ARTICLE

Pushing the boundaries of VCD spectroscopy in natural product chemistry

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Tom Vermeyen,[a,b] Andrea N. L. Batista,[c] Alessandra L. Valverde,[c] Wouter Herrebout[a]* and João M. Batista Jr.[d]*

Vibrational circular dichroism (VCD) is one of the most powerful techniques to assess stereochemistry of chiral molecules in solution state. The need for quantum chemical calculations to interpret experimental data, however, has precluded its widespread use by non-experts. 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 is used. Monoterpene mixtures are selected for this proof-of-concept study.

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 characterization 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

[a] *Department of Chemistry, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium. E-mail: wouter.herrebout@uantwerpen.be*
[b] *Department of Chemistry, Ghent University, Krijgslaan 281, B-9000 Ghent, Belgium.*
[c] *Institute of Chemistry, Fluminense Federal University, Outeiro de São João Batista s/n, 24020-141 Niterói-RJ, Brazil.*
[d] *Federal University of São Paulo, Institute of Science and Technology, R. Talim 330, 12231-280, São José dos Campos-SP, Brazil. E-mail: batista.junior@unifesp.br*
Electronic Supplementary Information (ESI) available: [Experimental details, additional IR/VCD spectra, GC-MS spectra, visual spectral markers, machine learning details]. See DOI: 10.1039/x0xx00000x

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 specialized 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. Monoterpene, either isolated or in mixtures, are selected as target molecules for this proof-of-concept study.

Vibrational Circular Dichroism (VCD)

VCD arises from the differential absorption for left- and right-circularly polarized 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]

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]

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 derivatization of *endo*-borneol,[39] the acetonization of 1,3-diols,[40] the derivatization of sphingosine with glutaraldehyde,[41] and the preparation of conformationally restrained cyclic carbodiimides.[42] Non-covalent derivatization 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 (HHDP) group in ellagitannins,[46] for the configuration of the 2(5*H*)-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]

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.

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. S1-S7), artificial mixtures of monoterpenes of each type were prepared and subjected to IR/VCD analysis (Figs. S8-S13). 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 S1). These procedures allowed us to identify the most discriminative

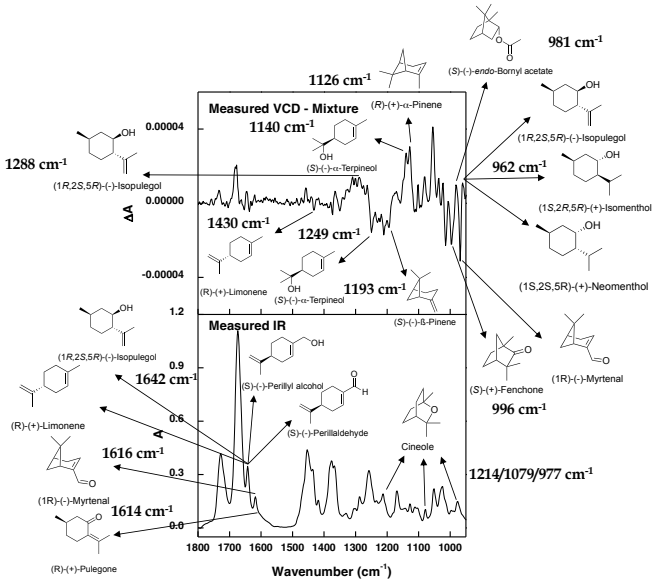

Figure 1. Monoterpene identified from an artificial mixture (J) of known composition by means of visual IR/VCD spectral markers. See ESI for detailed analysis of spectral markers and their vibrational origin.

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. For 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.

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 shifts and relative intensities were used to cluster different monoterpenes. Then, VCD spectra were analysed which, due to their bisignated

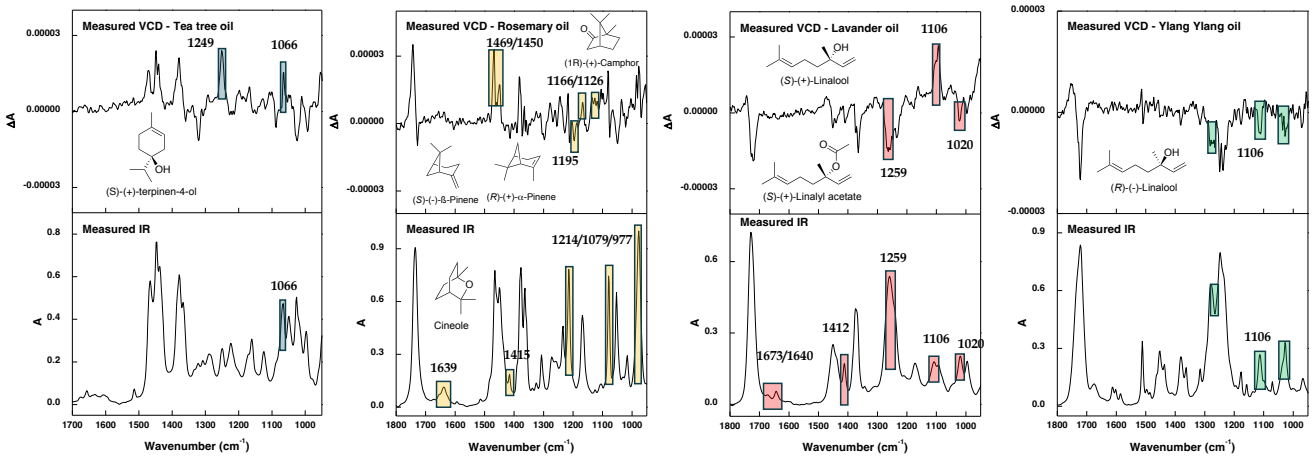

Figure 2. Monoterpene identified from natural mixtures of unknown composition (essential oils) by means of visual IR/VCD spectral markers. See text for discussion of individual bands. Identities of monoterpenes confirmed by GC-MS analysis.

nature, provide better resolution and discriminatory power. On the other hand, having bisignated bands may lead to attenuation or even cancelation 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. S8-S13). 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 1. As can be seen in Fig. 1, even in such a complex mixture, a combination of IR and VCD visual spectral markers were 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 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. S14-S17). It is important to emphasize that not only was 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 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, a ML protocol was idealized, developed and tested as described in the following section.

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. A 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 a 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. 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 S1, 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 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. S27-28) 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. S8-S12), a similar number of monoterpenes were correctly classified by the models (Figs. S21-24). 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 S1 and Figs. S25-26). When the models were applied to mixture J, the VCD model correctly identified the presence 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. 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 were 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 S2-S3. 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 summarize, 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.

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

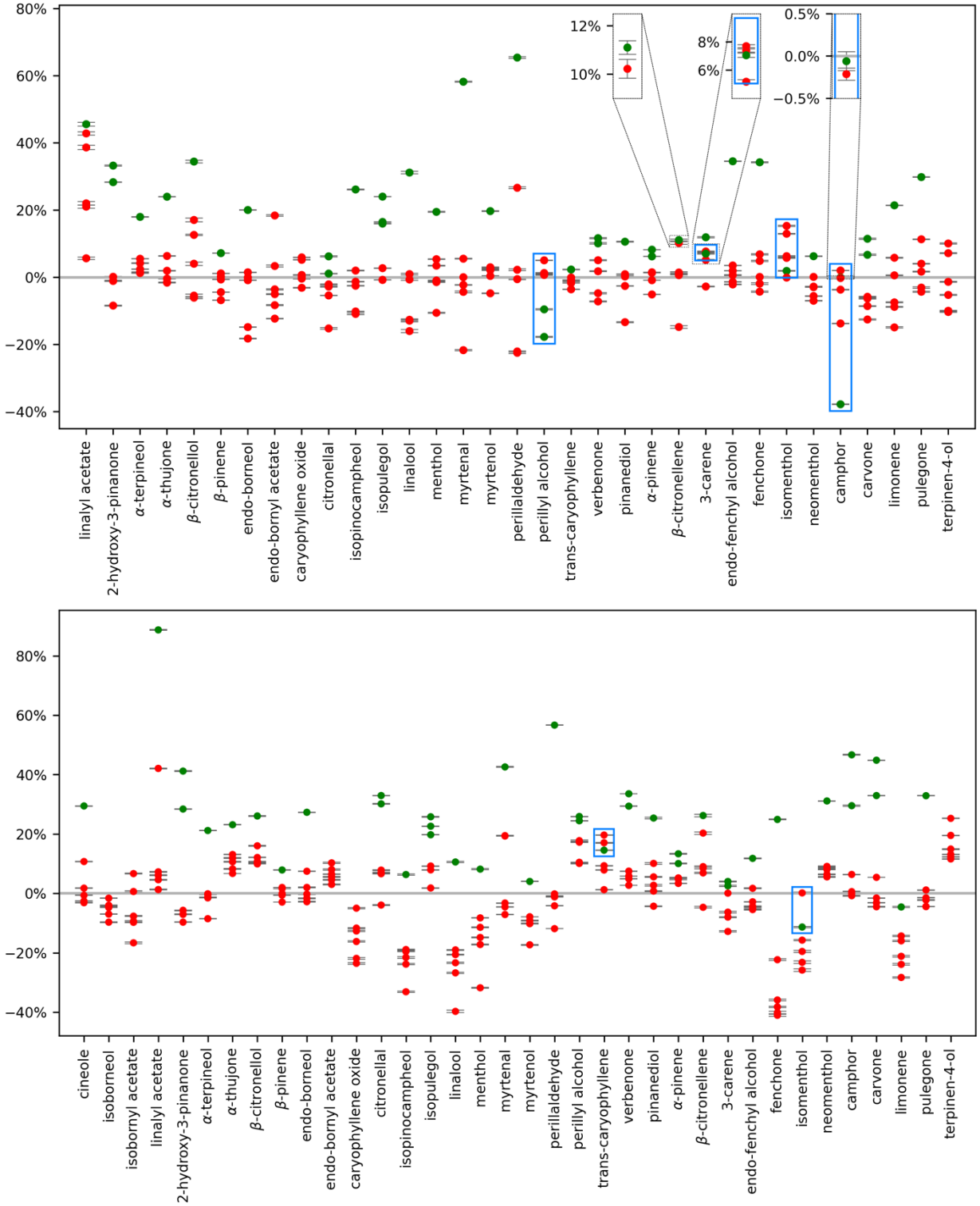

Figure 3. Predicted concentrations (in %) relative to the original concentrations of individual monoterpenes for mixtures A-F by the VCD based (top) and IR based model (bottom). The predicted concentration for each monoterpene is shown for each of the six mixtures and is colored according to whether the monoterpene is present (green) or absent (red) for a given mixture. The predicted concentrations are highlighted (in blue) 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 (dotted lines).

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. 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. Due to the number of false positives for natural mixtures, however, the suggested approach is not yet competitive with other classical methods such as GC-MS. 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.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by funds from the São Paulo Research Foundation (FAPESP grant #2019/22319-5) and by resources supplied by the Centre for Scientific Computing (NCC/GridUNESP) of São Paulo State University (UNESP). ANLB thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brazil (CAPES)-Finance Code 001 for a fellowship. TV thanks the Fund for Scientific Research-Flanders (FWO-Vlaanderen; grant number 1160419N) for a fellowship and the BOF research fund Antwerpen is acknowledged for financial support towards spectroscopic equipment.

Data availability

The IR/VCD spectra of pure terpenes, mixtures, and essential oils are available through a link in the ESI. The ML code will be available upon request.

Notes and references

1 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2020, **83**, 770.
2 A. Mandí and T. Kurtán, *Nat. Prod. Rep.*, 2019, **36**, 889.
3 R. Jwad, D. Weissberger and L. Hunter, *Chem. Rev.*, 2020, **120**, 9743.
4 US Food & Drug Administration. Development of New Stereoisomeric Drugs. Available online: 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.
6 A. N. L. Batista, F. M. Santos Jr., J. M. Batista Jr. and Q. B. Cass, *Molecules*, 2018, **23**, 492.
7 A. J. E. Novak and D. Trauner, *Trends Chem.*, 2020, **2**, 1052.
8 G. T. M. Bitchagno, V.-A. Nchiozem-Ngnitedem, D. Melchert and S. A. Fobofou, *Nat. Rev. Chem.*, 2022, **6**, 806.
9 A. N. L. Batista, B. R. P. Angrisani, M. E. D. Lima, S. M. P. Silva, V. H. Schettini, H. A. Chagas, F. M. Santos Jr., J. M. Batista Jr. and A. L. Valverde, *J. Braz. Chem. Soc.*, 2021, **32**, 1499.
10 S. Allenmark, *Nat. Prod. Rep*., 2000, **17**, 145.
11 J. M. Batista Jr., E. W. Blanch and V. S. Bolzani, *Nat. Prod. Rep*., 2015, **32**, 1280.
12 P. L. Polavarapu and E. Santoro, *Nat. Prod. Rep*., 2020, **37**, 1661.
13 G. Pescitelli and T. Bruhn, *Chirality*, 2016, **28**, 466.
14 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623.
15 T. B. Freedman, X. Cao, R. K. Dukor and L. A. Nafie, *Chirality*, 2003, **15**, 743.
16 L. A. Nafie, *Vibrational optical activity: principles and applications*, Wiley, 2011.
17 P. J. Stephens, F. J. Devlin and J. R. Cheeseman, *VCD spectroscopy for organic chemists*, CRC Press, 2012.
18 C. Merten, T. P. Golub and N. M. Kreienborg, *J. Org. Chem.*, 2019, **84**, 8797.
19 L. A. Nafie, *Chirality*, 2020, **32**, 667.
20 J. Bogaerts, R. Aerts, T. Vermeyen, C. Johannessen, W. Herrebout and J. M. Batista Jr., *Pharmaceuticals*, 2021, **14**, 877.
21 P. M. Dewick, *Medicinal natural products: a biosynthetic approach*, Wiley, 2009.
22 E. Debie, L. Jasper, P. Bultinck, W. Herrebout and B. Van der Veken, *Chem. Phys. Lett.*, 2008, **450**, 426.
23 E. Debie, P. Bultinck, W. Herrebout and B. Van der Veken, *Phys. Chem. Chem. Phys.*, 2008, **10**, 3498.
24 V. P. Nicu, E. Debie, W. Herrebout, B. Van der Veken, P. Bultinck and E. J. Baerends, *Chirality*, 2010, **21**, E287.
25 M. A. K. Koenis, Y. Xia, S. R. Domingos, L. Visscher, W. J. Buma and V. P. Nicu, *Chem. Sci.*, 2019, **10**, 7680.
26 L. A. Nafie, T. A. Keiderling and P. J. Stephens, *J. Am. Chem. Soc.*, 1976, **98**, 2715.
27 L. A. Nafie, M. Diem and D. W. Vidrine, *J. Am. Chem. Soc.*, 1979, **101**, 496.
28 L. A. Nafie, *Appl. Spectrosc.*, 2000, **54**, 1634.
29 C. Guo, R. D. Shah, R. K. Dukor, X. Cao, T. B. Freedman and L. A. Nafie, *Anal. Chem.*, 2004, **76**, 6956.
30 G. Longhi, R. Gengeni, F. Lebon, E. Castiglioni, S. Abbate, V. M. Pultz and D. A. Lightner, *J. Phys. Chem. A*, 2004, **108**, 5338.
31 X. Lu, H. Li, J. W. Nafie, T. Pazderka, M. Pazderková, R. K. Dukor and L. A. Nafie, *Appl. Spectrosc.*, 2017, **71**, 1117.
32 E. Burgueño-Tapia, L. G. Zepeda and P. Joseph-Nathan, *Phytochemistry*, 2010, **71**, 1158.
33 J. C. Pardo-Novoa, H. M. Arreafa-González, M. A. Gómez-Hurtado, G. Rodríguez-Gracía, C. M. Cerda-García-Rojas, P. Joseph-Nathan and R. E. del Río, *J. Nat. Prod.*, 2016, **79**, 2570.
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.
35 R.-Q. Gao, Q. Tan, D. Guo, T. Chen, R.-J. He, D. Li, H. Zhang and W.-G. Zhang, *Chirality*, 2017, **29**, 550.
36 L. F. Julio, E. Burgueño-Tapia, C. E. Díaz, N. Pérez-Hernández, A. González-Coloma and P. Joseph-Nathan, *Chirality*, 2017, **29**, 716.
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.
38 C. Guo, R. D. Shah, R. K. Dukor, T. B. Freedman, X. Cao and L. A. Nafie, *Vibr. Spectrosc.*, 2006, **42**, 254.

39 F. J. Devlin, P. J. Stephens and P. Besse, *J. Org. Chem.*, 2005, **70**, 2980.
40 F. Passareli, A. N. L. Batista, A. J. Cavalheiro, W. A. Herrebout and J. M. Batista Jr., *Phys. Chem. Chem. Phys.*, 2016, **18**, 30903.
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.
42 T. Taniguchi, T. Suzuki, H. satoh, Y. Shichibu, K. Knoish and K. Monde, *J. Am. Chem. Soc.*, 2018, **140**, 15577.
43 C. Grassin, E. Santoro and C. Merten, *Chem. Commun.*, 2022, **58**, 11527.
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.
45 J. M. Batista Jr., A. N. L. Batista, J. S. Mota, Q. B. Cass, M. J. Kato, V. S. Bolzani, T. B. Freedman, S. N. L ópez, M. Furlan and L. A. Nafie, *J. Org. Chem.*, 2011, **76**, 2603.
46 R. F. Sprenger, S. S. Thomasi, A. G. Ferreira, Q. B. Cass and J. M. Batista Jr., *Org. Biomol. Chem.*, 2016, **14**, 3369.
47 F. M. 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.
48 H. E. Ortega, J. M. Batista Jr., W. G. P. Melo, G. T. de Paula and M. T. Pupo, *J. Braz. Chem. Soc.*, 2019, **30**, 2672.
49 A. N. L. Batista, F. M. Santos Jr., A. L. Valverde and J. M. Batista Jr., *Org. Biomol. Chem.*, 2019, **17**, 9772.
50 M. E. 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.
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.
52 C. Holzwarth and I. Chabay, *J. Chem. Phys.*, 1972, **57**, 1632.
53 U. Narayanan and T. A. Keiderling, *J. Am. Chem. Soc.*, 1983, **105**, 6406.
54 S. S. Birke, I. Agbaje and M. Diem, *Biochemistry*, 1992, **31**, 450.
55 T. Taniguchi and K. Monde, *J. Am. Chem. Soc.*, 2012, **134**, 3695.
56 F. M. Santos Jr., C. L. Covington, A. C. F. Albuquerque, J. F. R. Lobo, R. M. Borges, M. B. Amorin and P. L. Polavarapu, *J. Nat. Prod.*, 2015, **78**, 2617.
57 T. Vermeyen, J. Brence, R. V. Echelpoel, R. Aerts. G. Acke, P. Bultinck and W. Herrebout, *Phys. Chem. Chem. Phys.*, 2021, **23**, 19781.