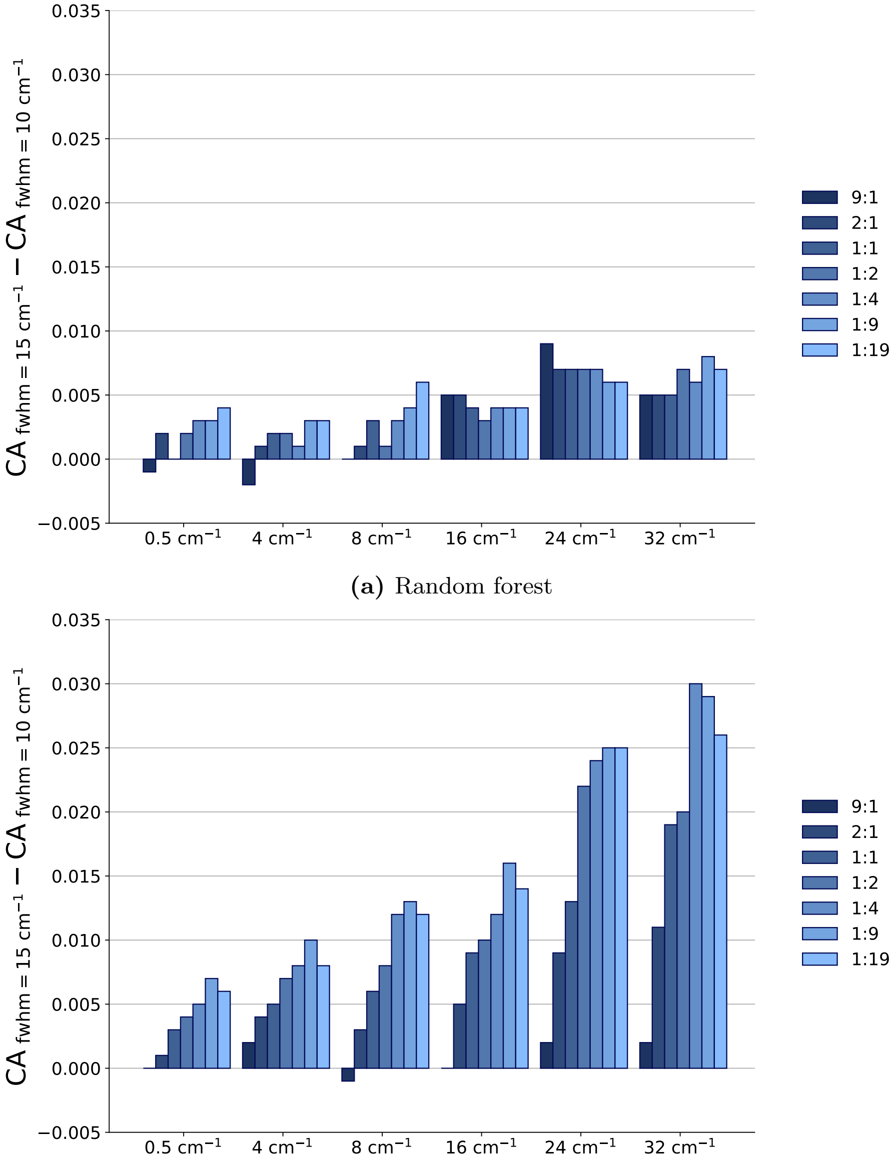
The origin for this exact density of the spectral information can be found in the Lorentzian broadening of the line spectra. Due to this broadening, bands are only indistinguishable when their maxima are separated by more than 10 cm^{-1} (the FWHM value) and wavenumbers separated by a smaller distance become strongly correlated. When the FWHM is increased to 15 cm^{-1} the performance remains more stable for spectra with a larger SI and the small CA drop for the 16 cm^{-1} SI disappears, as shown in Figure 4.17† and Figure 4.6. Thus, subsampling can be employed to such a degree that the spectral SI resembles the widths of the bands without experiencing any significant loss in accuracy.



(a) Random forest

(b) Feedforward neural network

Figure 4.5: Classification accuracy of the spectra for different sampling intervals for (a) random forest and (b) forward neural network. The different train-validation split ratios are coloured as described in the legend.



(a) Random forest

(b) Feedforward neural network

Figure 4.6: Difference in classification accuracy obtained between the spectra with bandwidth 15 cm^{-1} and 10 cm^{-1}, for (a) random forest and (b) feedforward neural network. The different train-validation split ratios are coloured as described in the legend.

4.3.4 Absolute configuration extraction with random forest

As mentioned earlier, the pattern that RFs employ to identify the AC can, in stark contrast with FNNs, to a certain extent be extracted using feature ranking and the scores associated with it. In Figure 4.7, the ranking score of all the spectral peaks in the entire dataset are illustrated for the different SIs. The larger the ranking score, the more important this specific wavenumber is for the AC determination.

The main spectral areas of interest remain similar across the different SIs, with the bands around 1180 cm^{-1} and between 1300 cm^{-1} and 1500 cm^{-1} dominating the AC determination. When comparing the median differential molar absorptivity $\Delta\epsilon$ for each wavenumber with the corresponding ranking values (Figure 4.8), we observe that the RF mainly focuses on the areas in which the median deviates from the zero line the strongest instead of focusing on areas containing the strongest intensities. This can be observed for instance in the area around 950 cm^{-1}, where despite both the central 50% and 95% quantiles containing strong intensities, the RF still considers it an area of low importance. However, the area around 1350 cm^{-1} appears, despite its near-zero median value, to be of significant importance to the AC determination. This is likely due to the central 95% quantile for decorated (+)-α-pinene structures containing strong positive intensities to weak negative intensities, making it easier to identify the AC using this band. It should be noted that these highly ranked areas are not the same as marker bands. Namely, the latter would imply that around a certain or several wavenumbers a specific VCD intensity and sign is directly indicative for the chirality of the compound, whereas in the former case the ranking indicates how important each wavenumber was during the identification of very complex patterns by the RF model to assign the AC.

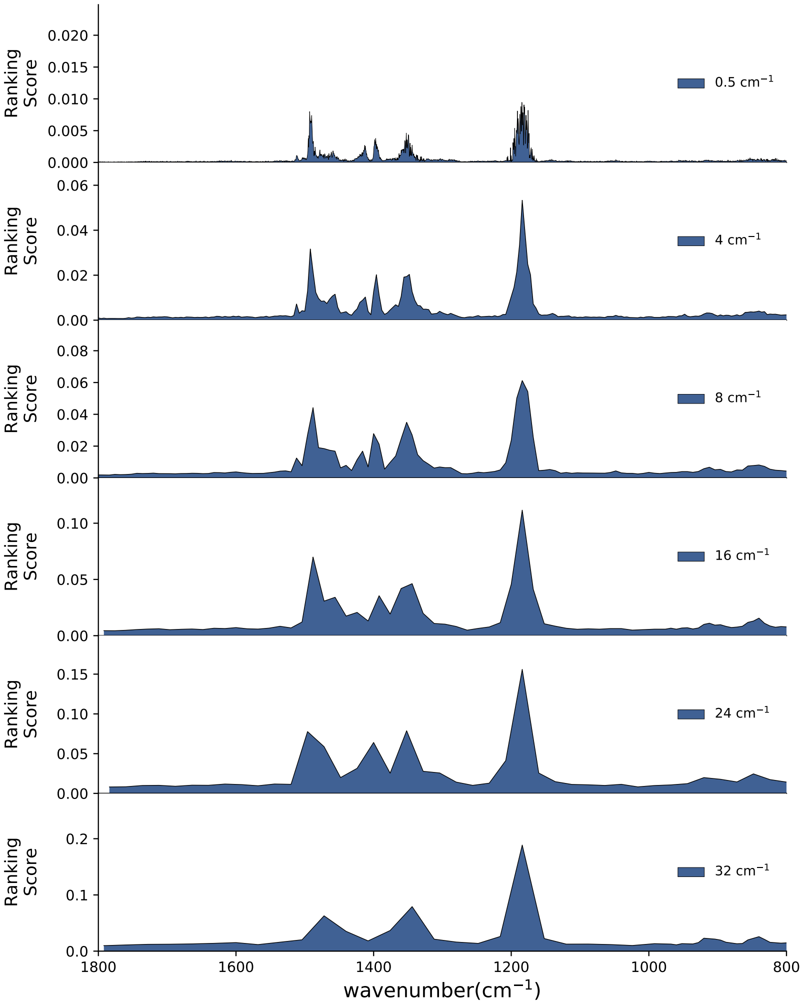


Figure 4.7: Random forest ranking score of the spectral features for the prediction of the chirality of the compounds for the different sampling intervals.

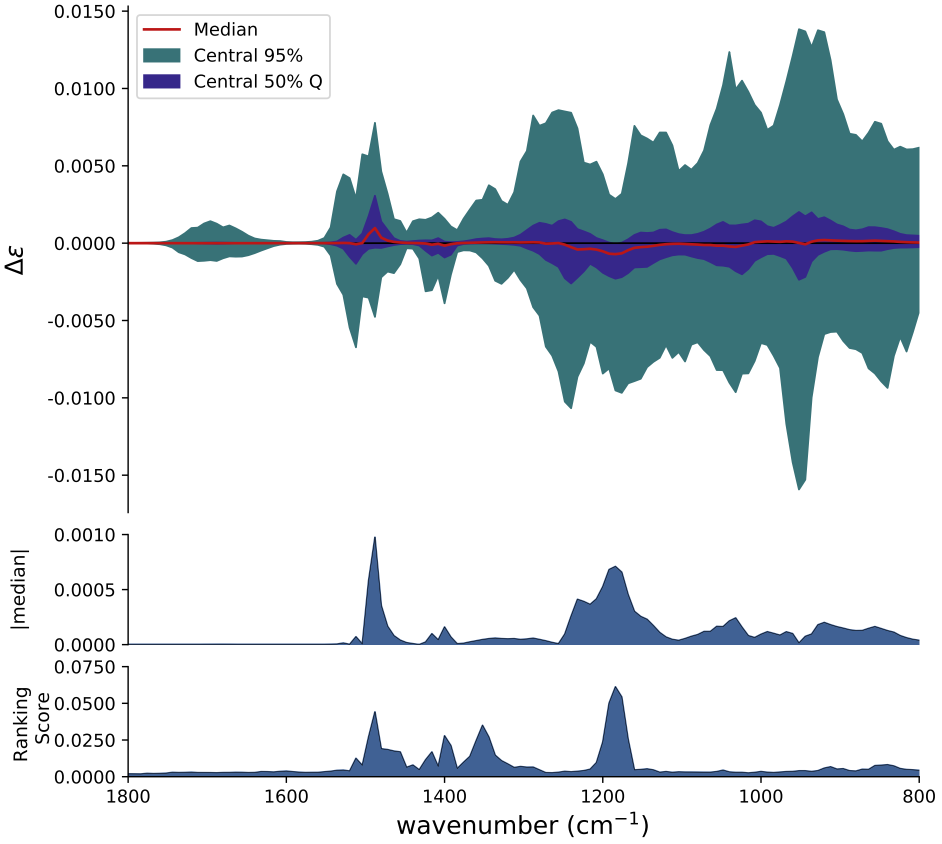


Figure 4.8: Top: The median value (red) and central 50% and 95% quantiles of the VCD spectra sharing the core structure of (+)-α-pinene, Middle: The absolute value of the median. Bottom: Random forest based ranking score for spectra with a sampling interval of 8 cm^{-1}.

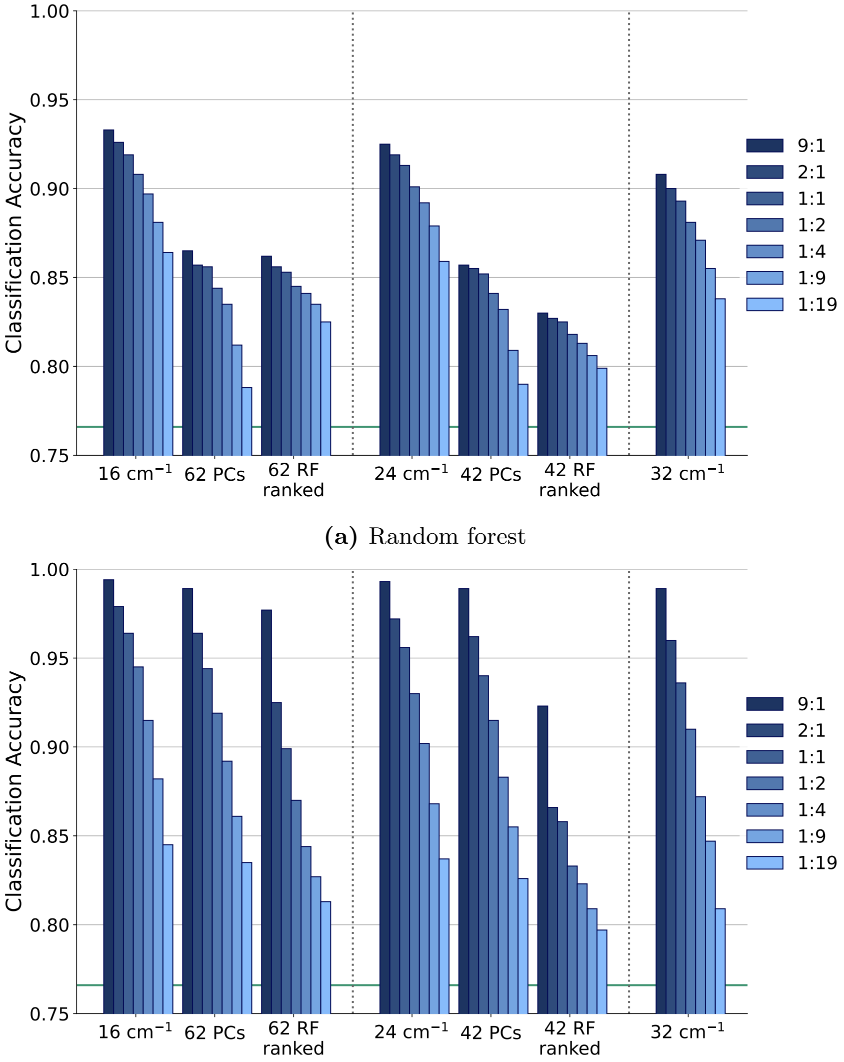
4.3.5 Dimensionality reduction with PCA and RF feature ranking

Up until now, only changing the SI was considered for reduction of the dimensionality of the input data for RF and FNN. However, both PCA and the RF based rankings discussed in the previous section can also be employed for this, using only the n most important components and wavenumbers, respectively. Comparing the performance of the dimensionality reduction methods, depicted in Figure 4.9, shows that the unbiased subsampling achieved by increasing the SI remains the better method. The biased subsampling based on RF ranking focuses on the most important spectral regions but does not take the high correlation between adjacent features into account. While increasing the SI still includes less important wavenumbers, the redundancy of the information is significantly lower. When this redundancy is removed with PCA, the CA still remains worse than obtained with unbiased subsampling. PCA includes most information in the spectra by focusing on the areas with the largest variance. However, as discussed in section 4.3.4 these areas do not necessarily contain the information most characteristic for the AC. Furthermore, this characteristic information will be encoded in a complex manner in the principal components.

4.3.6 Robustness and external validation of ML performance

Robustness of the results is an important issue. In the context of the present paper, robustness reflects the stability of the performance of ML methods with respect to changes in the spectra used as input. It is therefore not the same as robustness in the sense of peaks in a VCD spectrum being less or more affected by a change in a (DFT) computational parameter.[51,52] To gauge the robustness, we computed all VCD spectra for the entire database at other levels of theory, namely all remaining combinations of the B3LYP and B3PW91 functionals, with the 6-31G(d)/6-31++G(d,p)/ 6-311++G(2d,2p) basis sets, and trained ML models within each combination of functional and basis set in the same way as elaborately described above with the default functional and basis set. To retain a fair comparison of the performance, the hyperparameters of the ML models were not re-optimised (using the hyperparameters in Table 4.1†), while training the models on each combination of functional and basis set separately. Note that due to this workflow the data excluded from the training set becomes a test set, providing an even more reliable estimate of the performance.

The resulting similar performances (see Supporting Information†) demonstrate that using a different level of theory to generate input spectra has no significant influence on the ability of RF or FNN to establish the AC. Despite the similar performance, the ML models themselves are not internally the same. The models extract AC related information in a different manner for the different levels of theory (illustrated in the Supporting Information†). So, it is not due to a lack of influence of the functional and basis set that these ML methods perform equally well, but rather due to the robustness of the ML approach presented in this paper.



(a) Random forest

(b) Feedforward neural network

Figure 4.9: Comparison of subsampling techniques with Principal Component Analysis and only using the highest random forest ranked wavenumbers, for (a) random forest and (b) feedforward neural network. The different train-validation split ratios are coloured as described in the legend.

4.4 Conclusions

The value of Machine Learning (ML) methods for assigning the Absolute Configuration (AC) based on Vibrational Circular Dichroism (VCD) spectra has been demonstrated using a dataset of substituted α-pinene structure spectra. Random Forest (RF) and Feedforward Neural Networks (FNN) have proven to be the most performant among various ML methods for conducting the AC determination. At its best, a predictive accuracy up to 0.940 and 0.995 can be reached with RF and a shallow FNN, respectively. In stark contrast to the black box nature of FNN, the RF model allows the extraction of the spectral areas important for AC determination. Furthermore, the quality of AC determination remained unchanged, as long as the spectral sampling interval was comparable to or smaller than the width of the bands. Setting the sampling interval to a value comparable to the bandwidth, so-called subsampling, also proved to be the best dimensionality reduction method, outperforming PCA or methods exploiting supervised ranking. All conclusions made were validated by external validation.

This contribution emphasises the yet untapped potential of ML methods and deep learning in VCD spectroscopic application areas, as well as the added value that the creation of large experimental VCD databases in tandem with ML methods can provide in the future. Most importantly, once more databases are established, it becomes possible to speed up the AC determinations of particular molecular classes by not having to tackle every single compound in a case-by-case manner.

Supporting information

3D principal component analysis on VCD spectra

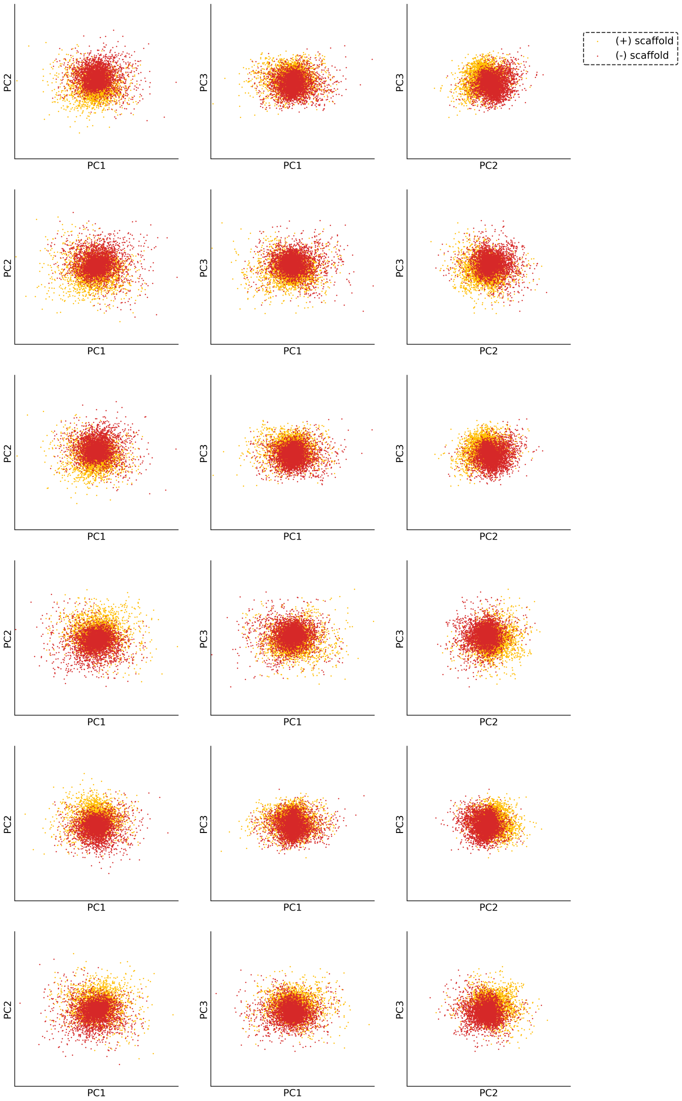


Figure 4.10: Comparison of the enantiomers' PCA transformed spectra, from top to bottom B3LYP/6-31G(d), B3PW91/6-31G(d), B3LYP/6-31++G(d,p), B3PW91/6-31++G(d,p), B3LYP/6-311++G(2d,2p), B3PW91/6-311++G(2d,2p).

Symmetry of PCA transformed enantiomer spectra

The VCD spectra of enantiomers are identical except for the inverted sign of the VCD intensity $\Delta\varepsilon(\tilde{\nu})$ at each wavenumber $\tilde{\nu}$. This implies that inverting all VCD intensities for a set of compounds sharing the (+)-scaffold ((+)-α-pinene) yields the spectra of the (−)-scaffold compounds. A VCD spectral dataset containing both enantiomers of each compound, such as the α-pinene dataset employed in this paper, holds for each unique VCD spectrum also the one with inverted sign, making the dataset centrosymmetric. This property of the dataset is illustrated in Figure 4.11 for the most characteristic wavenumbers (according to shallow decision tree; 1184 and 1352 cm^{-1}).

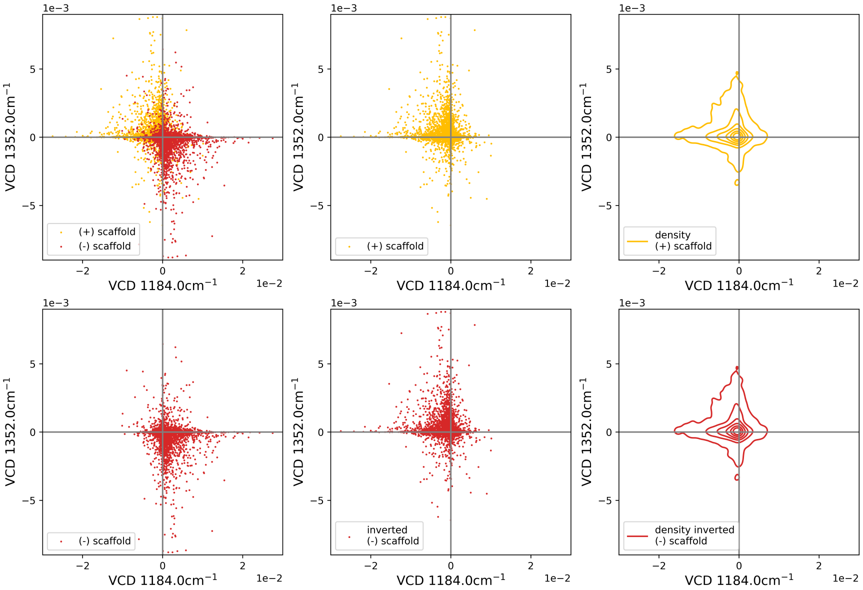


Figure 4.11: Symmetry of the pinene spectral data within the space defined by the VCD intensities for wavenumbers 1184 and 1352 cm^{-1}. The contour plots (right panels) show the distribution of the (−)-scaffold spectra with inverted sign and the (+)-scaffold spectra respectively.

The PCA technique rotates the axes within the space defined by the different wavenumbers, transforming the dataset into principal component space. The relationship between the spectral intensities of enantiomeric pairs remains unaffected by this PCA transformation. As illustrated in Figure 4.12, the dataset remains centrosymmetric within principal component space. Additionally, Figure 4.13 shows that the VCD intensities of enantiomers are identical but of opposite sign for the most relevant principal components.

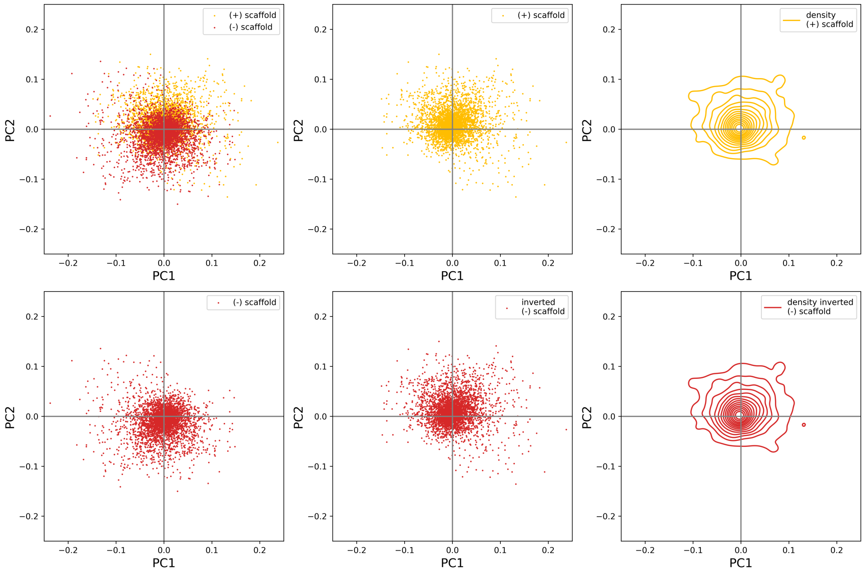


Figure 4.12: Symmetry of the pinene spectral dataset in 2D-PCA space. The inverted (−)-scaffold spectra are the (−)-scaffold spectra with inverted sign for both PC1 and PC2. The contour plots (right panels) show the distribution of the (+)-scaffold spectra and of the (-)-scaffold spectra with inverted sign respectively.

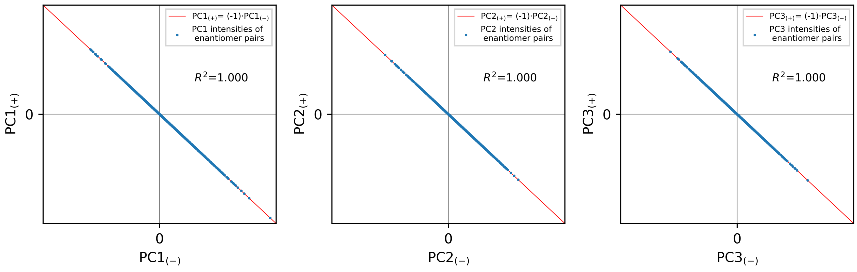


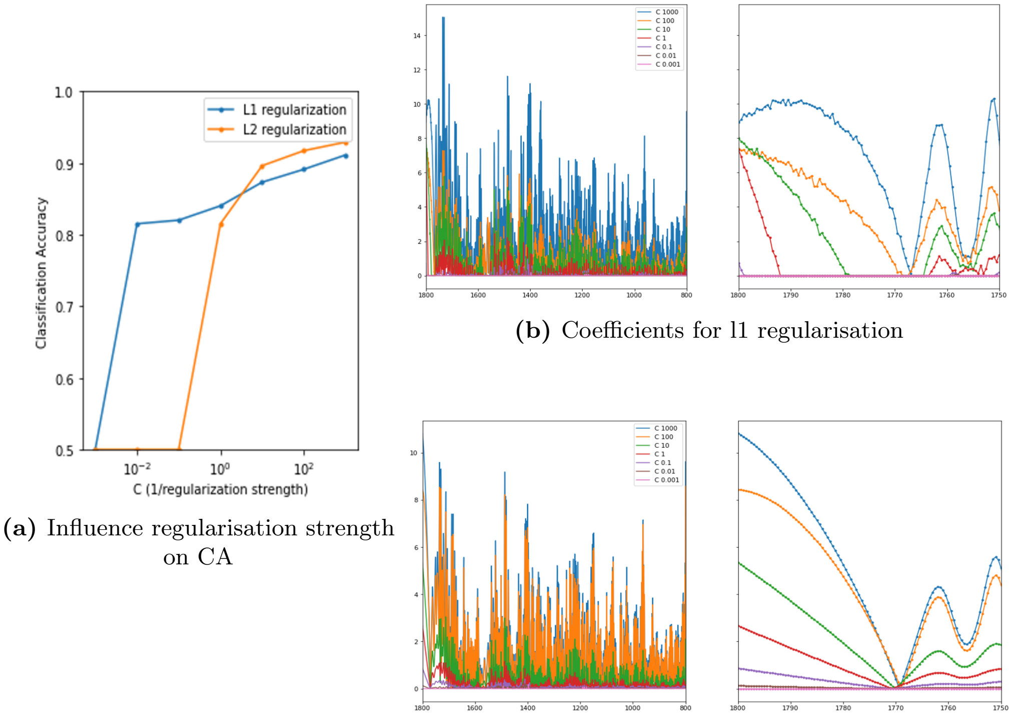
Figure 4.13: Relationship of VCD intensities for the principal components PC1-3. The intensity corresponding to a specific PC component for the (-)-enantiomer ($PC_{(-)}$) is compared with the value for the (+)-enantiomer ($PC_{(-)}$). R^2 is the pearson correlation between $PC_{(+)}$ and $(-1) \cdot PC_{(-)}$ averaged over all pairs of enantiomers.

Hyperparameters of the optimised models

LogReg	L2 regularisation, C 1000
NB	N.A.
SVM	Linear Kernel, tolerance 0.001, C 0.1
kNN	Neighbours 3, weighted Manhattan distance
RF	Trees 200, max tree depth 20
FNN	Hidden layers 2, neurons 100 and 20 respectively, optimiser Adam, L2 regularisation alpha 0.001, maximal iterations 500

Table 4.1: Optimised hyperparameter for the supervised machine learning models.

Logistic regression weights for weak & strong regularisation



(a) Influence regularisation strength on CA

(b) Coefficients for l1 regularisation

(c) Coefficients for l2 regularisation

Figure 4.14: Influence of regularisation strength and method for logistic regression on the classification accuracy and the coefficients.

Influence of database imbalance w.r.t substitutional populations

At this stage, it is interesting to see to what extent the predictive power is dependent on the exact substituents. The misclassified molecules of 10 separate RF training cycles using the same training method as before (9:1 split, 8 cm^{-1} sampling interval) were identified and the average misclassification for every substituent at every position was determined. This procedure was repeated for FNN (9:1 split, 8 cm^{-1} step size), but with 100 separate training cycles instead, in order to guarantee the values' statistical significance (as the misclassification is about 10 times smaller than that of RF). Through comparison of these misclassifications, depicted in Figure 4.15, a noticeable difference in predictability is manifested for the different substituents and positions; the general trend appears similar for both RF and FNN, which can be attributed to the difficult non-characteristic influences these substitutions have on the VCD spectrum and structural underrepresentation of certain groups/combinations in the dataset (depicted in Figure 4.2). However, it remains difficult to clearly reveal the extent to which one dominates over the other.

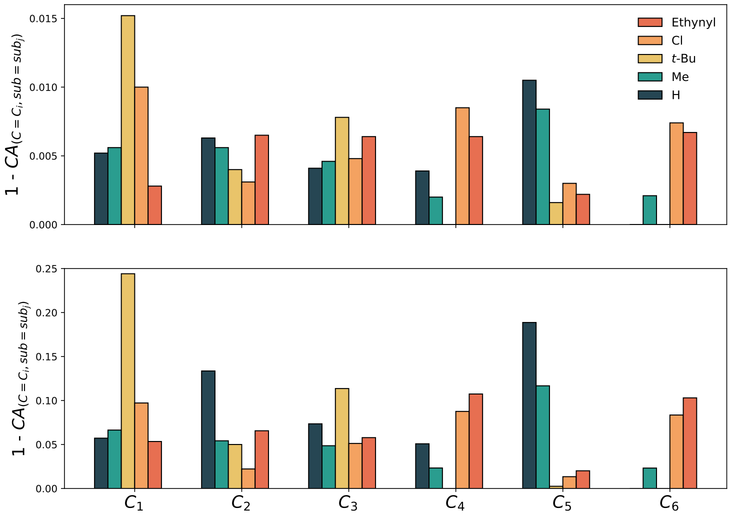
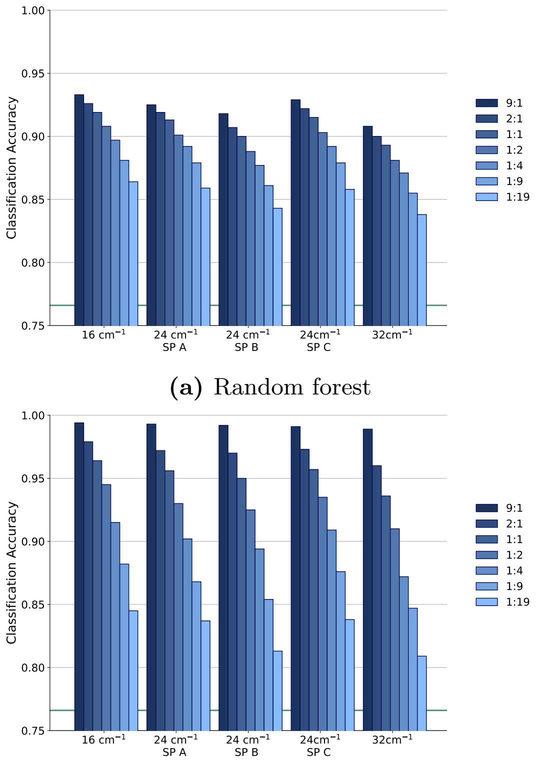


Figure 4.15: Relative misclassification of the spectra for a certain substituent at each position 1-6 separately for feedforward neural network(top) and fandom forest(bottom).

Influence of starting point on CA for 24 cm^{-1} sampling interval (B3PW91/ 6-31++G(d,p))

A different starting point or SI can lead to exclusion of a wavenumber characteristic for the AC. The drop in accuracy observed from an SI of 24 cm^{-1}could be caused by missing a specific wavenumber which was present in the spectra with an SI of 8 cm^{-1}, instead of a loss in information. We investigated this by training and evaluating on spectra of SI 24 cm^{-1} with three different starting point separately, after which their performances were compared to those obtained for SIs of 16 cm^{-1} and 32 cm^{-1}. As can be observed in Figure 4.16, the CA does depend on the exact starting point. However, the influence of changing the SI to 16 cm^{-1} or 32 cm^{-1} still remains larger than the starting point.

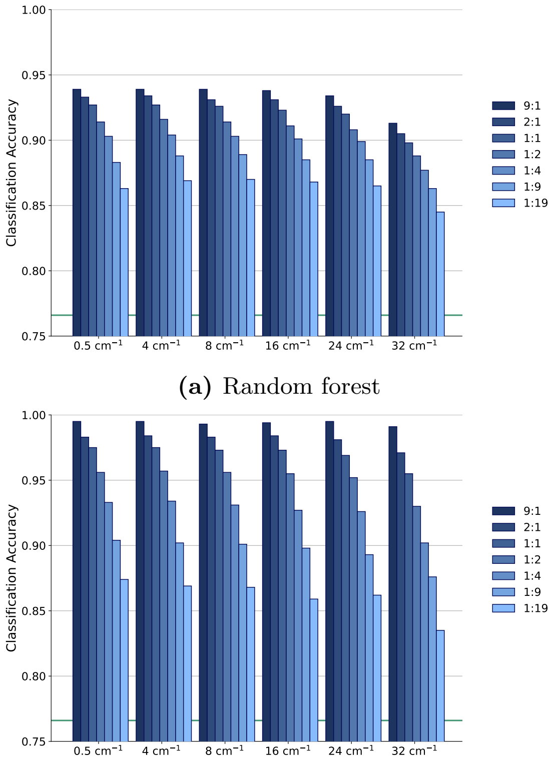


(a) Random forest

(b) Forward neural network

Figure 4.16: Influence of starting point (SP) on the classification accuracy for the 24 cm^{-1}sampling interval for (a) random forest and (b) feedforward neural network. Starting point A, B and C are 800, 808 and 816 cm^{-}1 respectively. The different train-validation split ratios are coloured as described in the legend.

CA for spectra with bandwidth of 15 cm^{-1}



(a) Random forest

(b) Forward neural network

Figure 4.17: Classification accuracy of the spectra with bandwidth 15 cm^{-1}, for (a) random forest and (b) feedforward neural network. The different train-validation split ratios are coloured as described in the legend.

External validation of all ML models with other functional/basis set for 0.5 cm^{-1} sampling interval

In order to evaluate the stability of the performance of the different ML models originally considered are with regards to the choice of functional and basis set, the mean CA and corresponding standard deviation over the different levels of theory are illustrated in Figure 4.18. We observe that the performance of LogReg, NB and, in particular, SVM is noticeably dependant on the level of theory, even when a large majority of the data is provided for training.

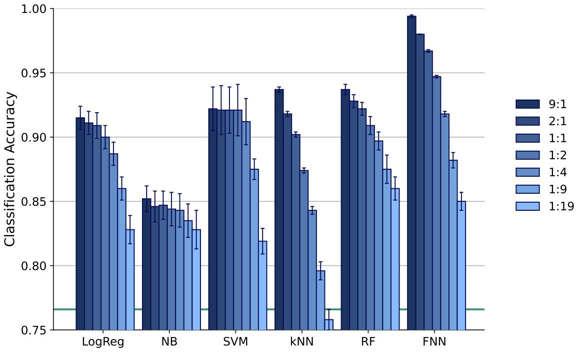
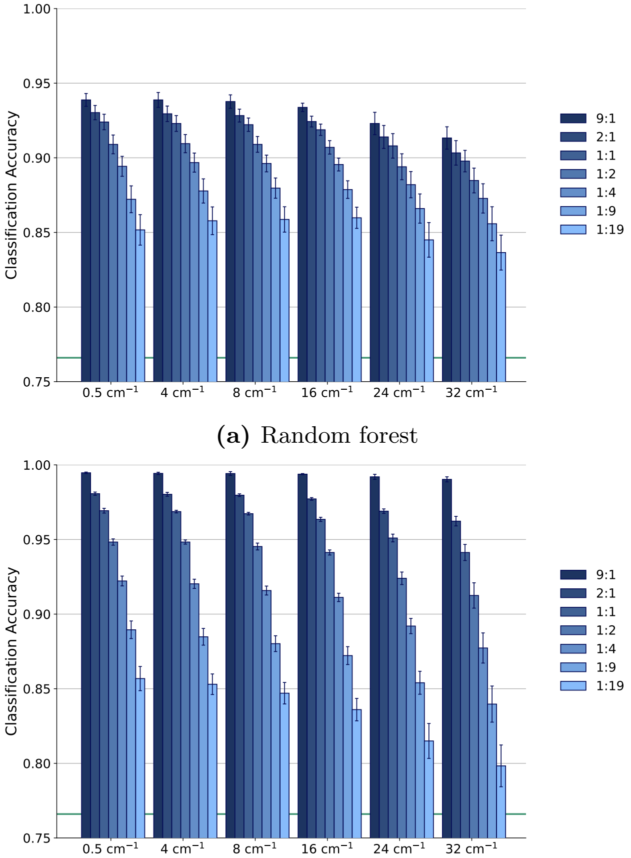


Figure 4.18: Mean Classification accuracy of the spectra for the different ML models over all combinations of the B3LYP and B3PW91 functionals, with the 6-31G(d)6-31++G(d,p)/ 6-311++G(2d,2p) basis sets. The different data split ratios are coloured as described in the legend.

External validation of performance for RF and FNN with other functional/basis set

To investigate to which degree the choice in functional and basis set will impact the performance of both RF and FNN, each model (with the same hyperparameters as described in Table 4.1) is trained on the spectra of the different levels of theory separately. This procedure is repeated for all the different SIs and data splits. Their mean performance and corresponding standard deviation over the six different levels of theory are determined and illustrated in Figure 4.19. As long as the SI remains similar or smaller than the FWHM and the majority of the data is provided for training, the standard deviation is negligible. As an example, the standard deviations for an SI of 8 cm^{-1} and a data split of 9:1, are 0.003 and 0.0004 for RF and FNN respectively. For an SI value of 24 cm^{-1} and 32 cm^{-1}, the standard deviation clearly increases, which strengthens our suggestion to keep the SI value similar to the FWHM. The standard deviation also increases when a smaller number of spectra is present in the training set. This is likely caused by the smaller reliability of the CA values for the individual levels of theory, as less training data with the same model complexity allows for more overfitting.



(a) Random forest

(b) Forward neural network

Figure 4.19: Mean Classification accuracy of the spectra for (a) random forest and (b) feedforward neural network over all combinations of the B3LYP and B3PW91 functionals, with the 6-31G(d)6-31++G(d,p)/ 6-311++G(2d,2p) basis sets.

Feature ranking for RF trained on various functional & basis set combinations

The question arises whether the similar performances discussed in previous two sections are due to the robustness of the ML methods or the ML models themselves are identical. In this section, the workflow described in section 4.3.4 is repeated for the aforementioned remaining combinations of functional and basis set. The resulting ranking scores of the spectral features (depicted in Figure 4.20) do differ for the different levels of theory, even when accounting for the horizontal shift of the vibrations' frequencies. Hence, the RF models extract AC related information in a different manner.

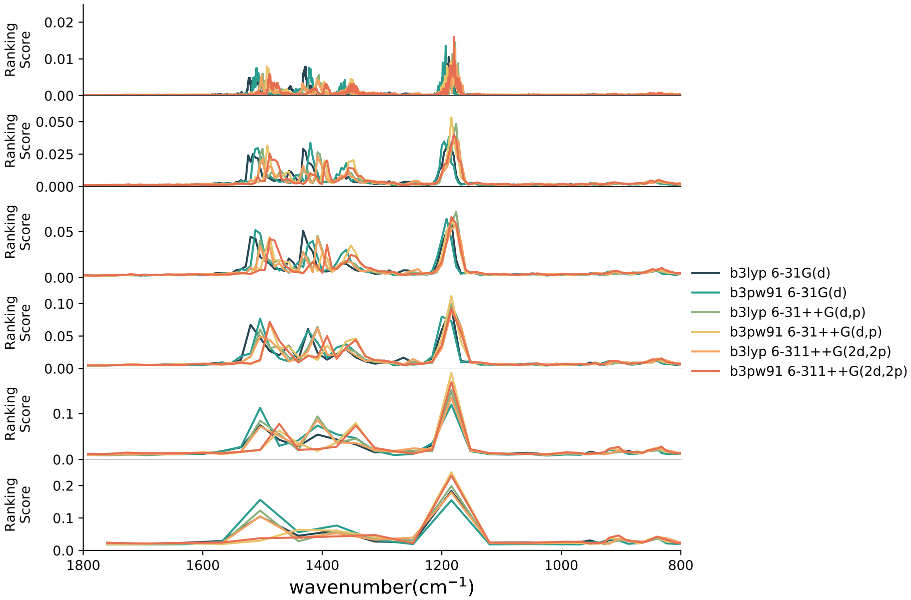


Figure 4.20: Random forest ranking score of the spectral features for the prediction of the chirality of the compounds for the different sampling intervals and combinations of functional and basis set. From top to bottom the sampling interval equals 0.5, 4, 8, 16, 24, 32 cm^{-1}.

Performance and structure of shallow decision trees trained on various functional & basis set combinations

To further exemplify the influence of the level of theory on how ML models extract AC related information from the spectra, shallow decision trees (depth 2) were trained on all spectra (SI 8 cm^{-1}) for a specific level of theory. As illustrated in Figure 4.21, the criteria (i.e. wavenumber and corresponding intensity) used for the criterion in each decision node vary, especially so for the second layer of decision nodes.

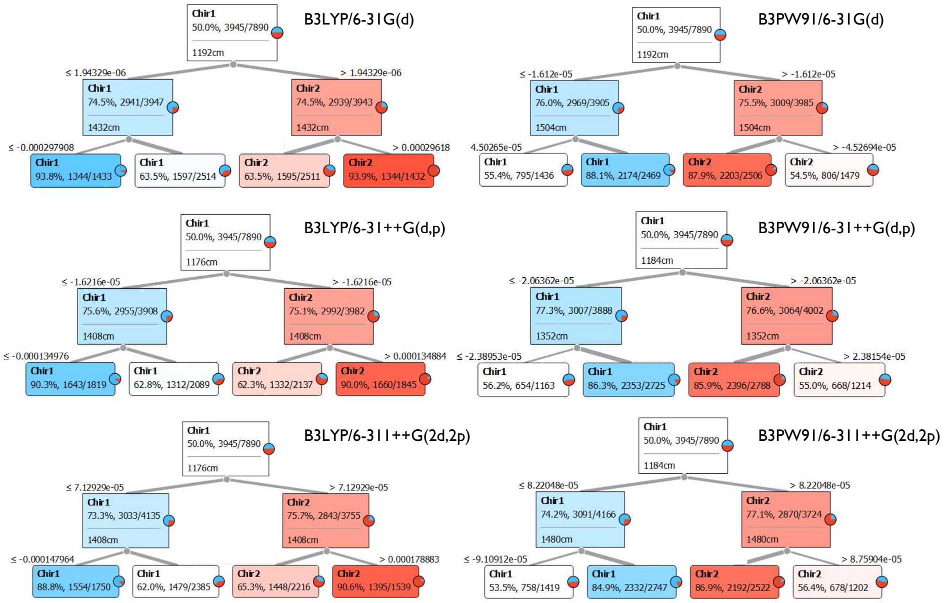


Figure 4.21: Shallow decision trees trained on VCD spectra (SI 8 cm^{-1}) of different levels of theory as denoted in the figure. The nodes are coloured according to their purity, with a blue-white-red gradient, with the dominant chirality class present in each node denoted as 1 ((+)-α-pinene) or 2 ((−)-α-pinene). For each node the absolute and relative population of the dominant class is given, along with the corresponding wavenumber and intensity criterion used in each decision node.

References

[1] M. Rouhi, *Chem. Eng. News*, 2003, **81**, 45–61.

[2] M. Rouhi, *Chem. Eng. News*, 2004, **82**, 47–62.

[3] J. M. Bijvoet, A. Peerdeman and A. van Bommel, *Nature*, 1951, **168**, 271–272.

[4] C. C. Hinckley, *J. Am. Chem. Soc.*, 1969, **91**, 5160–5162.

[5] T. R. Hoye and D. O. Koltun, *J. Am. Chem. Soc.*, 1998, **120**, 4638–4643.

[6] N. Kobayashi and A. Muranaka, *Circular Dichroism and Magnetic Circular Dichroism Spectroscopy for Organic Chemists*, The Royal Society of Chemistry, Cambridge, United Kingdom, 2012, pp. 1–199.

[7] J. a. M. Batista Jr., E. W. Blanch and V. d. S. Bolzani, *Nat. Prod. Rep.*, 2015, **32**, 1280–1302.

[8] C. Merten, T. Golub and N. Kreienborg, *J. Org. Chem.*, 2019, **84**, 8797–8814.

[9] L. A. Nafie, *Chirality*, 2020, **32**, 667–692.

[10] P. Stephens and F. Devlin, *Chirality*, 2000, **12**, 172–179.

[11] J. Kessler, V. Andrushchenko, J. Kapitán and P. Bouř, *Phys. Chem. Chem. Phys.*, 2018, **20**, 4926–4935.

[12] P. Bouř, J. Sopková, L. Bednárová, P. Maloň and T. A. Keiderling, *J. Comput. Chem.*, 1997, **18**, 646–659.

[13] J.-H. Choi, J.-S. Kim and M. Cho, *J. Chem. Phys.*, 2005, **122**, 174903.

[14] T. Giovannini, M. Olszòwka and C. Cappelli, *J. Chem. Theory Comput.*, 2016, **12**, 5483–5492.

[15] K. Bünnemann and C. Merten, *J. Phys. Chem. B*, 2016, **120**, 9434–9442.

[16] S. Yang and M. Cho, *J. Chem. Phys.*, 2009, **131**, 135102.

[17] J. Meiler, R. Meusinger and M. Will, *J. Chem. Inf. Comput. Sci*, 2000, **40**, 1169–1176.

[18] E. Jonas and S. Kuhn, *J. Cheminformatics*, 2019, **11**,.

[19] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari and P. Rinke, *Adv. Sci*, 2019, **6**, 1801367.

[20] J. A. Fine, A. A. Rajasekar, K. P. Jethava and G. Chopra, *Chem. Sci.*, 2020, **11**, 4618–4630.

[21] P. Kovács, X. Zhu, J. Carrete, G. Madsen and Z. Wang, *Astrophys. J.*, 2020, **902**, 100.

[22] M. Xu, C.-H. Wang, A. Terracciano, A. Masunov and S. Vasu, *Sci. Rep*, 2020, **10**, 13569.

[23] A. Mowat and G. Holmes, *Acta Hortic.*, 2003, **601**, 65–69.

[24] H.-Y. Chien, A.-T. Shih, B.-S. Yang and V. K. S. Hsiao, *Math Biosci Eng.*, 2019, **16**, 6874–6891.

[25] J. Houston, F. Glavin and M. Madden, *J. Chem. Inf. Model.*, 2020, **60**, 1936–1954.

[26] C. Cheng, J. Liu, C.-J. Zhang, M. Cai, H. Wan and W. Xiong, *Appl. Spectrosc. Rev.*, 2010, **45**, 148–164.

[27] M. McCann, M. Defernez, B. Urbanowicz, J. Tewari, T. Langewisch, A. Olek, B. Wells, R. Wilson and N. Carpita, *Plant Physiol.*, 2007, **143**, 1314–26.

[28] V. H. da Silva, F. Murphy, J. M. Amigo, C. Stedmon and J. Strand, *Anal. Chem.*, 2020, **92**, 13724–13733.

[29] X. Fan, W. Ming, H. Zeng, Z. Zhang and H. Lu, *Analyst*, 2019, **144**, 1789–1798.

[30] H. Bian, H. Yao, G. Lin, Y. Yu, R. Chen, X. Wang, R. Ji, X. Yang, T. Zhu and Y. Ju, *IEEE Photonics J.*, 2020, **12**, 1–9.

[31] K. Tanabe, T. Matsumoto, T. Tamura, J. Hiraishi, S. Saeki, M. Arima, C. Ono, S. Itoh, H. Uesaka, Y. Tatsugi, K. Yatsunami, T. Inaba, M. Mitsuhashi, S. Kohara, H. Masago, F. Kaneuchi, C. Jin and S. Ono, *Appl. Spectrosc.*, 2001, **55**, 1394–1403.

[32] M. Meyer and T. Weigelt, *Anal. Chim. Acta*, 1992, **265**, 183–190.

[33] Q. van Est, P. Schoenmakers, J. Smits and W. Nijssen, *Vib. Spectrosc.*, 1993, **4**, 263–272.

[34] Q.-Y. Zhang, G. Carrera, M. J. S. Gomes and J. a. Aires-de Sousa, *J. Org. Chem.*, 2005, **70**, 2120–2130.

[35] M. Kinalwa, E. Blanch and A. Doig, *Protein Sci.*, 2011, **20**, 1668–74.

[36] C. Guo, R. D. Shah, R. K. Dukor, X. Cao, T. B. Freedman and L. A. Nafie, *Anal. Chem*, 2004, **76**, 6956–6966.

[37] H. Li and L. Nafie, *J. Raman Spectrosc.*, 2012, **43**, 89–94.

[38] R. A. Lombardi and L. A. Nafie, *Chirality*, 2009, **21**, E277–286.

[39] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings,

B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision B.01*, 2016, Gaussian Inc. Wallingford CT.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

[41] K. Pearson, *Lond. Edinb. Dubl. Phil. Mag.*, 1901, **2**, 559–572.

[42] L. van der Maaten and G. Hinton, *J. Mach. Learn. Res*, 2008, **9**, 2579–2605.

[43] R. Tibshirani, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 1996, **58**, 267–288.

[44] A. E. Hoerl and R. W. Kennard, *Technometrics*, 1970, **12**, 55–67.

[45] D. D. Lewis, Machine Learning: ECML-98, Berlin, Heidelberg, 1998, pp. 4–15.

[46] C. Cortes and V. Vapnik, *Chem. Biol. Drug Des.*, 2009, **297**, 273–297.

[47] T. Cover and P. Hart, *IEEE Trans. Inf. Theory*, 1967, **13**, 21–27.

[48] L. Breiman, *Machine Learning*, 2001, **45**, 5–32.

[49] Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436–44.

[50] J. Demšar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Možina, M. Polajnar, M. Toplak, A. Staric, M. Stajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik and B. Zupan, *J. Mach. Learn. Res*, 2013, **14**, 2349–2353.

[51] V. P. Nicu and E. J. Baerends, *Phys. Chem. Chem. Phys.*, 2009, **11**, 6107–6118.

[52] V. P. Nicu, E. Debie, W. Herrebout, B. Van der Veken, P. Bultinck and E. J. Baerends, *Chirality*, 2010, **21**, E287–E297.

Chapter 5

Impact of conformation and intramolecular interactions on vibrational circular dichroism spectra identified with machine learning

5.1 Introduction

Chiroptical spectroscopic methods measure the difference in interaction between an optically active compound and left- or right-circularly polarized radiation.[1–4] The best known chiroptical method is Electronic Circular Dichroism (ECD), where one measures the difference in absorption of left- and right-handed circularly polarized visible and ultraviolet radiation by an optically active molecule. Vibrational Circular Dichroism (VCD) is an infrared chiroptical method where vibrational transitions are responsible for the difference in absorption. The main advantage of VCD compared to ECD is the richer information obtained from the former due to the much larger number of vibrational transitions compared to the number of accessible electronic transitions. Chiroptical methods find their main area of application in establishing the absolute configuration (AC) of molecules.[4–25] However, it also reveals a significant amount of information on the conformational properties of a molecule.[26–39] The link between conformation in the sense of its molecular geometry and its VCD spectrum is not easily established on the basis of e.g., some rules of thumb and one usually relies on the quantum chemically computed spectrum. The usual approach to establishing the AC of a compound is to choose a specific AC of the molecule, find all its conformers on the potential energy hypersurface and their energies and then combine all computed spectra using Boltzmann weighting in a simulated molecular spectrum for the chosen AC[1]. By repeating all these steps for each possible AC and comparing all computed spectra to the experimental one, one concludes what AC the experimental sample corresponded to. Said computed spectra are usually generated using Density Functional Theory (DFT) calculations requiring sufficient expertise and computational resources. Experience shows that the VCD spectra of individual conformers of the same molecule may differ significantly even if they belong to the same AC (see Supplementary Discussion 1), explaining why rules of thumb cannot be established.[26–34] The first hypothesis tested in this paper is that machine learning (ML) can be used to predict the VCD spectrum for a specific conformer using only its geometry, in this sense providing an alternative to the DFT procedure. The second hypothesis of this paper is that ML may help reduce significantly the total time cost needed to obtain a molecular spectrum. This entails that ML should allow skipping enough time normally spent in DFT calculations to more than compensate for the time it takes to establish the ML model. The third and final issue examined is the extent to which a ML model is transferable from one AC to another. Does it suffice to learn from one AC and use this for all other possible AC's? For instance, in a molecule with two stereocenters, does it suffice to establish a ML model for RR and to then use it also for RS, SR and SS?

ML methods are powerful methods for the extraction of complex patterns hidden in spectral data, speeding up conventional workflows, and accelerating computational methods.[40–51] We have recently shown that there is hope that ML can play a role in VCD spectroscopy[52]. More specifically, we have shown that for a large set of congeneric molecules adopting a single conformer, we can use ML to reveal to what AC a VCD spectrum of an unknown molecule corresponds. Now the ambitions are higher. We want to extract a VCD spectrum solely from a conformer geometry. Figure 5.1 contrasts the current work against our previous work[52] and other recent works[53–55] that address the link between an AC and an experimental spectrum or property. The present paper concentrates on the link between the structure of a conformer and its VCD spectrum within a given AC of a molecule. Success in establishing this link will then also strongly benefit the usual approach to VCD based AC assignments as it will allow circumpassing the quantum chemical calculation of all conformer VCD spectra.

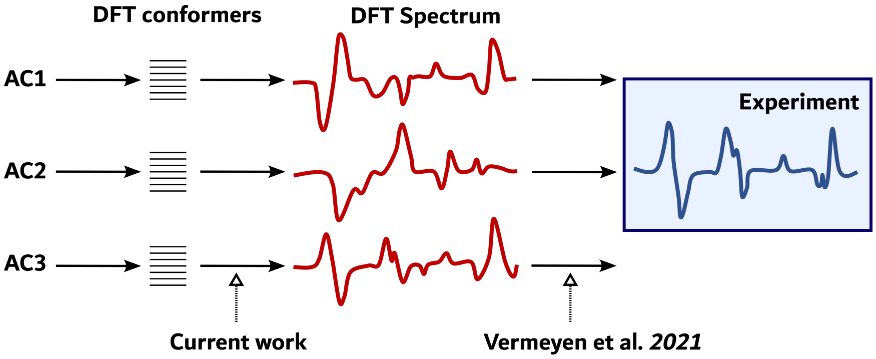


Figure 5.1: Scope of the current paper: can ML extract the link between the structure of a conformer for a specific AC of a compound and its corresponding DFT computed VCD spectrum? Note the difference with our previous ML work (in Chapter 4).

To prove or disprove the hypotheses set, a test bench of compounds must be established. Somewhat naively, one could think of any set of compounds for which VCD has been computed and/or measured but this is not useful. We namely wish to be able to control ourselves the degree of conformational flexibility of the molecules and the nature of their intramolecular interactions by changing a number of substituents. At the same time, both effects should not intercorrelate too much. This entails the use of admittedly somewhat peculiar molecules but the priority is given to stepwise understand and prove the hypotheses. This would not be possible using too diverse compounds while at the same time, error cancellation could play a much larger role there. We use a tetra-substituted naphthalene framework whose conformational flexibility and intramolecular interactions (such as hydrogen bonding) we can control by judicious selection of substituents. This allows us to test whether ML is sufficiently reliable over a range of chemical situations.

To be able to impact the conformational flexibility and the degree to which intramolecular interactions play a role without changing too many features simultaneously, we have chosen compounds that have the same backbone. A tetra-substituted naphthalene framework is chosen as backbone. To this substituents

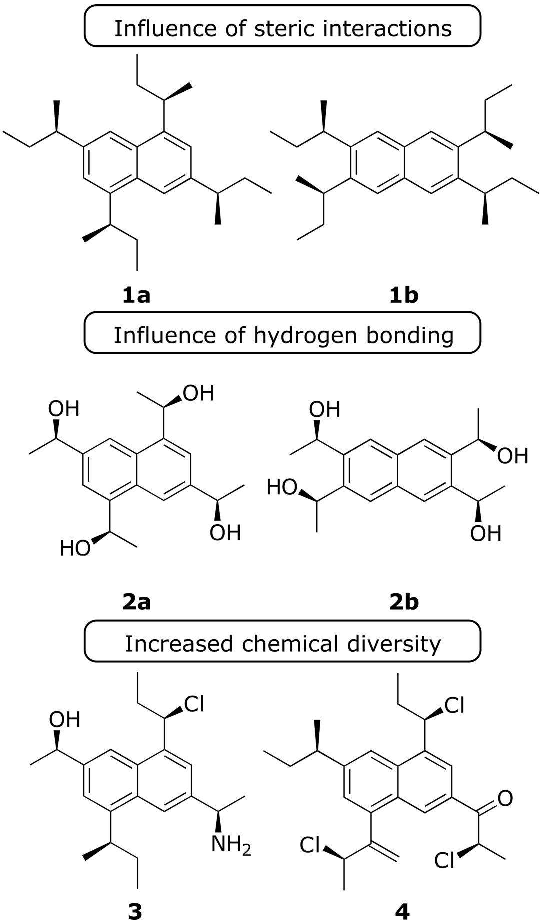

Influence of steric interactions
1a
1b
Influence of hydrogen bonding
OH
OH
HO
OH
OH
HO
OH
OH
2a
2b
Increased chemical diversity
OH
Cl
NH2
Cl
O
Cl
Cl
3
4

Figure 5.2: Overview of compounds for which the link between conformer and VCD spectrum is extracted with the ML workflow.

containing a chiral center in the *R*-configuration are added. Changing the substitution pattern enables to control the intramolecular interaction between the substituents. An overview of the compounds considered in this work is provided in Figure 5.2. In compounds **1a** and **2a**, the sidechains and their conformational properties can be expected to be largely independent from each other. For example, steric hindrance is limited thanks to the large distance between the sidechains. Vibrational mode coupling between the sidechain vibrations may still impact the vibrational frequencies and corresponding VCD intensities though.By changing the substitution pattern we impact the conformational freedom through specific intramolecular interactions. Hence, we may introduce steric interactions between the side chains when going from **1a** to **1b** and hydrogen bonding in going from **2a** to **2b**. Differences in the performance of the ML procedure can then be attributed to the interactions introduced. The influence of a wider variety in the functional groups present in the side chains, yielding more feature-rich spectra, is examined using compounds **3** and **4**. The absence of a C_2 axis in addition reveals the impact of the associated symmetry operation.

The obtained excellent quality of the spectral prediction suggests that ML can link the geometry of a conformation to its VCD spectrum. As such, ML can strongly reduce the effort spent in quantum chemically obtaining all VCD spectra provided the ML step has a much lower computational cost. This is indeed shown to be the case. On the other hand, unfortunately, the ML models are not transferable, not even within the same molecule but with different AC.

ML as well as DFT based prediction of VCD spectra are quite technical fields and every step needs to be very well thought of. Because of the highly technical nature, the precise methodology including all error checks and balances used are given in the methods section and supplementary material. The main lines of the approach taken are:

- Generate minimum energy conformations using a force field for all compounds in Figure 5.2 with chosen AC equal to RRRR
- Compute DFT geometries and VCD spectra using the B3PW91 functional and 6-31G(d) basis set
- Establish a training, validation and test set per compound to train an ML model to extract from solely a conformational geometry the VCD spectrum and test hypothesis 1 (see above)
- Repeat this for all conformations of a molecule in the chosen AC and establish the time gained by using ML (hypothesis 2)
- Test the ML model for a different AC of the same molecule or even a different molecule in the same AC or different (hypothesis 3)

Admittedly, in this study the entire usual approach involving elaborate DFT calculations is also still performed to serve as a comparison basis but the end goal is to strongly reduce the number of these calculations although some will always remain required to train the model.

5.2 Methodology

5.2.1 Conformational analysis and VCD DFT calculations

VCD spectra are very conformation dependent and so a molecular VCD spectrum for a chosen AC is composed of a set of conformer VCD spectra each weighted with their Boltzmann weight. Hence, to compute a proper VCD spectrum that takes into account all conformers and their Boltzmann weights, it is necessary to thoroughly sample the conformer ensemble within the chosen AC. The conformer geometries and VCD spectra also constitute the input for an ML model. In order to provide the model with an as diverse input as possible both low-energy and a substantial number of higher energy conformers are generated using a force-field based conformer generation algorithm. The geometry of the conformers is then optimized further using DFT and VCD spectra are calculated for each conformer. The details of each step are provided below.

- Conformer generation: A set of conformers is generated using the GMMX routine,[56] which implements a stochastic search mechanism. Conformational energies are computed with the MMFF94[57] force field as implemented in PcModel10[58]. During the stochastic search, a cut-off on the energy of the conformers equal to 40 kcal mol^{-1} is used. In practice, the generated conformers spread over a smaller range of force field energies. The high cut-off does not mean that we expect the high energy conformers to significantly impact the Boltzmann averaged spectrum but it may add to the diversity of the input for the ML stage. Second, experience shows that some interactions are not well handled at the force field stage and conformer energies may change significantly when moving to the DFT level.

- Geometry optimization and VCD spectrum generation: For each conformer, the geometry is optimized further and the VCD line spectrum is computed with the B3PW91[59] functional, the 6-31G(d) basis set and assuming the rigid rotor, ideal gas and harmonic approximation. These calculations are performed using Gaussian16[60].

- Spectrum broadening and representation: The computed conformer spectra are broadened using a Lorentzian band shape with a full width at half maximum (FWHM) of 10 cm^{-1}. The spectra were represented as vectors containing the molar absorbance difference $\Delta\varepsilon(\tilde{\nu}) = \epsilon_L(\tilde{\nu}) - \epsilon_R(\tilde{\nu})$ for wavenumbers $\tilde{\nu}$ ranging from 800 cm^{-1} to 1800 cm^{-1} using a sampling interval equal to the FWHM (10 cm^{-1}), so a 101-dimensional vector.

The distribution of the conformer DFT energies is discussed in Supplementary Methods 3.

5.2.2 ML model architecture, training and optimisation

A fully connected Feedforward Neural Network (FNN) is used in this work to extract the link between conformer geometries and their corresponding VCD spectra. The input is a vector containing molecular features describing the geometry of the conformer (for full description see section 5.2.3) and the output is the 101-dimensional vector representing its VCD spectrum. Layers of artificial neurons, so-called hidden layers, are inserted between the input and output layer. During training, the connections between the neurons establish the link between the VCD spectrum and the conformer geometry. An illustration of a FNN with two hidden layers is shown in Figure 5.3. Training a single FNN to predict VCD intensities for multiple $\tilde{\nu}$ simultaneously, improves the generalizability[61,62] of the connections between the layers.

The set of conformers for a specific AC of a single molecule is split randomly into three sets: a training, validation and test set. As mentioned earlier, the connections between the neurons are extracted from the training set. The validation set is used to optimize the so-called hyperparameters of the FNN such as its size and the algorithm used for training. The test set provides a final test of how well the FNN can predict the spectra of new conformers. Initially a default 80%:10%:10% (training:validation:test) split is used. Results of the ML approach for different splits are reported in Supplementary Discussion 5.

More technical details of the model are:

- Hyperparameter Optimization: for every application of the model the hy-

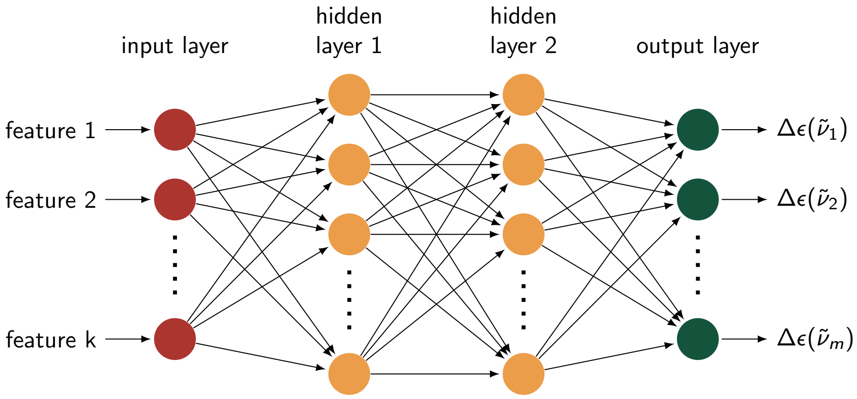

input layer
hidden layer 1
hidden layer 2
output layer
feature 1
feature 2
feature k
$\Delta\epsilon(\tilde{\nu}_1)$
$\Delta\epsilon(\tilde{\nu}_2)$
$\Delta\epsilon(\tilde{\nu}_m)$

Figure 5.3: Illustration of a FNN with two hidden layers. Molecular features such as dihedral angles are provided to the input neurons (red) and VCD intensities emerge from the output neurons (green).

perparameters are optimized using Bayesian optimization. Here, a tree-structured parzen estimator optimizes the hyperparameters within the search space shown in Table 5.1.[63] By reoptimizing the model for every application (such as compound, representation or training set size) we prevent data leaking from previous applications to the current model. The Bayesian optimization was implemented with Hyperopt 0.2.5[64].

- Dropout/Batch Normalization: during the Bayesian optimization the tree-structured parzen estimator can choose to introduce Dropout[65] for the hidden layers or batch normalization[66] layers to reduce overfitting.
- Loss function: the model is trained with the mean squared error as loss function. The exact implementation of the metric is explained in Supplementary Methods 1 and 2.

All models are built and trained on a Xeon E5-2680v4 processor using Tensorflow 2.2.0[67].

5.2.3 Molecular representation

The ML model is trained to predict the VCD conformer spectra from the conformer geometries of each molecule in turn and for a chosen AC. Ideally, a minimal set of intramolecular coordinates that fully describes the conformation is chosen

hyperparameter	values
number of hidden layers	$\{1, 2, 3, ..., 8\}$
number of neurons per hidden layer	$\{50, 60, 70, ..., 500\}$
dropout rate	$\{0, 0.05, 0.10, 0.15, 0.20\}$
activation function	$\{\mathrm{tanh}, \mathrm{elu}, \mathrm{relu}, \mathrm{selu}\}$
optimizer	$\{\mathrm{adam}, \mathrm{nadam}, \mathrm{rmsprop},$ $\mathrm{nesterov\ momentum}\}$
learning rate	$[10^{-5}, 10^{-2}]$
use of batch normalization	$\{True, False\}$
regularization type	$\{\mathrm{Lasso\ L}_1, \mathrm{Ridge\ L}_2\}$
regularization strength	$[10^{-9}, 10^{-4}]$
early stopping patience	5 epochs

Table 5.1: Hyperparameter space considered for Bayesian optimization of the FNN. For a more detailed description of the individual concepts we refer to the documentation of Tensorflow.[67]

as input for the model. For each of the six compounds, the major differences between conformers of the same compound lie in the geometrical arrangement of the sidechains. Therefore, the conformer geometry is presented to the ML model as the set of dihedral angles in the sidechains shown in Figure 5.4. Throughout this work we will refer to this set of dihedral angles as representation A. We expect this representation to capture most of the conformational flexibility. If the model cannot fully capture the link between conformer and spectrum with representation A, other geometric parameters are added to the representation and their influence is discussed.

Conformers of compounds **1a/1b/2a/2b** lacking a C_2 axis can arise in two different ways by rotating the sidechains internally, resulting in degenerate conformations which share the same VCD spectrum but are presented to the ML model as different conformations with this representation. Hence, we will teach the model that the predicted spectra need to be the same for both by explicitly including both members of such pairs.

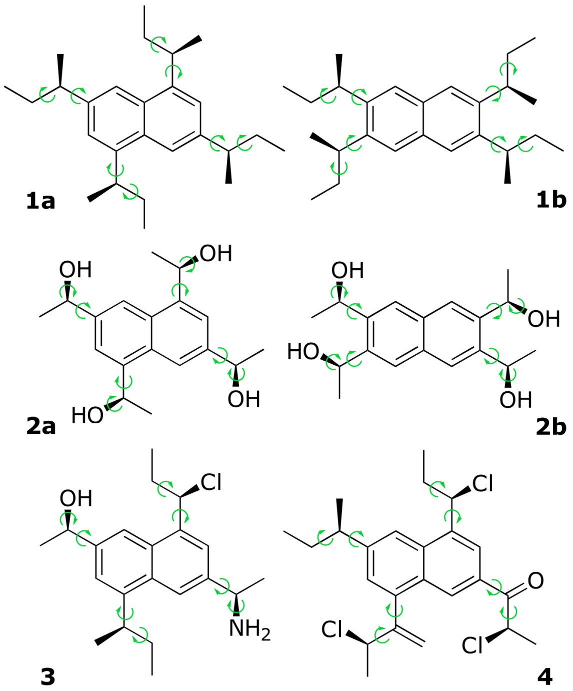


Figure 5.4: The set of dihedral angles used to describe the conformer geometries for each compound (representation A). The orientation of the NH_2 group of compound **3** was encoded as the bisector between the dihedral angles C-C-N(H_B)-H_A and C-C-N(H_A)-H_B to remove the influence of the numeric labels assigned to H_A and H_B.

5.3 Results and discussion

5.3.1 Hypothesis 1: Machine learning can predict conformer spectra solely from molecular geometry

The first hypothesis is that ML can learn from a dataset of conformer geometries and their VCD spectra the intricate link between both. To this end, for every compound, a training set of conformers and spectra is established so that an ML model can be obtained. This does -admittedly- mean that it is impossible to completely bypass all DFT calculations but the aim is to be able to limit the number to just enough to train a proper model (see supplementary material). For all compounds in Figure 5.2, an ML model was trained using different ratios of training, test and validation set and the hypothesis is examined by looking at the similarity between a DFT predicted conformer spectrum and one obtained using the trained ML model. The results are presented in Figure 5.5. For all applications, the molecular geometry is represented using only the sidechain dihedral angles.

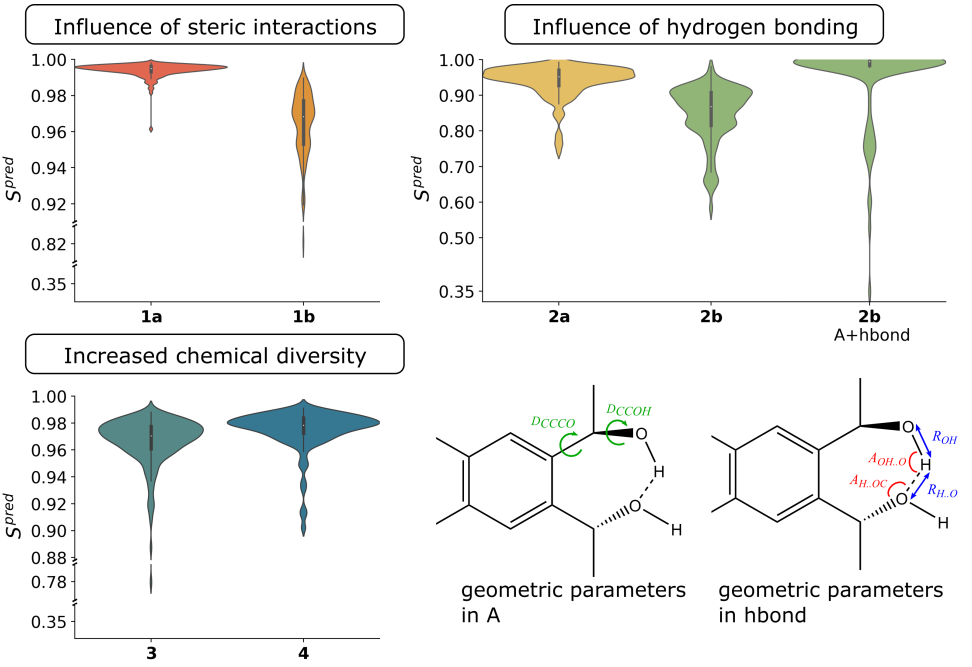


Figure 5.5: Performance of the ML approach. For each conformer of each compound, the similarity between the ML predicted spectrum and the DFT computed spectrum in the test set is shown. Separate plots are used per class of compounds in Figure 5.2.

As a similarity measure we use the cosine similarity measure S^{pred} which is the normalized overlap between the ML predicted and DFT computed spectrum (see Supplementary Methods 1 and 2 for details on the similarity measures). If it equals 1, the spectra are exactly the same. It can turn negative, meaning that the ML predicted spectrum would rather agree with the enantiomer of the DFT computed spectrum. This would be detrimental for the use of ML in VCD based AC assignment and it is gratifying that no conformers appear with negative similarities. Figure 5.5 are so-called violin plots. The width of the “blob” at every value of S^{pred} reflects how many conformers are binned within a small interval around that value. How wider the “blob” the more conformers have an S^{pred} in that bin.

Clearly, the procedure works very well in case of compound **1a**. The far majority of conformers comes with values around 0.99 and only a very small tail descends towards circa 0.96. To put this in perspective, the loss in exact similarity is of the order of or even better than the variation in spectrum if one compared DFT spectra for the same conformer obtained using a different basis set or functional. This shows that the ML procedure works very well. Compound **1b** is a structural isomer and there the results are somewhat less impressive. A vertically more spread out "blob" is obtained but the far majority of points still has an impressive similarity above 0.9. Two sets of conformers appear and one might be tempted to interpret this in terms of one collection of conformers with stronger steric hindrance and one with less, but no such connection is found (see Supplementary Discussion 2). Compound **2a** again shows that the majority of conformers exhibits very good agreement between the ML predicted and actual DFT computed spectrum although the similarities do go down to roughly 0.75. This is still more than sufficient in the context of AC determination.[68] One could suggest that conformers with higher energy lay lower in similarity, but this is not the case (see Supplementary Discussion 3). For compound **2b**, two plots are shown. The first is the result using ML training with only the sidechain dihedral angles as input. Hydrogen bonding is not well represented in this encapsulation of molecular geometry. When additional parameters are included (denoted as hbond, see bottom right panel and supplementary material), the violin plot shifts massively to higher similarities (see Supplementary Discussion 4). This means that sufficient attention must be paid to what is a proper representation of a conformer geometry. Compounds **3** and **4** introduce a wider range of substituents and it is clear that the agreement between DFT and ML predicted spectra is very good.

These results reveal that ML does allow to partially replace DFT calculations. Still, for many conformations the VCD spectrum needs to be calculated using DFT as one needs a training set for each compound but once an ML model is available, the spectra of all conformations for which no DFT calculation of the VCD spectrum was performed can be computed from the ML model. A detailed study of what fraction of conformers is required for DFT VCD calculations is given in the Supplementary Discussions 5 and 6.

5.3.2 Hypothesis 2: Machine learning can significantly reduce the computational cost for AC assignment

From a practical perspective, the scientifically already valuable results above, suggest that one could significantly reduce the effort to assign an AC to an experimental sample. In practice, assigning the AC of an experimental sample requires elaborate DFT calculations for all conformers in each possible AC, composing a Boltzmann averaged VCD spectrum and comparing it to an experimental measurement. Even if for the moment, we assume that a separate ML model needs to be trained for every assumed AC, there may be a significant time gain due to the use of ML. The most time consuming part in the usual approach lies in computing the VCD spectra, much less in the geometry optimization so for now we take for granted that the geometries and Boltzmann weights are DFT computed. One could envision to also skip the step of geometry optimization and use only geometries and energies from a force field calculation but this is subject of future work. At this exploratory stage, it is important not to reach too far in ambitions to avoid that conclusions could be based on partial error cancellation.

Compound	DFT cost classical approach	DFT cost ML-aided approach	Cost generation ML model	Time savings ML-aided approach
1a	7140 hours	5707 hours	7 hours	1426 hours
1b	5507 hours	4406 hours	4 hours	1097 hours
2a	2691 hours	2153 hours	6 hours	532 hours
2b	2025 hours	1619 hours	5 hours	401 hours
3	7293 hours	5828 hours	11 hours	1454 hours
4	4771 hours	3811 hours	6 hours	954 hours

Table 5.2: Comparison of the cost for the Boltzmann weighted spectrum with the classical approach (using the DFT computed spectra for all conformers) and the ML-aided approach where 80% of all conformer spectra come from DFT calculations and the remaining 20% are predicted with the ML model. Cost is reported in cpu time for a Intel Xeon E5-2680v4 processor.

Table 5.2 shows the total time cost for all compounds to compute a Boltzmann weighted VCD spectrum using the classical approach and using one where part of the DFT VCD calculations are replaced by the ML based prediction. To allow for a fair comparison, the time spent to train the ML model is also reported. The data in Table 5.2 is obtained using a very large fraction of DFT conformer VCD spectra (DFT spectra computed for 80% of all conformers). As the ML training step can be done quite efficiently, the relative time savings are mostly limited by the spectra generated to establish the ML model. Nonetheless, significant computer time is already being saved compared to the classical approach.

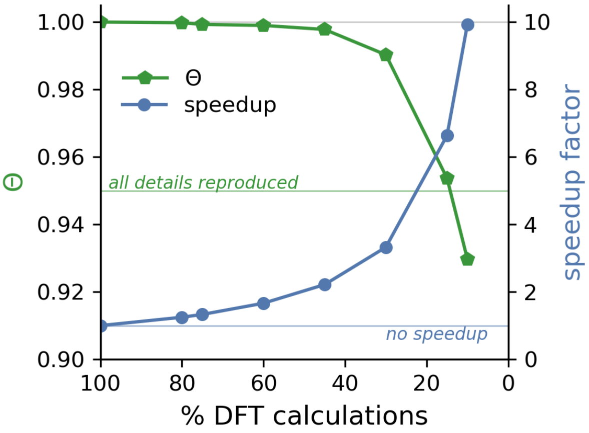


Figure 5.6: The relative speedup with the ML-aided approach (blue) using different percentages of DFT conformer spectra is shown for compound **4**. The similarity of the Boltzmann weighted spectrum obtained with the ML-aided approach and the one composed with all DFT conformer spectra (green) is determined for each percentage of DFT conformer spectra using the cosine similarity Θ (see Supplementary Methods 1 for details).

Computing DFT spectra for 80% of all conformers of course limits the possible time gain with ML. Hence, we also investigate the additional time savings possible if the ML model is generated using a smaller percentage of DFT computed spectra. Using fewer DFT spectra may adversely affect the similarity between the Boltzmann weighted spectrum composed with the DFT spectra of all conformers and one based on a combination of DFT spectra and ML predictions. Figure 5.6 shows the relative speedup for the Boltzmann weighted spectrum as a function of the percentage of DFT computed spectra, along with the similarity of the Boltzmann weighted spectrum and the one based on the DFT spectra of all conformers, for compound **4**. It is found that one can strongly reduce the percentage of DFT computed spectra without significantly affecting the resulting spectrum in the sense that the similarity to the spectrum composed with all DFT conformer spectra remains very high. At a similarity of 0.95, all details of the spectrum are still reproduced. With roughly 15% of the conformer spectra computed with DFT and used to generate the ML model, the ML-aided approach allows to retain a similarity above the threshold of 0.95 while providing a speedup with a factor of 6.6.

Similar speedup values as reported for compound **4** are also found for the other model compounds. The Boltzmann weighted spectra obtained for each compound with this approach, along with the associated speedup and similarity, are given in Supplementary Discussions 7 and 8.

5.3.3 Hypothesis 3: Machine learning can generate transferable models

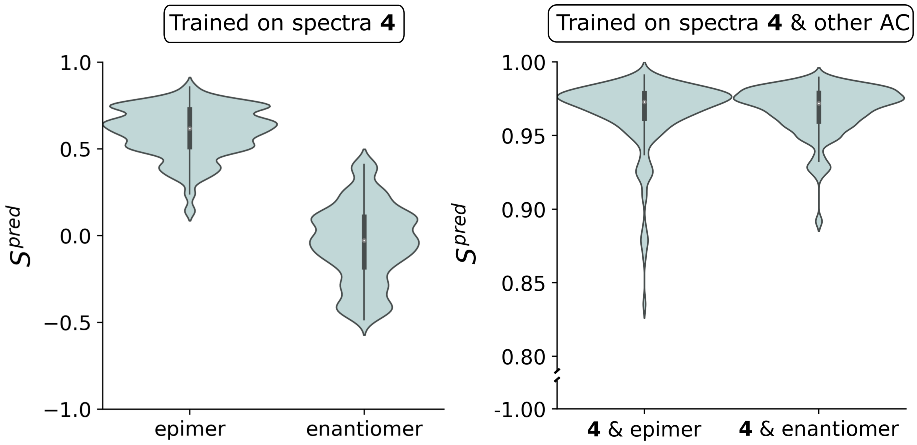


Figure 5.7: Transferability of the ML model to different stereoisomers. Left panel: similarity of the predictions for the epimer and enantiomer with the ML model trained on only compound **4** . Right panel: similarity of the predictions from the ML model trained on a combination of the conformers of compound **4** and the conformers of either the epimer or enantiomer.

All of the above is based on individually training an ML model for a specific AC of a specific molecule. The gratifying time savings reported above could be very strongly boosted if learning an ML model for a single AC would lead to a model that can also be used for a different stereoisomer. To test this we took compound **4** where ML works excellently for a single AC (see hypothesis 1 and Figure 5.5). We then switched the AC of compound **4** to both an epimer and the enantiomer, and used the ML model generated for compound **4** to predict conformer spectra for both. The results are presented in the left panel of Figure

5.7. The predictions for the new stereoisomers unfortunately do not resemble the DFT conformer spectra. The ML model is far from transferable to other AC's, especially if the conformer spectra differ significantly from the original AC. In an attempt to remedy this, one could suggest to train for multiple stereoisomers in one run. With this approach the conformer spectra of each stereoisomer are obtained with the same accuracy as for a single AC (right panel of Figure 5.7). With the current approach, the ML model can only faithfully reproduce spectra for stereoisomers it has been trained on.

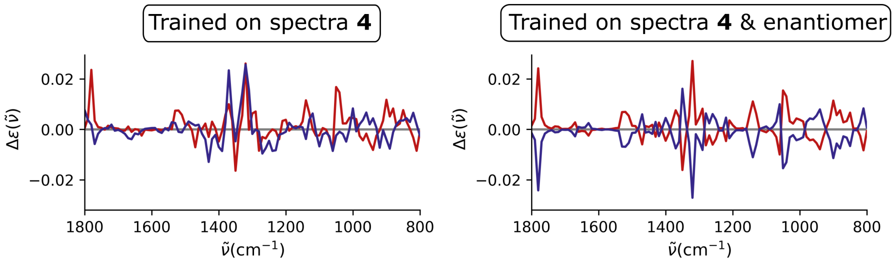


Figure 5.8: ML model predictions for enantiomeric conformer pairs. Left panel: prediction for a selected conformer of the test set of **4** (red) and its mirror image (blue) when the ML model is trained on conformers of **4**. Right panel: prediction for the conformer and its mirror image when the ML model is trained on conformers of both **4** and the enantiomer.

Figure 5.7 also reveals a particular feature. DFT spectra are always only computed for one enantiomer of the set of enantiomers as spectra of enantiomers are mirror images. It is clear that this was not picked up when training on only one of both enantiomers. When training on both enantiomers, the question is whether enough information was sourced such that for the two mirror images of the same conformation also a mirror image spectrum is obtained from the ML model. This is indeed the case as is shown in Figure 5.8 where in the left panel the spectra of an enantiomeric pair of conformers is compared when only one AC was used in training. In the right panel, the result is shown when both AC are included in training.

5.4 Conclusions

The potential of ML in VCD spectroscopy to (partially) replace DFT calculations was examined. Three hypotheses have been put forward, leading to the following conclusions:

- Hypothesis 1: Machine Learning can predict conformer spectra solely from molecular geometries.
  The similarity between the DFT computed spectrum of a conformer and the spectrum predicted with ML from its geometry is very high. ML can indeed learn the intricate and hidden connection between a conformer geometry and its VCD spectrum. Though, it is up to the user to make sure that the representation of the geometry in a practical form encapsulates all the necessary input to cover intramolecular interactions.

- Hypothesis 2: Machine Learning can significantly reduce the computational cost for AC assignment.
  The present results show that the ML training step may be done quite efficiently and as a result significant time savings are possible. Obviously, it remains up to the user to determine whether the time savings compensate for the learning curve associated with proper training in ML methods.

- Hypothesis 3: Machine Learning can generate transferable models.
  The current design architecture does not result in transferable ML models, neither between molecules nor among different AC's of the same molecule.

The current ML approach already satisfies 2 out of 3 hypotheses. Clearly, more development on the ML methodology is still needed to satisfy hypothesis 3. Nonetheless, ML shows promise as a tool for extracting the link between conformations and VCD spectra.

Supplementary discussions

1 S^{conf} distribution of IR & VCD spectra

Prior to deploying an ML model to predict conformer spectra, we need to establish the variability of spectra between conformers of the same compound. If the conformer spectra within the same AC of a single compound are all very similar, high S^{pred} values can be obtained without the ML model extracting a meaningful pattern between geometry and spectrum. If $\overline{S^{pred}}$ outclasses the mean similarity between DFT conformer spectra themselves, the model has successfully established a link between the geometry of a conformer and its spectrum. For each compound, the cosine similarity between all unique pairs of conformer spectra (S^{conf}; see equation 5.3) is determined. This analysis is performed for both the IR and VCD conformer spectra separately. Note that for the IR spectra S^{conf} can have values between 0 and 1 while for VCD spectra S^{conf} can have values between -1 and 1. The distribution of S^{conf} is depicted in Figure 5.9 for the IR conformer spectra and Figure 5.10 for the VCD conformer spectra. A violin plot is used to describe the distribution of S^{conf} for each compound. The width of the “blob” reflects the number of conformer pairs found within a small interval around that S^{conf} value and the added boxplot shows the median value and interquartile range.

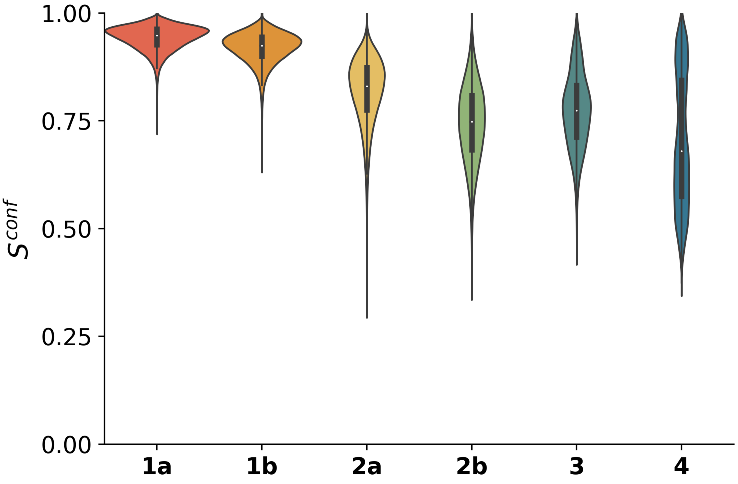


Figure 5.9: Similarity of conformer IR spectra. The cosine similarity between IR spectra corresponding to unique conformer pairs are reported for each compound. For the different compounds following $\overline{S^{conf}}$ values are obtained: 0.942 (**1a**), 0.919 (**1b**), 0.818 (**2a**), 0.744 (**2b**), 0.773 (**3**), 0.703 (**4**).

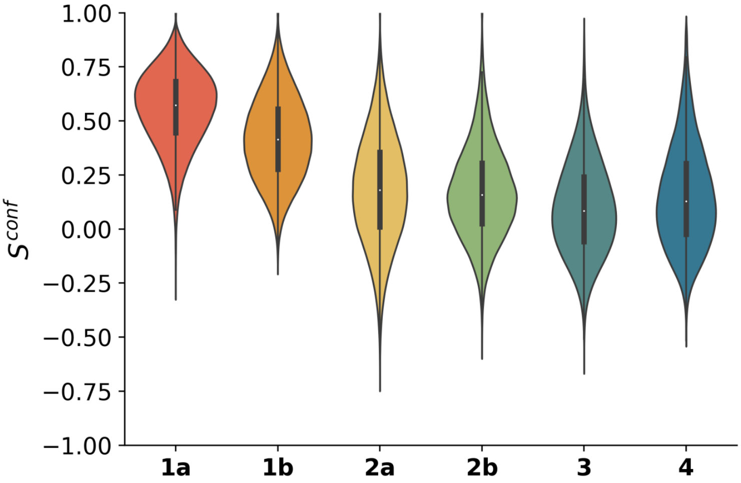


Figure 5.10: Similarity of conformer VCD spectra. The cosine similarity between VCD spectra corresponding to unique conformer pairs are reported for each compound. For the different compounds following $\overline{S^{conf}}$ values are obtained: 0.556 (**1a**), 0.418 (**1b**), 0.182 (**2a**), 0.170 (**2b**), 0.097 (**3**), 0.150 (**4**).

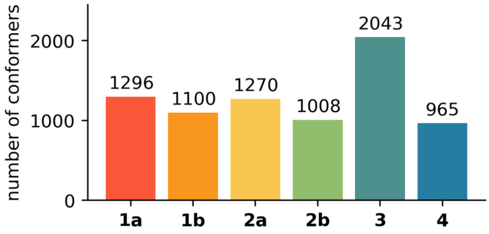


Figure 5.11: Number of conformer spectra for each compound.

The variability between IR conformer spectra is rather weak: For compound **1a** and **1b**, $\overline{S^{conf}}$ exceeds 0.9 and for compound **2a** $\overline{S^{conf}}$ is larger than 0.8. For these compounds any differences between the conformer IR spectra are subtle at best. For compounds **2b**, **3** and **4** lower $\overline{S^{conf}}$ values are obtained, though the overall sensitivity of the IR spectra to conformational differences remains limited.

The VCD conformer spectra are more sensitive to conformational differences than the IR conformer spectra. For each compound, the $\overline{S^{conf}}$ obtained for the VCD spectra is significantly lower than the $\overline{S^{conf}}$ for the corresponding IR spectra. For the VCD spectra, the largest $\overline{S^{conf}}$ is obtained for **1a** with a mean value of 0.556. Introduction of steric interactions increases the variability between conformer VCD spectra resulting in a lower $\overline{S^{conf}}$ of 0.418 (**1b**). When hydrogen bonding interactions are introduced, $\overline{S^{conf}}$ drops slightly from 0.182 (**2a**) to 0.170 (**2b**). The largest variability between conformer spectra is observed for the compounds with increased chemical diversity with $\overline{S^{conf}}$ values of 0.097 (**3**) and 0.150 (**4**). Interestingly, the lowest $\overline{S^{conf}}$ is obtained for **3** which has the largest number of possible conformers (Figure 5.11).

Altogether, the VCD spectra are clearly sensitive to conformational differences. The IR spectra are less sensitive to these conformational differences which limits the added value of the ML approach compared to VCD. The focus in this paper will therefore lie on assessing the ML approach for VCD.

2 Impact of steric interactions on S^{pred} for 1b

The steric interactions between adjacent *sec*-butyl side chains for compound **1b** decreases the $\overline{S^{pred}}$ compared to compound **1a**. One could, naively, expect that this results from lower S^{pred} values for conformers with larger steric hindrance. We test this notion using the lowest H..H distance between the CH_2 and the $(CH_2)CH_3$ groups of adjacent sidechains within each individual conformer (see Figure 5.12) as a metric for steric interaction. Figure 5.13 shows that the S^{pred} value obtained for a conformer is not linked to the steric hindrance within said conformer.

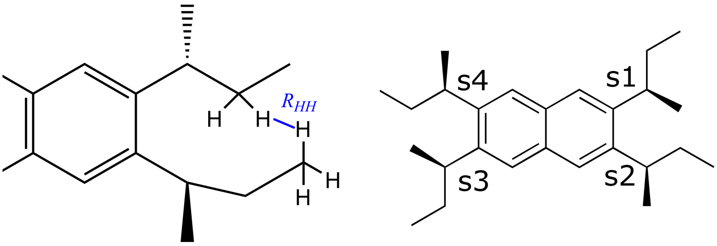


Figure 5.12: Minimum H..H distance used to describe steric interactions between the sidechains. The H..H distance shown in the left panel is calculated for each pair of adjacent sidechains (s1-s2, s2-s1, s3-s4, s4-s3; see right panel) and the minimum value of these H..H distances is determined for each individual conformer.

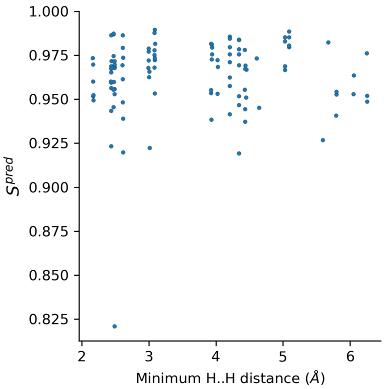


Figure 5.13: Influence of minimum H..H distance on S^{pred} for each conformer in the test set of **1b**.

3 Comparison of conformer $\Delta H^0_{298.15}$ and S^{pred}

In Figures 5.14-5.19 $\Delta H^0_{298.15}$ and S^{pred} are compared for every test set conformer of a compound. Figure dimensions were adapted for Figure 5.15 and 5.17 to accommodate for the larger range of $\Delta H^0_{298.15}$ for **1b** and of S^{pred} for **2b**. In general, there is no pattern showing that lower S^{pred} values are obtained for conformers with larger $\Delta H^0_{298.15}$. Thus, the accuracy of the ML model predictions is not biased towards conformers with lower $\Delta H^0_{298.15}$.

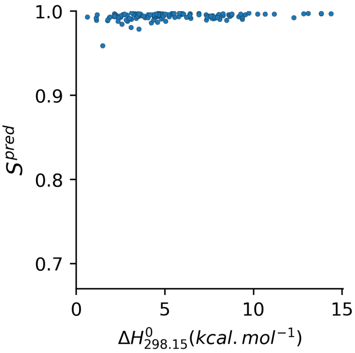


Figure 5.14: Comparison of $\Delta H^0_{298.15}$ and S^{pred} for the test set of compound **1a**.

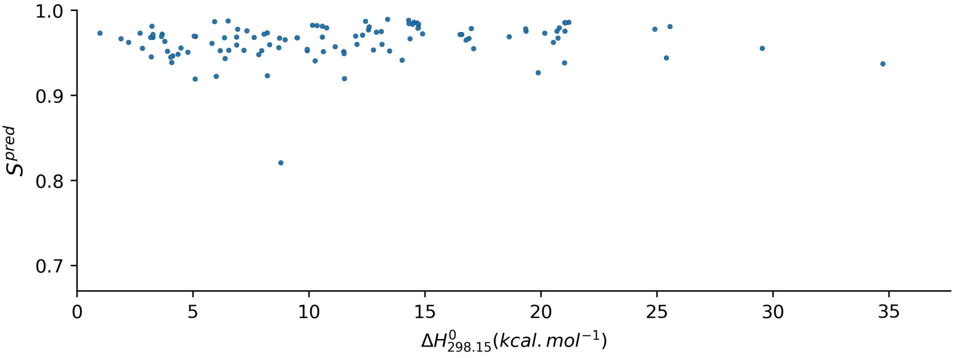


Figure 5.15: Comparison of $\Delta H^0_{298.15}$ and S^{pred} for the test set of compound **1b**.

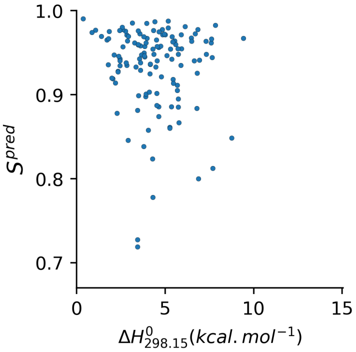


Figure 5.16: Comparison of $\Delta H^0_{298.15}$ and S^{pred} for the test set of compound **2a**.

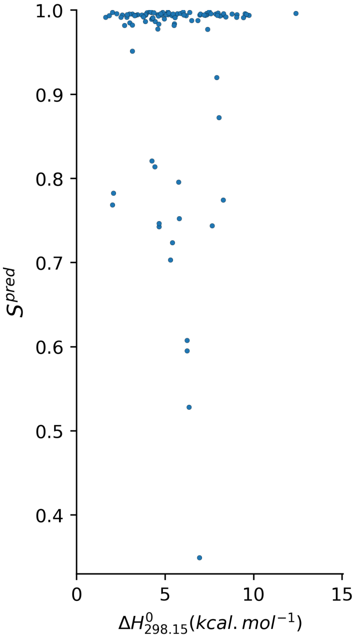


Figure 5.17: Comparison of $\Delta H^0_{298.15}$ and S^{pred} for the test set of compound **2b**.

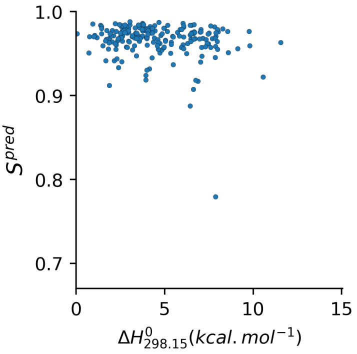


Figure 5.18: Comparison of $\Delta H^0_{298.15}$ and S^{pred} for the test set of compound **3**.

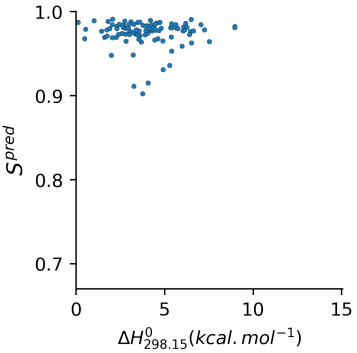


Figure 5.19: Comparison of $\Delta H^0_{298.15}$ and S^{pred} for the test set of compound **4**.

4 Influence of representation on performance for individual conformers of 2b

For compound **2b** lower S^{pred} values are obtained with representation A compared to the other compounds. Representation A might not sufficiently capture the hydrogen bonding in a way understandable to the ML model. To improve the accuracy of the ML predicted spectra, the additional parameters shown in Figure 5.5 are used to describe the geometry of the conformers. Doing so, the S^{pred} shifts to higher values as a result (see Figure 5.20). For a large majority of the conformers in the test set S^{pred} increases with the additional parameters (see Figure 5.21), though for a handful of conformers lower S^{pred} values are obtained with the new representation. Including these hbond parameters does not negatively impact generalization as indicated by the generalization factor reported in Supplementary Discussion 6. To summarize: including the hydrogen bond angles and distances allows the ML model to extract a more detailed link between the conformer and its VCD spectrum though a small fraction of the conformers do not fit in the pattern obtained with this new representation.

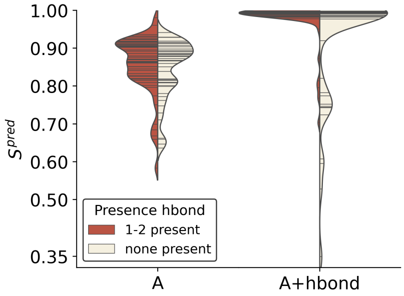


Figure 5.20: Distribution of the S^{pred} values for conformers in the test set of compound **2b** with (red) and without (white) intramolecular hydrogen bonds for both representations. Individual S^{pred} values are shown as horizontal bars within the violin plots.

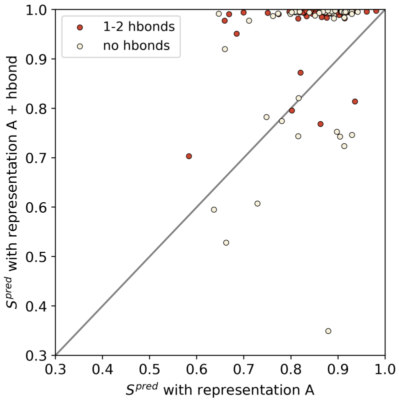


Figure 5.21: S^{pred} values obtained with representation A and A+hbond for each conformer in the test set of compound **2b**. Presence of an intramolecular hydrogen bond within a conformer is denoted by colour. Conformers lying above the diagonal have increased S^{pred} values upon adding the hbond parameters.

5 $\overline{S^{pred}}$ on VCD conformer spectra with different data splits

As demonstrated in the section covering hypothesis 1 (see Figure 5.5), excellent performance can be obtained for the different compounds with the default data split of 80:10:10 (training:validation:test). The ML model can clearly establish the link between conformer and spectrum. As the cost of the entire ML workflow is equal to computing 1-3 DFT spectra for this data split, the test set spectra are obtained at a fraction of the computational cost of the conventional DFT procedure. With a large majority of the conformers used as training samples, the relative speed-up for obtaining all DFT conformer spectra for a single compound remains rather limited. To establish the relative speed-up obtainable whilst maintaining good predictive quality, the performance of the ML approach is examined for different data splits in this section.

The training and optimization procedure used in the previous sections, is repeated for all compounds while incrementally relocating conformers from the training set to the validation set. Repeating the optimization of the ML model ensures that no data leakage occurs (i.e. all information from the original training set is erased) and allows to decrease model complexity to accommodate for smaller training sets. The same test set is used across the different splits. Doing so, $\overline{S^{pred}}$ is established in a consistent manner and the influence of the training set size is isolated. Representation A+hbond is used for compound **2b** and representation A for the other compounds. The $\overline{S^{pred}}$ values obtained for the different splits and compounds are shown in Figure 5.22.

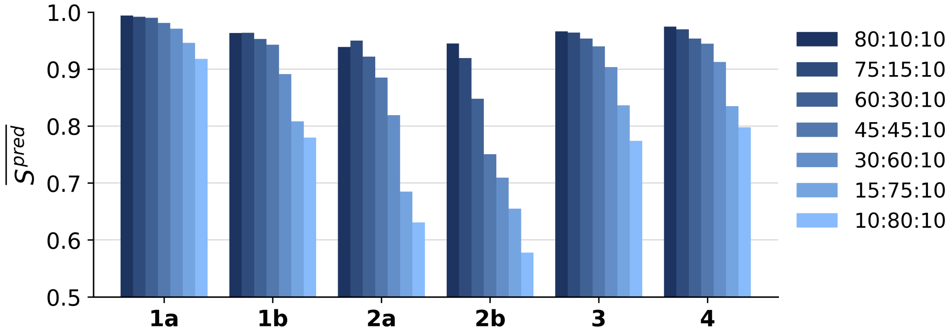


Figure 5.22: Performance of the ML approach for different training set sizes. For each compound and data split, $\overline{S^{pred}}$ is shown. Training set size is denoted by colour.

For **1a** excellent $\overline{S^{pred}}$ is retained across the different splits. Even when only 10% of the conformers is used for training, the test set spectra are reproduced with a $\overline{S^{pred}}$ of 0.918. For **1b** the conformer spectra are predicted with excellent $\overline{S^{pred}}$ when at least 45% of the conformers reside in the training set. For **2a** and, to a larger extent, **2b** $\overline{S^{pred}}$ drops quickly when less training samples are provided. While a training set size of 60% still yields excellent $\overline{S^{pred}}$ for **2a**, with the same training set size a $\overline{S^{pred}}$ of 0.848 is obtained for **2b**. For **3** and **4**, the decrease in $\overline{S^{pred}}$ with smaller training sets is less steep: both compounds retain excellent $\overline{S^{pred}}$ if at least 30% of the spectra are included in the training set. These $\overline{S^{pred}}$ values are very impressive as for **3** and **4** the conformer spectra varied the most (see Supplementary Discussion 1).

The influence of vibrational mode delocalization provides an explanation for the steeper performance drop for **2a** compared to **3** and **4**. Whereas the majority of the vibrational modes within the considered wavenumber range involve only a single sidechain (along with the naphthalene moiety) for **3** and **4**, for **2a** the vibrational modes are delocalized over the entire compound. As discussed in the main paper, the sidechains are largely independent from each other in conformer space. However, they do correlate through the delocalized vibrational modes, along with the corresponding vibrational frequencies and VCD intensities. Providing a large majority of the spectra for training, this complex pattern can still be extracted by the ML model. With a small training set, too few examples are provided to the ML model to properly learn these correlations. While the same degree of delocalization is expected for **1a**, the conformer VCD spectra are more similar (see Supplementary Discussion 1). Thus, obtaining good performance with a smaller training set is relatively less challenging.

The influence of intramolecular interactions is more pronounced for smaller training sets. Correlation between sidechains in conformer space complicates the spectral prediction problem. In case of hydrogen bonding $\overline{S^{pred}}$ drops more sharply. As intramolecular hydrogen bonding has a strong influence on the vibrational modes, this correlation is even stronger and more complex in **2b**. Thus, when less examples of these correlations are provided, properly accounting for the different correlations becomes more challenging for the ML model.

6 Training, validation and test set MSE for different data splits

In this section the MSE metrics are reported for the training (MSE_{train}), validation (MSE_{val}) and test set (MSE_{test}). While the performance on the test set is the ultimate test of the ML approach, the ratios between the MSE metrics (see equation 5.1 and 5.2) can also be of interest for future reference. The use of these ratios was inspired on the work of Röbel[69,70]. The ratios are shown in Figure 5.23 and the values of the MSE metrics are reported in Table 5.3.

$$\rho_{train} = \frac{MSE_{test}}{MSE_{train}} \tag{5.1}$$

$$\rho_{val} = \frac{MSE_{test}}{MSE_{val}} \tag{5.2}$$

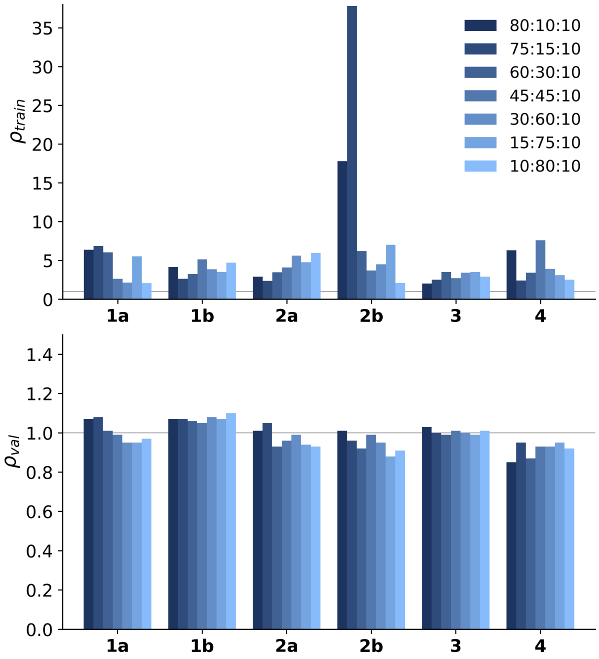


Figure 5.23: Bar plot of ρ_{train} (top) and ρ_{val} (bottom) for each compound and training set size (denoted by colour). Representation A+hbond was used for compound **2b** and representation A for the other compounds.

compound	representation	training set size (%)	ρ_{train}	ρ_{val}	MSE_{train}	MSE_{val}	MSE_{test}	$\overline{S^{pred}}$
1a	A	80	6.4	1.07	0.002	0.011	0.012	0.994
1a	A	75	6.9	1.08	0.002	0.015	0.016	0.992
1a	A	60	6.0	1.01	0.003	0.020	0.020	0.990
1a	A	45	2.6	0.99	0.014	0.038	0.037	0.981
1a	A	30	2.1	0.95	0.027	0.060	0.057	0.971
1a	A	15	5.5	0.95	0.019	0.113	0.107	0.946
1a	A	10	2.1	0.97	0.077	0.164	0.160	0.918
1b	A	80	4.1	1.07	0.017	0.065	0.070	0.964
1b	A	75	2.6	1.07	0.027	0.067	0.072	0.964
1b	A	60	3.2	1.06	0.029	0.088	0.093	0.953
1b	A	45	5.1	1.05	0.021	0.103	0.108	0.943
1b	A	30	3.9	1.08	0.054	0.194	0.210	0.891
1b	A	15	3.5	1.07	0.104	0.339	0.364	0.808
1b	A	10	4.7	1.10	0.086	0.368	0.405	0.780
2a	A	80	2.9	1.01	0.033	0.095	0.096	0.939
2a	A	75	2.4	1.05	0.036	0.081	0.085	0.950
2a	A	60	3.5	0.93	0.036	0.134	0.125	0.922
2a	A	45	4.1	0.96	0.047	0.201	0.192	0.885
2a	A	30	5.6	0.99	0.054	0.307	0.303	0.819
2a	A	15	4.8	0.94	0.103	0.521	0.490	0.685
2a	A	10	6.0	0.93	0.094	0.599	0.559	0.631
2b	A+hbond	80	17.8	1.01	0.006	0.106	0.107	0.945
2b	A+hbond	75	37.8	0.96	0.004	0.158	0.151	0.919
2b	A+hbond	60	6.2	0.92	0.042	0.282	0.260	0.848
2b	A+hbond	45	3.7	0.99	0.112	0.419	0.413	0.751
2b	A+hbond	30	4.5	0.95	0.106	0.499	0.472	0.709
2b	A+hbond	15	7.0	0.88	0.076	0.605	0.534	0.655
2b	A+hbond	10	2.1	0.91	0.302	0.701	0.635	0.578
3	A	80	2.0	1.03	0.032	0.065	0.067	0.966
3	A	75	2.5	1.00	0.029	0.070	0.070	0.964
3	A	60	3.5	0.99	0.026	0.091	0.090	0.954
3	A	45	2.7	1.01	0.043	0.113	0.114	0.940
3	A	30	3.4	1.00	0.054	0.182	0.182	0.904
3	A	15	3.5	0.99	0.086	0.302	0.300	0.836
3	A	10	2.9	1.01	0.136	0.389	0.393	0.774
4	A	80	6.3	0.85	0.007	0.056	0.048	0.975
4	A	75	2.4	0.95	0.024	0.059	0.057	0.970
4	A	60	3.4	0.87	0.025	0.098	0.084	0.953
4	A	45	7.6	0.93	0.014	0.111	0.103	0.945
4	A	30	3.9	0.93	0.041	0.172	0.159	0.913
4	A	15	3.1	0.95	0.093	0.304	0.288	0.835
4	A	10	2.5	0.92	0.139	0.379	0.347	0.798

Table 5.3: Mean squared errors on the train, validation and test set (standard scaled) spectra for the different compounds and splits.

7 Construction of Boltzmann weighted spectrum with ML predictions

In the section covering hypothesis 1 we have established that VCD conformer spectra can be accurately predicted with an ML model. A remaining question is the extent to which the Boltzmann weighted spectrum obtained with the classical approach used in most applications of VCD, can be approximated using the predictions of the ML model. We address this question for each compound by constructing a Boltzmann weighted spectrum $\Delta\varepsilon^{ML}(\tilde{\nu})$ that uses all DFT enthalpies along with the DFT spectra of the training set and ML predicted spectra for the validation and test sets. The Boltzmann weighted spectrum obtained with the classical approach i.e. using DFT enthalpies and DFT spectra for all conformers, is referred to as $\Delta\varepsilon^{DFT}(\tilde{\nu})$. The similarity of $\Delta\varepsilon^{DFT}(\tilde{\nu})$ and $\Delta\varepsilon^{ML}(\tilde{\nu})$, referred to as Θ, is determined for each compound (using equation 5.8) and shown in Figure 5.24. For the 80:10:10 split, $\Delta\varepsilon^{ML}(\tilde{\nu})$ and $\Delta\varepsilon^{DFT}(\tilde{\nu})$ are completely indistinguishable ($\Theta > 0.999$). So, by replacing 20% of the conformer DFT spectra with ML predictions, we introduce the significant time savings reported in Table 5.2 without losing even the tiniest details of $\Delta\varepsilon^{DFT}(\tilde{\nu})$.

The results in Supplementary Discussion 5 indicate that with fewer conformers in the training set the accuracy of the ML predicted conformer spectra does decrease. By constructing $\Delta\varepsilon^{ML}(\tilde{\nu})$ with a larger fraction of ML predicted spectra, we test whether a similar drop is observed for Θ. Note that $\Delta\varepsilon^{ML}(\tilde{\nu})$ is constructed for each of the smaller training sets using the ML model obtained with said training set. Again, $\Delta\varepsilon^{ML}(\tilde{\nu})$ reproduces $\Delta\varepsilon^{DFT}(\tilde{\nu})$ with excellent accuracy across all different splits. When at least 30% of all conformers (or 45% for **2b**) are included in the training set, $\Delta\varepsilon^{ML}(\tilde{\nu})$ replicates even the tiniest details of $\Delta\varepsilon^{DFT}(\tilde{\nu})$ ($\Theta \geq 0.99$). A slight decrease in Θ is noted for even smaller training sets but Figure 5.25 shows that $\Delta\varepsilon^{DFT}(\tilde{\nu})$ and $\Delta\varepsilon^{ML}(\tilde{\nu})$ remain very similar. So, for these compounds $\Delta\varepsilon^{DFT}(\tilde{\nu})$ is approximated very well by leveraging the ML predicted conformer spectra.

Curiously, Θ decreases more steeply for **4** compared to **3** despite the larger $\overline{S^{conf}}$ and larger/similar $\overline{S^{pred}}$ for **4**. A possible explanation for this lies in the cancellation of VCD intensities during Boltzmann averaging of conformer spectra. VCD intensities of opposite sign for different conformers can partially cancel each other. Errors in the predicted VCD intensities can be averaged out in a similar manner. The degree of error compensation will depend on the sign and (Boltzmann weighted) magnitude of the errors and thus differ between compounds.

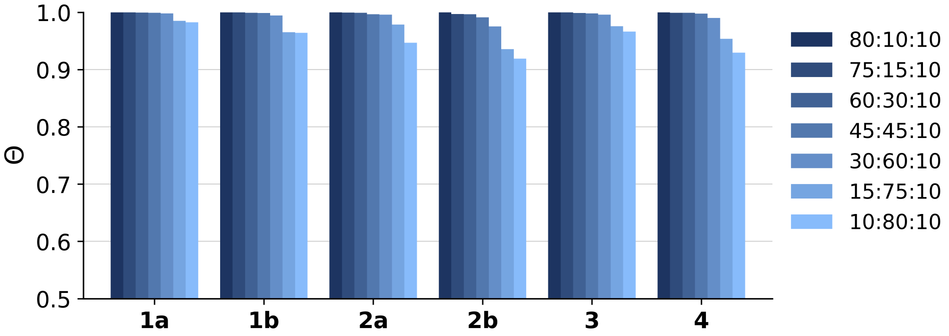


Figure 5.24: Similarity Θ of the Boltzmann weighted spectra $\Delta\varepsilon^{DFT}(\tilde{\nu})$ and $\Delta\varepsilon^{ML}(\tilde{\nu})$ for each compound and split. Training set size is denoted by colour.

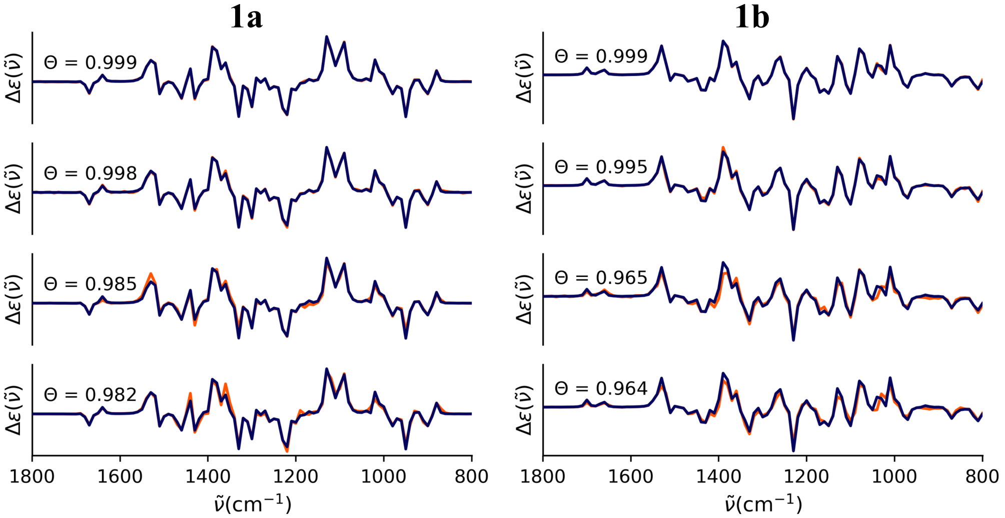


Figure 5.25: Comparison of $\Delta\varepsilon^{DFT}(\tilde{\nu})$ (blue) and $\Delta\varepsilon^{ML}(\tilde{\nu})$ (orange) for the different compounds (as denoted) and smaller training set sizes. For each compound $\Delta\varepsilon^{ML}(\tilde{\nu})$ is shown as obtained for the following training set sizes (from top to bottom): 45%, 30%, 15% and 10%. Values of Θ are reported for each split and compound in the corresponding figures.

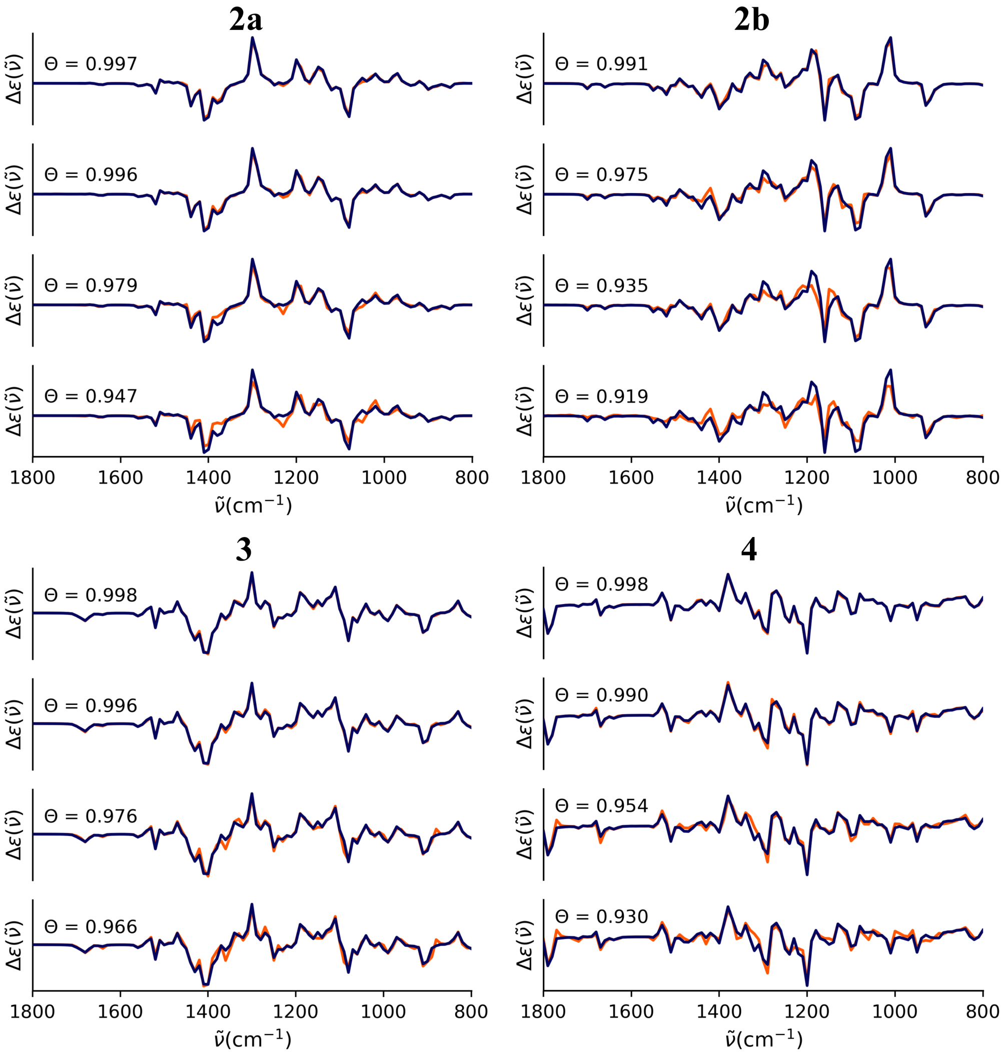


Figure 5.25: Comparison of $\Delta\varepsilon^{DFT}(\tilde{\nu})$ (blue) and $\Delta\varepsilon^{ML}(\tilde{\nu})$ (orange) for the different compounds (as denoted) and smaller training set sizes. For each compound $\Delta\varepsilon^{ML}(\tilde{\nu})$ is shown as obtained for the following training set sizes (from top to bottom): 45%, 30%, 15% and 10%. Values of Θ are reported for each split and compound in the corresponding figures.(*cont.*)

8 Relative speedup for Boltzmann weighted spectrum

The ratio in the cost of the classical approach (using only DFT spectra) and the ML-aided approach (including time spent in the ML model generation), referred to as the relative speedup, is determined for each compound and split. The result is shown in Figure 5.26 for all compounds along with the corresponding similarity Θ previously reported in Figure 5.24. As the ML training step can be done very efficiently, the speedup is inversely proportional to the training set size. Very significant speedup may be obtained by limiting the percentage of conformations for which the spectrum is computed by DFT and used for training the ML model. As this percentage grows smaller, some similarity loss is found but putting the limit at 0.95, it is clear that one may strongly lower this percentage before the spectrum starts to deviate substantially from the spectrum obtained using only DFT conformer spectra (see Figure 5.25).

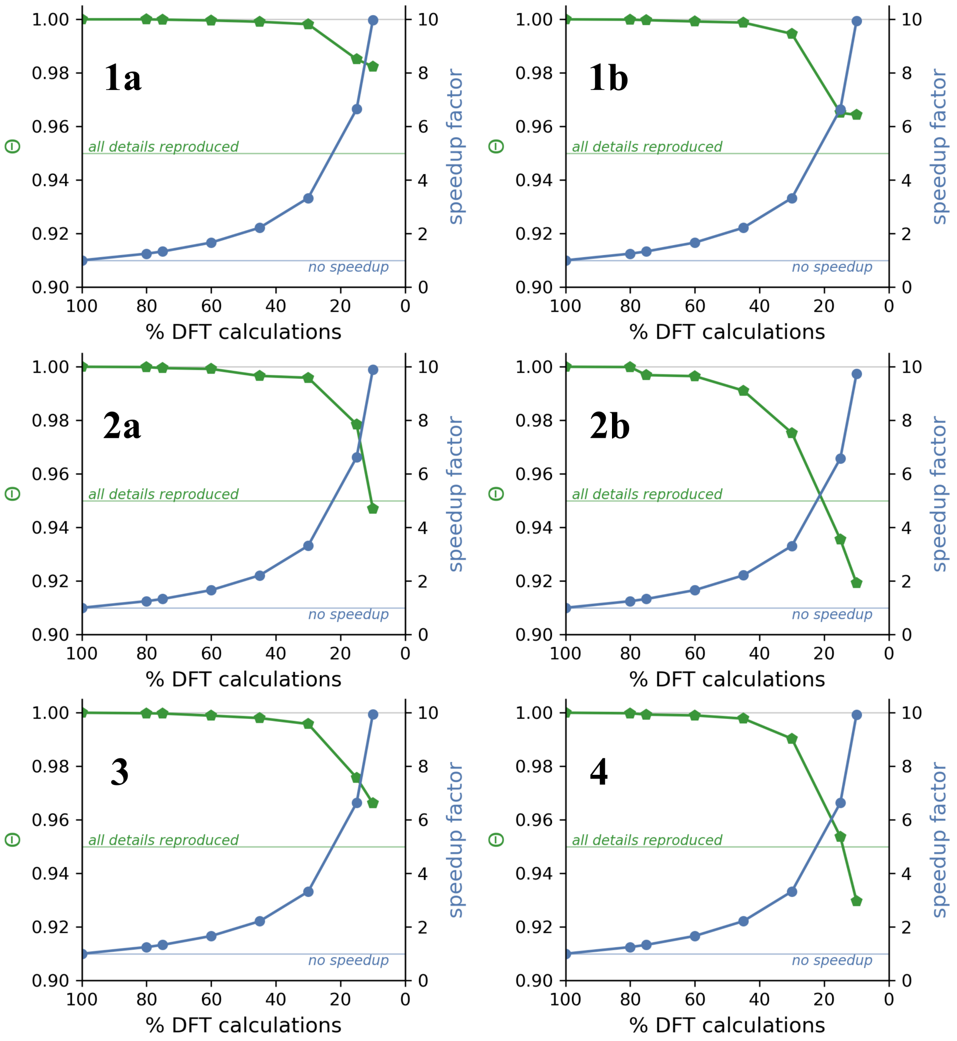


Figure 5.26: Comparison of the relative speedup obtained with the ML-aided approach (blue) and Θ (green) for each compound and training set size.

Supplementary methods

1 Spectral similarity and model performance

In this paper we probe the ability of ML to predict conformer spectra directly from the geometry of the conformers. The metrics used to express the similarity of different spectra and to train the ML model play a pivotal role and hence a more detailed discussion on these metrics is warranted. The cosine similarity (also known as the overlap integral or Carbó index) is often used within the VCD community to define the relative similarity of two spectra.[68] The similarity metric adopts values between -1 and 1, where 1 indicates the spectra to be identical and -1 identifies the spectra as perfect mirror images. The similarity metrics used within this paper are detailed below and a summary is provided in Table 5.4 for future reference. Their use will become clear in later sections.

- A first similarity measure reflects the relative similarity of DFT calculated conformer spectra. Here, the cosine similarity S_{ij}^{conf} of two calculated spectra $\Delta\varepsilon_i^{calc}(\tilde{\nu})$ and $\Delta\varepsilon_j^{calc}(\tilde{\nu})$ is determined for a pair of conformers i and j (equation 5.3). The mean value of S^{conf} over all unique pairs of the M conformers is denoted $\overline{S^{conf}}$ (equation 5.4).

$$S_{ij}^{conf} = \frac{\sum\limits_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_i^{calc}(\tilde{\nu}) \cdot \Delta\varepsilon_j^{calc}(\tilde{\nu})\right)}{\sqrt{\sum\limits_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_i^{calc}(\tilde{\nu})\right)^2} \sqrt{\sum\limits_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_j^{calc}(\tilde{\nu})\right)^2}} \tag{5.3}$$

$$\overline{S^{conf}} = \frac{2}{M(M-1)} \sum_{i}^{M-1} \sum_{j>i}^{M} S_{ij}^{conf} \tag{5.4}$$

- The ML model is trained using the mean squared error (MSE) between the scaled DFT conformer spectra $\Delta\varepsilon^{calc,sc}(\tilde{\nu})$ and the predictions for the scaled spectra $\Delta\varepsilon^{pred,sc}(\tilde{\nu})$ (equation 5.5) for all N conformers in a set (training, validation or test). A detailed explanation of the scaling methodology is provided in Supplementary Methods 2. After training and optimization, the predictions for the scaled spectra are transformed to the same scale as $\Delta\varepsilon^{calc}(\tilde{\nu})$, resulting in $\Delta\varepsilon^{pred}(\tilde{\nu})$.

$$MSE = \frac{1}{101 \cdot N} \sum_{i}^{N} \sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_i^{calc,sc}(\tilde{\nu}) - \Delta\varepsilon_i^{pred,sc}(\tilde{\nu}) \right)^2 \tag{5.5}$$

We do not train the ML model using a cosine similarity as it is a relative similarity metric. If trained with a cosine similarity, the ML model only learns to recreate the shape of a conformer spectrum but not the overall VCD intensity $\sum_{\tilde{\nu}=800}^{1800}(\Delta\varepsilon_i^{calc}(\tilde{\nu}))^2$ for each conformer i. The mismatch in this intensity between the DFT and ML predicted spectra will be different for each conformer. As a result, the ML predictions can no longer be used to build a Boltzmann weighted spectrum.

- Once the ML model is trained and optimized, the relative similarity of a DFT calculated spectrum $\Delta\varepsilon_i^{calc}(\tilde{\nu})$ of conformer i and the spectrum predicted by the ML model for said conformer $\Delta\varepsilon_i^{pred}(\tilde{\nu})$ is expressed using the cosine similarity S_i^{pred} (equation 5.6). The mean value of S^{pred} over all N conformers within the test set is denoted $\overline{S^{pred}}$ (equation 5.7).

$$S_i^{pred} = \frac{\sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_i^{calc}(\tilde{\nu}) \cdot \Delta\varepsilon_i^{pred}(\tilde{\nu}) \right)}{\sqrt{\sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_i^{calc}(\tilde{\nu}) \right)^2} \sqrt{\sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon_i^{pred}(\tilde{\nu}) \right)^2}} \tag{5.6}$$

$$\overline{S^{pred}} = \frac{1}{N} \sum_{i}^{N} S_i^{pred} \tag{5.7}$$

In the context of this paper we see the ML prediction as excellent for S^{pred} values exceeding 0.9 as any remaining errors in the spectrum are hardly visible to the human eye. In Figure 5.27 we show this using some examples for different S^{pred} values.

- Eventually, the Boltzmann weighted spectrum $\Delta\varepsilon^{DFT}(\tilde{\nu})$ computed entirely from DFT spectra is compared with the Boltzmann weighted spectrum $\Delta\varepsilon^{ML}(\tilde{\nu})$ where DFT spectra of the training set are combined with the ML predicted spectra for the remaining conformers. The similarity of both

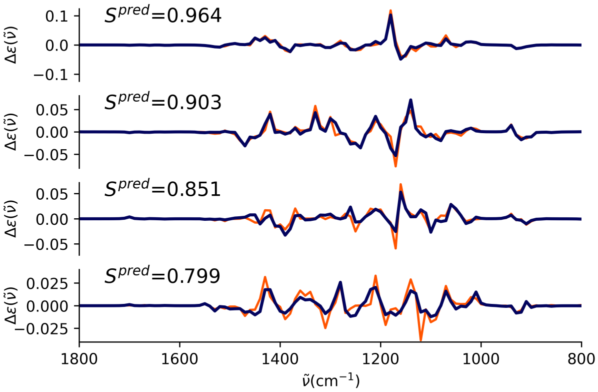


Figure 5.27: Comparison of DFT conformer spectra (orange) and corresponding ML predicted spectra (blue) for different S^{pred} values. The DFT spectra and ML predictions are taken from Supplementary Discussion 4.

Boltzmann weighted spectra is determined with Θ (equation 5.8).

$$\Theta = \frac{\sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon^{DFT}(\tilde{\nu}) \cdot \Delta\varepsilon^{ML}(\tilde{\nu})\right)}{\sqrt{\sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon^{DFT}(\tilde{\nu})\right)^2} \sqrt{\sum_{\tilde{\nu}=800}^{1800} \left(\Delta\varepsilon^{ML}(\tilde{\nu})\right)^2}} \tag{5.8}$$

2 Data scaling and influence on MSE

It is common practice in ML applications to standardize the features in the dataset such that each feature (e.g. a dihedral angle of a compound) has a mean value of 0 and a standard deviation of 1. This feature scaling typically improves the training of the ML model and is even required for many ML algorithms.[66,71,72] To prevent issues during training as a result of the low intensity of the VCD conformer spectra, we also scale the conformer spectra $\Delta\varepsilon^{calc}(\tilde{\nu})$ for each compound during the training of the ML model. Using the scaling methodology detailed below we ensure that the MSE obtained for the scaled conformer spectra $\Delta\varepsilon^{calc,sc}(\tilde{\nu})$ (see equation 5.5) remains proportional to the mean squared error if conformer spectra of original scale $\Delta\varepsilon^{calc}(\tilde{\nu})$ were used instead. The scaling methodology is applied to each compound separately.

First, we determine the mean value of $\Delta\varepsilon^{calc}(\tilde{\nu})$ over all M conformers of a

metric	description	use
S^{conf}	similarity of DFT conformer spectra for all unique conformer pairs of a single compound.	conformational sensitivity of VCD.
$\overline{S^{conf}}$	mean value of S^{conf}.	
MSE	mean squared error for the ML predicted spectra for the conformers in a set.	training and optimization of the ML model.
S^{pred}	similarity of DFT conformer spectra and corresponding ML predictions for each conformer in the test set.	describes ML model performance for new conformers.
$\overline{S^{pred}}$	mean value of S^{pred}.	
Θ	similarity of the Boltzmann weighted spectrum obtained with only DFT conformer spectra and the one obtained with DFT conformer spectra for the training set and ML predicted spectra for the validation and test set.	describes the accuracy of the Boltzmann weighted spectrum when a portion of the conformer spectra are replaced with ML predictions.

Table 5.4: Overview of similarity metrics used throughout this paper.

compound for each $\tilde{\nu}$ separately and refer to it as $\mu(\tilde{\nu})$ (equation 5.9).

$$\mu(\tilde{\nu}) = \frac{1}{M} \sum_{i}^{M} \Delta\varepsilon_i^{calc}(\tilde{\nu}) \tag{5.9}$$

Next, we define s as the standard deviation of $\Delta\varepsilon_i^{calc}(\tilde{\nu})$ over all conformers and all $\tilde{\nu}$ (equation 5.10), with ω as the mean value of $\Delta\varepsilon_i^{calc}(\tilde{\nu})$ over all conformers and all $\tilde{\nu}$ (equation 5.11).

$$s = \sqrt{\frac{1}{101 \cdot M - 1} \sum_{\tilde{\nu}=800}^{1800} \sum_{i}^{M} \left(\omega - \Delta\varepsilon_i^{calc}(\tilde{\nu})\right)^2} \tag{5.10}$$

$$\omega = \frac{1}{101 \cdot M} \sum_{\tilde{\nu}=800}^{1800} \sum_{i}^{M} \Delta\varepsilon_i^{calc}(\tilde{\nu}) \tag{5.11}$$

The scaled DFT spectrum for a conformer i, denoted as $\Delta\varepsilon_i^{calc,sc}(\tilde{\nu})$, is obtained by subtracting $\mu(\tilde{\nu})$ from $\Delta\varepsilon_i^{calc}(\tilde{\nu})$ and dividing the result by s.

$$\Delta\varepsilon_i^{calc,sc}(\tilde{\nu}) = \frac{\Delta\varepsilon_i^{calc}(\tilde{\nu}) - \mu(\tilde{\nu})}{s} \tag{5.12}$$

During training, the ML model learns to predict $\Delta\varepsilon^{calc,sc}(\tilde{\nu})$ from the conformer geometries, so the predictions made by the ML model (denoted $\Delta\varepsilon^{pred,sc}(\tilde{\nu})$) will be of similar scale as $\Delta\varepsilon^{calc,sc}(\tilde{\nu})$. These predictions are brought back to the same scale as $\Delta\varepsilon^{calc}(\tilde{\nu})$ using equation 5.13 and the resulting $\Delta\varepsilon^{pred}(\tilde{\nu})$ are the ML predicted spectra used in equation 5.6 and in Supplementary Discussion 7 to construct a Boltzmann weighted spectrum.

$$\Delta\varepsilon_i^{pred}(\tilde{\nu}) = s \cdot \Delta\varepsilon_i^{pred,sc}(\tilde{\nu}) + \mu(\tilde{\nu}) \tag{5.13}$$

3 Energy distribution of conformers

As discussed in section 5.2.1, the conformers are generated for each compound with a force field using a maximum energy window of 40 kcal mol^{-1}. This does not necessarily mean that the conformer energies actually span such a range. Additionally, the relative energies obtained for the conformers with a force field or using DFT (after geometry optimization) are likely different. As the final Boltzmann weighted spectrum discussed in later sections will be based on the DFT enthalpies, we are mainly interested in the range of DFT-based enthalpies that the conformers occupy. For each compound, the enthalpy values discussed are relative to the lowest-enthalpy conformer of said compound. The enthalpy distributions in Figure 5.28 show that nearly all conformers are found within a 10 kcal mol^{-1} window for all compounds but **1b** and the distributions are centered around 5 kcal mol^{-1}. For **1b** half of the conformers are found within a 10 kcal mol^{-1} window and most of the remaining conformers lie between between 10 and 20 kcal mol^{-1}. The broader enthalpy distribution for the conformers of **1b**, compared to **1a**, can be attributed to the steric interactions between the sidechains. The influence of the steric interactions on the performance of the ML approach is discussed in Supplementary Discussion 2.

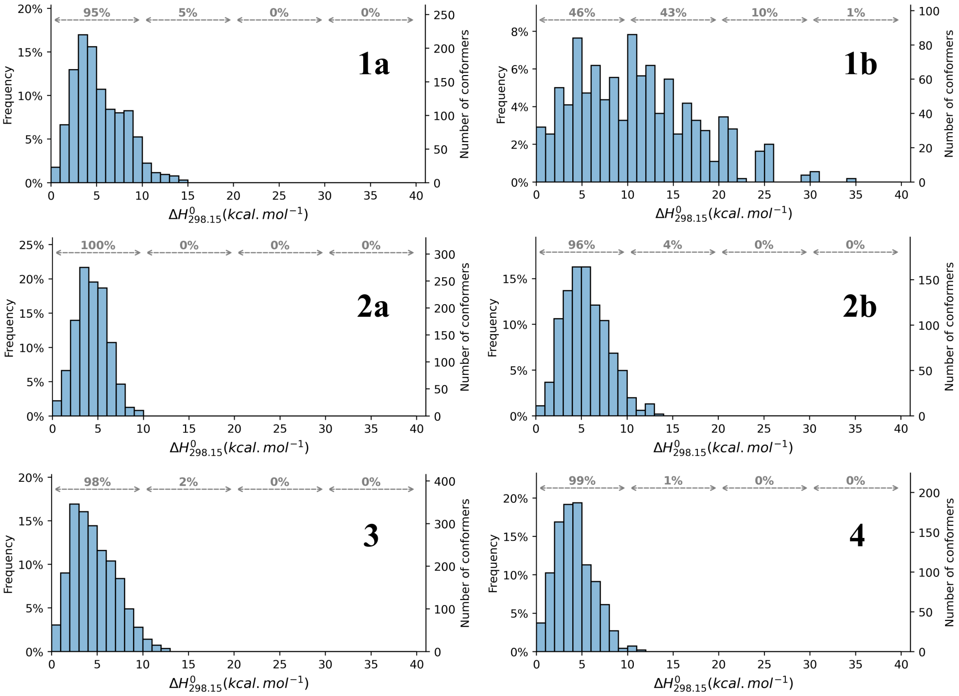


Figure 5.28: Distribution of the enthalpy, relative to the lowest conformer enthalpy for each compound, for the different conformers of the same compound. For each 10 kcal mol^{-1} bin (i.e. 0-10, 10-20, 20-30 and 30-40 kcal mol^{-1}), the fraction of conformers present within this bin is reported.

References

[1] L. A. Nafie, *Vibrational Optical Activity: Principles and Applications*, Wiley, Hoboken, NJ, 2011.

[2] N. Kobayashi and A. Muranaka, *Circular Dichroism and Magnetic Circular Dichroism Spectroscopy for Organic Chemists*, The Royal Society of Chemistry, Cambridge, United Kingdom, 2012, pp. 1–199.

[3] P. Stephens and F. Devlin, *Chirality*, 2000, **12**, 172–179.

[4] J. M. Batista Jr., E. W. Blanch and V. d. S. Bolzani, *Nat. Prod. Rep.*, 2015, **32**, 1280–1302.

[5] C. Merten, T. P. Golub and N. M. Kreienborg, *J. Org. Chem.*, 2019, **84**, 8797–8814.

[6] E. C. Sherer, C. H. Lee, J. Shpungin, J. F. Cuff, C. Da, R. Ball, R. Bach, A. Crespo, X. Gong and C. J. Welch, *J. Med. Chem.*, 2014, **57**, 477–494.

[7] J. Bogaerts, F. Desmet, R. Aerts, P. Bultinck, W. Herrebout and C. Johannessen, *Phys. Chem. Chem. Phys.*, 2020, **22**, 18014–18024.

[8] D. Rossi, R. Nasti, S. Collina, G. Mazzeo, S. Ghidinelli, G. Longhi, M. Memo and S. Abbate, *J. Pharm. Biomed.*, 2017, **144**, 41–51.

[9] Y. Zhang, M. R. Poopari, X. Cai, A. Savin, Z. Dezhahang, J. Cheramy and Y. Xu, *J. Nat. Prod.*, 2016, **79**, 1012–1023.

[10] M. Górecki, *Org. Biomol. Chem.*, 2015, **13**, 2999–3010.

[11] E. Santoro, G. Mazzeo, A. G. Petrovic, A. Cimmino, J. Koshoubu, A. Evidente, N. Berova and S. Superchi, *Phytochemistry*, 2015, **116**, 359–366.

[12] S. Qiu, E. De Gussem, K. Abbaspour Tehrani, S. Sergeyev, P. Bultinck and W. Herrebout, *J. Med. Chem.*, 2013, **56**, 8903–8914.

[13] D. E. Pivonka and S. S. Wesolowski, *Appl. Spectrosc.*, 2013, **67**, 365–370.

[14] S. S. Wesolowski and D. E. Pivonka, *Bioorganic Med. Chem. Lett.*, 2013, **23**, 4019–4025.

[15] J. Shen, J. Yang, W. Heyse, H. Schweitzer, N. Nagel, D. Andert, C. Zhu, V. Morrison, G. A. Nemeth, T.-M. Chen, Z. Zhao, T. A. Ayers and Y.-M. Choi, *J. Pharm. Anal.*, 2014, **4**, 197–204.

[16] S. Abbate, G. Longhi, F. Lebon and M. Tommasini, *Chem. Phys.*, 2012, **405**, 197–205.

[17] N. Vanthuyne, C. Roussel, J.-V. Naubron, N. Jagerovic, P. M. Lázaro, I. Alkorta and J. Elguero, *Tetrahedron Asymmetry*, 2011, **22**, 1120–1124.

[18] P. J. Stephens, J. J. Pan, F. J. Devlin, K. Krohn and T. Kurtán, *J. Org. Chem.*, 2007, **72**, 3521–3536.

[19] L. A. Caldas, M. T. Rodrigues, A. N. L. Batista, J. M. Batista, J. H. G. Lago, M. J. P. Ferreira, I. G. S. Rubio and P. Sartorelli, *Molecules*, 2020, **25**, 3005.

[20] K. Knippen, B. Bredenkötter, L. Kanschat, M. Kraft, T. Vermeyen, W. Herrebout, K. Sugimoto, P. Bultinck and D. Volkmer, *Dalton Trans.*, 2020, **49**, 15758–15768.

[21] Z.-Q. Wang, C.-J. Wu, Z.-H. Wang, C. Huang, J. Huang, J.-H. Wang and T.-M. Sun, *J. Mol. Struct.*, 2017, **1146**, 484–489.

[22] J. Kong, L. A. Joyce, J. Liu, T. M. Jarrell, J. C. Culberson and E. C. Sherer, *Chirality*, 2017, **29**, 854–864.

[23] M. A. Aparicio-Cuevas, I. Rivero-Cruz, M. Sánchez-Castellanos, D. Menéndez, H. A. Raja, P. Joseph-Nathan, M. d. C. González and M. Figueroa, *J. Nat. Prod.*, 2017, **80**, 2311–2318.

[24] G. Mazzeo, E. Santoro, A. Andolfi, A. Cimmino, P. Troselj, A. G. Petrovic, S. Superchi, A. Evidente and N. Berova, *J. Nat. Prod.*, 2013, **76**, 588–599.

[25] J. C. Pardo-Novoa, H. M. Arreaga-González, M. A. Gómez-Hurtado, G. Rodríguez-García, C. M. Cerda-García-Rojas, P. Joseph-Nathan and R. E. del Río, *J. Nat. Prod.*, 2016, **79**, 2570–2579.

[26] D. P. Demarque and C. Merten, *Chem. Eur. J.*, 2017, **23**, 17915–17922.

[27] D. P. Demarque, S. Heinrich, F. Schulz and C. Merten, *Chem. Commun.*, 2020, **56**, 10926–10929.

[28] D. P. Demarque, M. Kemper and C. Merten, *Chem. Commun.*, 2021, **57**, 4031–4034.

[29] P. Fagan, L. Kocourková, M. Tatarkovič, F. Králík, M. Kuchař, V. Setnička and P. Bouř, *ChemPhysChem*, 2017, **18**, 2258–2265.

[30] F. Králík, P. Fagan, M. Kuchař and V. Setnička, *Chirality*, 2020, **32**, 854–865.

[31] T. Vermeyen and C. Merten, *Phys. Chem. Chem. Phys.*, 2020, **22**, 15640–15648.

[32] M. R. Poopari, Z. Dezhahang and Y. Xu, *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, 2015, **136**, 131–140.

[33] K. D. R. Eikås, M. T. P. Beerepoot and K. Ruud, *J. Phys. Chem. A*, 2022, **126**, 5458–5471.

[34] B. Légrády, E. Vass and G. Tarczay, *J. Mol. Spectrosc.*, 2018, **351**, 29–38.

[35] S. Ma, X. Cao, M. Mak, A. Sadik, C. Walkner, T. B. Freedman, I. K. Lednev, R. K. Dukor

and L. A. Nafie, *J. Am. Chem. Soc.*, 2007, **129**, 12364–12365.

[36] T. A. Keiderling, *Chem. Rev.*, 2020, **120**, 3381–3419.

[37] T. Hongen, T. Taniguchi, S. Nomura, J.-I. Kadokawa and K. Monde, *Macromolecules*, 2014, **47**, 5313–5319.

[38] R.-M. Ho, M.-C. Li, S.-C. Lin, H.-F. Wang, Y.-D. Lee, H. Hasegawa and E. L. Thomas, *J. Am. Chem. Soc.*, 2012, **134**, 10974–10986.

[39] J. Kessler, V. Andrushchenko, J. Kapitán and P. Bouř, *Phys. Chem. Chem. Phys.*, 2018, **20**, 4926–4935.

[40] L. Zhao, J. Zhang, Y. Zhang, S. Ye, G. Zhang, X. Chen, B. Jiang and J. Jiang, *JACS Au*, 2021, **1**, 2377–2384.

[41] C. Sun, Y. Tian, L. Gao, Y. Niu, T. Zhang, H. Li, Y. Zhang, Z. Yue, N. Delepine-Gilon and J. Yu, *Sci. Rep.*, 2019, **9**, 11363.

[42] J. Meiler, R. Meusinger and M. Will, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1169–1176.

[43] S. Ye, K. Zhong, J. Zhang, W. Hu, J. D. Hirst, G. Zhang, S. Mukamel and J. Jiang, *J. Am. Chem. Soc.*, 2020, **142**, 19071–19077.

[44] R. Mamede, F. Pereira and J. A. de Sousa, *Sci. Rep.*, 2021, **11**, 23720.

[45] E. Jonas and S. Kuhn, *J. Cheminformatics*, 2019, **11**,.

[46] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari and P. Rinke, *Adv. Sci.*, 2019, **6**, 1801367.

[47] J. A. Fine, A. A. Rajasekar, K. P. Jethava and G. Chopra, *Chem. Sci.*, 2020, **11**, 4618–4630.

[48] P. Kovács, X. Zhu, J. Carrete, G. Madsen and Z. Wang, *Astrophys. J.*, 2020, **902**, 100.

[49] M. McCann, M. Defernez, B. Urbanowicz, J. Tewari, T. Langewisch, A. Olek, B. Wells, R. Wilson and N. Carpita, *Plant Physiol.*, 2007, **143**, 1314–26.

[50] V. H. da Silva, F. Murphy, J. M. Amigo, C. Stedmon and J. Strand, *Anal. Chem.*, 2020, **92**, 13724–13733.

[51] K. Tanabe, T. Matsumoto, T. Tamura, J. Hiraishi, S. Saeki, M. Arima, C. Ono, S. Itoh, H. Uesaka, Y. Tatsugi, K. Yatsunami, T. Inaba, M. Mitsuhashi, S. Kohara, H. Masago, F. Kaneuchi, C. Jin and S. Ono, *Appl. Spectrosc.*, 2001, **55**, 1394–1403.

[52] T. Vermeyen, J. Brence, R. Van Echelpoel, R. Aerts, G. Acke, P. Bultinck and W. Herrebout, *Phys. Chem. Chem. Phys.*, 2021, **23**, 19781–19789.

[53] R. Mamede, B. S. o. de Almeida, M. Chen, Q. Zhang and J. Aires-de Sousa, *J. Chem. Inf. Model.*, 2021, **61**, 67–75.

[54] K. Adams, L. Pattanaik and C. W. Coley, *Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations*, 2021, arXiv: 2110.04383.

[55] O.-E. Ganea, L. Pattanaik, C. W. Coley, R. Barzilay, K. F. Jensen, W. H. Green and T. S. Jaakkola, *GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles*, 2021, arXiv: 2106.07802.

[56] Kevin E. Gilbert, *GMMX (version 1.5)*, 2011, Serena Software Bloomington IN.

[57] T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

[58] Kevin E. Gilbert, *Pcmodel (version 10.0)*, 2013, Serena Software Bloomington IN.

[59] A. D. Becke, *J. Chem. Phys*, 1993, **98**, 5648–5652.

[60] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Revision C.01*, 2016, Gaussian Inc. Wallingford CT.

[61] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong and X. Shen, *IEEE Trans. Neural Netw. Learn. Syst.*, 2020, **31**, 2409–2429.

[62] R. Caruana, *Mach. Learn.*, 1997, **28**, 41–75.

[63] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, Proceedings of the 24th International Conference on Neural Information Processing Systems, Red Hook, NY, 2011, pp. 2546–2554.

[64] J. Bergstra, D. Yamins and D. Cox, Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013, pp. 115–123.

[65] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *J. Mach. Learn. Res.*, 2014, **15**, 1929–1958.

[66] S. Ioffe and C. Szegedy, Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015, pp. 448–456.

[67] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015, `https://www.tensorflow.org/`.

[68] E. Debie, E. De Gussem, R. K. Dukor, W. Herrebout, L. A. Nafie and P. Bultinck, *ChemPhysChem*, 2011, **12**, 1542–1549.

[69] A. Röbel, Dynamic pattern selection for faster learning and controlled generalization of neural networks, 1994.

[70] A. Engelbrecht, *Computational Intelligence*, John Wiley & Sons Ltd, Chichester, United Kingdom, 2nd edn, 2007, pp. 95–97.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

[72] D. Singh and B. Singh, *Appl. Soft Comput.*, 2020, **97**, 105524.

Chapter 6

Pushing the boundaries of VCD spectroscopy in natural product chemistry

The results of the paper ‘Pushing the boundaries of VCD spectroscopy in natural product chemistry’ have been recreated in this chapter. This project is the result of a collaboration with the Federal University of São Paulo and Fluminense Federal University. Visual inspection of the mixture spectra was performed by Andrea N. L. and Batista Junior, João M., while the ML-based analysis of the mixture spectra was performed by the author of this thesis.

6.1 Introduction

Natural product molecules from land, marine and/or microbial sources continue to play a crucial role in drug discovery and development.[1] The biological potential of natural small molecules, known as secondary (or special) metabolites, stems from the fact that they are designed to interact with biological chiral targets, such as proteins, either inside or outside of the producing organisms. These compounds are commonly involved in chemically mediated defence, growth in competitive environments, signalling, and reproduction. These functions are closely correlated to their structural and stereochemical diversity, which are made possible by intricate biosynthetic machinery.[2] Natural products are produced from a variety of building blocks and are subjected to several post-biosynthetic modifications. These molecules commonly incorporate distinct chiral elements (point and axial chirality) within a single chemical structure and are found in complex mixtures. The combination of the structural and stereochemical features of natural compounds provides the physicochemical and topological requirements for proper membrane permeation and selective receptor interactions.[3] Despite the potential biological applications of natural products, their efficient incorporation into the drug discovery pipeline has a high price tag. Current regulatory affairs require full pharmacological and toxicological characterisation of each enantiomer for approval of chiral drugs,[4] which makes the determination of the exact three-dimensional arrangement of the atoms in isolated compounds an important bottleneck. Additionally, the enantiomeric purity of secondary metabolites adds another layer of complexity to natural product chemistry. Although natural products are commonly believed to be enantiomerically pure or enriched, a great number of enantiomeric mixtures or even racemates have been described for secondary metabolites.[5–8] Based on the challenges described above, it is not uncommon to find in the literature incorrect assignments of both structure and stereochemistry of natural compounds. This is particularly worrisome since the use of empirical correlations of spectral data for structurally related compounds is a common practice in natural product chemistry, which increases the risks of error amplifications. A recent survey has demonstrated an increase in the number of stereochemical reassignments of natural products over the last decade.[9] The most used methods to reassign absolute configuration were organic synthesis, followed by chiroptical methods, mainly associated with DFT calculations, and NMR. Chiroptical methods, especially optical rotation (OR) and electronic circular dichroism (ECD), have a longstanding history of successful applications to secondary metabolites.[10] Vibrational methods, such as vibrational circular dichroism (VCD) and Raman optical activity (ROA), on the other hand, underwent a growth in their use by natural product chemists only over the last two decades.[11,12] Historically, the application of the classic chiroptical spectroscopic methods OR and ECD has been based on empirical correlations of structurally-related molecules for which the absolute configuration was known. Unfortunately, empirical rules commonly present exceptions leading to frequent misassignments.

Current best practice guidelines recommend the comparison of observed ECD spectra with quantum chemically simulated data.[13] In the case of VCD for small molecule stereochemical investigations, widespread empirical correlations were not observed, and the technique came of age after the development of the magnetic field perturbation method by Stephens *et al.* that allowed the calculation of VCD intensities at DFT level to be incorporated into commercial software.[14] Due to the more complex spectral patterns in the IR fingerprint region and higher sensitivity to structural features, finding VCD spectral markers for similar structures was more challenging than for ECD and a greater dependence on DFT calculations soon followed. Although the development of accurate quantum chemical calculations has led to the renaissance[15] of chiroptical spectroscopy with a great increase in the number of natural product molecules being investigated, unfortunately, it has not been translated into a similar expansion on the number of research groups using the techniques. Most of the VCD assignments of absolute configuration of natural products published in the literature come from just a handful of research groups, which are commonly specialised in chiroptical spectroscopy but not necessarily in natural product chemistry. This situation indicates that VCD has not yet been included in the natural product chemist toolbox. We believe that one of the main difficulties in attracting more natural product chemists to use chiroptical spectroscopy for stereochemical elucidation is the aforementioned need for DFT calculations to interpret experimental spectra.[6,8] Therefore, herein, we propose the search and validation of IR and VCD spectral markers to circumvent the requirement of DFT calculations allowing for absolute configuration assignments even in complex mixtures. To that end, a combination of visual inspection and machine-learning based methods will be used. Monoterpenes, either isolated or in mixtures, are selected as target molecules for this proof-of-concept study.

6.1.1 Vibrational circular dichroism

VCD arises from the differential absorption for left- and right-circularly polarised infrared (IR) radiation by a chiral (non-racemic) molecule during a vibrational transition. It is the expansion of the electronic CD phenomenon into the IR spectral region where vibrational transitions occur. One of the main advantages of VCD over other techniques is the possibility of analysis directly in the solution-state, without requiring either single-crystals or suitable UV-vis chromophores. Since it is based on IR spectroscopy, a large number of transitions is commonly available that are sensitive to both structure (functional groups/connectivity) and stereochemistry. Additionally, like for other chiroptical methods, the final VCD spectrum reflects quantitively the conformational population of the target chiral molecule in a given solvent. Therefore, IR/VCD represents an ideal tool to simultaneously study composition and stereochemistry of chiral molecules in complex mixtures. Deep discussions on VCD history, theory, instrumentation, and applications are beyond the scope of this manuscript. Further information can be found elsewhere.[15–20]

6.1.2 Monoterpenes

Monoterpenes (C_{10}) are members of the large and structurally diverse natural product family of terpenoids. Monoterpenes derive from the condensation of two C_5 isoprene units, joined in a head-to-tail fashion.[21] Based on the dominance of carbocation chemistry for the formation of terpenoids in general, which commonly involves rearrangements, monoterpenes are found in nature in a huge variety of structures (strained/unstrained cyclic, bicyclic, and linear forms) and stereochemical outcomes. Most monoterpenes are optically active, with enantiomers of a given compound being produced either by the same or different organisms. These compounds are also commonly found in complex mixtures i.e., essential oils. Due to the chiral nature, availability in suitable enantiomeric purity, and conformational rigidity of some bicyclic monoterpenes, which result in high-quality vibrational spectra in the mid-IR region, compounds such as α-pinene and camphor have been used as standards for VCD intensity calibration.[16] Historically, monoterpenes have also been used in important VCD technological advancements, both in theory[22–25] and instrumentation.[26–31] Regarding applications, VCD has been used to assign the absolute configuration of a series of isolated monoterpenes,[32–36] with a single study attempting to establish VCD chiral signatures of essential oils.[37] A compilation of IR/VCD spectral standards for terpenes was published in 2006.[38]

6.1.3 Spectral markers

In order to facilitate the application of VCD for stereochemical assignments of complex chiral molecules, some efforts have been made to reduce the dependency on DFT calculations. One of the most used approaches involve molecule rigidification and/or the search for spectral markers. Some examples of rigidification include the derivatisation of *endo*-borneol,[39] the acetonisation of 1,3-diols,[40] the derivatisation of sphingosine with glutaraldehyde,[41] and the preparation of conformationally restrained cyclic carbodiimides.[42] Non-covalent derivatisation methods to simplify calculations of carboxylic acids have been recently devised,[43] along with the covalent introduction of a suitable deuterated VCD chromophore with absorption removed from the IR fingerprint region for the C-1 configuration of sugar molecules.[44] Our group has been particularly interested in finding IR/VCD spectral signatures for conformation and configuration of chiral natural products. Examples include VCD markers for the configuration of esterified chromane and monoterpene moieties,[45] for the configuration of the hexahydroxydiphenoyl group in ellagitannins,[46] for the configuration of the 2(5H)-furanone moiety in acetogenins,[47] for the configuration at C-9 of both strepchazolin A and B,[48] as well as the IR marker for the E/Z double bond configuration of spongosoritins[49] and the VCD marker for the stacking of the pyrrolidine ring of proline and the aromatic ring of tyrosine in pohlianin A.[50] These searches of spectral markers are related to the concept of inherently dissymmetric VCD chromophores.[51] Finally, following important historical developments,[52–54] a non-empirical VCD method that does not require DFT calculations was proposed in 2012 for absolute configuration assignments.[55] The VCD exciton chirality method, however, requires the presence of two infrared chromophores (e.g. carbonyl groups) close in space, to allow for their coupling, and chirally disposed. The existence of further carbonyl groups on the other hand, complicates the exciton coupling analysis, hampering its application without the aid of DFT calculations.[56]

6.1.4 Proposed approach

As discussed above, one of the main reasons why few natural product chemists use VCD as a standard method to assign the absolute configurations of chiral secondary metabolites is the requirement of quantum chemical calculations to interpret experimental data. Since the search and validation of IR/VCD spectral markers have proven to be a viable approach for a series of structurally diverse molecules, herein, we decided to investigate monoterpene molecules (37 + 2 sesquiterpenes) both isolated and in mixtures in a search for spectral signatures that can be used to both identify and assign their stereochemistry directly in mixtures and without requiring further DFT calculations. Visual comparison will be explored in a search of either similar or discriminative vibrational bands for individual molecules. Then, inspired by a recent proof-of-concept study using machine learning (ML) to extract absolute configurations from VCD spectra of decorated α-pinene derivatives,[57] we will extend the application of the ML methodology to identify monoterpenes in complex mixtures, such as essential oils which, to the best of our best knowledge, has not been tested for VCD. In this way, we will assess the feasibility of such an approach and identify possible pitfalls for its future development. This concept, if successful, will allow the determination of composition, stereochemistry, and enantiomeric excesses of essential oil components from IR/VCD spectra not only without requiring DFT calculations, but also bypassing the need for chiral GC analysis. The main methodology to study terpene mixtures has been chiral GC, however, it commonly requires the availability of both enantiomers of a given target for identification purposes.

6.2 Results and discussion

IR and VCD spectra of commercially available individual monoterpenes were recorded in $CDCl_3$ solution in the region of 950-1800 cm^{-1} and compared visually. They were grouped first based on their cyclic skeleton types,[21] namely, menthane, pinane, bornane and fenchane types. The isocamphane type had no representative, while carene and thujane types had a single representative each. The linear compounds were grouped as geraniol derivatives. Achiral compounds, such as cineole, as well as some racemic monoterpenes (isoborneol and isobornyl acetate) were also included for the IR analysis. After the spectra of individual molecules were obtained (Figs. 6.4-6.10†), artificial mixtures of monoterpenes of each type were prepared and subjected to IR/VCD analysis (Figs. 6.11-6.16†). These mixtures were used to investigate possible band overlaps and cancelations from similar structures thus aiding the spectral marker validation procedure. Other mixtures with increasing complexity were then prepared and subjected to the same type of analysis (A-J, Table 6.1†). These procedures allowed us to identify the most discriminative spectral regions for each molecule type. Once the visual inspection on mixtures of know composition was finished, the accuracy of the spectral markers identified was tested on natural mixtures of unknown composition. To that end, tea tree, rosemary, lavender, and ylang-ylang essential oils were employed. The compounds identified in the essential oils by the IR/VCD analysis were then confronted with GC-MS results on the same samples. Following the visual inspection approach, ML methods were applied. The following sections will present the specific results of both approaches with their potential and limitations.

6.2.1 Visual inspection

The monoterpenes investigated at this stage included the pinane type (1*R*)-(−)-myrtenol, (1*R*)-(−)-myrtenal, (1*R*)-(−)-myrtenyl acetate, (*S*)-(−)-β-pinene, (*R*)-(+)-α-pinene, (1*R*,2*R*,3*S*,5*R*)-(−)-pinanediol, (1*S*)-(−)-verbenone, (1*S*,2*S*,5*S*)-(−)-2-hydroxy-3-pinanone, and (1*R*,2*R*,3*R*,5*S*)-(−)-isopinocampheol; the menthane type 1 (*R*)-(−)-terpinen-4-ol, (*S*)-(−)-perillaldehyde, (*S*)-(−)-α-terpineol, (*S*)-(−)-perillyl alcohol, (*R*)-(−)-carvone, and (*R*)-(+)-limonene; the menthane type 2 (1*S*,2*S*,5*R*)-(+)-neomenthol, (1*R*,2*S*,5*R*)-(−)-isopulegol, (1*R*,2*S*,5*R*)-(−)-menthol, (1*S*,2*R*,5*R*)-(+)-isomenthol, and (*R*)-(+)-pulegone; the bornane type (1*R*)-(+)-camphor, (1*S*)-(−)-camphor, (*S*)-(−)-*endo*-borneol, (*S*)-(−)-*endo*-bornyl acetate, (±)-isobornyl acetate, (±)-isoborneol, the fenchane type (*S*)-(+)-fenchone, and (1*R*)-(+)-*endo*-fenchyl alcohol; the geraniol type (*S*)-(−)-β-citronellol, (*R*)-(−)-linalool, (*R*)-(−)-linalyl acetate, (*R*)-(−)-linalool, (*S*)-(+)-β-citronellene, and (*S*)-(−)-citronellal. Cineole, (1*S*)-(+)-3-carene, and (1*S*,4*R*)-(−)-α-thujone were also included in more complex mixtures. Inspections were first carried out on IR spectra in a search for either similar or discriminatory bands. Both frequency

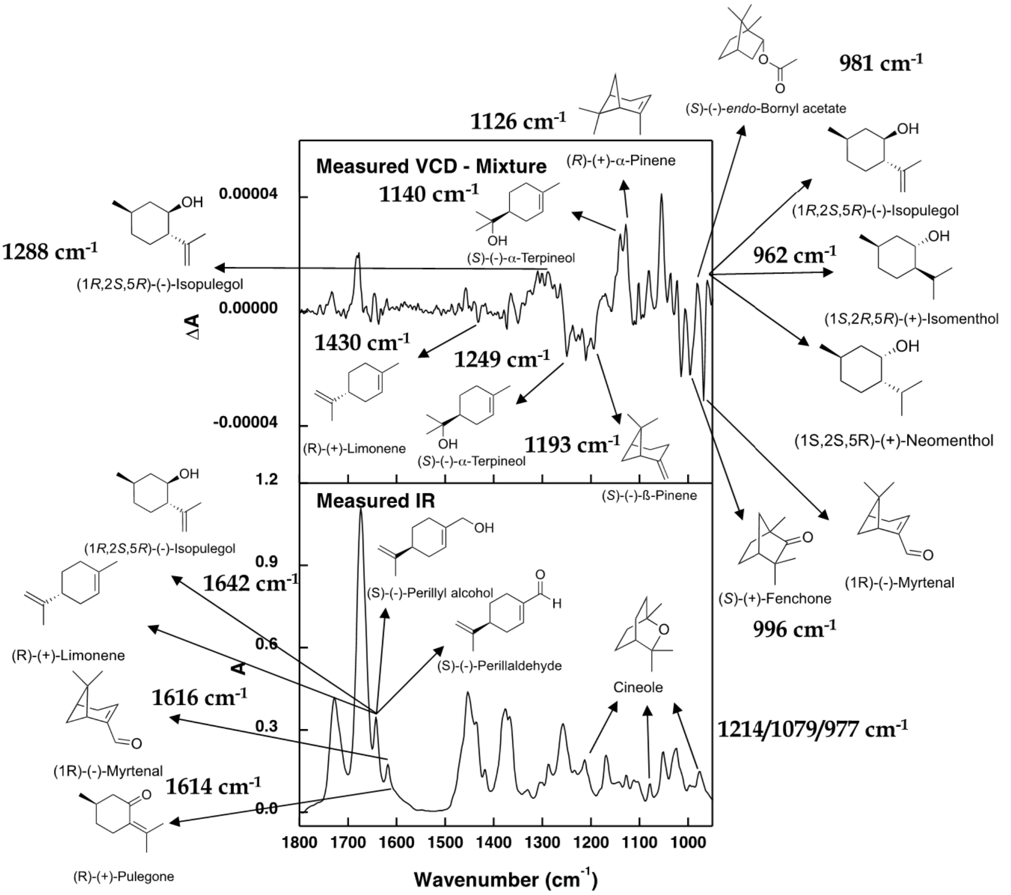


Figure 6.1: Monoterpene identified from an artificial mixture (J) of known composition by means of visual IR/VCD spectral markers. See ESI† for detailed analysis of spectral markers and their vibrational origin

shifts and relative intensities were used to cluster different monoterpenes. Then, VCD spectra were analysed which, due to their bisignated nature, provide better resolution and discriminatory power. On the other hand, having bisignated bands may lead to attenuation or even cancellation of oppositely signed bands of particular monoterpenes when present in mixtures. Detailed analyses of individual terpene types are provided in the ESI. Once the markers for each class of monoterpenes were identified visually for individual compounds, their utility was tested in complex mixtures. Analyses of mixtures of compounds belonging to the same molecule type are presented in the ESI (Figs. 6.11-6.16†). This approach allowed us to verify possible intermolecular interactions, spectral correlations and VCD band cancellations. Then, the visual IR and VCD spectral markers were tested on an artificial mixture (mixture J) containing molecules of different types, which included (1*R*)-(–)-myrtenal, (*S*)-(–)-β-pinene, (*R*)-(+)-α-pinene, (*S*)-(–)-perillaldehyde, (*S*)-(–)-α-terpineol, (*S*)-(–)-perillyl alcohol, (*R*)-(–)-carvone, (*R*)-(+)-limonene, (1*S*,2*S*,5*R*)-(+)-neomenthol, (1*R*,2*S*,5*R*)-(–)-isopulegol, (1*S*,2*R*,5*R*) -(+)-isomenthol, (*R*)-(+)-pulegone, (*S*)-(+)-fenchone, (*S*)-(–)-*endo*-borneol, (*S*)-(–)-*endo*-bornyl acetate, and cineole. These results are presented in Figure 6.1. As can be seen in Figure 6.1, even in such a complex mixture, a combination of IR and VCD visual spectral markers was able to tell apart most of the compounds. Please refer to ESI† for specific vibrational frequencies as well as molecular origin of the selected bands.

Following the analysis of the artificial complex mixture of known composition, natural mixtures (essential oils) were analysed. Figure 6.2 presents the IR and VCD spectra of tea tree, rosemary, lavender, and ylang-ylang essential oils from which the main components were identified by means of the spectral markers described above. The presence of the monoterpenes in question was confirmed by GC-MS analysis (Figs. 6.17-6.20†). It is important to emphasise that not only were monoterpene identities secured but also their absolute configuration, simultaneously. Regarding tea tree oil, the IR band at 1066 cm^{-1} and the corresponding positive VCD bands indicated the presence of (*S*)-(+)-terpinen-4-ol, which was confirmed by GC-MS with abundance of 57.88 (area%). The broad positive VCD band at around 1250 cm^{-1} confirmed the presence of the menthane type skeleton. As for rosemary oil, the IR band 1639 cm^{-1} indicated the presence of β-pinene, while those at 1214, 1079 and 977 cm^{-1} were markers for the presence of the achiral monoterpene cineole. Additionally, the IR band at 1415 cm^{-1} indicated the presence of camphor. Regarding VCD, the (+)-1469 and (–)-1195 cm^{-1} bands led to the identification of (*S*)-(–)-β-pinene, while the positive bands at 1450/1126 cm^{-1} indicated the presence of (*R*)-(+)-α-pinene. The positive VCD band at 1166 cm^{-1} showed the occurrence of (1*R*)-(+)-camphor. The GC-MS analysis (see ESI†) confirmed the presence of β-pinene (5.21 area%), α-pinene (7.28 area%), camphor (7.18 area%), and cineole (70,9 area%). It is noteworthy that in the case of rigid bicyclic monoterpenes with large VCD intensities, the present approach is capable of detecting them and assigning their absolute configurations when present in abundances as low as 5%. Analysis of the IR spectrum of lavender oil showed bands at 1640 and 1412 cm^{-1}, which indicated

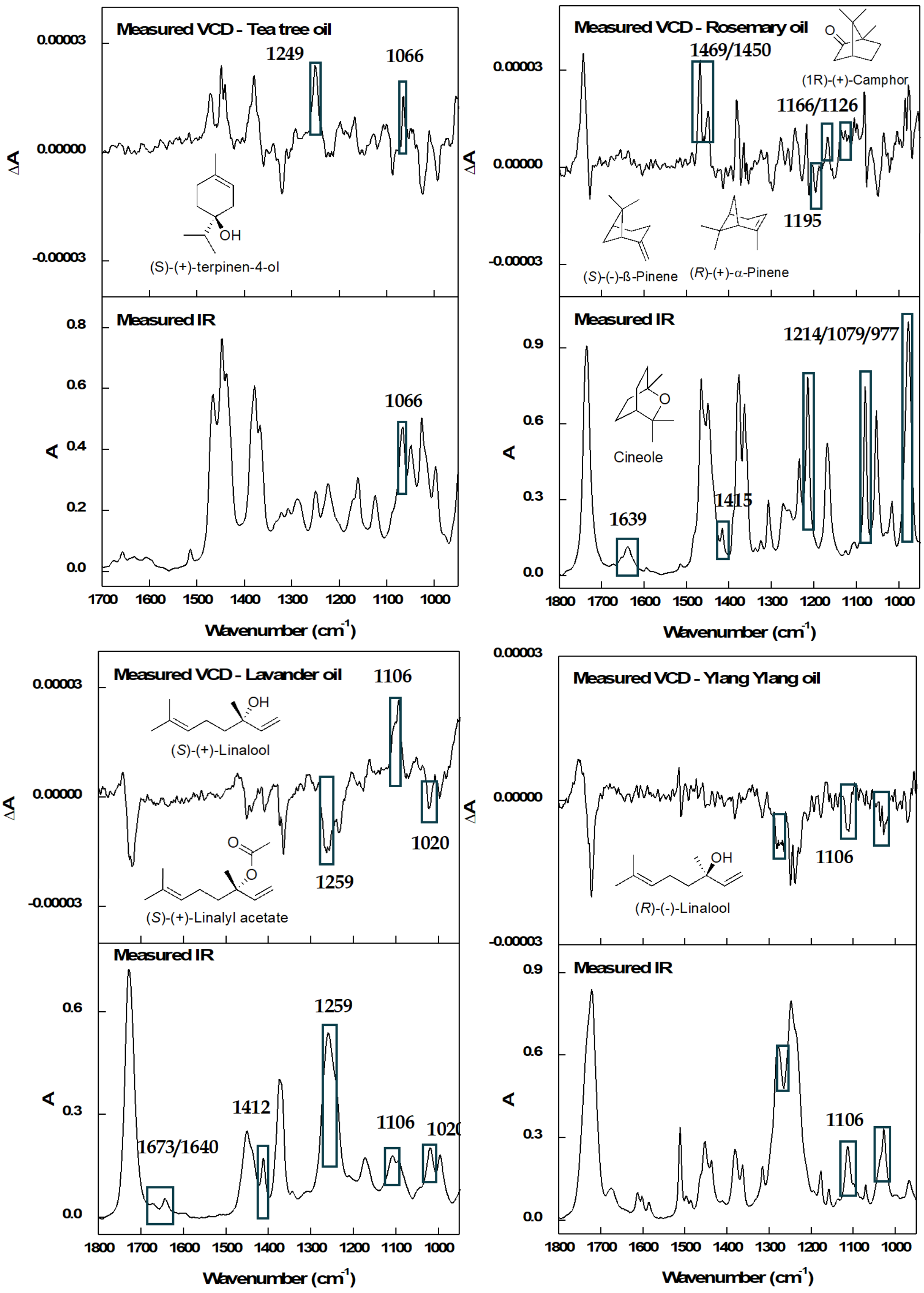


Figure 6.2: Monoterpene identified from natural mixtures of unknown composition (essential oils) by means of visual IR/VCD spectral markers. See text for discussion of individual bands. Identities of monoterpenes confirmed by GC-MS analysis.

the presence of compounds with terminal double bonds that, combined with the band at 1672, led to the identification of acyclic monoterpenes. The presence of the band at 1106 cm^{-1} confirmed the presence of linalool, while the bands 1720, 1259 and 1020 cm^{-1} confirmed the presence of linalyl acetate. VCD investigation indicated their assignment as (S)-(+)-linalool (positive band at 1106 cm^{-1}) and (S)-(+)-linalyl acetate (negative bands at 1259 and 1020 cm^{-1}). GC-MS spectra confirmed these monoterpenes as the most abundant in the essential oil: 44.6 area% for linalool and 42.66 area% for linalyl acetate (see ESI†). Finally, for ylang-ylang oil, the same IR/VCD bands described for lavender oil were identified, with the main difference being the sign of the 1106 cm^{-1} VCD band, which indicated the presence of (R)-(–)-linalool. GC-MS analysis, on the other hand, confirmed the linalool (19.48 area%), but did not confirm linalyl acetate.

Despite successful, the use of visually identified spectral markers requires painstaking analysis which may be subjected to user bias. Additionally, many IR/VCD bands remained unassigned. In order to circumvent such drawbacks and expedite analysis, an ML protocol was idealised, developed and tested as described in the following section.

6.2.2 Machine learning

As mentioned in the previous section, the use of visually identified spectral markers is laborious. Additionally, marker bands in VCD can be attenuated or even cancelled in a mixture due to opposite intensities arising from other components. An ML model can leverage the intensities in other spectral regions to detect components even if their marker bands are cancelled. Therefore, we were interested in testing whether an ML model could identify the monoterpenes present in different mixtures. If successful, one would no longer need to manually identify spectral markers and the accuracy of the detection would be improved. In the absence of a large monoterpene and mixture spectral dataset, the ML model was trained on a set of *in silico* mixtures (noisy linear combinations of monoterpenes), yielding an IR- and a VCD-based model. The VCD-based model generates the monoterpene composition as output using the VCD spectrum of the mixture as input, whereas the IR-based model predicts the composition with the IR spectrum as input. A detailed description of the ML model and the training procedure is presented in the ESI†.

A set of six artificial mixtures containing each up to 8 monoterpenes of different types was prepared (Table 6.1†, mixtures A-F) to evaluate and finetune the monoterpene detection. The current dataset covers representative compounds for most of the common monoterpene types. An essential oil, on the other hand, likely contains one or more compounds that are still absent from the present dataset. We mimic such a situation by excluding myrtenyl acetate from the *in silico* training mixtures, while actually including it in the artificial mixture A. By doing so, we test the stability of the model in the presence of a 'new' component. The predicted relative concentrations obtained for mixtures A-F are shown in Figure 6.3. As the decision boundary still needed to be fine-tuned, we were mainly interested in whether the largest predicted concentrations were obtained for mixtures containing each said monoterpene. A detailed analysis of the predictions and the patterns leveraged by the models is provided in the ESI†.

The VCD based model successfully extracted the presence of 26 out of the 30 chiral monoterpenes present throughout mixtures A-F. The VCD model also demonstrated chiral sensitivity: while (1*R*)-(+)-camphor in mixture C was not detected, a strong negative (1*R*)-(+)-camphor concentration was obtained for mixture A that contains (1*S*)-(–)-camphor. The IR based model properly classified 29 of the 31 monoterpenes present in mixtures A-F. The patterns learned from the *in silico* mixtures (Figs. 6.24-6.25†) clearly performed well on these mixtures. As the presence of myrtenyl acetate in mixture A did not hamper the accuracy, the patterns showed robustness to small external influences. These patterns also translated well to other mixtures of similar complexity. When the models were applied to artificial mixtures of monoterpenes of a single type (Figs. 6.11-6.16†), a similar number of monoterpenes were correctly classified by the models (Figs. 6.26-6.29†). For each of these mixtures, the VCD model correctly classified on average 25 chiral monoterpenes and the IR model did so for 29 monoterpenes. Thus, even if a mixture contained structurally similar compounds, its composition can still be extracted. The ML methodology provides a viable new approach for determining the composition of monoterpene mixtures.

Next, we tested the model on the artificial mixtures containing a larger number of monoterpenes (mixtures H-J, Table 6.1† and Figs. 6.30-6.31†). When the models were applied to mixture J, the VCD model correctly identified the pres-

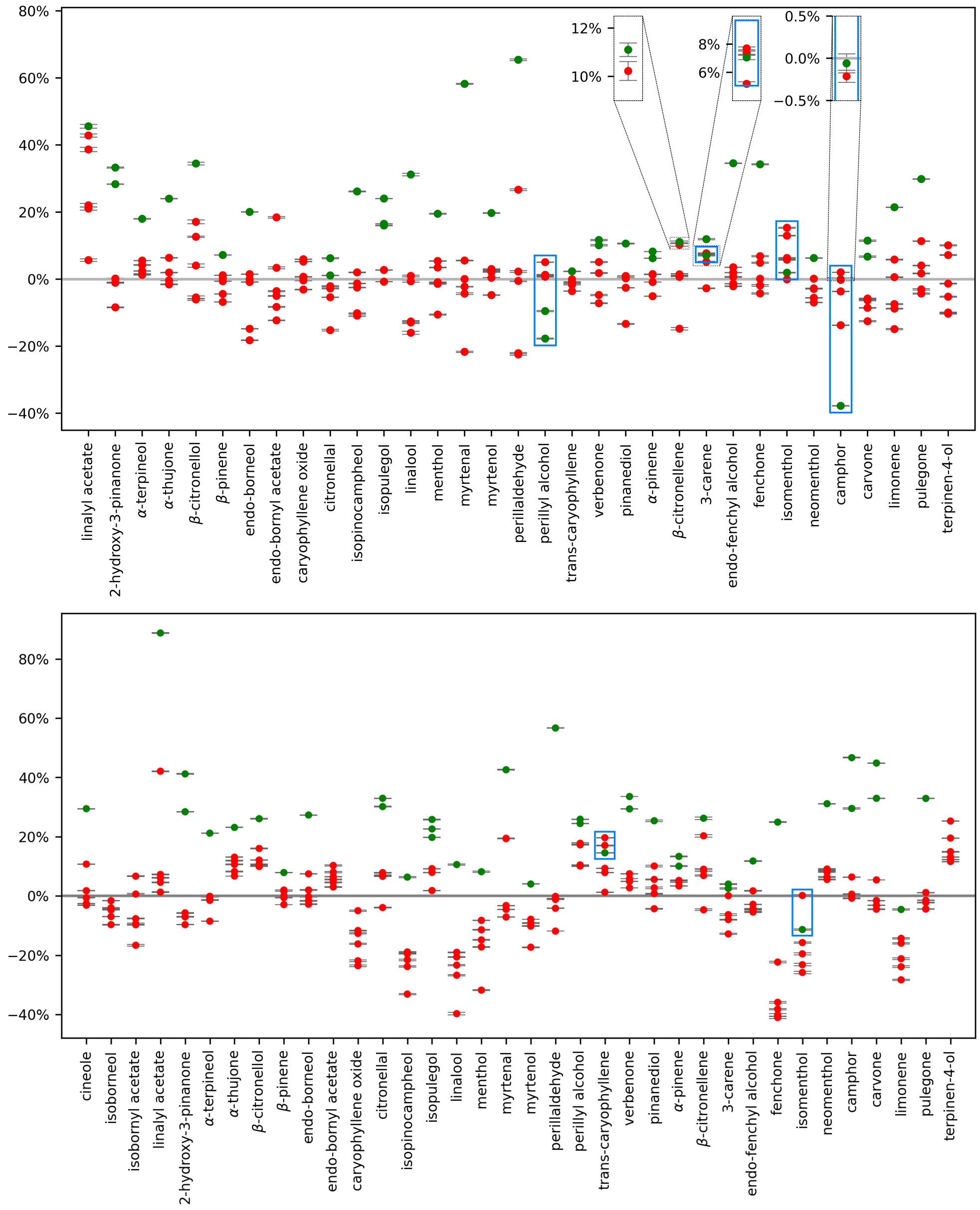


Figure 6.3: Predicted concentrations (in %) relative to the original concentrations of individual monoterpenes for mixtures A-F by the VCD based (top) and IR based model (bottom). The predicted concentration for each monoterpene is shown for each of the six mixtures and is colored according to whether the monoterpene is present (green) or absent (red) for a given mixture. The predicted concentrations are highlighted for a monoterpene when no correct decision boundary can be drawn for this monoterpene (for a correct decision boundary, all mixtures that contain said monoterpene need to lie above it and all mixtures that do not contain said monoterpene below it). The error margin (bars) is the standard deviation upon the predicted value during cross-validation (see ESI† for more details). Some regions are zoomed in for clarity.

ence or absence of 24 chiral monoterpenes and the IR model correctly classified 28 monoterpenes (true positives and true negatives; see ESI† for methodology details). Compared to the visual inspection (Fig. 6.1), the ML model enabled to extract more information from the marker and non-marker bands in the spectrum. As a result, a larger number of the monoterpenes present in the mixture was detectable. The VCD model correctly classified 22 chiral monoterpenes for mixture H and 20 for mixture I. With the IR model, 22 monoterpenes from mixture H and 23 monoterpenes from mixture I were correctly identified as either present or absent. With the lower individual contributions of each single terpene in more complex mixtures, extracting their composition was more challenging. Nonetheless, the models could still perform well depending on the exact mixture composition, as demonstrated for mixture J.

Subsequently, the models were asked to predict the terpenes present in the 4 essential oils and the results are reported in Tables 6.2-6.3†. The content of the essential oils was unknown prior to these predictions, removing any potential user bias. Lavender oil is largely made up of linalool and linalyl acetate which were both detected by the IR model, whereas the VCD model mainly detected (*R*)-(−)-linalool. In ylang-ylang oil both models confirmed the presence of (*R*)-(−)-linalool. The major component of rosemary oil, cineole, was clearly detected by the IR model. The presence of (*R*)-(+)-α-pinene and (*S*)-(−)-β-pinene was additionally detected by both models. For the final extract, tea tree oil, the IR model correctly detected terpinen-4-ol and the tiny fraction of limonene; neither of which was detected by the VCD model. Even so, the IR model succeeded in correctly detecting these terpenes. It is important to note that for each of these oils a non-negligible number of false positives (terpenes absent from the oil which are detected by the model) was obtained. When only a small number of components in the oil is included in the dataset, the mixture spectra contain contributions which the model has not been taught to handle, resulting in an increased number of false positives. The transparency of VCD to achiral compounds, on the other hand, limits the number of new components capable of contributing to the mixture spectrum, which could result in fewer false positives.

To summarise, with the dataset of terpenes presented in this article, we could build ML models to determine the terpene composition of mixtures with moderate complexity. For mixtures of high complexity, the models begin to struggle to accurately predict the presence of the terpenes, especially if the major contributions are not accounted for in the dataset. The current models are not ready yet to tackle analysis of essential oils in general due to the limited number of compounds in the spectral database. The approach, however, shows promise in its ability to detect those compounds indeed represented. We believe that continuing to build this dataset, with spectra of either pure compounds or mixtures, will enable researchers to push the boundaries of VCD applications to secondary metabolites.

6.3 Conclusions

Despite advances over the last decade, VCD spectroscopy remains an untapped resource for the determination of the absolute configuration by the natural product community. One of the reasons is the requirement of quantum chemical calculations to interpret experimental data. In this perspective, we present an approach to simultaneously detect and assign absolute configuration of natural products even in mixtures, and without the need of DFT calculations. The proposed approach focuses on the search of IR and VCD spectral markers/regions of individual molecules to be applied in complex mixtures. As a proof-of-concept, monoterpenes were chosen as target molecules. The spectral marker/regions searches were undertaken both by visual inspection and by means of machine learning. Visual inspection is a viable procedure for monoterpenes; however, it is time-consuming and prone to user bias. Machine learning methods, on the other hand, renders itself as a promising tool for detection and stereochemical analysis of complex mixtures. Due to the number of false positives for natural mixtures, the suggested approach is not yet competitive with other classical methods such as GC-MS. Although the results obtained for natural mixtures could have been better, the good performance for artificial mixtures indicates that ML is a promising tool provided the number of molecules/spectra included in the dataset is expanded. Consequently, further IR/VCD spectra need to be recorded for structurally diverse molecules, both aquiral/racemic and chiral, that commonly compose essential oils and other important mixtures. Once the number of IR/VCD spectra available is increased, we expect ML-based methods to be able to tackle mixtures of increasing complexity, such as essential oils, crude extracts, as well as reaction media of stereoselective chemical transformations.

Supporting information

Experimental details

All monoterpenes and essential oils used were purchased from Sigma-Aldrich and used without further purification. The artificial mixtures were prepared by mixing equal amounts of each compound. IR and VCD spectra were recorded simultaneously with a BioTools Chiral*IR*-2x FT-VCD spectrometer with either single or dual-PEM setups using a resolution of 4 cm^{-1} and a collection time of 10-12 hours. The optimum retardation of the ZnSe photoelastic modulator(*S*) (PEM) was(ere) set at 1400 cm^{-1} . The IR and VCD spectra were recorded in $CDCl_3$ solutions (0.2-0.8M) in a BaF_2 cell with a 100 μm path length. Minor instrumental baseline offsets were eliminated from the final VCD spectrum by subtracting the VCD spectrum each compound from that obtained for the solvent under the same conditions. The database of VCD and IR spectra is publicly available and can be retrieved using the following DOI [10.5281/zenodo.7875469]. The absolute configuration of each monoterpene when applicable was secured by DFT calculations at the B3PW91/PCM(CHCl3)/6-311G(d,p) level (data not shown). These calculations also allowed the assignment of the vibrational origin of specific bands.

Results of IR/VCD visual inspection

Figures 6.4-6.10 present the superposition of the IR/VCD spectra of individual monoterpenes within a given molecule type, namely, pinane, menthane 1 and 2, bornane, fenchane and geraniol type as well as the spectra of the single representatives of carene and thujane types along with cineole. Then, IR/VCD spectra of the mixtures of compounds of each type are presented in Figures 6.11-6.16. The following discussions about spectral markers are focused on transitions able to tell apart compounds within the same molecule type.

Regarding pinane type monoterpenes, the main discriminatory IR bands observed (Fig. 6.11) were those at 1639 cm^{-1} present in (*S*)-(−)-β-pinene (exocyclic double bond stretching); 1616 cm^{-1} present in (*R*)-(−)-myrtenal and (1*S*)-(−)-verbenone (α,β-unsaturated double bond stretching); 1250 cm^{-1} present in

(1*R*)-(–)-myrtenol and (1*R*,2*R*,3*S*,5*R*)-(–)-pinanediol (C-O stretching); and 1035 cm^{-1} present in (1*R*,2*R*,3*R*,5*S*)-(–)-isopinocampheol (C-O stretching coupled to C-H bendings of the whole molecular framework). The VCD marker bands included those at (–)-1195 cm^{-1} present in (*S*)-(–)-β-pinene (C-H bendings of the whole molecular framework); (+)-1126 cm^{-1} present in (*R*)-(+)-α-pinene (C-H bendings of the whole molecular framework); (+)-1035 cm^{-1} present in (1*R*,2*R*,3*R*,5*S*)-(–)-isopinocampheol (C-O stretching coupled to C-H bendings of the whole molecular framework); and (–)-967 cm^{-1} present in (1*R*)-(–)-myrtenal (C-sp^3-C-sp^2stretching coupled to C-H bendings of the whole molecular framework).

For menthane type 1 molecules (Fig. 6.12), the IR discriminative bands were those at 1643 cm^{-1} present in (*S*)-(–)-perillaldehyde, (*S*)-(–)-perillyl alcohol, (*R*)-(–)-carvone (broader shoulder) and (*R*)-(+)-limonene (stretching terminal double bond); 1415 cm^{-1} present in (*S*)-(–)-perillaldehyde (CH_2 scissoring), 1045 cm^{-1} present in (*S*)-(–)-α-terpineol (C-sp^3-C-sp^2stretching coupled to C-H bendings of the whole molecular framework); and 975 cm^{-1} present in (*S*)-(–)-perillyl alcohol (Coupled C-C stretchings and C-H bending of the whole molecular framework). As for VCD marker bands, the band at (–)-1434 cm^{-1} (asymmetric CH3 bending and C-H2 scissoring modes) was present in all molecules, except (*R*)-(–)-terpinen-4-ol, while that at (–)-1250 cm^{-1} (C-H bendings of the whole molecular framework and C-H2 twisting modes) was present in all molecules, except (*S*)-(–)-perillaldehyde. The band (–)-1045 cm^{-1} was present only in (*S*)-(–)-α-terpineol (C-sp^3-C-sp^2stretching coupled to C-H bendings of the whole molecular framework).

For the menthane type 2 monoterpenes (Fig. 6.13) important IR bands include those at 1677, 1614 (broad) (α,β-unsaturated carbonyl stretchings), and 1286 cm^{-1} (C-sp^2-C-sp^2stretch) present in (*R*)-(+)-pulegone; 1642 (terminal double bond stretch), 1394 (double bond scissoring), and 1286 cm^{-1} (coupled O-H and C-H bending modes) present in (1*R*,2*S*,5*R*)-(–)-isopulegol. As for VCD, at around 1286 cm^{-1}, both (*R*)-(+)-pulegone and (1*R*,2*S*,5*R*)-(–)-isopulegol presented a positive band, however, in contrast to IR, these bands were better resolved due to their different vibrational origins. At 1103 cm^{-1} (C-CH3 and C-O stretchings) a positive VCD band was characteristic of (1*R*,2*S*,5*R*)-(–)-menthol, while a +,–band (low to high wavenumbers) centered at 1070 cm^{-1} (same C-C strecthings coupled to bendings of the whole molecular framework) was observed for (1*S*,2*R*,5*R*)-(+)-isomenthol. A negative 1012 cm^{-1} band (C-C stretchings and C-H isopropyl bending) was observed for both (1*R*,2*S*,5*R*)-(−)-isopulegol and (1*S*,2*S*,5*R*)-(+)-neomenthol, while a positive band at 962 cm^{-1} was present in the spectra of (1*R*,2*S*,5*R*)-(−)-isopulegol, (1*S*,2*R*,5*R*)-(+)-isomenthol, and (1*S*,2*S*,5*R*)-(+)-neomenthol. While the band at (–)-1012 cm^{-1} band seems to be selective of menthane molecules with trans relationship between the isopropyl and methyl groups, the (+)-962 cm^{-1} band arise from C-C stretchings and C-H bendings of the whole molecular framework, being representative of the menthane type 2 scaffold.

Considering bornane type molecules (Fig. 6.14) the IR marker bands identified include that at 1415 cm^{-1} observed for (1*R*)-(+)-camphor (CH_2 scissoring in the vicinity of carbonyl group); those at 998 and 1068 cm^{-1} present in (±)-isoborneol (C-C-O stretching coupled to CH_2 rocking vibrations), and those at 1012, 1229 and 1253 characteristic of (*S*)-(–)-*endo*-borneol (C-C-O stretching coupled to CH_2 rocking vibrations). These latter vibrations reflect the *endo* and exo orientations of the OH group in these stereoisomers. Distinctive VCD bands in bornane type molecules were observed at (+)-1320 and (+)-1166 cm^{-1} for (1*R*)-(+)-camphor (C-C stretch of quaternary bridgehead carbon coupled to CH_2 wagging and C-C stretch of quaternary bridge carbon coupled to methyne bending, respectively); at 1259 cm^{-1} a negative couplet-like band (from low to high wavenumbers) was observed for (*S*)-(–)-*endo*-borny acetate (C-sp^2-O stretching coupled to CH_2 wagging and methyne bending modes); centered at 1125 cm^{-1} a negative couplet-like band (from low to high wavenumbers) was observed for (*S*)-(–)-*endo*-borneol and (*S*)-(–)-*endo*-borny acetate (C-O stretching coupled to C-C stretches of the whole molecular framework and methyne bending modes); at 1070 and 981 cm^{-1} two positive VCD bands were observed for (*S*)-(–)-*endo*-borny acetate (C-C stretching coupled to C-H bendings involving the whole molecular framework), while positive VCD bands at 1053 and 981 cm^{-1} were present for (*S*)-(–)-*endo*-borneol. Interestingly, the bands at 1135, 1070 and 981 cm^{-1} (fundamentals 124, 116, and 101, respectively in the original publication) could have been used to assign the absolute configuration of the monoterpenic portion of the monoterpene chromane esters isolated from *Peperomia obtusifolia* in 2011[58]. At that time, the stereochemistry of the bornyl moieties tethered to the 3,4-dihydro-5-hydroxy-2,7-dimeth-yl-8-(3”-methyl-2”-butenyl)-2-(4’-methyl-1’,3’-pentadienyl)-2H-1-benzopyran-6-carboxylic acid were determined using arithmetic operations on experimental and calculated spectra for diastereomeric compounds.

In the case of fenchane type molecules (Fig. 6.15), the IR bands at 1080, 1064 and 1010 cm^{-1} (C-C stretchings coupled to C-H bendings involving the whole molecular framework) were present in (1*R*)-(+)-*endo*-fenchyl alcohol, while the band at 1023 cm^{-1} (C-sp^3-C-sp^{2}stretching coupled to C-H bendings involving the whole molecular framework) was characteristic of (*S*)-(+)-fenchone in this region. In the VCD spectra, the positive band at 1080 cm^{-1} was observed for (1*R*)-(+)-*endo*-fenchyl alcohol, while the (+)-1023 cm^{-1} and (–)-996 cm^{-1} (C-H bendings involving the whole molecular framework), were characteristic of (*S*)-(+)-fenchone.

Finally, considering the linear terpenes (geraniol type) (Fig. 6.16), the IR band at 1672 cm^{-1} was observed for all molecules since it involved the stretching of the trisubstituted double bound from the terminal isoprene unit. Bands at 1637 and 1412 cm^{-1} were observed for (*R*)-(–)-linalyl acetate, (*R*)-(–)-linalool and (*S*)-(+)-β-citronellene and involved stretching and scissoring modes of their terminal double bond; at 1477 cm^{-1} a shoulder band was present only in the spectrum of (*S*)-(–)-β-citronellol (scissoring of CH_2-OH); at 1106 cm^{-1} a band was observed for (*R*)-(–)-linalool arising C-O stretching and O-H bending of the tertiary alcohol, while the same vibration modes were observed at 1054 cm^{-1} for the primary alcohol (*S*)-(–)-β-citronellol. Despite the lower intensities and noisier VCD spectra observed for linear monoterpenes, some discriminatory VCD bands were identified, such as the (+)-1089 cm^{-1} observed for (*S*)-(+)-β-citronellene (C-H bendings involving the whole molecular framework); the (–)-1075 cm^{-1} band (C-C stretches coupled to C-H and O-H bending modes) observed for (*R*)-(–)-linalool, and the (+)-1054 cm^{-1} observed for (*S*)-(–)-β-citronellol.

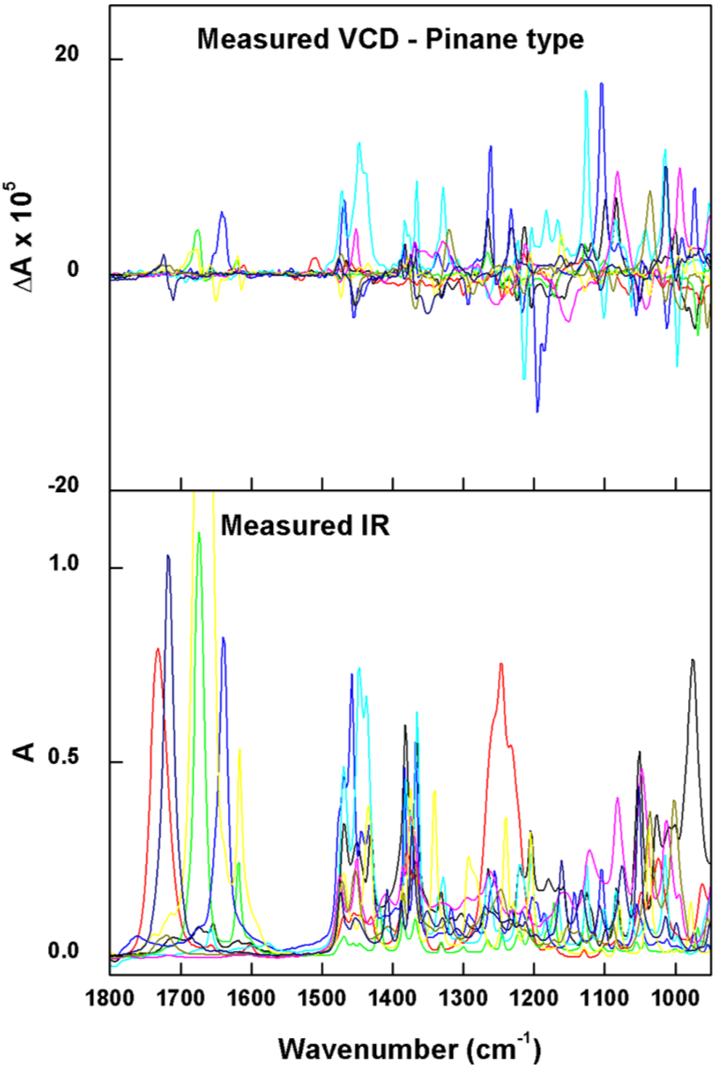


Figure 6.4: Superposition of IR/VCD experimental spectra in $CDCl_3$ of pinane type monoterpenes. (Black) (1*R*)-(−)-myrtenol; (Green) (1*R*)-(−)-myrtenal; (Red) (1*R*)-(−)-myrtenyl acetate; (Blue) (*S*)-(−)-β-pinene; (Cyan) (*R*)-(+)-α-pinene; (Magenta) (1*R*,2*R*,3*S*,5*R*)-(−)-pinanediol; (Yellow) (1*S*)-(−)-verbenone; (Navy) (1*S*,2*S*,5*S*)-(−)-2-hydroxy-3-pinanone; (Dark Yellow) (1*R*,2*R*,3*R*,5*S*)-(−)-isopinocampheol

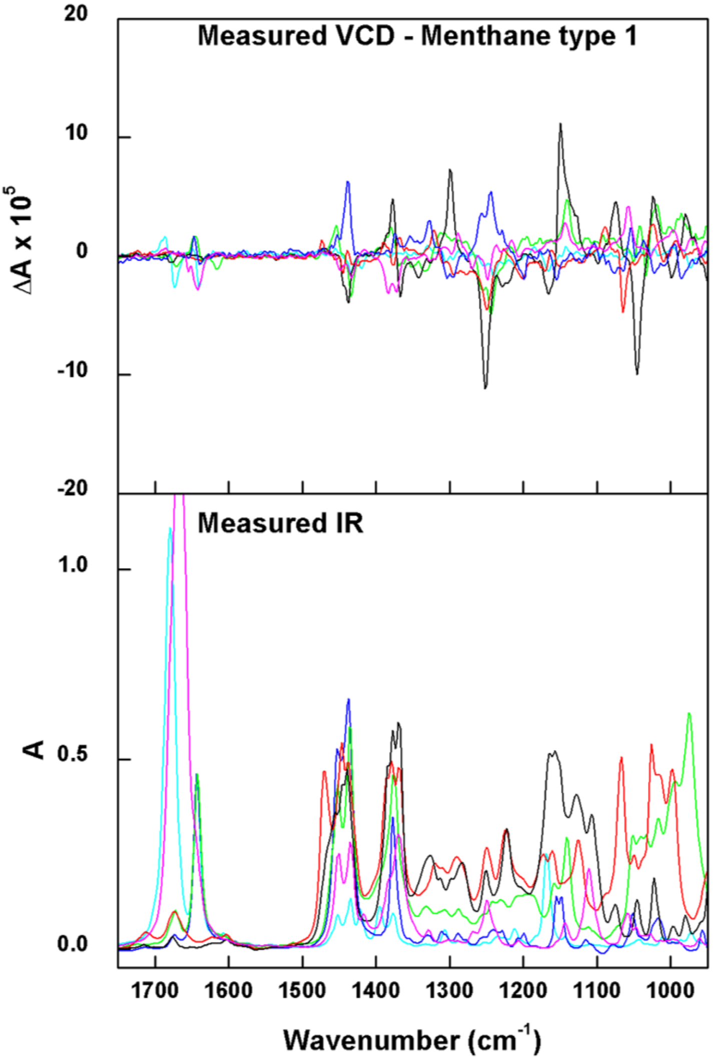


Figure 6.5: Superposition of IR/VCD experimental spectra in $CDCl_3$ of menthane type 1 monoterpenes. (Black) (*S*)-(−)-α-terpineol; (Green) (*S*)-(−)-perillyl alcohol; (Red) (*R*)-(−)-terpinen-4-ol; (Blue) (*R*)-(+)-limonene; (Cyan) (*S*)-(−)-perillaldehyde; (Magenta) (*R*)-(−)-carvone.

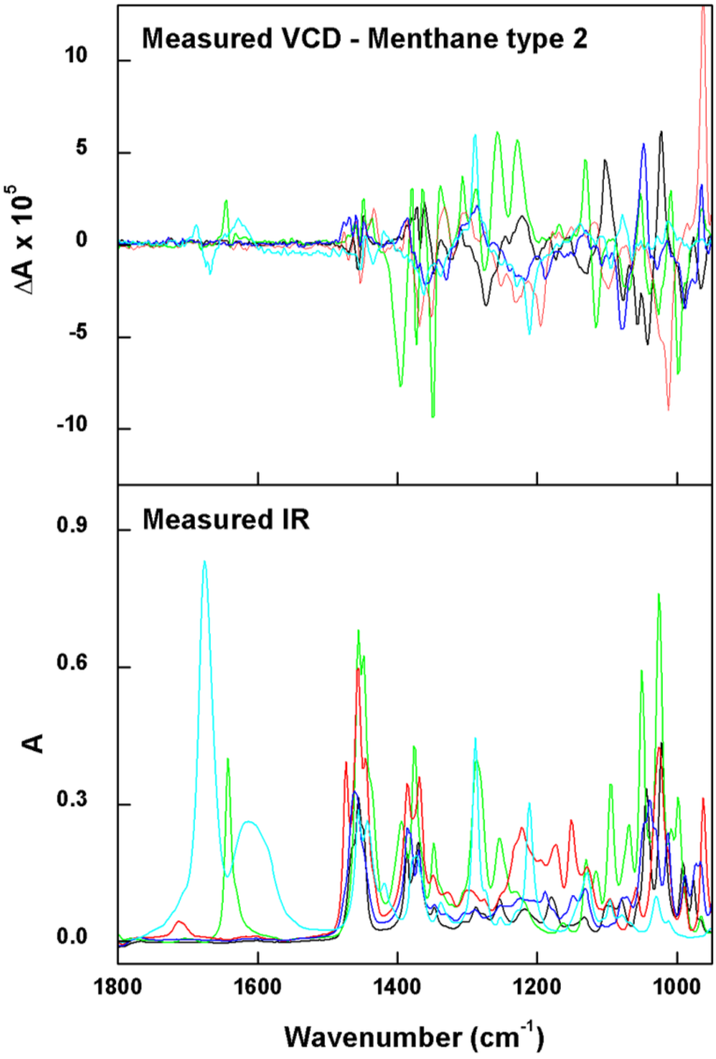


Figure 6.6: Superposition of IR/VCD experimental spectra in $CDCl_3$ of menthane type 2 monoterpenes. (Black) (1*R*,2*S*,5*R*)-(−)-menthol; (Green) (1*R*,2*S*,5*R*)-(−)-isopulegol; (Red) (1*S*,2*S*,5*R*)-(+)-neomenthol; (Blue) (1*S*,2*R*,5*R*)-(+)-isomenthol; (Cyan) (*R*)-(+)-pulegone.

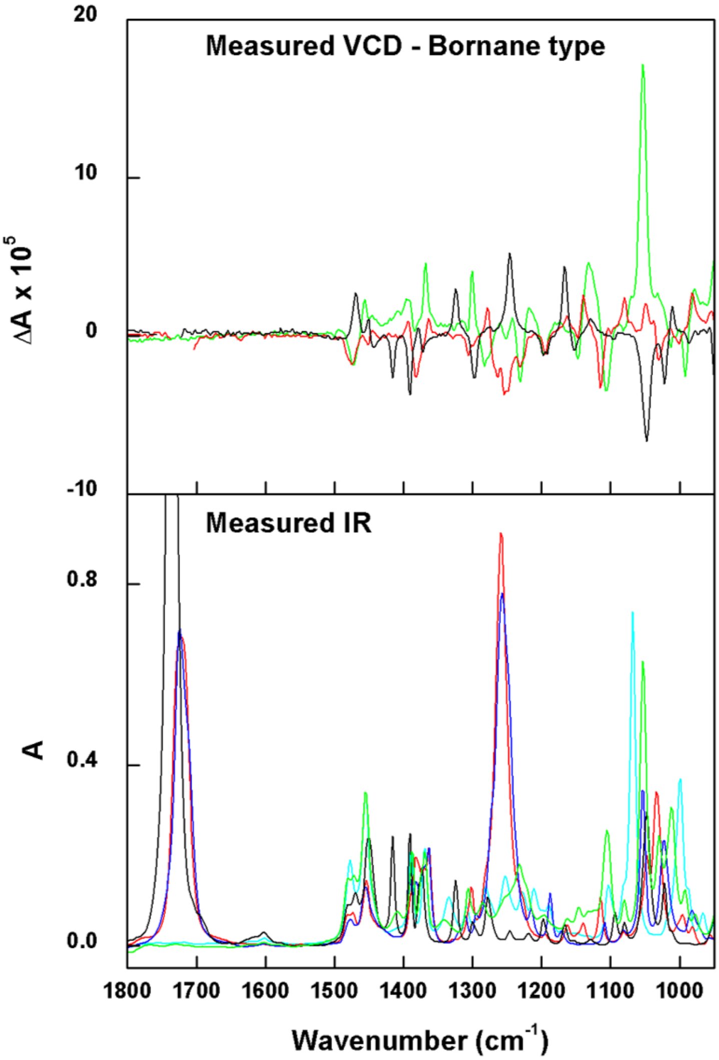


Figure 6.7: Superposition of IR/VCD experimental spectra in $CDCl_3$ of bornane type monoterpenes. (Black) (1*R*)-(+)-camphor; (Green) (*S*)-(−)-*endo*-borneol; (red) (*S*)-(−)-*endo*-borny acetate; (Blue) (±)-isobornyl acetate (IR only); (Cyan) (±)-isoborneol (IR only). Gap in the carbonyl region due to high noise level.

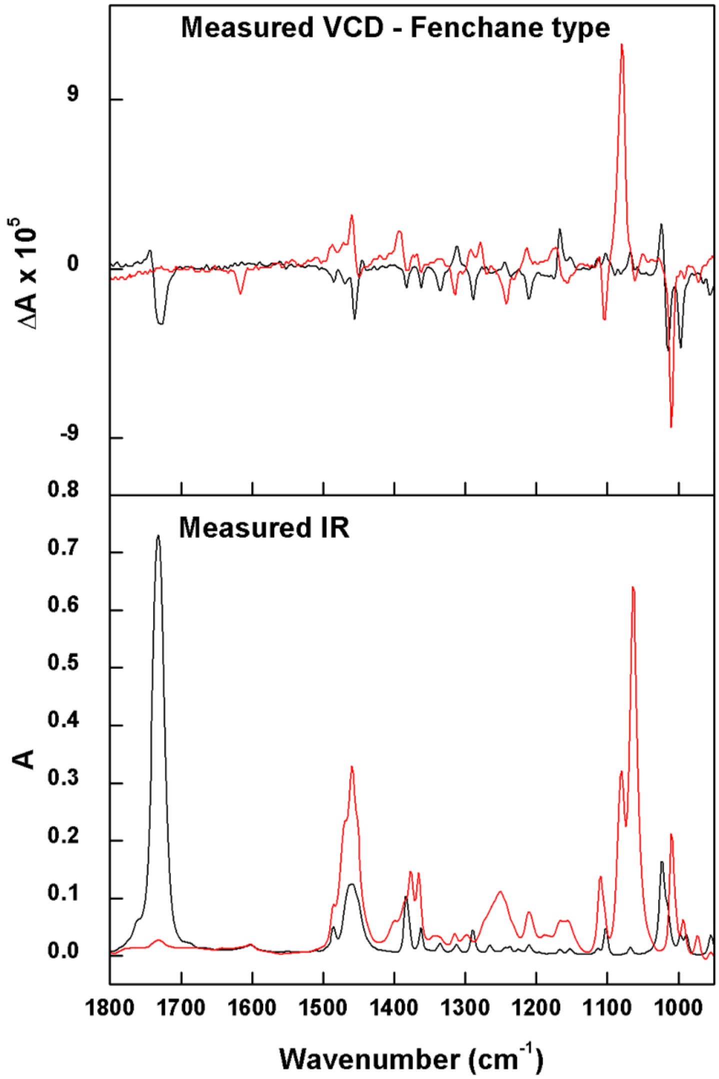


Figure 6.8: Superposition of IR/VCD experimental spectra in $CDCl_3$ of fenchane type monoterpenes. (Black) (S)-(+)-fenchone; (Red) 1R)-(+)-*endo*-fenchyl alcohol.

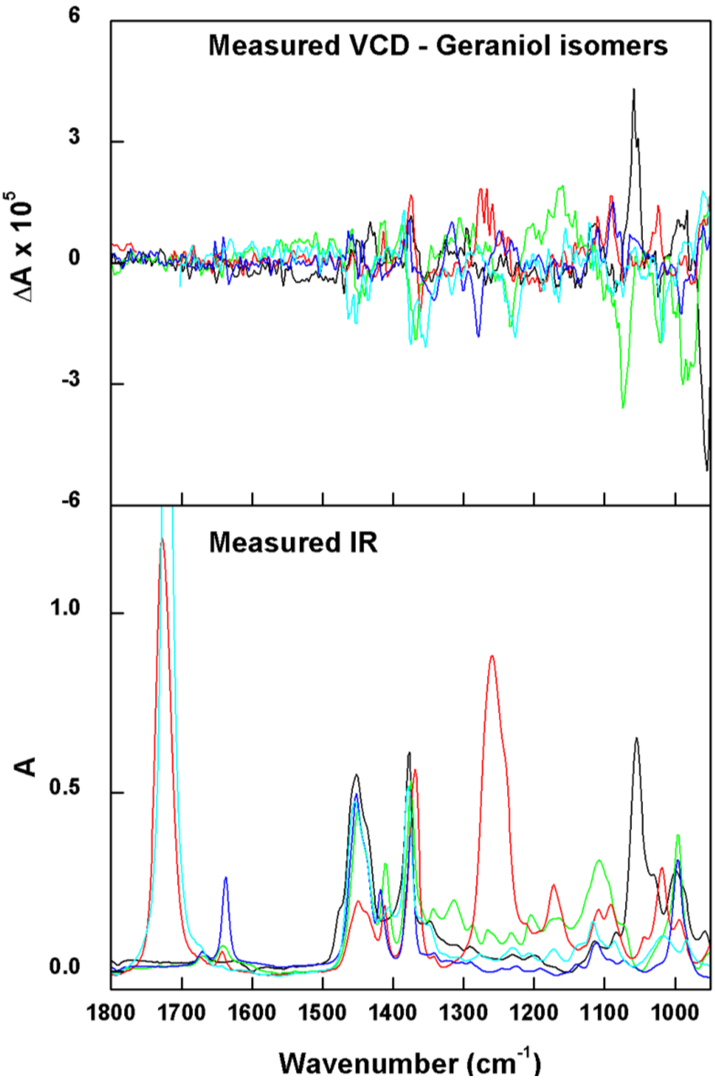


Figure 6.9: Superposition of IR/VCD experimental spectra in $CDCl_3$ of geraniol type monoterpenes. (Black) (*S*)-(−)-β-citronellol; (Green) (*R*)-(−)-linalool; (Red) (*R*)-(−)-linalyl acetate; (Blue) (*S*)-(+)-β-citronellene; (Cyan) (*S*)-(−)-β-citronellal. Gap in the carbonyl region due to high noise level.

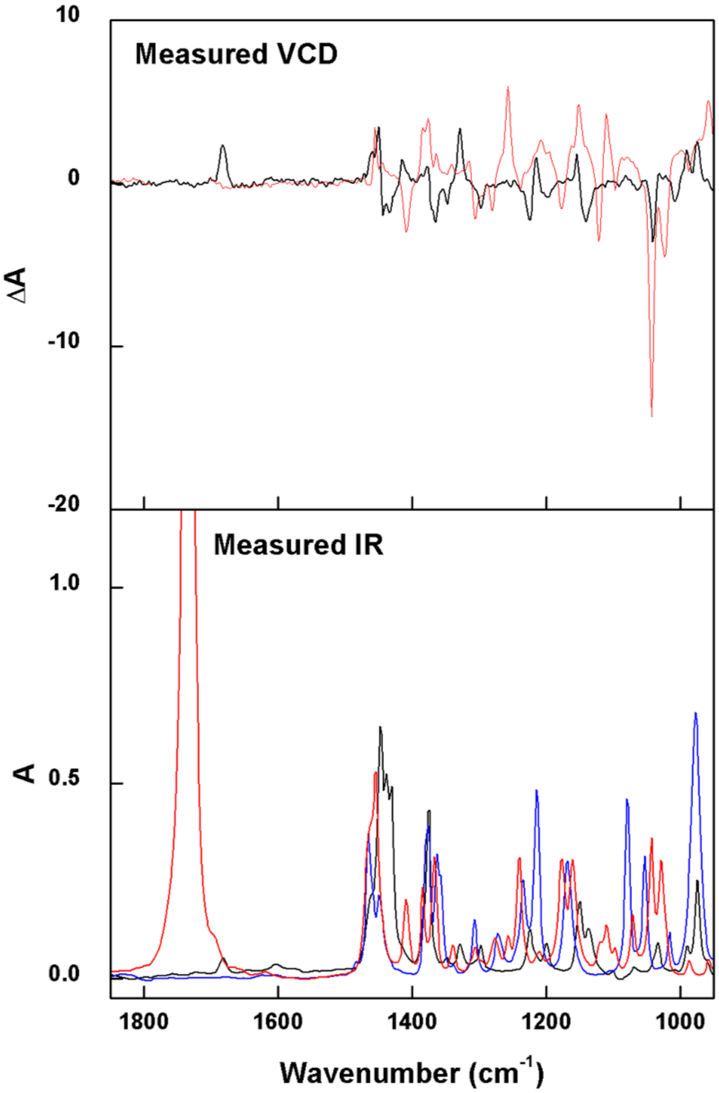


Figure 6.10: Superposition of IR/VCD experimental spectra in $CDCl_3$ of: (Black) (1*S*)-(+)-3-carene; (Red) ((1*S*,4*R*)-(−)-α-thujone; (Blue) cineole (IR only). Gap in the carbonyl region due to high noise level.

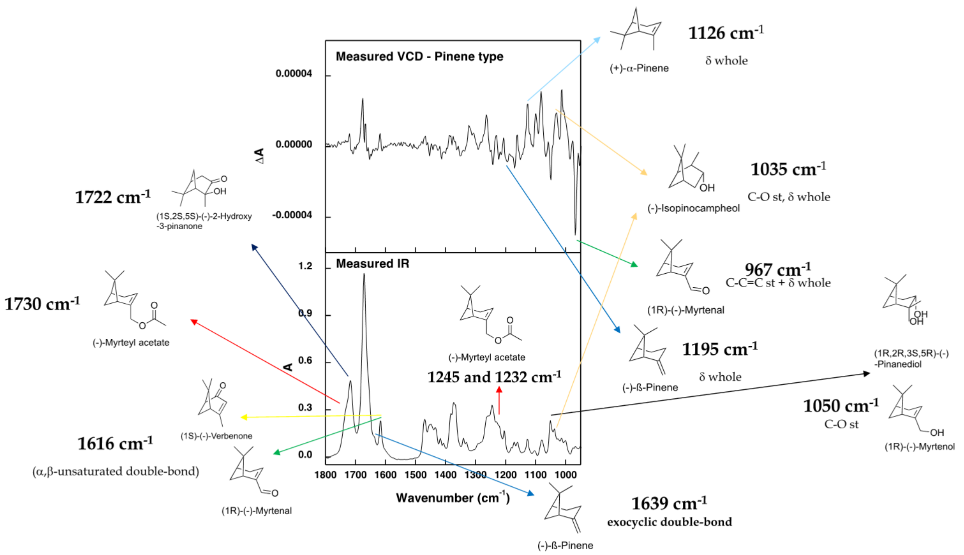


Figure 6.11: Monoterpenes identified from the artificial mixture of pinane type molecules by means of visual IR/VCD spectral markers. Selected vibrational frequencies and molecular origin also provided.

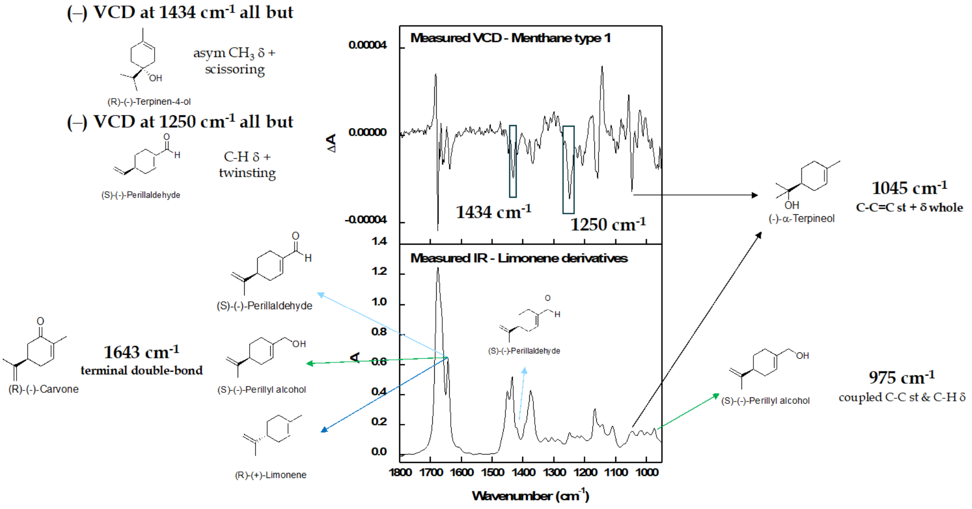


Figure 6.12: Monoterpenes identified from the artificial mixture of menthane type 1 molecules by means of visual IR/VCD spectral markers. Selected vibrational frequencies and molecular origin also provided. Shaded areas represent common bands.

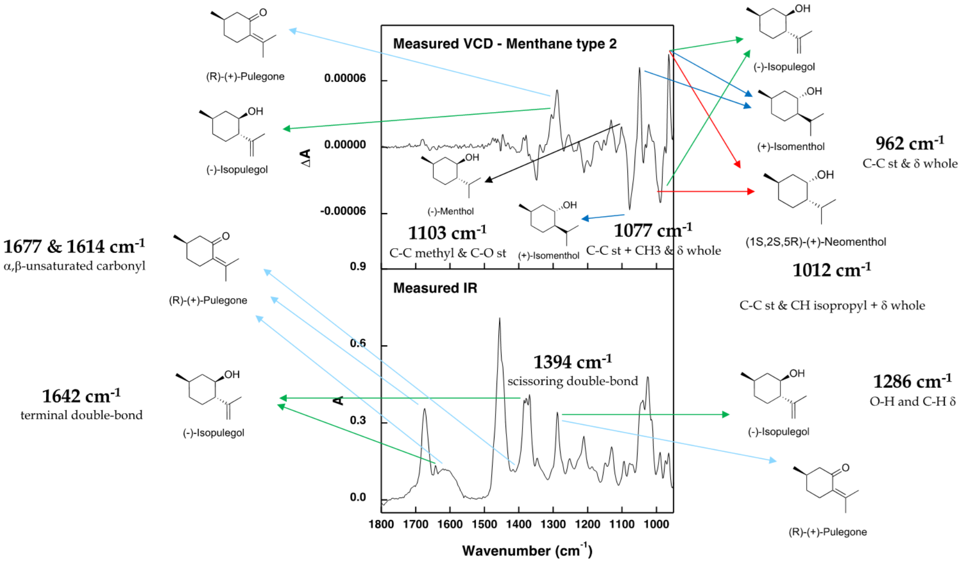


Figure 6.13: Monoterpenes identified from the artificial mixture of menthane type 2 molecules by means of visual IR/VCD spectral markers. Selected vibrational frequencies and molecular origin also provided.

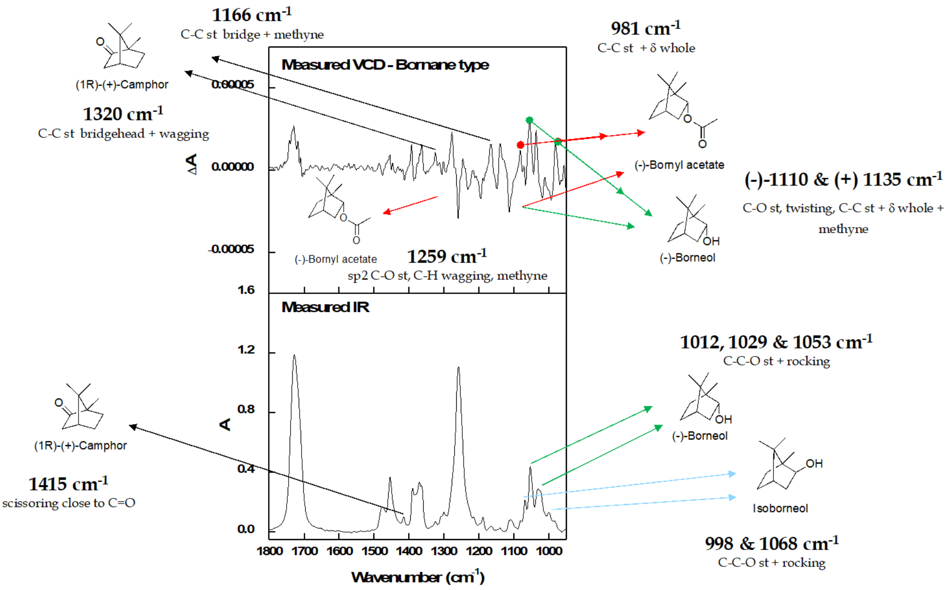


Figure 6.14: Monoterpenes identified from the artificial mixture of bornane type molecules by means of visual IR/VCD spectral markers. Selected vibrational frequencies and molecular origin also provided. Shaded areas indicate couplet signals.

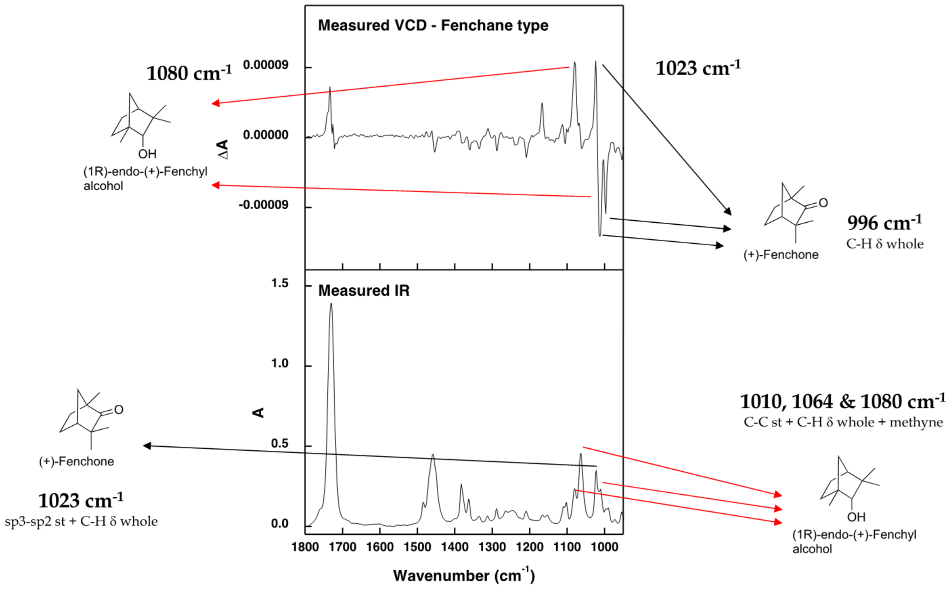


Figure 6.15: Monoterpenes identified from the artificial mixture of fenchane type molecules by means of visual IR/VCD spectral markers. Selected vibrational frequencies and molecular origin also provided.

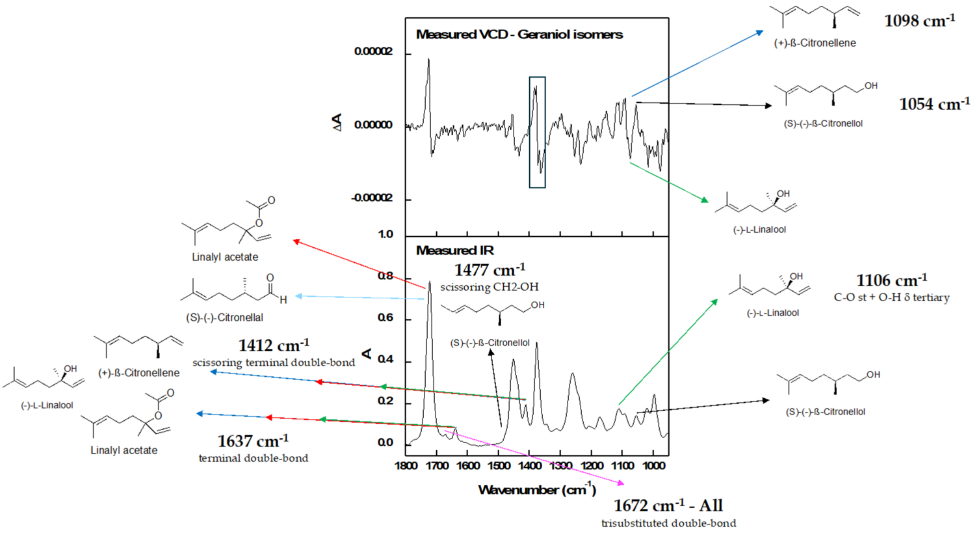
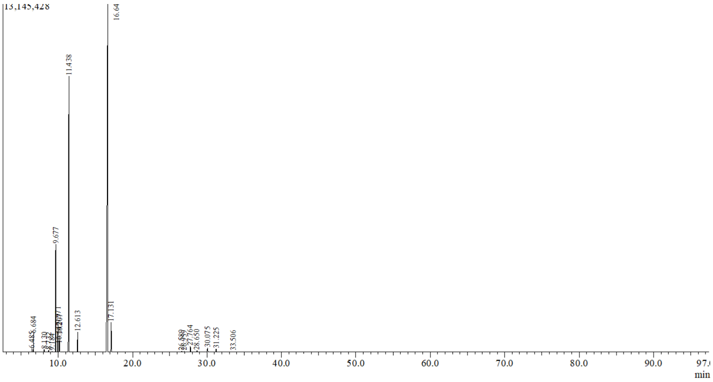


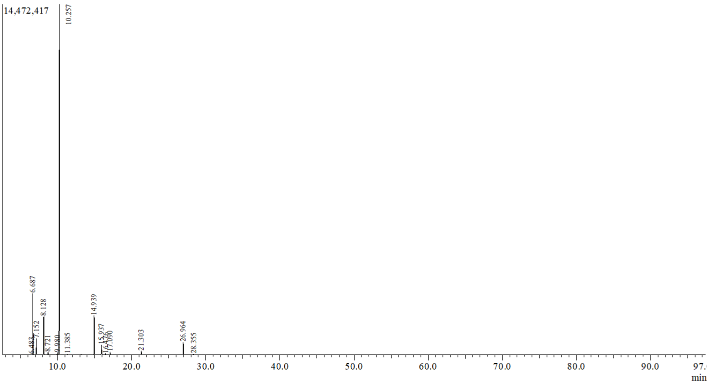
Figure 6.16: Monoterpenes identified from the artificial mixture of geraniol type molecules by means of visual IR/VCD spectral markers. Selected vibrational frequencies and molecular origin also provided. Shaded areas represent common bands.



Peak Report

Peak#	R.Time	I.Time	F.Time	Area	Area%	Height	Height%	A/H	Mark	Name
1	6.485	6.445	6.540	278560	0.20	124421	0.38	2.24		(1S)-2,6,6-Trimethylbicyclo[3.1.1]hept-2-e
2	6.684	6.635	6.760	1536672	1.12	693524	2.10	2.22		(1S)-2,6,6-Trimethylbicyclo[3.1.1]hept-2-e
3	8.130	8.080	8.195	249591	0.18	94333	0.29	2.65		.beta.-Pinene
4	8.732	8.690	8.815	198223	0.14	73506	0.22	2.70		4-(4-Methylpent-3-enyl)-3,6-dihydro-1,2-d
5	9.184	9.140	9.240	116093	0.08	43286	0.13	2.68		Tricyclo[4.1.0.0(2,4)]heptane, 5-(phenylthi
6	9.677	9.590	9.765	9670618	7.06	4019617	12.18	2.41		(+)-2-Carene
7	9.971	9.915	10.050	2648156	1.93	1048314	3.18	2.53		o-Cymene
8	10.142	10.050	10.170	826375	0.60	305256	0.93	2.71	V	Limonene
9	10.207	10.170	10.305	1777173	1.30	623122	1.89	2.85	V	5-Hepten-2-one, 6-methyl-
10	11.438	11.325	11.510	32100436	23.44	10301473	31.23	3.12		.gamma.-Terpinene
11	12.613	12.550	12.710	2185972	1.60	753982	2.29	2.90		Cyclohexene, 4-methyl-3-(1-methylethylid
12	16.648	16.415	16.730	79271621	57.88	13141442	39.83	6.03		3-Cyclohexen-1-ol, 4-methyl-1-(1-methylet
13	17.131	17.040	17.240	3664831	2.68	1115823	3.38	3.28		Terpineol
14	26.589	26.535	26.635	109585	0.08	37215	0.11	2.94		1H-Cycloprop[e]azulene, 1a,2,3,4,4a,5,6,7
15	26.959	26.905	27.020	89193	0.07	29558	0.09	3.02		1,3,6,10-Dodecatetraene, 3,7,11-trimethyl-
16	27.764	27.695	27.890	803104	0.59	216642	0.66	3.71		Aromandendrene
17	28.650	28.600	28.715	178451	0.13	57728	0.17	3.09		Aromandendrene
18	30.075	30.005	30.170	632058	0.46	156728	0.48	4.03		Naphthalene, 1,2,3,5,6,7,8,8a-octahydro-1,
19	31.225	31.145	31.330	549862	0.40	127522	0.39	4.31		Naphthalene, 1,2,3,5,6,8a-hexahydro-4,7-d
20	33.506	33.455	33.570	80527	0.06	26583	0.08	3.03		(2S,4R)-p-Mentha-[1(7),8]-diene 2-hydrop
				136967101	100.00	32990075	100.00			

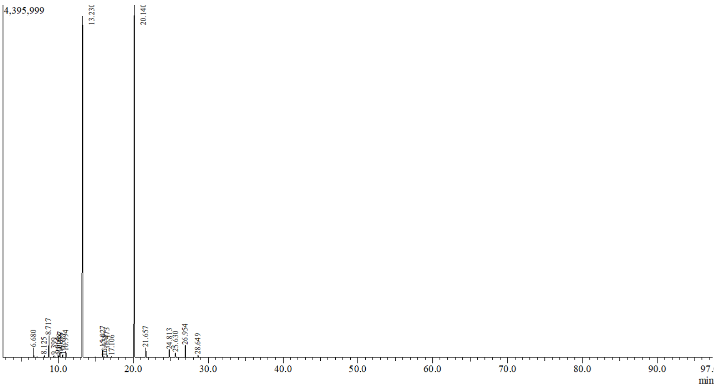
Figure 6.17: GC-MS analysis of tea tree oil.



Peak Report

Peak#	R.Time	I.Time	F.Time	Area	Area%	Height	Height%	A/H	Mark	Name
1	6.483	6.455	6.525	60210	0.08	27601	0.12	2.18		Tricyclo[4.1.0.0(2,4)]heptane, 5-(phenylthi
2	6.687	6.630	6.760	5264602	7.28	2544402	11.32	2.07		.alpha.-Pinene
3	7.152	7.095	7.220	1650321	2.28	688891	3.06	2.40		Camphene
4	8.128	8.065	8.205	3763574	5.21	1585198	7.05	2.37		.beta.-Pinene
5	8.721	8.680	8.790	286812	0.40	107838	0.48	2.66		.beta.-Myrcene
6	9.980	9.935	10.065	313009	0.43	100371	0.45	3.12		o-Cymene
7	10.257	10.105	10.340	51305089	70.97	14471035	64.37	3.55		Eucalyptol
8	11.385	11.340	11.445	102189	0.14	35917	0.16	2.85		.alpha.-Phellandrene
9	14.939	14.855	15.020	5190140	7.18	1620916	7.21	3.20		(+)-2-Bornanone
10	15.937	15.865	16.000	1285373	1.78	403026	1.79	3.19		Isoborneol
11	16.476	16.430	16.560	172460	0.24	55785	0.25	3.09		Terpinen-4-ol
12	17.090	17.040	17.200	443227	0.61	126225	0.56	3.51		Terpineol
13	21.303	21.240	21.385	574823	0.80	162311	0.72	3.54		Acetic acid, 1,7,7-trimethyl-bicyclo[2.2.1]l
14	26.964	26.880	27.040	1775099	2.46	517018	2.30	3.43		Aromandendrene
15	28.355	28.300	28.410	108366	0.15	33887	0.15	3.20		trans-.beta.-Ocimene
				72295294	100.00	22480421	100.00			

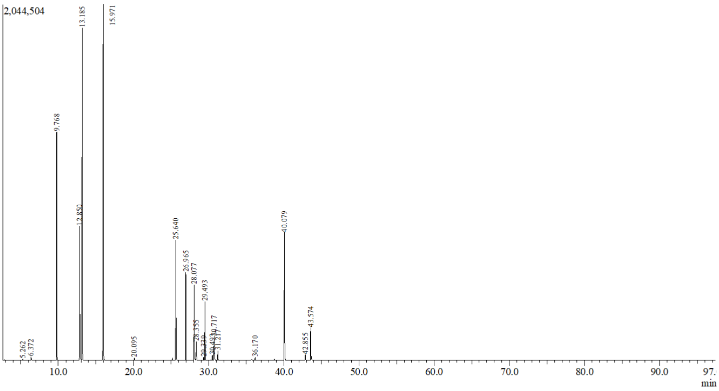
Figure 6.18: GC-MS analysis of rosemary oil.



Peak Report

Peak#	R.Time	I.Time	F.Time	Area	Area%	Height	Height%	A/H	Mark	Name
1	6.680	6.635	6.735	289075	0.83	124599	1.24	2.32		(1S)-2,6,6-Trimethylbicyclo[3.1.1]hept-2-e
2	8.125	8.080	8.180	87946	0.25	36724	0.36	2.39		.beta.-Pinene
3	8.717	8.670	8.815	731104	2.11	274715	2.72	2.66		.beta.-Myrcene
4	9.399	9.360	9.445	81045	0.23	30661	0.30	2.64		(1S)-2,6,6-Trimethylbicyclo[3.1.1]hept-2-e
5	9.964	9.925	10.025	99883	0.29	39574	0.39	2.52		Benzene, tert-butyl-
6	10.138	10.100	10.155	59888	0.17	26581	0.26	2.25		Cyclopropane, 1,1'-ethenylidenebis-
7	10.197	10.155	10.270	189289	0.55	60663	0.60	3.12	V	5-Hepten-2-one, 6-methyl-
8	10.577	10.545	10.645	102684	0.30	39530	0.39	2.60		(1S)-2,6,6-Trimethylbicyclo[3.1.1]hept-2-e
9	10.994	10.955	11.090	233603	0.67	77250	0.77	3.02		1,3,7-Octatriene, 3,7-dimethyl-
10	13.230	13.100	13.305	15465248	44.65	4246477	42.12	3.64		Linalool
11	15.927	15.870	16.000	394582	1.14	122968	1.22	3.21		Isoborneol
12	16.135	16.085	16.190	44451	0.13	15860	0.16	2.80		3-Methyl-3-nitrobut-1-ene
13	16.473	16.410	16.550	340825	0.98	106100	1.05	3.21		Terpinen-4-ol
14	17.106	17.070	17.165	43573	0.13	16647	0.17	2.62		4-Pentene-2-ol, 2-methyl
15	20.140	20.020	20.225	14775784	42.66	4393637	43.58	3.36		Linalyl acetate
16	21.657	21.595	21.750	453468	1.31	124372	1.23	3.65		.beta.-Myrcene
17	24.813	24.750	24.925	375972	1.09	99104	0.98	3.79		2,6-Octadien-1-ol, 2,7-dimethyl-
18	25.630	25.585	25.725	217842	0.63	59977	0.59	3.63		4-Hexen-2-one, 3-methyl-
19	26.954	26.880	27.025	546320	1.58	154090	1.53	3.55		(Z,Z)-.alpha.-Farnesene
20	28.649	28.600	28.725	107149	0.31	32973	0.33	3.25		(E)-.beta.-Famesene
				34639731	100.00	10082502	100.00			

Figure 6.19: GC-MS analysis of lavender oil.



Peak Report

Peak#	R.Time	I.Time	F.Time	Area	Area%	Height	Height%	A/H	Mark	Name
1	5.262	5.240	5.290	10898	0.04	7759	0.08	1.40		4-Penten-2-one
2	6.372	6.350	6.425	48402	0.16	20769	0.22	2.33		4-Hexen-2-one, 3-methyl-
3	9.768	9.715	9.890	3275963	10.85	1312823	14.06	2.50		Benzene, 1-methoxy-4-methyl-
4	12.850	12.795	12.970	2206648	7.31	768246	8.22	2.87		Ethanone, 2-amino-1-phenyl-
5	13.185	13.090	13.295	5882877	19.48	1906601	20.41	3.09		Linalool
6	15.971	15.875	16.065	7020640	23.25	2039082	21.83	3.44		Acetic acid, phenylmethyl ester
7	20.095	20.060	20.160	40694	0.13	12503	0.13	3.25		3-Hexen-2-one, 3-methyl-
8	25.640	25.555	25.735	2418264	8.01	690800	7.40	3.50		2,6-Octadien-1-ol, 2,7-dimethyl-
9	26.965	26.885	27.045	1802430	5.97	506270	5.42	3.56		Aromandendrene
10	28.077	28.010	28.180	1524163	5.05	432764	4.63	3.52		Acetic acid, cinnamyl ester
11	28.355	28.275	28.425	396319	1.31	108371	1.16	3.66		1,3,7-Octatriene, 3,7-dimethyl-
12	29.339	29.310	29.385	33439	0.11	15100	0.16	2.21		1,3,6,10-Dodecatetraene, 3,7,11-trimethyl-
13	29.493	29.420	29.570	1252966	4.15	337623	3.61	3.71		Germacrene D
14	30.493	30.435	30.545	95576	0.32	30396	0.33	3.14		1,3,6,10-Dodecatetraene, 3,7,11-trimethyl-
15	30.717	30.650	30.790	457426	1.51	132673	1.42	3.45		1,3,6,10-Dodecatetraene, 3,7,11-trimethyl-
16	31.217	31.160	31.280	183260	0.61	54964	0.59	3.33		Naphthalene, 1,2,3,5,6,8a-hexahydro-4,7-d
17	36.170	36.120	36.220	45350	0.15	18008	0.19	2.52		Propane, 1-bromo-
18	40.079	39.985	40.165	2696199	8.93	729424	7.81	3.70		Benzyl Benzoate
19	42.855	42.800	42.910	99644	0.33	32486	0.35	3.07		2,6-Octadien-1-ol, 3,7-dimethyl-, acetate
20	43.574	43.500	43.670	709612	2.35	183881	1.97	3.86		Benzoic acid, 2-hydroxy-, phenylmethyl es

Figure 6.20: GC-MS analysis of ylang ylang oil.

Machine learning model structure and development

Due to the absence of a large monoterpene mixture dataset, a set of *in silico* mixture IR and VCD spectra was generated. These spectra were constructed as random linear combinations of the monoterpene spectra, upon which gaussian noise is added. As the *in silico* mixture spectra are linear combinations, a (L2-regularised) linear model was chosen as the basis for the ML model. The model was trained on the VCD and IR *in silico* mixture spectra separately to predict the concentration of each monoterpene. During training, the model teaches itself the marker bands for each terpene, identifies which compounds can attenuate their intensity and from which areas of the spectrum non-marker bands can improve detection. The added noise guided the model to ignore spectral features with intensities close to the noise level and the regularisation implored the model to focus on the wavenumber most important for detecting the specified terpene. By doing so, we limited the overfitting of the model to the *in silico* spectra. The noise level was based on the noise level found in the experimental IR and VCD spectra. The strength of regularisation was increased as much as possible without significantly decreasing the accuracy of the *in silico* concentration predictions (R^2 of approximately 0.98 for unseen *in silico* spectra). Technical details on the training and optimisation procedure are provided in the next section.

Prior to evaluating the obtained results on the experimental mixtures, we discuss the differences in diversity of the IR and VCD spectra for the monoterpenes. As shown in Figure 6.21, the IR spectra are less diverse and grouped into 3 clusters: compounds containing a non-conjugated carbonyl group, a conjugated carbonyl group or lacking any carbonyl group moiety. The spectra within each cluster are strongly similar, increasing the difficulty in separating the contributions of individual terpenes. The low noise level of IR can compensate for the lower diversity, as small contributions can be more easily discerned. In contrast to IR, the VCD spectra are much less correlated as shown in Figure 6.22. The individual contributions of different chiral terpenes are, therefore, expected to be more easily separated from each other. The higher noise level of VCD and baseline uncertainties could increase the difficulty of detecting all contributions, though. The VCD-based model holds two additional advantages for analysis of complex mixtures. The transparency of VCD to achiral compounds improves the stability of the model towards the presence of achiral compounds absent from the dataset. Also, the high sensitivity of VCD to molecular chirality introduces said sensitivity in the model, enabling future use of the model for determination of stereochemistry of essential oil components.

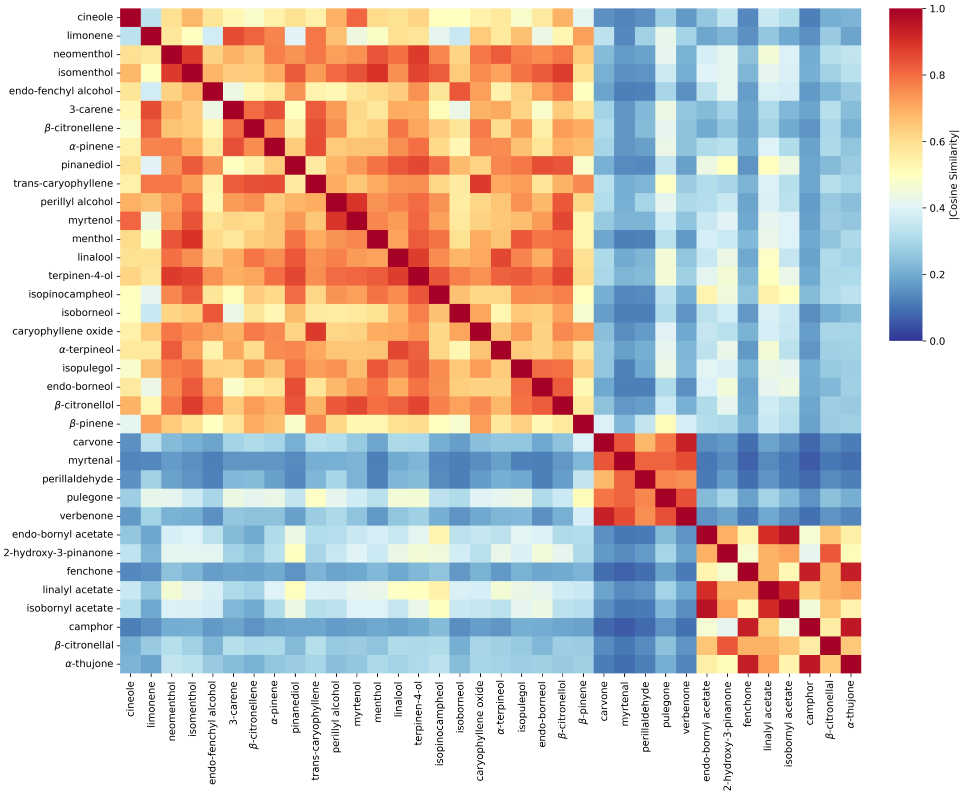


Figure 6.21: Similarity of IR spectra for each pair of monoterpenes. Similarity is expressed as the absolute cosine similarity. The order of the monoterpenes is based on hierarchical clustering on the IR similarity values.

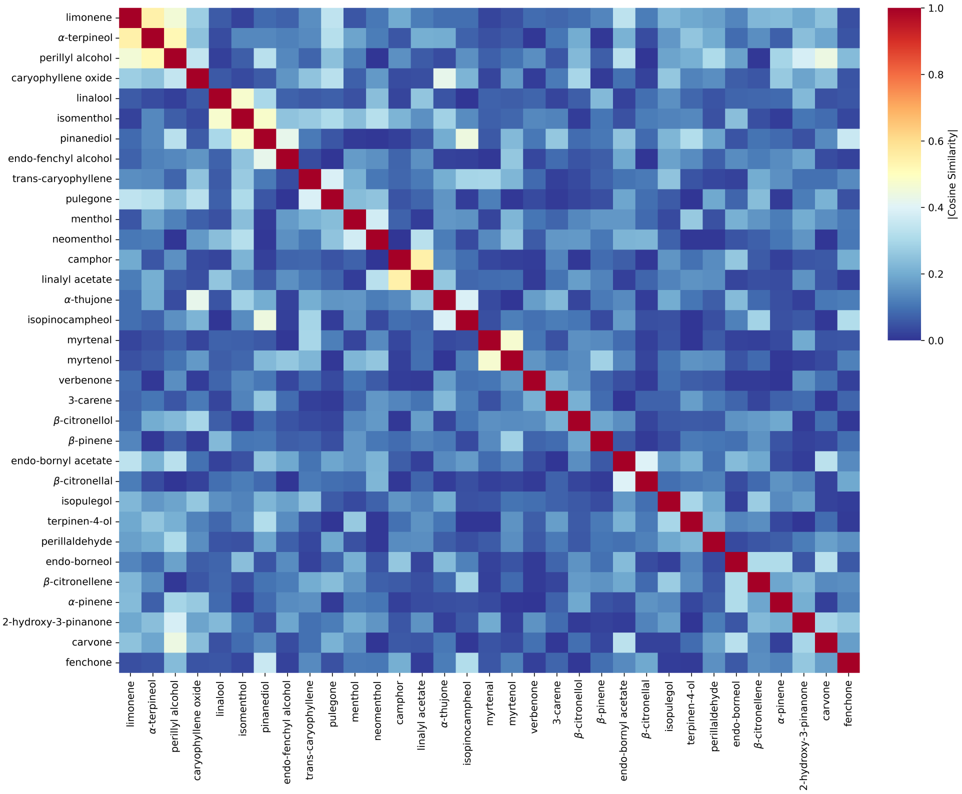


Figure 6.22: Similarity of VCD spectra for each pair of chiral monoterpenes. Similarity is expressed as the absolute cosine similarity. The order of the monoterpenes is based on hierarchical clustering on the VCD similarity values.

Technical details of machine learning model and hyperparameter optimisation

As mentioned in the previous section, the basis of the ML model is a L2-regularised linear regression (also known as Ridge regression) and the model is trained to predict the concentrations of each terpene from the noisy *in silico* mixture spectra. The output of the VCD model is a 33-dimensional vector containing the concentration of each chiral terpene. For the IR model the output is a 36-dimensional vector containing the concentrations of all chiral and achiral terpenes. The spectral intensities (IR or VCD) for each of the 441 wavenumbers between 950 and 1800 cm^{-1} constitute the input of the model. The model was built and trained using the scikit-learn library (version 0.24.2)[59] and default settings were used unless specified otherwise. For the model, the strength of the regularisation, referred to as α, is an important hyperparameter requiring optimisation. Both the VCD and IR model were trained with a range of α values using 10-fold cross validation. The resulting performance for the *in silico* training and validation sets are shown in Fig 6.23. For smaller α values the VCD model is overfitted to the training set and larger α values result in underfitting. By setting α to 1.10^{-9}, both influences are balanced and the resulting model predicts the terpene concentrations with a R^2 of ± 0.98 for *in silico* mixtures. For the IR model, we chose the largest αvalue (1.10^{-1}) that resulted in a similar accuracy (R^2 of ± 0.98). By doing so, we keep the relative level of regularisation consistent for the VCD and IR models.

After the hyperparameter optimisation of the VCD model, the VCD models arising from each fold are combined into an ensemble where the predicted concentration for a single terpene is the mean value of the predicted concentrations of each model and the standard deviation is used to quantify the error upon the mean value. This approach is known as bagging[60] and can improve the robustness of the predictions while providing a notion for the uncertainty upon the predicted values. The approach is repeated for the IR model using the IR models of each fold.

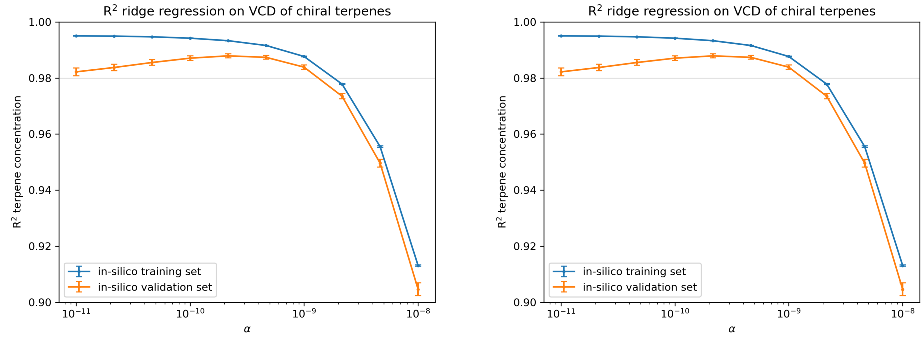


Figure 6.23: Optimisation of regularisation strength α for the VCD (left panel) and IR (right panel) *in silico* mixture spectra. The reported R^2 values are the averages of R^2 for each cross-validation fold and the error on this average is the standard deviation for the R^2 values.

Predictions by machine learning model on mixtures of known composition

The contents of the different experimental mixtures are provided in Table 6.1 and the predicted relative concentrations for the IR and VCD models are shown in Figure 6.3 and Figures 6.26-6.31. Performance of the L2-regularised model on the VCD spectra of mixtures A-F is very promising. For most terpenes, the presence of a specific terpene in a mixture was linked with a higher predicted concentration for that terpene. The VCD-based model could not properly detect the presence/absence of (*S*)-(–)-perillyl alcohol, (1*S*)-(+)-carene, (1*S*,2*R*,5*R*)-(+)-isomenthol and camphor in mixtures A-F (see Figure 6.3). Camphor was the only terpene for which both enantiomers are present in a mixture: (1*S*)-(–)-camphor in mixture A and (1*R*)-(+)-camphor in mixture C. The camphor concentration was expressed in terms of (1*R*)-(+)-camphor for the VCD model so the strong negative prediction should indicate the presence of (1*S*)-(–)-camphor, as enantiomers have mirror image VCD spectra. So the large negative concentration predicted for mixture A shows that the model has identified (1*S*)-(–)-camphor. However, no clear detection of (1*R*)-(+)-camphor was obtained for mixture C. For (*R*)-(–)-linalyl acetate, the largest predicted concentration out of the mixtures corresponds to mixture E. For two other mixtures void of (*R*)-(–)-linalyl acetate, however, rather large concentrations were predicted. This is likely a consequence of its low VCD intensity. Detecting such low contributions in a mixture spectrum will require large coefficients and the prediction quality will be more easily affected by noise. If training is performed with L1-regularisation instead of L2, invoking sparsity in the model, the main difference on model performance lies in that the largest predicted (1*S*)-(+)-3-carene concentration is obtained for mixture D, while the presence of trans-caryophyllene cannot be detected reliably. While the largest predicted concentrations for each terpene correctly reflects its presence in a mixture, the gap between predicted concentrations when the terpene is present or absent was small for (*S*)-(−)-citronellal, (*S*)-(+)-β-citronellene, (1*S*)-(+)-3-carene (for L1) and (*R*)-(−)-linalyl acetate. The VCD patterns arising from the carbonyl vibration are particularly sensitive to the molecular environment. In complex mixtures, a mixture spectrum could therefore deviate from the linear approximation for the mixture spectra. We trained the linear model again while omitting signals above 1500 cm^{-1}, but performance did not improve.

The IR spectra contain less noise and intensities cannot partially cancel each other, but they are more strongly correlated. The balance between these differences determines the performance of an IR-based model. We trained a L2-regularised model trained on *in silico* IR mixture spectra and assessed its performance on the IR spectra of mixtures A-F. The model was trained and validated on detecting the presence of chiral and achiral/racemic terpenes (i.e. cineole, isoborneol and isobornyl acetate). The model could not detect the presence of two terpenes: isomenthol and trans-caryophyllene (sesquiterpene). Also the gap between predicted concentrations for when a terpene is present or absent was small for α-pinene, 3-carene and β-citronellene (see Figure 6.3). The linear approach suggested in this work worked slightly better for IR than for VCD. A combination of the higher noise level, higher uncertainty on the baseline or the possibility of cancelling intensities is likely the reason for this.

The question now remained whether the linear model for a single terpene used only the marker bands or also leveraged the other regions in the spectra to improve its predictions. The coefficients of the L2 linear model are plotted for each terpene in Figures 6.24-6.25. To address this question, we investigated the coefficients from the linear model for a few selected terpenes. For VCD, the model clearly used marker bands to detect some terpenes: e.g. the positive band at 1290 cm^{-1} for (*R*)-(+)-pulegone, the positive band at 1149 cm^{-1} for

(*S*)-(–)-α-terpineol, the positive band at 1052 cm^{-1} for (*S*)-(–)-*endo*-borneol, the negative band at 1718 cm^{-1} for (*S*)-(–)-citronellal, and the negative band at 1738 cm^{-1} for (1*R*)-(+)-camphor were all heavily used by the respective linear models. The linear model used these marker bands but did not completely rely on them; for many terpenes, numerous non-zero coefficients were found to contribute to their detection. In IR, the model weighed the carbonyl region as important for more terpenes compared to VCD. The coefficients of the IR-based model for carvone provided a clear example of how non-marker bands supplement the marker bands for its detection. For carvone, a strong positive and negative coefficient was observed at 1660 and 1620 cm^{-1}, respectively. Carvone has a strong marker IR band at 1660 cm^{-1}, however so do myrtenal and verbenone. The IR spectra of myrtenal and verbenone both contain a smaller IR band at 1620 cm^{-1}, while carvone does not absorb at this frequency. Thus, the model leveraged the IR intensities at 1620 cm^{-1} that detected the false positives of myrtenal and verbenone for detecting carvone with the 1660 cm^{-1} marker band. For pulegone, the most intense IR band at 1677 cm^{-1} was mainly ignored as multiple terpenes absorb at a similar frequency. The 1614 cm^{-1} band is notably broad, with the 1610-1560 cm^{-1} section of the band overlapping only partially with fenchone. The model leveraged all intensities between 1610 and 1560 cm^{-1} to detect pulegone along with the 1210 and relatively isolated 1288 cm^{-1} bands.

Next, we tested the performance of the models on the mixtures of pinane type, menthane type 1, menthane type 2, bornane type and fenchane type (Figures 6.26-6.29). For each of these mixtures, we bundled its predictions with those for mixtures A-F and observed whether the presence of a specific terpene was still linked with a higher predicted concentration. The terpenes for which mismatches between predicted concentrations and their presence were already obtained with A-F will not be discussed, but will still be highlighted in the figures.

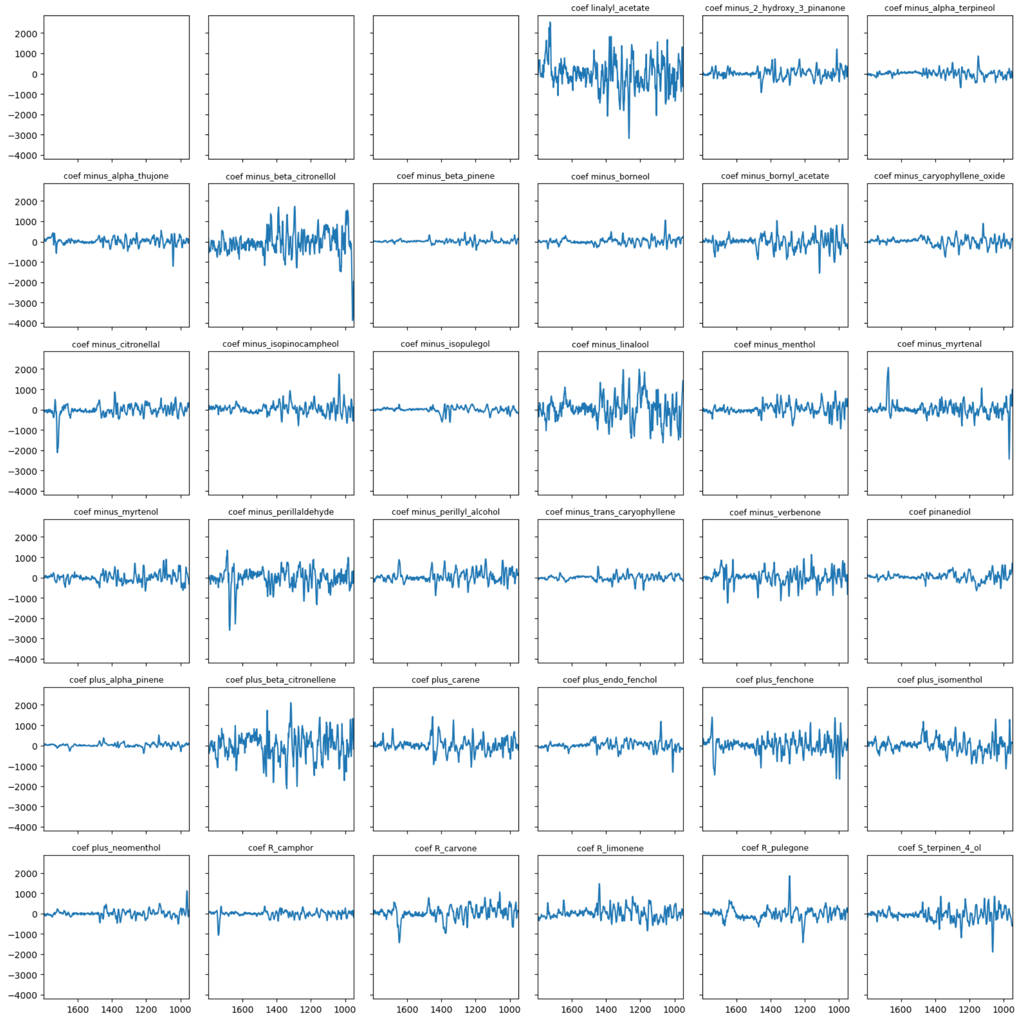


Figure 6.24: Coefficients for the L2-regularised VCD model for each chiral monoterpene

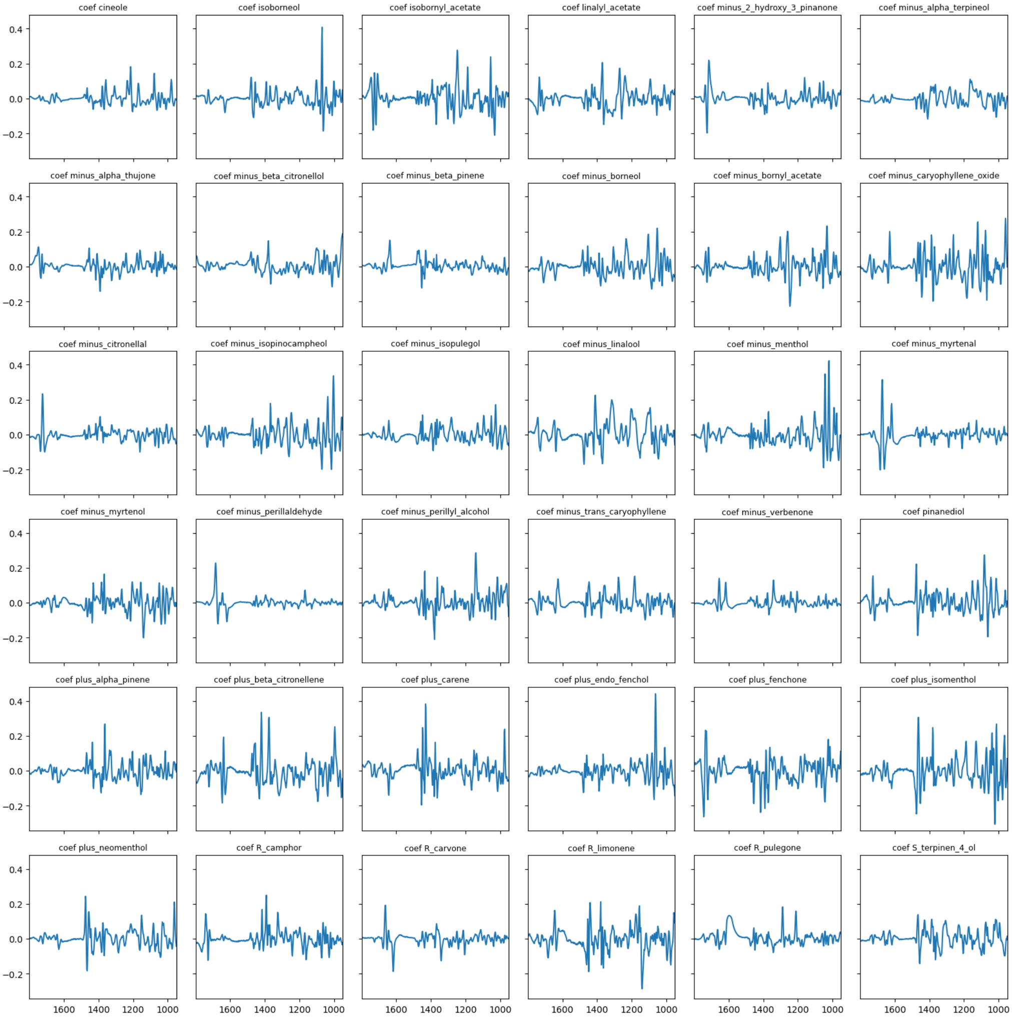


Figure 6.25: Coefficients for the L2-regularised IR model for each monoterpene.

Mixture	Terpene content
A	(1*S*,4*R*)-(−)-α-thujone, (1*R*,2*R*,3*R*,5*S*)-(−)-isopinocampheol, (*R*)-(+)-α-pinene, (1*R*)-(+)-*endo*-fenchyl alcohol, (1*S*,2*S*,5*R*)-(+)-neomenthol, (1*S*)-(−)-camphor, (1*R*)-(−)-myrtenyl acetate (not present in *in silico* mixtures)
B	(*S*)-(−)-β-pinene, (*S*)-(−)-*endo*-borneol, (1*R*,2*S*,5*R*)-(−)-isopulegol, (*R*)-(+)-α-pinene, (*S*)-(+)-fenchone, (*R*)-(+)-limonene, (*R*)-(+)-pulegone
C	(1*S*,2*S*,5*S*)-(−)-2-hydroxy-3-pinanone, (1*R*,2*S*,5*R*)-(−)-isopulegol, (1*R*,2*S*,5*R*)-(−)-menthol, (1*R*)-(−)-myrtenal, (*S*)-(−)-perillyl alcohol, (1*R*,2*R*,3*S*,5*R*)-(−)-pinanediol, (1*R*)-(+)-camphor, (*R*)-(−)-carvone
D	Cineole, (*S*)-(−)-α-terpineol, (1*R*)-(−)-myrtenol, (*S*)-(−)-perillaldehyde, (1*S*)-(−)-verbenone, (1*S*)-(+)-3-carene, (1*S*,2*R*,5*R*)-(+)-isomenthol
E	(*R*)-(−)-linalyl-acetate, (*S*)-(−)-β-citronellol, (*S*)-(−)-citronellal, (*R*)-(−)-linalool, (*S*)-(+)-β-citronellene
F	(1*S*,2*S*,5*S*)-(−)-2-hydroxy-3-pinanone, (*S*)-(−)-citronellal, (1*R*,2*S*,5*R*)-(−)-isopulegol, (*S*)-(−)-perillyl-alcohol, (−)-trans-caryophyllene, (1*S*)-(−)-verbenone, (1*S*)-(+)-3-carene, (*R*)-(−)-carvone
H	(1*S*,2*S*,5*S*)-(−)-2-hydroxy-3-pinanone, (*S*)-(−)-α-terpineol, (*S*)-(−)-β-citronellol, (*S*)-(−)-β-pinene, (*R*)-(−)-linalool, (1*R*)-(−)-myrtenal, (1*R*)-(−)-myrtenol, (*S*)-(−)-perillaldehyde, (*S*)-(−)-perillyl-alcohol, (*R*)-(+)-α-pinene, (*S*)-(+)-β-citronellene, (1*S*)-(+)-carene, (*S*)-(+)-fenchone, (1*S*,2*S*,5*R*)-(+)-neomenthol, (*R*)-(−)-carvone, (*R*)-(+)-limonene, (*R*)-(+)-pulegone, cineole
I	(±)-isobornyl-acetate, (*R*)-(−)-linalyl-acetate, (1*S*,2*S*,5*S*)-(−)-2-hydroxy-3-pinanone, (*S*)-(−)-α-terpineol, (*S*)-(−)-β-citronellol, (*S*)-(−)-β-pinene, (*S*)-(−)-*endo*-bornyl acetate, (*R*)-(−)-linalool, (1*R*)-(−)-myrtenal, (1*R*)-(−)-myrtenol, (*S*)-(−)-perillaldehyde, (*S*)-(−)-perillyl-alcohol, (*R*)-(+)-α-pinene, (*S*)-(+)-β-citronellene, (1*S*)-(+)-3-carene, (*S*)-(+)-fenchone, (1*S*,2*S*,5*R*)-(+)-neomenthol, (*R*)-(−)-carvone, (*R*)-(+)-limonene, (*R*)-(+)-pulegone, (*R*)-(−)-terpinen-4-ol, cineole
J	Cineole, (*S*)-(−)-α-terpineol, (*S*)-(−)-β-pinene, (*S*)-(−)-*endo*-borneol, (*S*)-(−)-bornyl acetate, (1*R*,2*S*,5*R*)-(−)-isopulegol, (1*R*)-(−)-myrtenal, (*S*)-(−)-perillaldehyde, (*S*)-(−)-perillyl alcohol, (*R*)-(+)-α-pinene, (*S*)-(+)-fenchone, (1*S*,2*R*,5*R*)-(+)-isomenthol, (1*S*,2*S*,5*R*)-(+)-neomenthol, (*R*)-(−)-carvone, (*R*)-(+)-limonene, (*R*)-(+)-pulegone

Table 6.1: Content of the experimental mixtures added for evaluation of the linear models.

For the mixture of pinane derivatives, the presence of (S)-(−)-citronellal and (−)-trans-caryophyllene was wrongly predicted with the VCD model. The IR model wrongly detected pinanediol and limonene. For the menthane 1 mixture, both models could not properly detect (R)-(−)-terpinen-4-ol. The IR spectrum of the menthane 2 mixture allowed to identify all terpenes. The VCD spectrum identified all terpenes present, but a mismatch was obtained for (R)-(−)-carvone and (S)-(−)-citronellal. Interestingly, large positive concentrations were predicted on both spectra for (1S,2R,5R)-(+)-isomenthol for which mismatches were obtained on A-F. On the bornane type mixture no additional mismatches were noted for IR and a single wrong detection for (R)-(−)-linalyl acetate in VCD was noted. For the mixture of fenchone and fenchol, a single wrong prediction was obtained for citronellal on the IR spectrum. The VCD-based model wrongly detected the presence of (R)-(−)-linalyl acetate, (S)-(−)-citronellal and (S)-(+)-β-citronellene. For this set of mixtures, each composed of structurally similar terpenes, the accuracy of the ML approach remained similar to the accuracy obtained for A-F, with on average 1 and 2 additional wrong detections for IR and VCD respectively. The models show clear potential for the analysis of terpene mixtures. Now the question remained how far the application area can be pushed. Therefore, we increased the complexity of the mixtures even further and tested whether the models could still identify the terpenes present.

The three mixtures of increased complexity (H-J) are composed of 16-22 terpenes each. The same methodology was repeated, combining the A-F predictions along with each of these three mixtures separately, and the obtained results are shown in Figures 6.30-6.31. The predictions for J remained fairly accurate, with 3 additional mismatches for (S)-(−)-*endo*-bornyl acetate, (S)-(−)-perillaldehyde and (R)-(+)-pulegone on the VCD spectrum and two mismatches for borneol and limonene on the IR spectrum. For mixture H, the accuracy of the model decreased with 7 additional mismatches for the IR and 4 for VCD. Similarly, a lower accuracy was obtained for mixture I with 9 additional mismatches for IR and 8 for VCD. For mixtures of such complexity, where each terpene provides only a tiny contribution to the mixture spectrum, accurately detecting the terpenes present becomes more challenging. Depending on the exact mixture composition, the models can still extract the presence of the monoterpenes as demonstrated with mixture J.

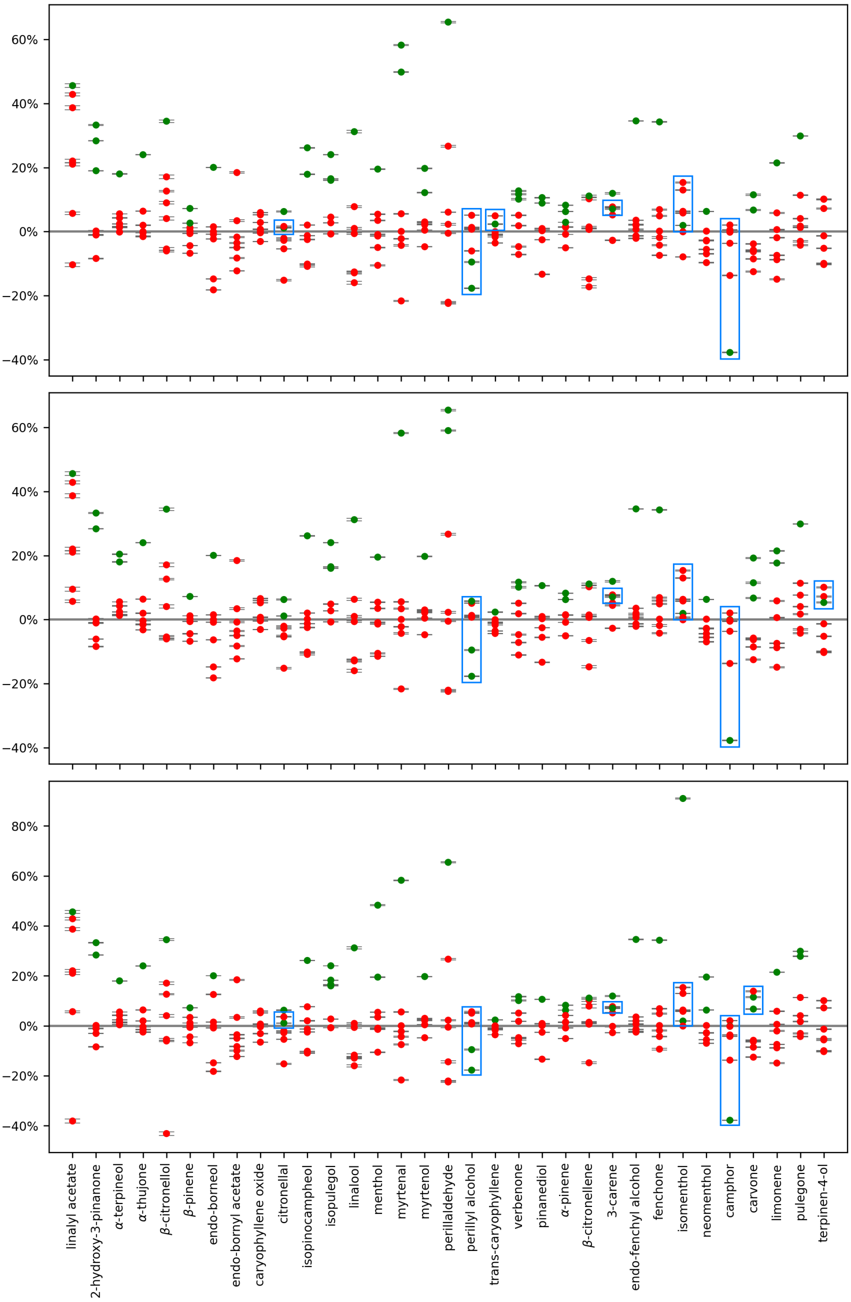


Figure 6.26: Predicted concentrations for the chiral terpenes by the VCD-based model on the combination of mixtures A-F with, from top to bottom, pinane type, menthane 1 type, menthane 2 type respectively. The predicted concentration is colored according to whether the terpene is present (green) or absent (red) for a mixture. The predicted concentrations are highlighted for terpenes when no correct decision boundary can be drawn.

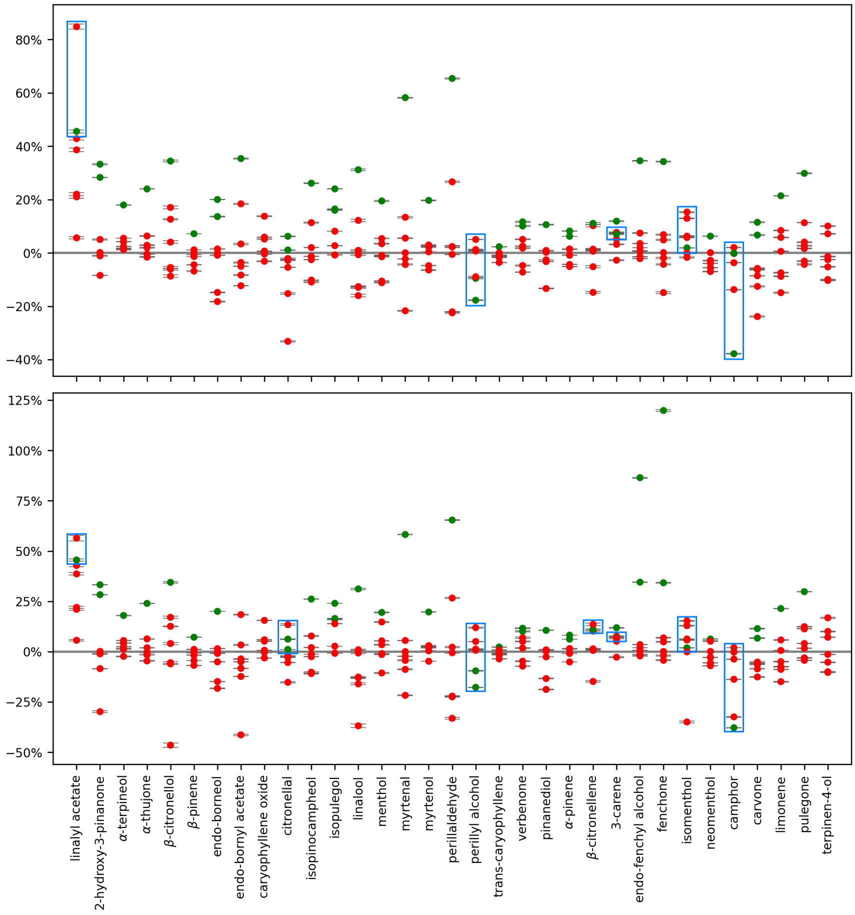


Figure 6.27: Predicted concentrations for the chiral terpenes by the VCD-based model on the combination of mixtures A-F with, from top to bottom, bornane type, fenchane type, respectively. The predicted concentration is colored according to whether the terpene is present (green) or absent (red) for a mixture. The predicted concentrations are highlighted for terpenes when no correct decision boundary can be drawn.

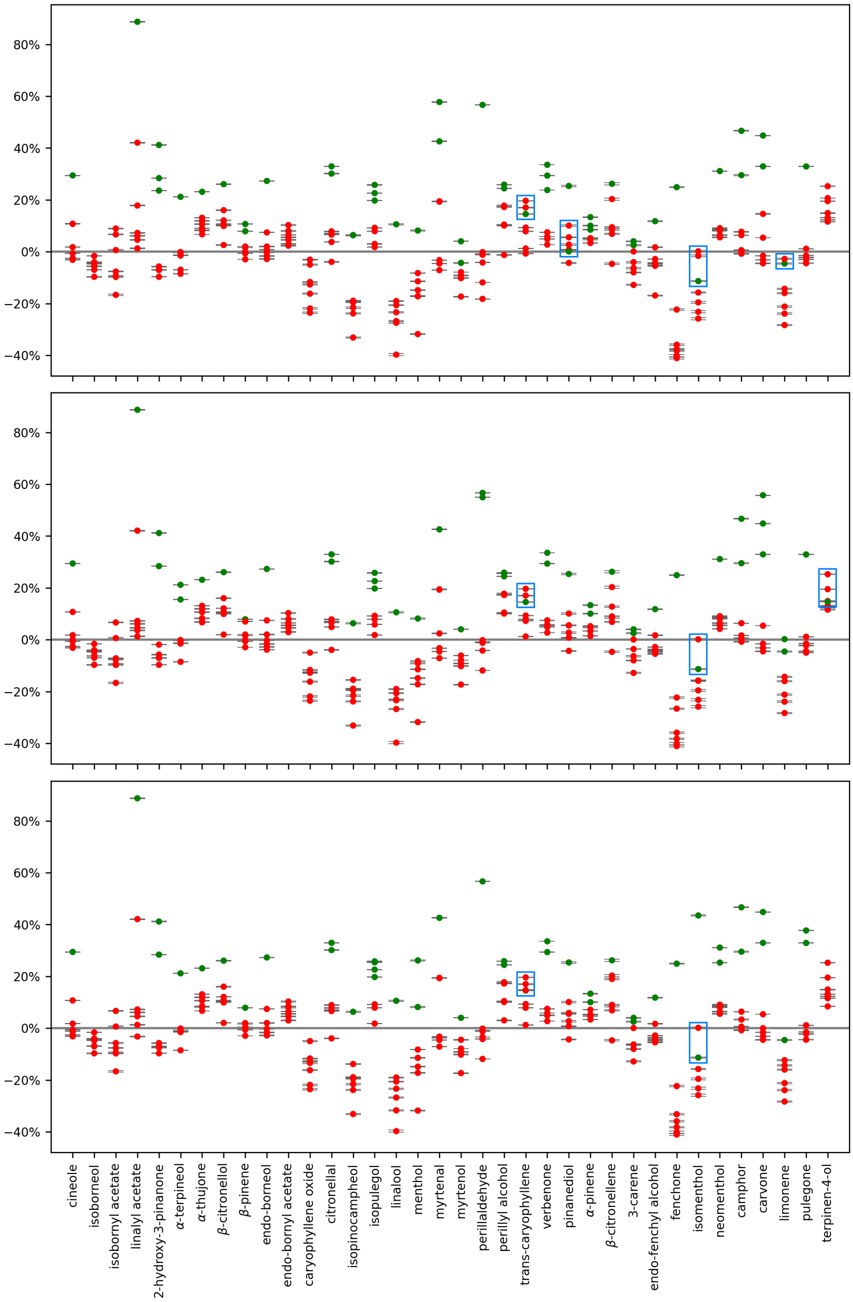


Figure 6.28: Predicted concentrations for the chiral terpenes by the IR-based model on the combination of mixtures A-F with, from top to bottom, pinane type, menthane 1 type, menthane 2 type respectively. The predicted concentration is colored according to whether the terpene is present (green) or absent (red) for a mixture. The predicted concentrations are highlighted for terpenes when no correct decision boundary can be drawn.

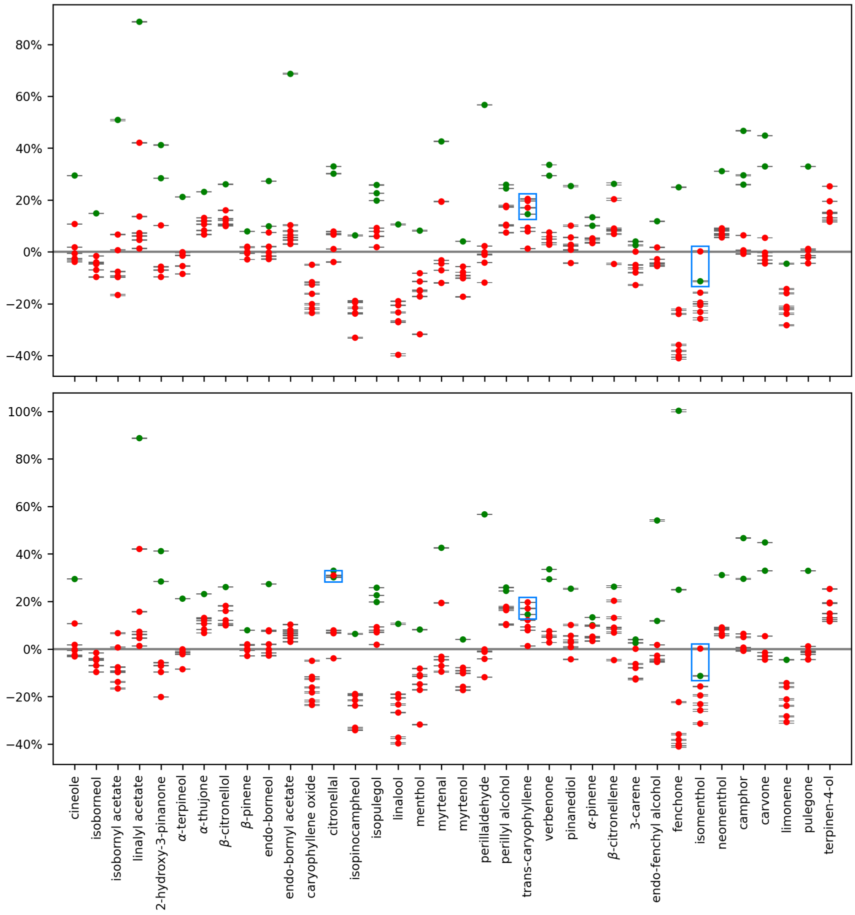


Figure 6.29: Predicted concentrations for the chiral terpenes by the IR-based model on the combination of mixtures A-F with, from top to bottom, bornane type, fenchane type, respectively. The predicted concentration is colored according to whether the terpene is present (green) or absent (red) for a mixture. The predicted concentrations are highlighted for terpenes when no correct decision boundary can be drawn.

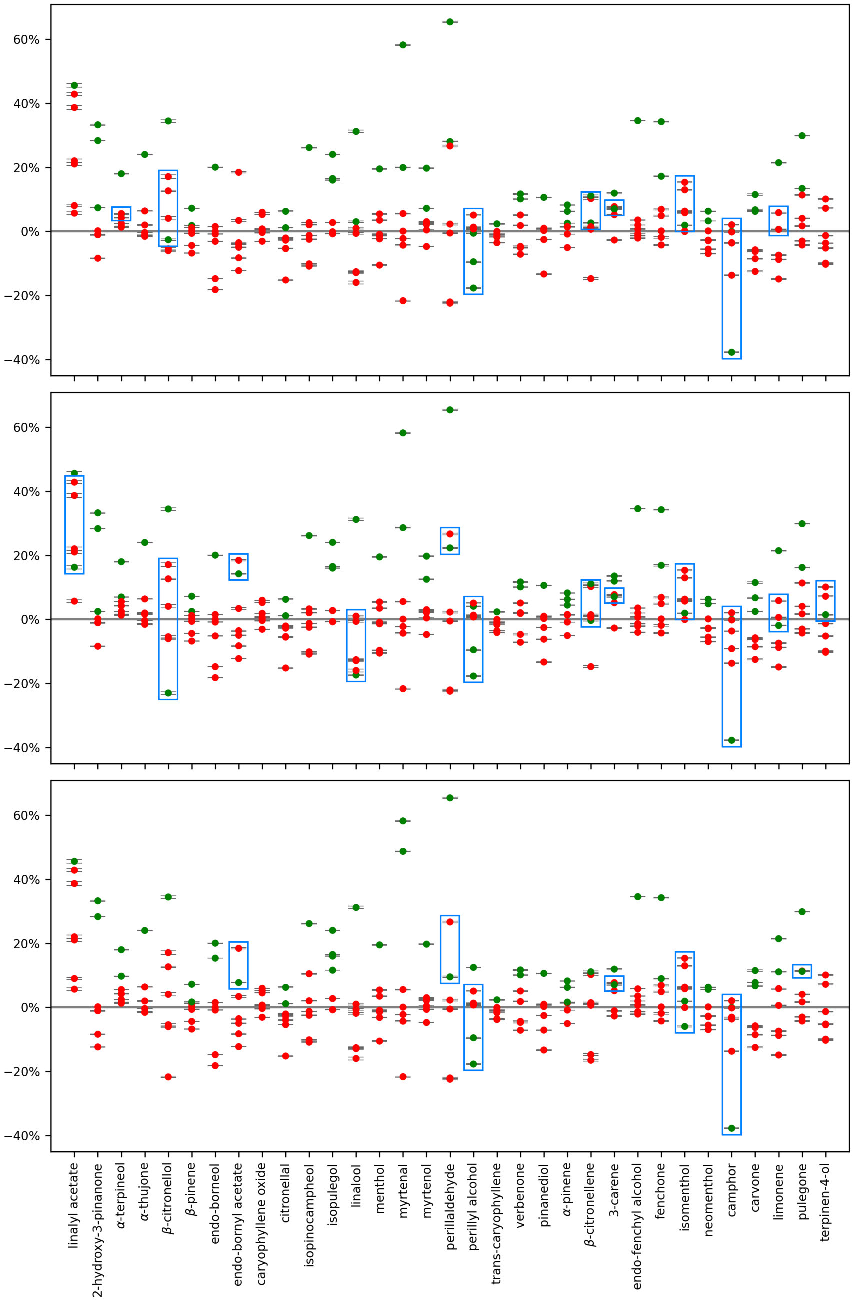


Figure 6.30: Predicted concentrations for the chiral terpenes on the combination of mixtures A-F with mixtures H (top), I (middle) and J (bottom) by the VCD-based model. The predicted concentration is colored according to whether the terpene is present (green) or absent (red) for a mixture. The predicted concentrations are highlighted for terpenes when no correct decision boundary can be drawn.

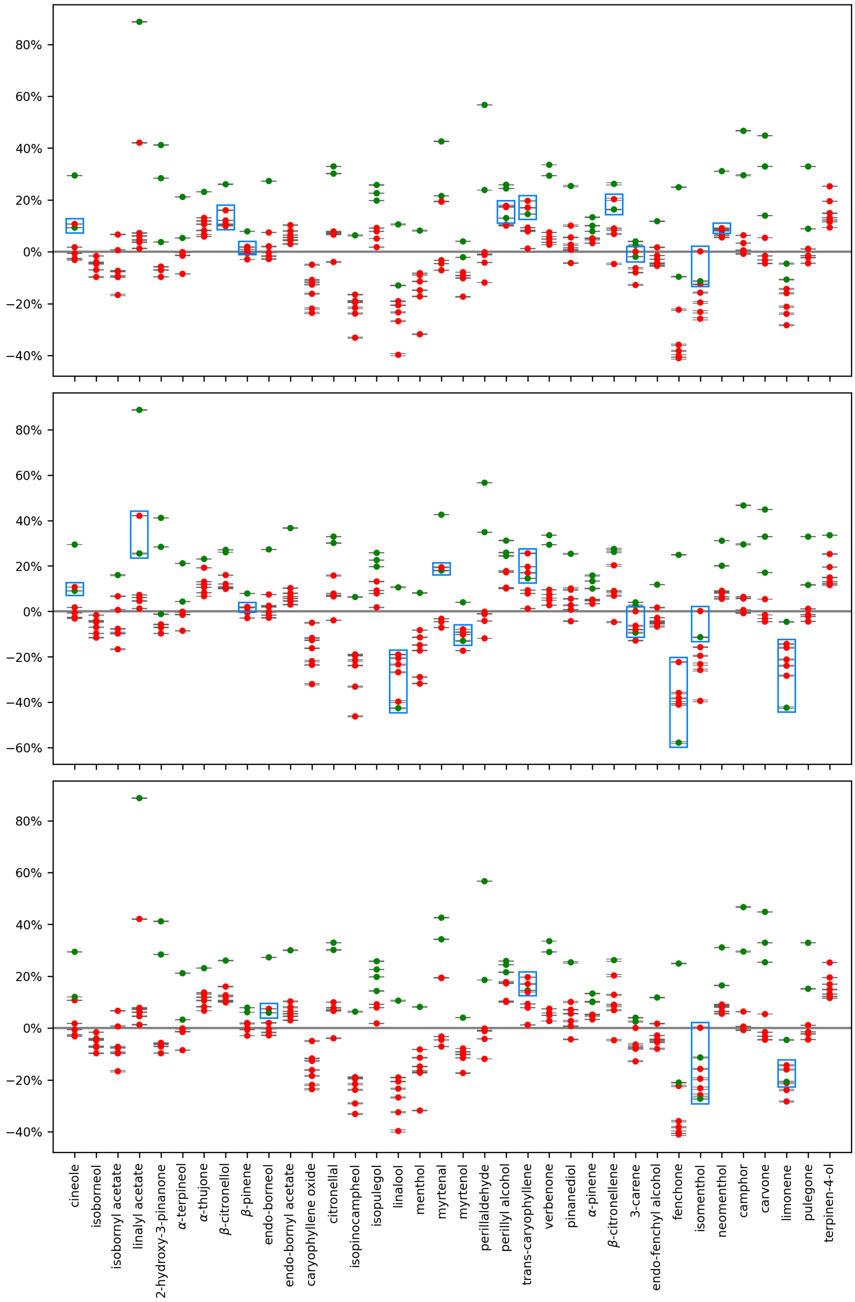


Figure 6.31: Predicted concentrations for the chiral terpenes on the combination of mixtures A-F with mixtures H (top), I (middle) and J (bottom) by the IR-based model. The predicted concentration is colored according to whether the terpene is present (green) or absent (red) for a mixture. The predicted concentrations are highlighted for terpenes when no correct decision boundary can be drawn.

Predictions by machine learning model on oils

Both models were then applied to the 4 essential oils using the decision boundaries fine-tuned with mixtures A-F. Predictions for the terpenes for which each model yielded unreliable predictions on the mixtures A-F were omitted. An overview of the true positives, false positives and false negatives is provided in Table 6.2 for the VCD model and Table 6.3 for the IR model.

For tea tree oil, the VCD model wrongly detected the presence of 3 terpenes (*S*)-(–)-citronellal, (1*R*,2*R*,3*S*,5*R*)-(–)-pinanediol and (*R*)-(+)-α-pinene and did not detect (*S*)-(+)-terpinen-4-ol or the tiny fraction of limonene and β-pinene. The IR model clearly detected terpinen-4-ol, along with limonene. However, the model detected the absent compounds β-citronellol, menthol, α-pinene and 3-carene. In rosemary oil the VCD model correctly identified the presence of (*R*)-(+)-α-pinene and (*S*)-(–)-β-pinene present in the oil, whereas the IR model detected α-pinene, β-pinene and α-terpineol present in the oil. Both models remained undecisive concerning the tiny fraction of (1*R*)-(+)-camphor present. Regarding lavender oil, the VCD model detected the presence of (*R*)-(–)-linalool, along with the tiny fraction of (*R*)-(+)-α-pinene, and the IR model identified both linalool and linalyl acetate. The VCD model also predicted the presence of (*S*)-(–)-citronellal, (1*R*,2*R*,3*S*,5*R*)-(–)-pinanediol, (*R*)-(–)-carvone and (1*S*,2*S*,5*R*)-(+)-neomenthol. The IR model generated a false positive for limonene, and isopinocampheol (and remains indecisive for menthol). Neither model detected the ±1% of terpinen-4-ol present in the oil. The tiny fraction of β-pinene was not detected by either model. The tiny fraction of α-pinene was barely detected by the VCD model but not by the IR model. The IR model also potentially detected the presence of isoborneol as the predicted concentration exceeds those for mixtures A-F (in which it was absent). For ylang-ylang oil, (*R*)-(–)-linalool was detected by both models. False positives were obtained for (*S*)-(–)-β-citronellal and (*R*)-(–)-carvone with the VCD model. The IR model wrongly predicted the presence of following compounds: borneol, isopinocampheol, isopulegol, carene, limonene, pulegone. A large concentration was also predicted for isomenthol, for which the presence could not be properly detected for mixtures A-F. While no decision boundaries could be drawn for isoborneol and isobornyl acetate (due to their absence from A-F), large relative concentrations were obtained for them.

Essential oil	True positives	False positives	False negatives
Tea tree oil	\	(*S*)-(−)-citronellal, (1*R*,2*R*,3*S*,5*R*)-(−)-pinanediol, (*R*)-(+)-α-pinene	(*S*)-(+)-terpinen-4-ol. Tiny fractions: limonene and β-pinene
Rosemary oil	(*R*)-(+)-α-pinene, (*S*)-(−)-β-pinene	(*R*)-(−)-linalyl acetate, (1*R*,2*R*,3*S*,5*R*)-(−)-pinanediol, (1*S*)-(+)-3-carene and (*R*)-(−)-carvone	Tiny fractions: α-terpineol
Lavender oil	(*R*)-(−)-linalool* Tiny fractions: (*R*)-(+)-α-pinene	(*S*)-(−)-β-citronellal, (1*R*,2*R*,3*S*,5*R*)-(−)-pinanediol, (*R*)-(−)-carvone, (1*S*,2*S*,5*R*)-(+)-neomenthol	terpinen-4-ol, (*S*)-(+)-linalyl acetate. Tiny fractions: β-pinene
Ylang-ylang oil	(*R*)-(−)-linalool	(*S*)-(−)-citronellal and (*R*)-(−)-carvone	\

Table 6.2: Accuracy of the predictions on the essential oils by the VCD model. Results conflicting in the chirality of terpene with visual inspection are indicated with an asterisk.

Essential oil	True positives	False positives	False negatives
Tea tree oil	terpinen-4-ol. Tiny fractions: limonene	β-citronellol, menthol, α-pinene, carene	Tiny fractions: β-pinene
Rosemary oil	α-pinene, cineole, β-pinene.	α-thujone, perillyl alcohol, β-citronellol, camphor.	Isoborneol. Tiny fractions: α-terpineol.
Lavender oil	linalool, linalyl acetate.	limonene, isopinocampheol	terpinen-4-ol. Tiny fractions: α-pinene and β-pinene.
Ylang-ylang oil	linalool	borneol, isopinocampheol, isopulegol, carene, limonene, pulegone	\

Table 6.3: Accuracy of the predictions on the essential oils by the IR model.

References

[1] D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2020, **83**, 770–803.

[2] A. Mándi and T. Kurtán, *Nat. Prod. Rep.*, 2019, **36**, 889–918.

[3] R. Jwad, D. Weissberger and L. Hunter, *Chem. Rev.*, 2020, **120**, 9743–9789.

[4] *US Food & Drug Administration: Development of New Stereoisomeric Drugs*, `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/development-new-stereoisomeric-drugs`, accessed on 24 Oct 2022.

[5] J. M. Finefield, D. H. Sherman, M. Kreitman and R. M. Williams, *Angew. Chem. Int. Ed.*, 2012, **51**, 4802–4836.

[6] A. N. L. Batista, F. M. dos Santos, J. a. M. Batista and Q. B. Cass, *Molecules*, 2018, **23**, 492.

[7] A. J. E. Novak and D. Trauner, *Trends Chem.*, 2020, **2**, 1052–1065.

[8] G. T. M. Bitchagno, V.-A. Nchiozem-Ngnitedem, D. Melchert and S. A. Fobofou, *Nat. Rev. Chem.*, 2022, **6**, 806–822.

[9] A. Batista, B. Angrisani, M. E. Lima, S. da Silva, V. Schettini, H. Chagas, F. dos Santos Jr., J. Batista and A. Valverde, *J. Braz. Chem. Soc.*, 2021, **32**, 1499–1518.

[10] S. G. Allenmark, *Nat. Prod. Rep.*, 2000, **17**, 145–155.

[11] J. a. M. Batista Jr., E. W. Blanch and V. d. S. Bolzani, *Nat. Prod. Rep.*, 2015, **32**, 1280–1302.

[12] P. L. Polavarapu and E. Santoro, *Nat. Prod. Rep.*, 2020, **37**, 1661–1699.

[13] G. Pescitelli and T. Bruhn, *Chirality*, 2016, **28**, 466–474.

[14] P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.

[15] T. B. Freedman, X. Cao, R. K. Dukor and L. A. Nafie, *Chirality*, 2003, **15**, 743–758.

[16] L. Nafie, *Vibrational Optical Activity: Principles and Applications*, Wiley, Hoboken, NJ, 2011.

[17] P. Stephens, F. Devlin and J. Cheeseman, *VCD Spectroscopy for Organic Chemists*, CRC Press, Boca Raton, FL, 2012.

[18] C. Merten, T. P. Golub and N. M. Kreienborg, *J. Org. Chem.*, 2019, **84**, 8797–8814.

[19] L. A. Nafie, *Chirality*, 2020, **32**, 667–692.

[20] J. Bogaerts, R. Aerts, T. Vermeyen, C. Johannessen, W. Herrebout and J. M. Batista, *Pharmaceuticals*, 2021, **14**, 877.

[21] P. Dewick, *Medicinal Natural Products: A Biosynthetic Approach*, Wiley, Chichester, United Kingdom, 2009.

[22] Debie, Elke and Jaspers, L and Bultinck, Patrick and Herrebout, W and Van der Veken, Bert, *Chem. Phys. Lett.*, 2008, **450**, 426–430.

[23] E. Debie, P. Bultinck, W. Herrebout and B. van der Veken, *Phys. Chem. Chem. Phys.*, 2008, **10**, 3498–3508.

[24] V. P. Nicu, E. Debie, W. Herrebout, B. Van der Veken, P. Bultinck and E. J. Baerends, *Chirality*, 2010, **21**, E287–E297.

[25] M. A. J. Koenis, Y. Xia, S. R. Domingos, L. Visscher, W. J. Buma and V. P. Nicu, *Chem. Sci.*, 2019, **10**, 7680–7689.

[26] L. A. Nafie, T. A. Keiderling and P. J. Stephens, *J. Am. Chem. Soc.*, 1976, **98**, 2715–2723.

[27] L. A. Nafie, M. Diem and D. W. Vidrine, *J. Am. Chem. Soc.*, 1979, **101**, 496–498.

[28] L. A. Nafie, *Appl. Spectrosc.*, 2000, **54**, 1634–1645.

[29] C. Guo, R. D. Shah, R. K. Dukor, X. Cao, T. B. Freedman and L. A. Nafie, *Anal. Chem.*, 2004, **76**, 6956–6966.

[30] G. Longhi, R. Gangemi, F. Lebon, E. Castiglioni, S. Abbate, V. M. Pultz and D. A. Lightner, *J. Phys. Chem. A*, 2004, **108**, 5338–5352.

[31] X. Lu, H. Li, J. W. Nafie, T. Pazderka, M. Pazderková, R. K. Dukor and L. A. Nafie, *Appl. Spectrosc.*, 2017, **71**, 1117–1126.

[32] E. B. no Tapia, L. G. Zepeda and P. Joseph-Nathan, *Phytochemistry*, 2010, **71**, 1158–1161.

[33] J. C. Pardo-Novoa, H. M. Arreaga-González, M. A. Gómez-Hurtado, G. Rodríguez-García, C. M. Cerda-García-Rojas, P. Joseph-Nathan and R. E. del Río, *J. Nat. Prod.*, 2016, **79**, 2570–2579.

[34] M. E.-A. Said, I. Bombarda, J.-V. Naubron, P. Vanloot, M. Jean, A. Cheriti, N. Dupuy and C. Roussel, *Chirality*, 2017, **29**, 70–79.

[35] R.-Q. Gao, J. Fan, Q. Tan, D. Guo, T. Chen, R.-J. He, D. Li, H. Zhang and W.-G. Zhang, *Chirality*, 2017, **29**, 550–557.

[36] L. F. Julio, E. B. no Tapia, C. E. Díaz, N. Pérez-Hernández, A. González-Coloma and

P. Joseph-Nathan, *Chirality*, 2017, **29**, 716–725.

[37] M. E.-A. Said, P. Vanloot, I. Bombarda, J.-V. Naubron, E. M. Dahmane, A. Aamouche, M. Jean, N. Vanthuyne, N. Dupuy and C. Roussel, *Anal. Chim. Acta*, 2016, **903**, 121–130.

[38] C. Guo, R. D. Shah, R. K. Dukor, T. B. Freedman, X. Cao and L. A. Nafie, *Vibr. Spectrosc.*, 2006, **42**, 254–272.

[39] F. J. Devlin, P. J. Stephens and P. Besse, *J. Org. Chem.*, 2005, **70**, 2980–2993.

[40] F. Passareli, A. N. L. Batista, A. J. Cavalheiro, W. A. Herrebout and J. M. Batista Junior, *Phys. Chem. Chem. Phys.*, 2016, **18**, 30903–30906.

[41] A. Nakahashi, A. K. C. Siddegowda, M. A. S. Hammam, S. G. B. Gowda, Y. Murai and K. Monde, *Org. Lett.*, 2016, **18**, 2327–2330.

[42] T. Taniguchi, T. Suzuki, H. Satoh, Y. Shichibu, K. Konishi and K. Monde, *J. Am. Chem. Soc.*, 2018, **140**, 15577–15581.

[43] C. Grassin, E. Santoro and C. Merten, *Chem. Commun.*, 2022, **58**, 11527–11530.

[44] M. Z. M. Zubir, N. F. Maulida, Y. Abe, Y. Nakamura, M. Abdelrasoul, T. Taniguchi and K. Monde, *Org. Biomol. Chem.*, 2022, **20**, 1067–1072.

[45] J. a. M. Batista Jr., A. N. L. Batista, J. S. Mota, Q. B. Cass, M. J. Kato, V. S. Bolzani, T. B. Freedman, S. N. López, M. Furlan and L. A. Nafie, *J. Org. Chem.*, 2011, **76**, 2603–2612.

[46] R. F. Sprenger, S. S. Thomasi, A. G. Ferreira, Q. B. Cass and J. M. Batista Junior, *Org. Biomol. Chem.*, 2016, **14**, 3369–3375.

[47] F. M. dos Santos Jr., K. U. Bicalho, I. H. Calisto, G. S. Scatena, J. B. Fernandes, Q. B. Cass and J. M. Batista Jr., *Org. Biomol. Chem.*, 2018, **16**, 4509–4516.

[48] H. Ortega, J. M. Batista, W. Melo, G. de Paula and M. Pupo, *J. Braz. Chem. Soc.*, 2019, **30**, 2672–2680.

[49] A. N. L. Batista, F. M. dos Santos, A. L. Valverde and J. M. Batista, *Org. Biomol. Chem.*, 2019, **17**, 9772–9777.

[50] M. A. S. Yokomichi, H. R. L. Silva, L. E. V. N. Brandao, E. F. Vicente and J. M. Batista Jr., *Org. Biomol. Chem.*, 2022, **20**, 1306–1314.

[51] L. Laux, V. Pultz, S. Abbate, H. A. Havel, J. Overend, A. Moscowitz and D. A. Lightner, *J. Am. Chem. Soc.*, 1982, **104**, 4276–4278.

[52] G. Holzwarth and I. Chabay, *J. Chem. Phys.*, 1972, **57**, 1632–1635.

[53] U. Narayanan and T. A. Keiderling, *J. Am. Chem. Soc.*, 1983, **105**, 6406–6411.

[54] S. S. Birke, I. Agbaje and M. Diem, *Biochemistry*, 1992, **31**, 450–455.

[55] T. Taniguchi and K. Monde, *J. Am. Chem. Soc.*, 2012, **134**, 3695–3698.

[56] F. M. Santos Jr., C. L. Covington, A. C. F. de Albuquerque, J. F. R. Lobo, R. M. Borges, M. B. de Amorim and P. L. Polavarapu, *J. Nat. Prod.*, 2015, **78**, 2617–2623.

[57] T. Vermeyen, J. Brence, R. Van Echelpoel, R. Aerts, G. Acke, P. Bultinck and W. Herrebout, *Phys. Chem. Chem. Phys.*, 2021, **23**, 19781–19789.

[58] J. a. M. J. Batista, A. N. L. Batista, J. S. Mota, Q. B. Cass, M. J. Kato, V. S. Bolzani, T. B. Freedman, S. N. López, M. Furlan and L. A. Nafie, *The Journal of Organic Chemistry*, 2011, **76**, 2603–2612.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

[60] L. Breiman, *Machine Learning*, 1996, **24**, 123–140.

Chapter 7

Summary

Many natural products and pharmaceutical compounds are chiral and hence their mirror images (enantiomers) interact differently with a chiral environment. In organisms, important chemical receptors built from amino acids and sugars are also chiral and, as a result, enantiomers of the aforementioned compounds can lead to different biological activities in the human body. Thus, determining the so-called Absolute Configuration (AC) of a compound is important for research areas like drug development and agrochemistry. The AC of a chiral compound can be identified by its interaction with chiral fields, including circularly polarized light. A chiral compound will interact differently with left- and right- handed circularly polarized light. In Vibrational Circular Dichroism (VCD), the difference in absorption of both forms of circularly polarized Infrared Radiation (IR) by a chiral compound is recorded. With the molecular vibrations as chromophores, VCD combines chiral sensitivity with the wealth of conformational information of IR spectroscopy. VCD has established itself as a reliable tool to distinguish enantiomers and other stereoisomers. Unfortunately, there are no general empirical rules to link a VCD spectrum to a specific AC. Typically, Density Functional Theory (DFT) is used to predict the VCD spectrum for each possible relative configuration after which comparing the DFT spectra with the experimental spectrum allows to establish the AC of a compound. As DFT calculations need to be performed for each conformer of a specific relative configuration, the computational cost of AC determination increases with the compound's conformational flexibility. The work presented in this thesis explores the added value of super-

vised Machine Learning (ML) for the existing AC determination workflow.

As the combination of supervised ML methods with VCD has not yet been explored, their compatibility is completely unknown. The potential of this combination is investigated by addressing the following questions: Are ML models able to extract the AC from a spectrum directly? Is it possible for an ML model to predict DFT conformer spectra? Can an ML model retrieve the composition of mixtures from their VCD spectra?

The first question focuses on the central application of VCD applications on small molecules: AC determination. The chirality of a compound is encapsulated in its VCD spectrum in a rather opaque way. An ML model could extract the chirality from its VCD spectrum when trained on a series of structurally similar compounds. To answer the first question, a VCD dataset of ± 4k enantiomer pairs sharing an α-pinene core structure is generated. A researcher could extract the AC from this dataset without ML techniques with an accuracy up to 75-80% using optimised empirical rules, whereas a Feedforward Neural Network (FNN) can do so up to 99.5% accuracy. While AC extraction is more difficult for a Random Forest (RF) model (up to 94.5% accuracy) than an FNN, RF enables to identify the spectral areas containing the crucial chiral information. Thus, ML models are clearly able to extract the AC directly from the VCD spectrum. Once more spectral databases are established, AC determination of particular molecular classes could be performed without DFT.

The second question relates to the high sensitivity of VCD towards the conformations a single compound can adapt. This sensitivity enables researchers to study the conformational population of chiral compounds in solution. The impact of the conformation on the VCD spectrum is, however, not easily established for small flexible molecules. The VCD spectra of conformers are therefore computed with DFT and averaged according to their Boltzmann weights. For compounds with a high degree of conformational flexibility, the computational cost of VCD applications increases significantly. This computational cost could be decreased if the link between conformer and spectrum can be extracted with an ML model. For a set of congener compounds, an FNN is trained on a subset of conformers and its ability to predict the spectra of the remaining conformers is gauged. The FNN predicted spectra match the DFT conformer spectra to a very high extent when ample conformers are provided to train on. When stronger intramolecular interactions occur, such as intramolecular hydrogen bonds, the FNN predicted spectra are less accurate, especially if the molecular representation used to encode the conformations fails to reflect them properly. Nonetheless, with the FNN workflow the molecular VCD spectrum is obtained at a fraction of the cost without sacrificing its accuracy.

With both questions answered, a final question is raised: Can an ML model discern the different compounds present in a mixture based on its VCD and/or IR spectrum? This question does not tackle an existing application area of VCD like the previous two, instead, it introduces a previously unexplored application. As a proof of concept, a linear ML model was trained using a dataset of monoterpenes to detect their presence in mixtures. For the monoterpenes mixtures, the model successfully identifies the monoterpenes present. Predicting the content of more complex mixtures does become more challenging for the model. For natural oils, the model is able to detect from the IR spectrum most of the monoterpenes present. Nonetheless, the current method results in too many false positives, hindering its practical applications for natural oils.

In summary, this thesis demonstrates the value of combining VCD spectroscopy with ML methods. By tackling these questions, a foundation is provided to the community to build future applications of ML in VCD. Finally, it is the hope of the author that this work will motivate the construction of a general VCD database by the community. Increased availability of spectral data will increase the scope and possibilities of ML applications within the field likewise.

Chapter 8

Samenvatting

Veel natuurlijke producten en farmaceutische verbindingen zijn chiraal en hun spiegelbeelden (enantiomeren) interageren bijgevolg anders met een chirale omgeving. Belangrijke chemische bouwstenen van receptoren in organismen - aminozuren en suikers - zijn ook chiraal en als gevolg daarvan kunnen enantiomeren een verschillende biologische activiteit vertonen in het menselijk lichaam. Daarom is het bepalen van de zogenaamde Absolute Configuratie (AC) van een verbinding belangrijk voor onderzoeksgebieden zoals geneesmiddelen en agrochemie. De AC van een verbinding kan worden geïdentificeerd op basis van de interactie met chirale velden zoals circulair gepolariseerde licht. Een chirale verbinding zal anders interageren met de links- en rechtshandige vormen van het circulair gepolariseerde licht. Bij Vibrationeel Circulair Dichroïsme (VCD) wordt het verschil in absorptie van InfraRood (IR) straling tussen beide vormen van polarisatie door een chirale verbinding geregistreerd. Met de moleculaire vibraties als chromoforen, combineert VCD de chirale gevoeligheid met de overvloed aan conformationele informatie van IR-spectroscopie. VCD is uitgegroeid tot een betrouwbare analysetechniek om enantiomeren (en andere stereoisomeren) te onderscheiden. Helaas zijn er geen algemene empirische regels om een VCD-spectrum te koppelen aan een specifieke AC. Doorgaans wordt de DensiteitsFunctionaalTheorie (DFT) gebruikt om het VCD-spectrum te voorspellen voor elke mogelijke relatieve configuratie. Door deze DFT-spectra te vergelijken met het experimentele spectrum kan de AC van een verbinding correct worden bepaald. Aangezien DFT-berekeningen moeten worden uitgevoerd voor elk conformeer van een specifieke relatieve configuratie, neemt de benodigde computerkracht voor AC-bepaling toe met de conformationele flexibiliteit van de verbinding. Het werk dat in deze thesis wordt gepresenteerd, onderzoekt de toegevoegde waarde van Machine Learning (ML) voor de bestaande workflow voor AC-bepaling.

Voor het begin van dit onderzoek was er nog geen onderzoek gepubliceerd over het gebruik van ML voor VCD, waardoor hun compatibiliteit volledig onbekend was. Het potentieel van deze combinatie wordt onderzocht met behulp van de volgende vragen: Zijn ML-modellen in staat om de AC uit een VCD spectrum te halen? Is een ML-model in staat om DFT-spectra van conformeren te voorspellen? Kan een ML-model de samenstelling van mengsels vaststellen op basis van hun VCD-spectra?

De eerste vraag richt zich op de hoofdtoepassing van VCD voor kleine moleculen: de bepaling van de Absolute Configuratie (AC). De chiraliteit van een verbinding zit vervat in het VCD-spectrum, maar deze informatie kan alleen via complementaire kwantumchemische berekeningen geëxtraheerd worden. Een ML-model zou de chiraliteit uit het VCD-spectrum kunnen halen na training op een reeks structureel vergelijkbare verbindingen. Om de eerste vraag te beantwoorden, wordt een VCD-dataset van ongeveer vierduizend enantiomeerparen met eenzelfde kernstructuur gegenereerd. Een onderzoeker zou de AC uit deze dataset kunnen halen, zonder gebruik van ML-technieken, met een nauwkeurigheid van maximaal 75-80% met behulp van geoptimaliseerde empirische regels. Echter, een Feedforward Neural Network (FNN) kan de AC uit het spectrum halen met een nauwkeurigheid van maximaal 99,5%. Hoewel AC-extractie moeilijker is voor een Random Forest (RF) model (tot 94,5% nauwkeurigheid) dan voor een FNN, laat een RF toe om de spectrale gebieden te identificeren die de meest karakteristieke chirale informatie bevatten. Deze resultaten bevestigen dat de ML-modellen de AC direct uit een VCD spectrum kunnen halen. Zodra er meer spectrale databases zijn opgebouwd, zou de bepaling van de AC van specifieke moleculaire klassen kunnen worden uitgevoerd zonder DFT.

De tweede vraag richt zich op de hoge gevoeligheid van VCD voor de conformeren die een molecule kan aannemen. Deze gevoeligheid stelt onderzoekers in staat om de conformationele eigenschappen van chirale verbindingen in oplossing te bestuderen. De invloed van de conformatie op het VCD-spectrum is echter niet gemakkelijk vast te stellen voor kleine flexibele moleculen. De VCD-spectra van conformeren worden daarom doorgaans berekend met behulp van DFT en het moleculair spectrum is dan het gewogen gemiddelde van de conformeer spectra (met de Boltzmannfactoren als gewicht). Voor verbindingen met een hoge flexibiliteit nemen de rekenkosten van VCD-toepassingen aanzienlijk toe. Deze rekenkosten verminderen indien een ML-model de link tussen conformeer en spectrum kan leren. Om dit te testen, wordt een FNN getraind op een deel van de conformeren en wordt beoordeeld in hoeverre het model in staat is om de spectra van de overige conformeren te voorspellen. De voorspelde spectra komen in zeer grote mate overeen met de DFT-conformeerspectra wanneer voldoende conformeren worden aangeboden om het model te trainen. Wanneer sterkere intramoleculaire interacties (bv. intramoleculaire waterstofbruggen) optreden, zijn de voorspelde spectra minder accuraat, vooral als de gekozen moleculaire representatie deze interacties onvoldoende beschrijft. Niettemin kan men met de huidige FNN-workflow het moleculaire VCD-spectrum verkrijgen tegen een fractie van de oorspronkelijke rekenkosten, zonder de nauwkeurigheid negatief te beïnvloeden.

Tenslotte, wordt er nog laatste vraag gesteld die, in tegenstelling tot de vorige twee vragen, zich niet richt op een bestaand toepassingsgebied van VCD. In plaats daarvan behandelt het een toepassing die totnogtoe niet verkend is: Kunnen ML-modellen de moleculen aanwezig in een mengsel detecteren via het VCD- en/of IR-spectrum van het mengsel? Als proof-of-concept werd een ML-model getraind met behulp van een dataset van monoterpenen om hun aanwezigheid in mengsels te detecteren. Het model identificeert succesvol welke monoterpenen aanwezig zijn in de monoterpeenmengsels. Echter, bij complexere mengsels wordt het moeilijker voor het model om alle bestanddelen te identificeren. Voor natuurlijke oliën kan het model de meeste monoterpenen detecteren op basis van het IR-spectrum. Desondanks leidt de huidige methode tot te veel valse positieven, wat de praktische toepassingen ervan voor natuurlijke oliën belemmert.

Samengevat toont deze scriptie de meerwaarde van ML-methoden voor VCD-spectroscopie. Door deze vragen te beantwoorden, wordt een basis gelegd voor de wetenschappelijke gemeenschap om toekomstige toepassingen van ML in VCD te ontwikkelen. Tenslotte hoop ik dat dit werk de aanmaak van een algemene VCD-database door de gemeenschap zal stimuleren. Een grotere beschikbaarheid van spectrale gegevens zal de reikwijdte en mogelijkheden van ML-toepassingen binnen het vakgebied vergroten.

Chapter 9

Future perspectives

9.1 AC determination with supervised models

With the main application of VCD being AC determination, our first chosen project was to extract the AC from VCD spectra with ML methods. Here, we chose a well-defined application domain, being the substituted pinene structures, and tested the capabilities of the ML methods within this domain. Doing so, it was clear that ML methods identified the molecular chirality with impressive accuracy. The chirality of structurally similar compounds can clearly be extracted from VCD spectra much better with ML than with chiral fingerprints. The models barely needed optimization, no deep learning techniques were needed and the models did not require 2D structural data of the compounds. A possible avenue to explore is to test the current methodology on compounds with flexible structures and of increased diversity. Here, the generation of more VCD data for training ML models is essential. Doing so would broaden the application domain, encompassing more of the commonly found structural motifs in molecules. However, given the absence of a VCD dataset of considerable size covering chemically diverse compounds, developing a ‘general AC determination’ model is not possible at the moment. To develop such a model, a significant effort into sharing spectral data and assembling them into a coherent dataset is highly needed.

Another possible avenue to explore would be to adjust the current methodology to include predicting the AC of compounds with multiple chiral centres from a single 2D structure. Of course, given the dataset size requirement, a compound with a large number of stereoisomers is needed to generate enough training data. For this reason, we propose to test this idea on cholic acid. Cholic acid contains 11 chiral centers, as shown in Figure 9.1, which results in 2048 (2^{11}) possible stereoisomers. The compound is also rather flexible thanks to the hydroxyl moieties, acidic sidechain and possible ring flips. As a result, the computational cost of computing all DFT spectra is significant (the spectra were only obtained after running calculations half a year on the VSC infrastructure).

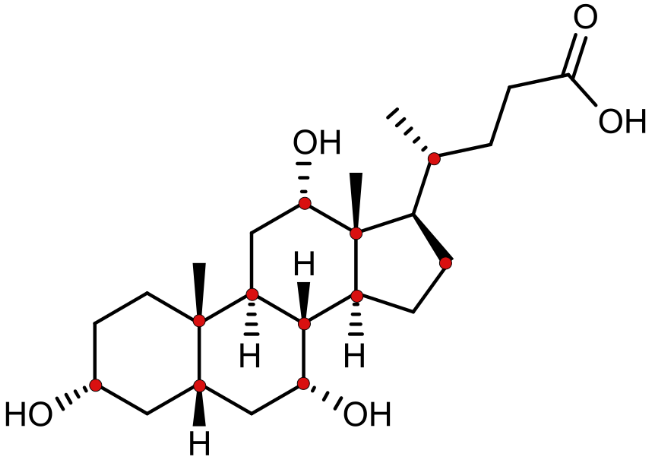


Figure 9.1: Cholic acid with chiral carbon atoms highlighted in red.

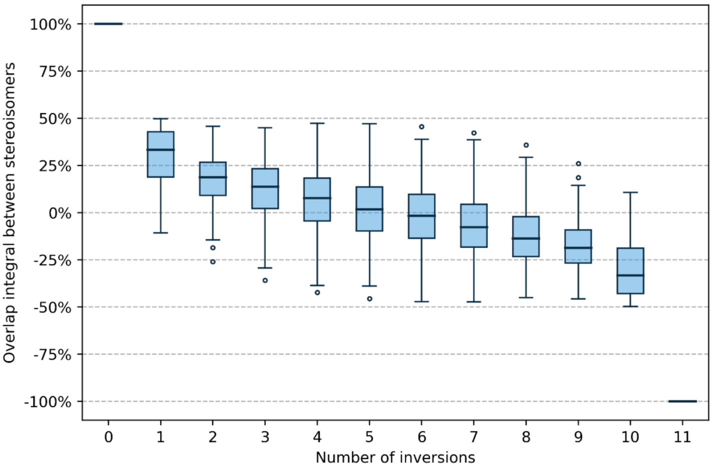


Figure 9.2: Influence of chiral centers' configurations on VCD spectrum.[a]

As part of a master thesis project, simple ML methods (excluding neural networks) were tested to predict the configuration of each of the chiral centers from the VCD spectrum. As expected, the ML task proved more complex than the pinene AC prediction. The spectra are very sensitive to the inversion of a single center (Figure 9.2), with epimer spectra having an median similarity of +- 0.33. These models could identify the AC of an individual chiral center with an accuracy of ± 90%, indicating the potential of this approach. However, the resulting full AC (i.e. the 11 chiral labels) of the compound was only correct for a third of the stereoisomers. For the full AC label, a single-label accuracy of 90% results in a full AC accuracy of $(90\%)^{11}$ (± 31 %) if the label predictions of different centers are assumed to be independent. The problem with this approach for many chiral centers lies in balancing the need to create sufficient data and the increasing complexity of AC determination with the number of chiral centers. Therefore, other models should be chosen where the predictions of different chiral centers are not independent of each other. In other words, the patterns learned by the model should include information on multiple chiral centers or model chains have to be used where the predicted AC of one chiral center is included as a feature for AC prediction for other chiral centers. Efforts to solve this problem with these ML methods with such a model chain, illustrated in Figure 9.3, did not improve the full AC accuracy. The introduction of multioutput neural networks would be a next step in solving this issue. The patterns extracted from the spectrum with these networks can be geared towards multiple chiral centers. Adding the IR spectra alongside the VCD spectra may also improve the ability of the models to discern epimers and diastereomers.

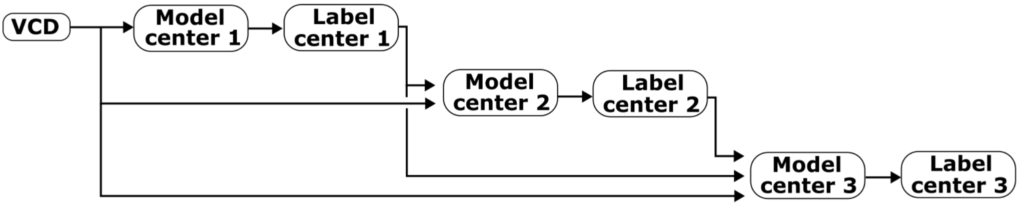


Figure 9.3: Illustration of a model chain for AC determination for 3 chiral centers.

[a] Figure taken from master thesis "Putting Machine Learning Applications for Vibrational Circular Dichroism to the Test.", J. Vanhove, University of Antwerp, 2022-2023.

9.2 Predicting conformer spectra

The aim of this project was to identify whether an FNN could learn the link between conformers and their VCD spectra. The choice of the six model compounds allowed to identify which sources may hinder such a workflow. From this, it was clear that the ML task became more difficult when the dihedral angles of neighbouring sidechains started interacting (i.e. become correlated). Therefore, it would be interesting to see how the workflow performs on macrocycles or linear molecules. For such compounds, the rotation of a dihedral angle influences the values the other dihedral angles can adopt. Preliminary results for the compounds in Figure 9.4, show that the differences between conformer spectra are even stronger here (lower $\overline{S^{conf}}$). This reinforces the proposed idea that the conformer spectra differ more strongly when the dihedral angles are less independent from each other. For (S)-3-ethoxynonane, the predicted conformer spectra are less accurate, though the resulting Boltzmann spectrum remains accurate (Figure 9.5). For (S,S)-1,2-dichlorocyclotetradecane, the ML model fails to learn the link between conformation and VCD spectrum. Therefore, more sophisticated models and molecular representations, describing their 3D geometries, have to be tested. Here, more attention should be paid to properly describing the weak intramolecular interactions that influence the molecular flexibility.

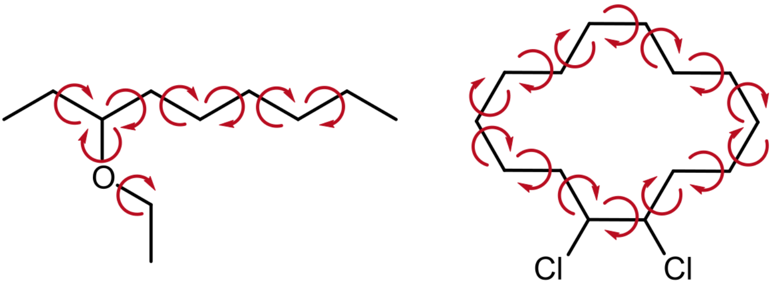


Figure 9.4: Flexible compounds of interest. Red arrows indicate the dihedral angles describing the molecular flexibility.

As demonstrated in Chapter 5, the influence of sidechains on the molecular VCD spectrum can clearly be predicted with neural networks. Sidechains substantially increase the flexibility of a compound, resulting in an increase of the number of conformers. For this reason, VCD applications on very flexible com-

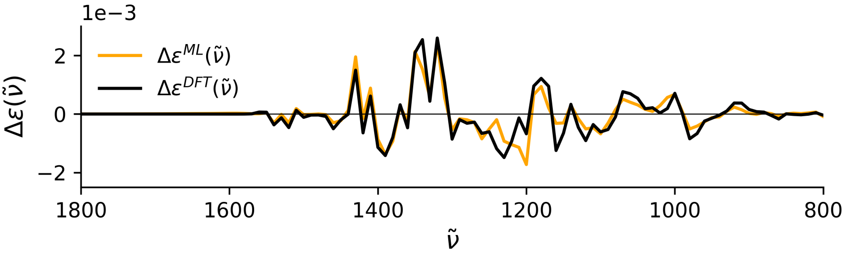


Figure 9.5: Boltzmann weighted spectrum of (S)-3-ethoxynonane and ML predicted one with the 30% of the conformers in the training set. The cosine similarity Θ of the spectra is equal to 0.90.

pounds may require ignoring these sidechains, resulting in lower quality computed spectra. Compounds with both flexible sidechains and a flexible core structure will have an even larger number of conformers, making calculating VCD spectra for such compounds impractical. Therefore, another avenue for further research is to replace the rigid naphthalene core by a flexible core structure. Recent work in our group by Dr. Aerts showed that the VCD conformer spectra of a macrocyclic glycopeptide are very sensitive towards the conformational changes. Here, the molecular spectrum could not be recreated with a representative subset of conformers, indicating that many more conformers need to be included for the Boltzmann weighted spectrum. It would be interesting to see if a similar ML workflow could be used to account for the influence of the sidechains. However, the current challenges with macrocyclic compounds have to be tackled first.

The current workflow still uses the DFT geometries as input for the FNN and the energies for the Boltzmann weighting. One way to alleviate this problem is to replace the quantitative conformer representation into a qualitative one e.g. discretising the input features. Each dihedral angle is then bagged under g/G/T or c/t and the model is trained to predict conformer spectra from this qualitative representation. As a result, the exact dihedral angles of a conformer are no longer needed for prediction. Tests on compounds **1a**, **2a** and **3** show that an FNN trained on a qualitative representation predicts conformer VCD spectra with the same accuracy. Hence, a qualitative representation provides sufficient information to learn VCD conformer spectra. Nonetheless, we could not generate qualitative representations for **1b** and **2b**, so more research into designing alternative qualitative representations is needed. The current workflow still relies on the DFT energies for the Boltzmann weighting of the conformer spectra. In future applications, the conformer energies could be predicted from these approximate inputs with the same model or transfer learning with pre-trained models. Alternatively, models from literature could be used to generate conformers and predict the energies.

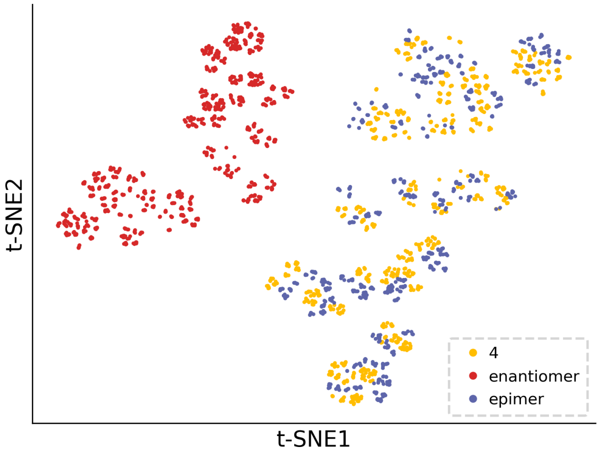


Figure 9.6: The conformer spectra of compound **4**, the epimer and enantiomer (see Chapter 5) projected into 2D space with t-SNE.

Of course, the next step would be to further mature the workflow to make it transferable towards different stereoisomers. Figure 9.6 identifies a pitfall for this further development: the application domain for the conformer does not include conformer spectra of the enantiomer. Therefore, it is unlikely that a model can automatically learn conformer spectra of enantiomers unless specifically trained to do so. One way to counter this issue, is to use transfer learning and hope that the chemical patterns within it include chirality. However, few models with large application domains use representations that discriminate between enantiomers. I believe that a more fruitful avenue lies in predicting the molecular VCD spectra of stereoisomers (with the same 2D structure). For example, the workflow suggested for multi-center AC determination in the previous section could be flipped around. In this, cholic acid would be a nice case study to try this idea on, as it has 11 chiral centres which results in 2048 possible ACs.

9.3 Natural oil extracts

The main goal for furthering the monoterpene project is to broaden the scope of the dataset. The model worked perfectly on the in-silico mixtures and the extracted patterns seemed to work decently on monoterpene mixtures. Some minor tweaking of the in-silico mixture generation and the feature engineering part might be beneficial, but is not expected to substantially influence model performance at this point. The largest shortcoming of the model lies in the transference to mixtures containing samples absent from the dataset. Here, I believe including more monoterpenes and other compounds commonly found in natural oils can push the applicability of the workflow. Again, here the importance of furthering the general VCD database shines through. The crux of the issues regarding ML applications lies in the general lack of availability of VCD spectral data.

Chapter A

Appendix

A.1 Example of FNN prediction for wine dataset

The purpose of this section is to illustrate how a neural network predicts a label for a given sample. Here, we consider a neural network which accepts a wine sample i and predicts whether the sample belongs to a specific cultivator ($y_i = 1$ if True, else $y_i = 0$) using 3 features (proline content, flavanoids content and color intensity). The neural network has 2 hidden layers, each with 2 hidden neurons using the ReLu activation. The output layer has a single neuron using the Sigmoid activation. Figure A.1 shows the current structure of the FNN.

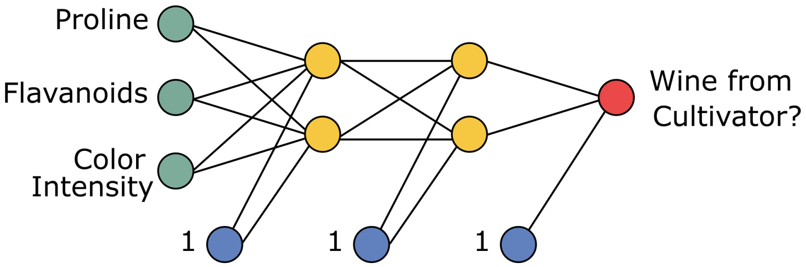


Figure A.1: General structure of the FNN for the wine classification example.

With the current structure, each of the 3 non-input layers will have a weight matrix and a bias vector associated with them. As the output layer contains a single neuron, the output vector is of length 1 and, therefore, the bias vector $\boldsymbol{b}^{(3)}$

is of length 1 and $\boldsymbol{W}^{(3)}$ is a (2×1) matrix.

After training the model on multiple wine samples, the following weight matrices and bias vectors are obtained:

$$\boldsymbol{W}^{(1)} = \begin{bmatrix} 0.092 & 2.690 \\ 0.170 & 1.237 \\ -0.440 & 0.420 \end{bmatrix} \qquad \boldsymbol{b}^{(1)} = \begin{bmatrix} -1.340 \\ 4.371 \end{bmatrix}$$

$$\boldsymbol{W}^{(2)} = \begin{bmatrix} 0.132 & 0.363 \\ -1.349 & 1.640 \end{bmatrix} \qquad \boldsymbol{b}^{(2)} = \begin{bmatrix} -0.763 \\ 0.968 \end{bmatrix}$$

$$\boldsymbol{W}^{(3)} = \begin{bmatrix} -0.427 \\ -1.116 \end{bmatrix} \qquad \boldsymbol{b}^{(3)} = \begin{bmatrix} 11.08 \end{bmatrix}$$

To more easily differentiate the trained parameters from the numbers that depend on the chosen sample, the latter ones will be highlighted in red. Consider sample a, which has the following input vector:

$$\boldsymbol{x}_a = \begin{bmatrix} -0.802 \\ -0.431 \\ -1.344 \end{bmatrix}$$

The first hidden layer takes a weighted average of this input:

$$\begin{aligned} \boldsymbol{z}_a^{(1)} &= \boldsymbol{W}^{(1)\top} \boldsymbol{x}_a + \boldsymbol{b}^{(1)} \\ &= \begin{bmatrix} 0.092 & 0.170 & -0.440 \\ 2.690 & 1.237 & 0.420 \end{bmatrix} \begin{bmatrix} -0.802 \\ -0.431 \\ -1.344 \end{bmatrix} + \begin{bmatrix} -1.340 \\ 4.371 \end{bmatrix} \\ &= \begin{bmatrix} 0.444 \\ -3.255 \end{bmatrix} + \begin{bmatrix} -1.340 \\ 4.371 \end{bmatrix} \\ &= \begin{bmatrix} -0.896 \\ 1.116 \end{bmatrix} \end{aligned}$$

which is then passed along to the ReLU activation function:

$$\boldsymbol{h}_a^{(1)} = \begin{bmatrix} ReLU(-0.896) \\ ReLU(1.116) \end{bmatrix} = \begin{bmatrix} 0 \\ 1.116 \end{bmatrix}$$

The second hidden layer uses the vector $\boldsymbol{h}_a^{(1)}$ as input and takes a weighted average of its components:

$$\begin{aligned}
\boldsymbol{z}_a^{(2)} &= \boldsymbol{W}^{(2)\top}\boldsymbol{h}_a^{(1)} + \boldsymbol{b}^{(2)} \\
&= \begin{bmatrix} 0.132 & -1.349 \\ 0.363 & 1.640 \end{bmatrix} \begin{bmatrix} 0 \\ 1.116 \end{bmatrix} + \begin{bmatrix} -0.763 \\ 0.968 \end{bmatrix} \\
&= \begin{bmatrix} -1.505 \\ 1.830 \end{bmatrix} + \begin{bmatrix} -0.763 \\ 0.968 \end{bmatrix} \\
&= \begin{bmatrix} -2.268 \\ 2.798 \end{bmatrix}
\end{aligned}$$

which is then passed along to the ReLU activation function:

$$\boldsymbol{h}_a^{(2)} = \begin{bmatrix} ReLU(-2.268) \\ ReLU(2.798) \end{bmatrix} = \begin{bmatrix} 0 \\ 2.798 \end{bmatrix}$$

Finally, the output layer uses $\boldsymbol{h}_a^{(2)}$ as input and takes a weighted average of its components. As the output layer only contains a single neuron, z_a^{pred} and the intercept $b^{(3)}$ are scalars:

$$\begin{aligned}
z_a^{pred} &= \boldsymbol{W}^{(3)\top}\boldsymbol{h}_a^{(1)} + b^{(3)} \\
&= \begin{bmatrix} -0.427 & -1.116 \end{bmatrix} \begin{bmatrix} 0 \\ 2.798 \end{bmatrix} + 11.08 \\
&= -3.123 + 11.08 \\
&= 7.957
\end{aligned}$$

The sigmoid activation function is then given z_a^{pred} to produce the output:

$$y_a^{pred} = \sigma(7.957) = 1.000$$

So, the model predicts that wine sample a originates from the cultivator. An overview of these calculations, along with the matrix and vectors involved, is provided in Figure A.2.

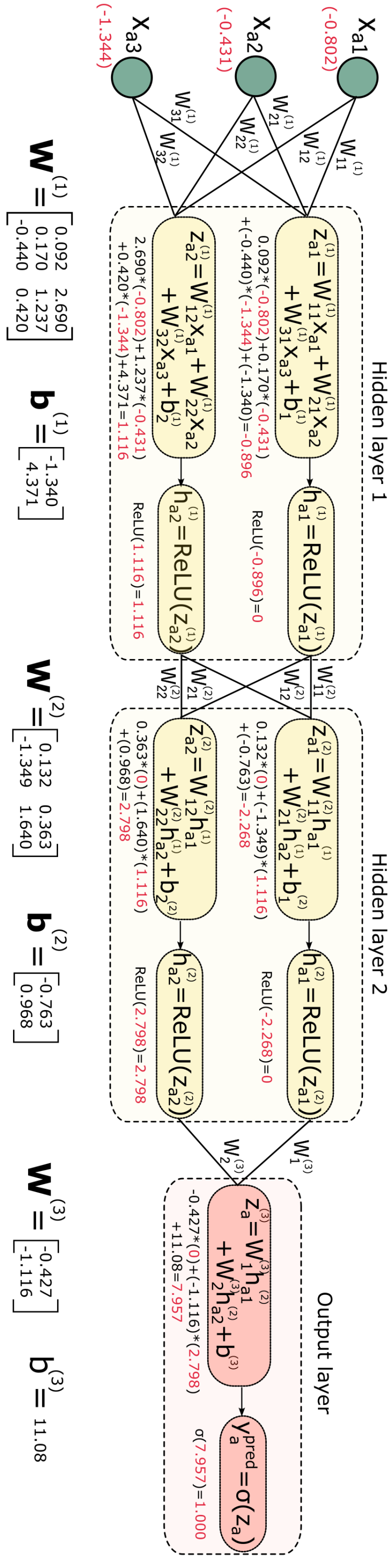


Figure A.2: Prediction of y_a^{pred} for sample a of the wine dataset by the FNN.

A.2 Violin plots

Violin plots provide an easy-to-read visualisation of the data distribution and the key statistical descriptors. As they are not commonly used outside of the data science field, a more detailed description of their makeup is warranted and is provided in this section.

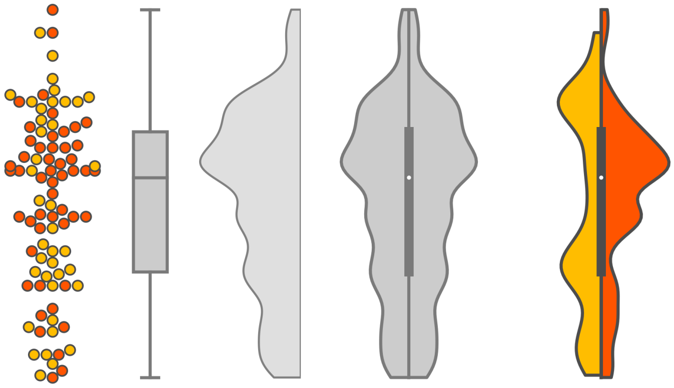


Figure A.3: Makeup of a violin plot. For a set of data instances, both a kernel density and a boxplot are combined into a single violin plot. The kernel density part of the violin plot can be split using a boolean criterion (orange = True, yellow = False), where each side of the violin plot shows the kernel density of one of the two subsets.

A violin plot consists of 2 overlapping plots. The first part is a box plot, which contains the key statistical descriptors of the data. The line in the box shows the mean value, whereas the box edges indicate the values of Q25 and Q75. The whiskers indicate the full range of the distribution (excluding the points deemed to be outliers). The second part of the violin plot is a kernel density plot rotated 90 degrees, with the density in the horizontal direction and the data values in the vertical one. This kernel density is then mirrored along the vertical axis. The width of the violin represents the density of the data instances around that value. For split violin plots, the data is split using a boolean condition and the two kernel density distributions are shown at either side of the box plot. Such violin plots are used in the supplementary information of Chapter 5 to compare values for conformers with (left) and conformers without (right) a hydrogen bond.

Within this work, the Seaborn library is used to create the violin plots.

This library can plot violin plots upon pre-existing matplotlib objects via the `ax = matplotlib.axes.Axes` option. Additional details on violin, kernel density and box plots can be found in the documentation provided by Seaborn.