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

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

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

Tom Vermeyen,[a,b] Jure Brence,[c,d] Robin Van Echelpoel,[a] Roy Aerts,[a] Guillaume Acke,[b] Patrick Bultinck*[b] and Wouter Herrebout*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

The added value of supervised Machine Learning (ML) methods to determine the Absolute Configuration (AC) of compounds from their Vibrational Circular Dichroism (VCD) spectra was explored. Among all ML methods considered, Random Forest (RF) and Feedforward Neural Network (FNN) yield the best performance for identification of the AC. At its best, FNN allows near-perfect AC determination, with accuracy of prediction up to 0.995, while RF combines good predictive accuracy (up to 0.940) with the ability to identify the spectral areas important for the identification of the AC. No loss in performance of either model is observed as long as the spectral sampling interval used does not exceed the spectral bandwidth. Increasing the sampling interval proves to be the best method to lower the dimensionality of the input data, thereby decreasing the computational cost associated with the training of the models.

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

[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. E-mail: patrick.bultinck@ugent.be
[c] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia.
[d] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia.
† Electronic Supplementary Information (ESI) available: Additional figures, tables and analysis referred to in the main text. See DOI: 00.0000/00000000.

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.

2 Methodology

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.

2.1 Database design

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 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 Fig 1, were substituted on six different carbon atoms in the scaffold, generating all possible substitution pattern combinations.

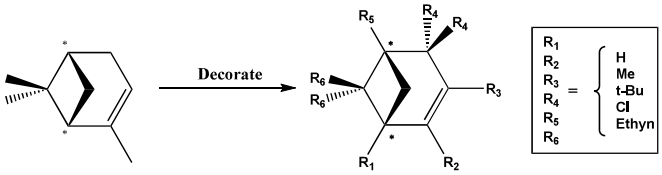


Fig. 1 Decoration of the core structure of (-)-α-pinene. The carbon atoms involved are similarly to R_{1-6} defined as C_{1-6}.

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 intro-

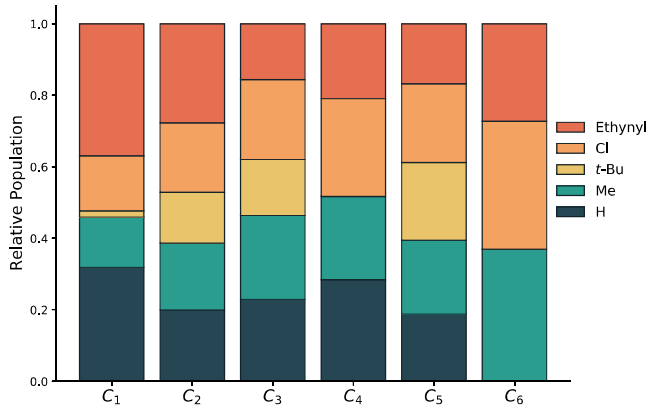


Fig. 2 Relative representation of the substituents on the six different sites C_{1-6}.

duced due to the abovementioned omission of certain structures based on steric clashes, as illustrated in Figure 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 Section D†.

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}.

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 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 represent 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.

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 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 S1†.

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.

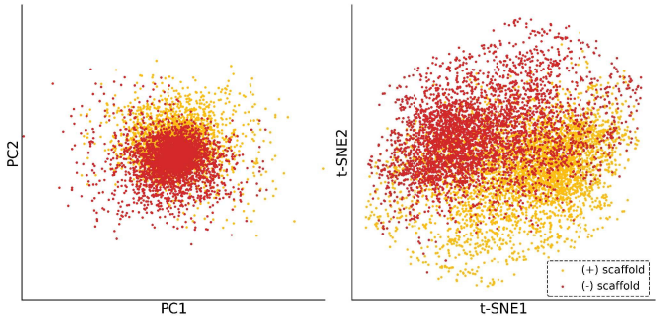


Fig. 3 Visualisation of the spectra after dimensionality reduction with (left) 2D-PCA and (right) 2D-t-SNE, with yellow and red dots corresponding to the VCD spectra of decorated (+)- and (-)-α-pinene structures respectively.

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.

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.

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, classified 0.785 of the spectra 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 S1†).

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 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 3, only a small part of the 2D-PCA plot remains charac-

teristic 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.

3.2 Identification of best performing ML models

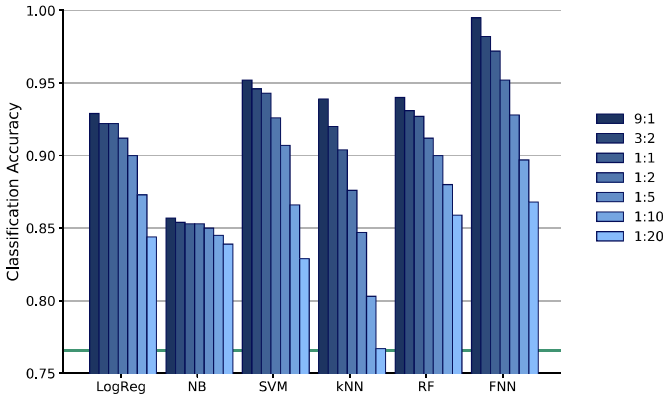
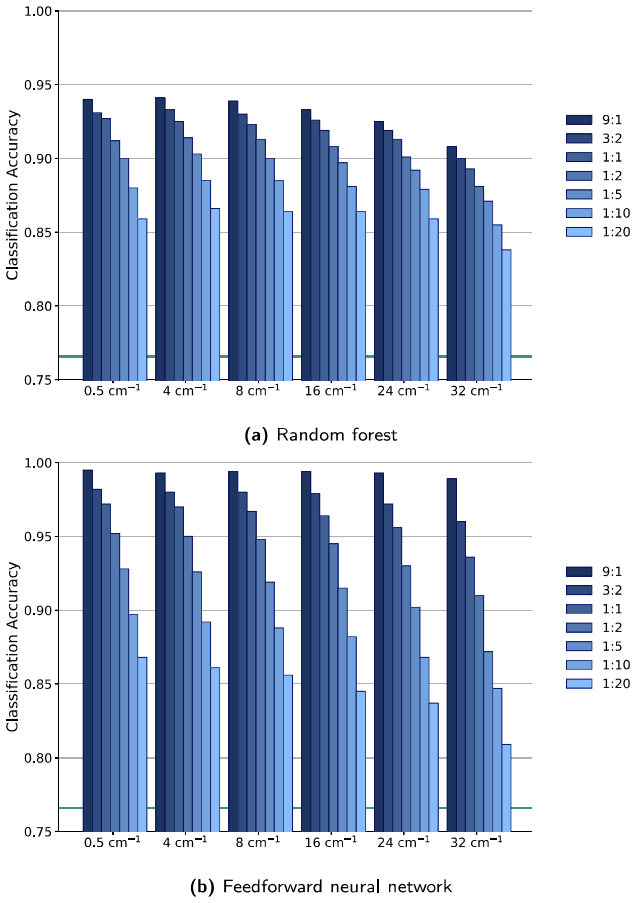


Fig. 4 Classification accuracy of the spectra for several supervised ML models. The different train-validation split ratios are coloured as described in the legend.

The CA for the various ML methods is summarised in Figure 4 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 S2†. 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 stronger regularisation, the accuracy dropped significantly (see Figure S2†). 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 S6†). 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.

3.3 Influence of spectral sampling interval



(a) Random forest

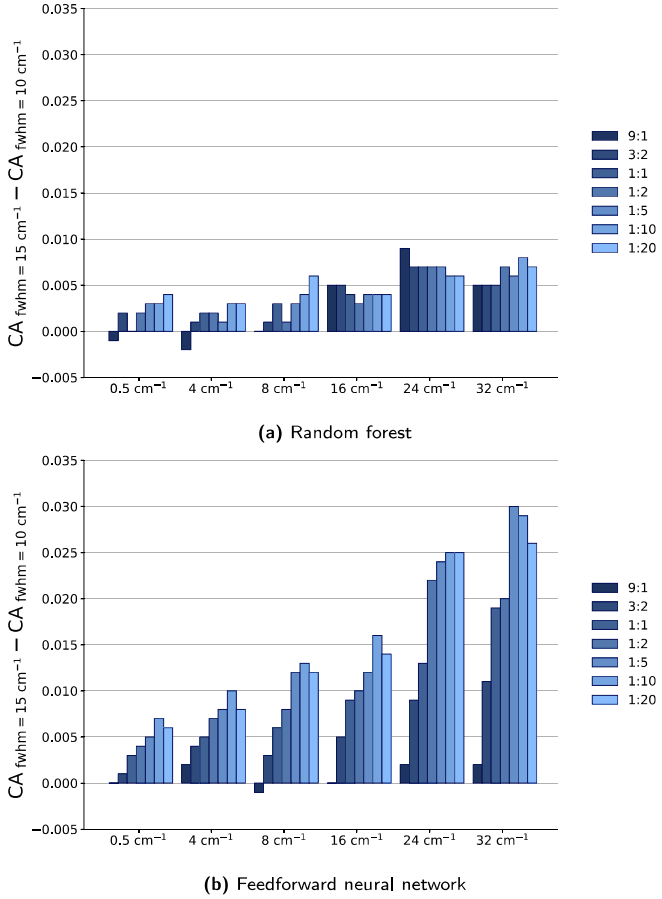
(b) Feedforward neural network

Fig. 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.

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 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 S4†) 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 S5† and Figure 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

Fig. 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.

3.4 AC pattern extraction with RF

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 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 $\Delta\varepsilon$ for each wavenumber with the corresponding ranking values (Figure 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.

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 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

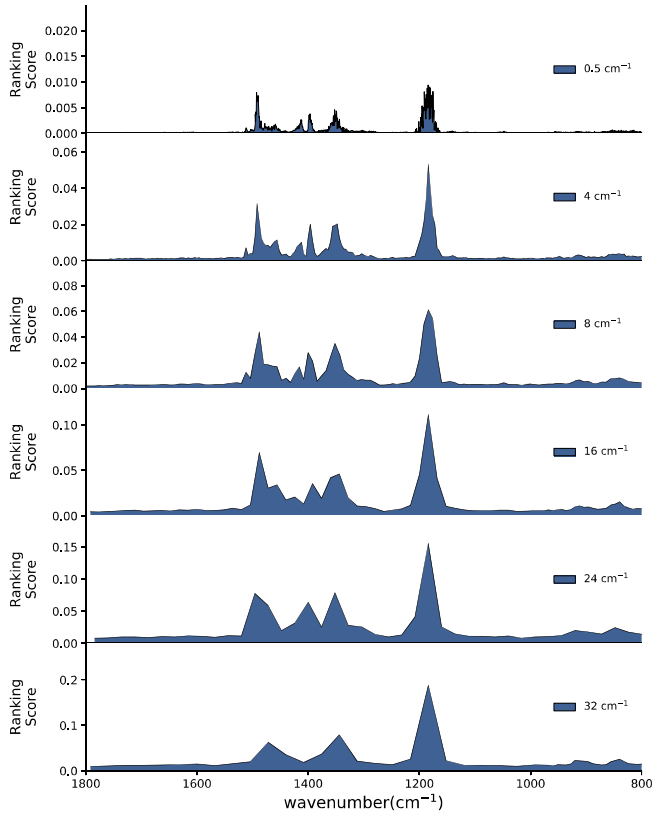


Fig. 7 Random forest ranking score of the spectral features for the prediction of the chirality of the compounds for the different sampling intervals.

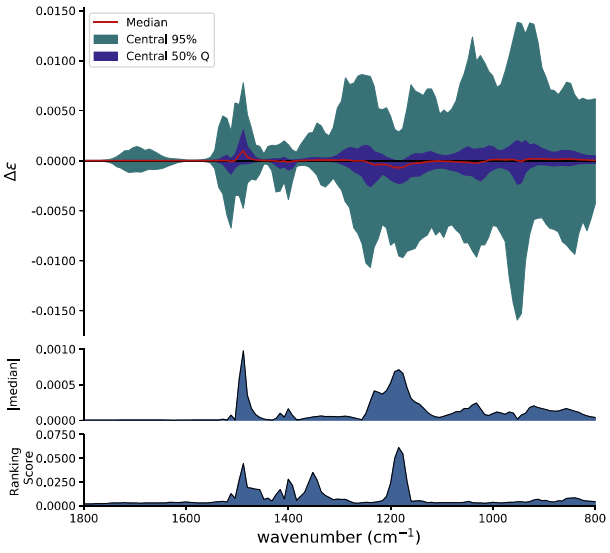
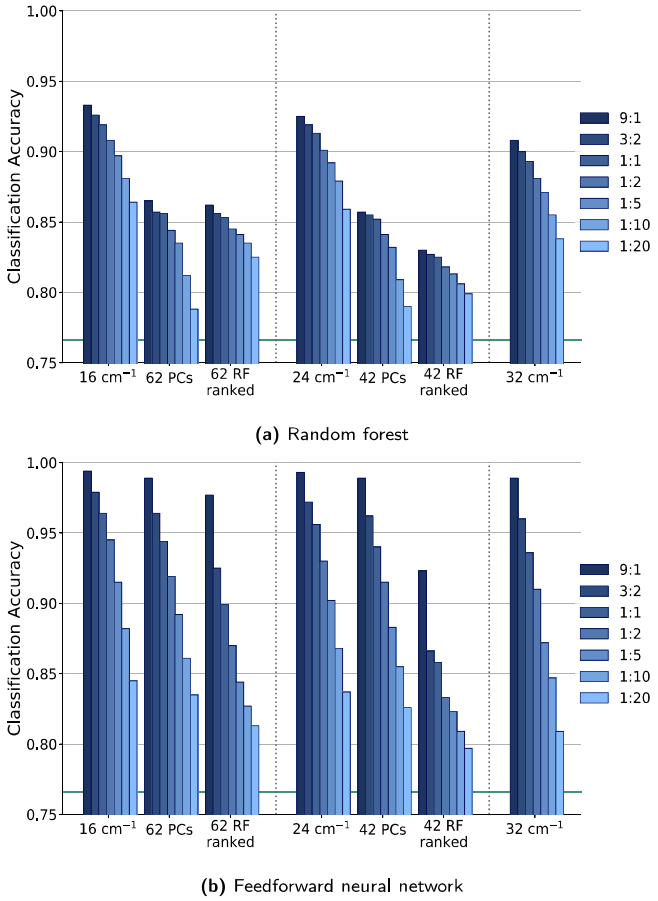


Fig. 8 Top: The median value (red) and central 50% and 95% quantiles of the VCD spectra sharing the core structure of (+)-α-pinene, Middle: The absolute value of the median. Bottom: Random forest based ranking score for spectra with a sampling interval of 8 cm^{-1}.

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 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.



(a) Random forest

(b) Feedforward neural network

Fig. 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.

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 S1†), 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 section G† and H†) 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 section I† and J†). 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.

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.

5 Author Contributions

Conceptualisation: T.V. and J.B., Generation of database: T.V. and R.V.E., Model training, optimisation and validation: T.V., Analysis: T.V. and J.B., Supervision: W.H. and P.B., Writing - original draft: T.V., Writing - review & editing: T.V., R.A., G.A., P.B., W.H.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work has been funded by the Fund for Scientific Research-Flanders (FWO-Vlaanderen; grant number 1160419N). The Flemish Supercomputer Centre (VSC) is acknowledged for providing computational resources and support. The University of Antwerp (BOF-NOI) is acknowledged for the pre-doctoral scholarship of R.A.

Notes and 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, 2012, pp. 1–199.
7 J. 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. 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.

Supporting Information

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

Tom Vermeyen,[a,b] Jure Brence,[c,d] Robin Van Echelpoel,[a] Roy Aerts,[a] Guillaume Acke,[b] Patrick Bultinck*[b] and Wouter Herrebout*[a]

[a] *Department of Chemistry, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium.*

[b] *Department of Chemistry, Ghent University, Krijgslaan 281, B-9000 Ghent, Belgium.*

[c] *Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia.*

[d] *Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia.*

E-mail: wouter.herrebout@uantwerpen.be, patrick.bultinck@ugent.be

Contents

A 3D Principal Component Analysis on VCD spectra

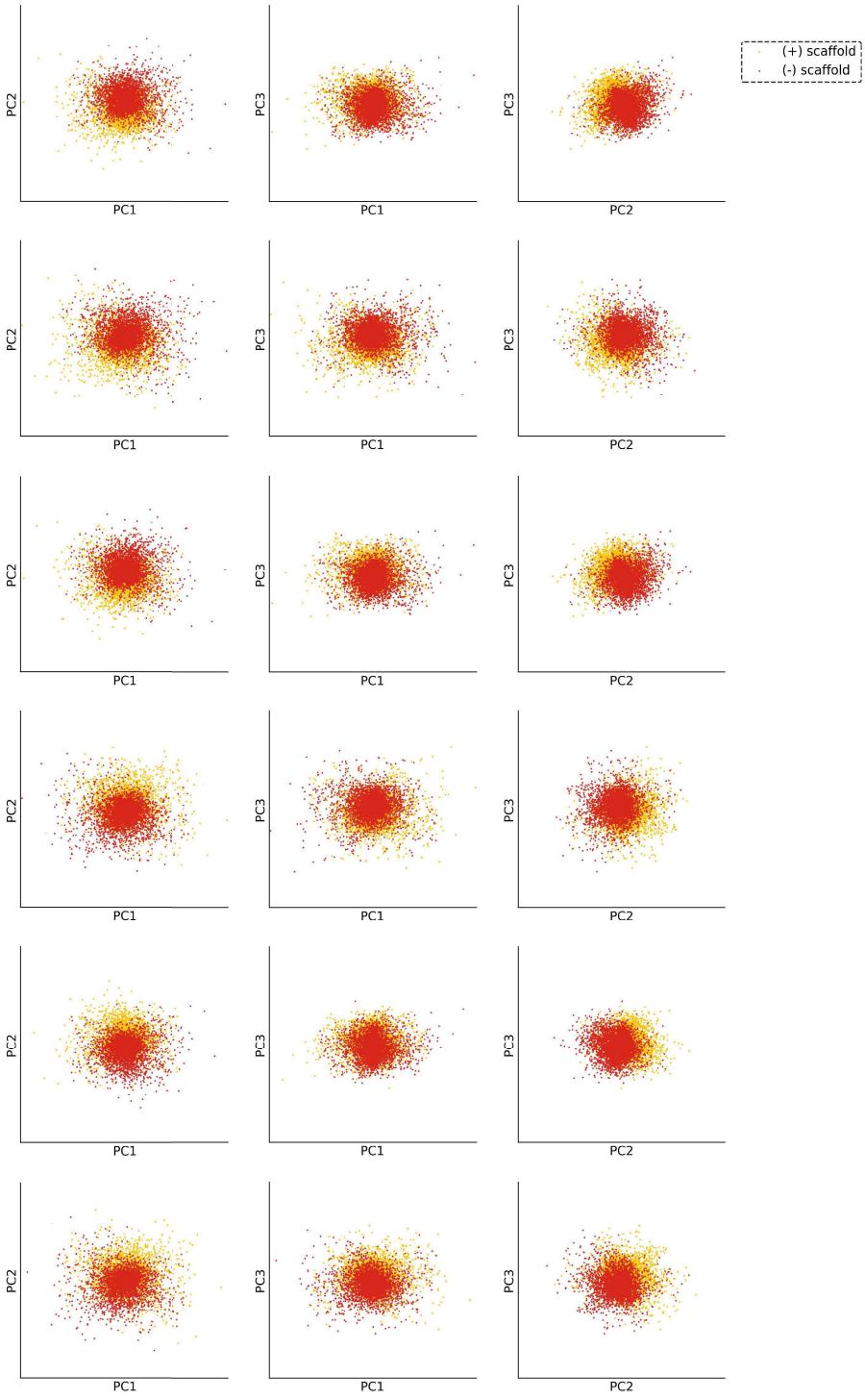


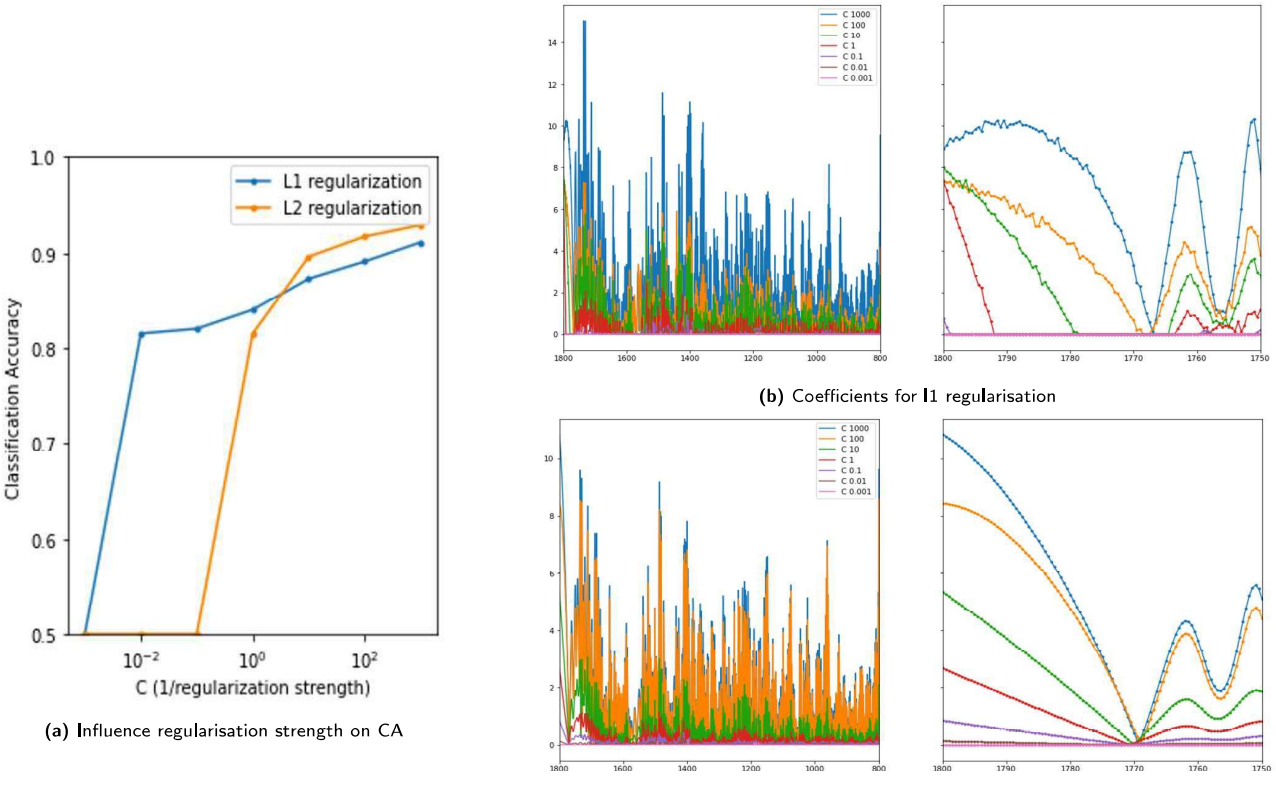
Fig. S1 Comparison of the enantiomers' PCA transformed spectra, from top to bottom B3LYP/6-31G(d), B3PW91/6-31G(d), B3LYP/6-31++G(d,p), B3PW91/6-31++G(d,p), B3LYP/6-311++G(2d,2p), B3PW91/6-311++G(2d,2p).

B 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 S1 Optimised hyperparameter for the supervised machine learning models.

C Logistic regression weights for weak & strong regularisation



(a) Influence regularisation strength on CA

(b) Coefficients for l1 regularisation

(c) Coefficients for l2 regularisation

Fig. S2 Influence of regularisation strength and method for logistic regression on the classification accuracy and the coefficients.

D 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 S3, 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 2). However, it remains difficult to clearly reveal the extent to which one dominates over the other.

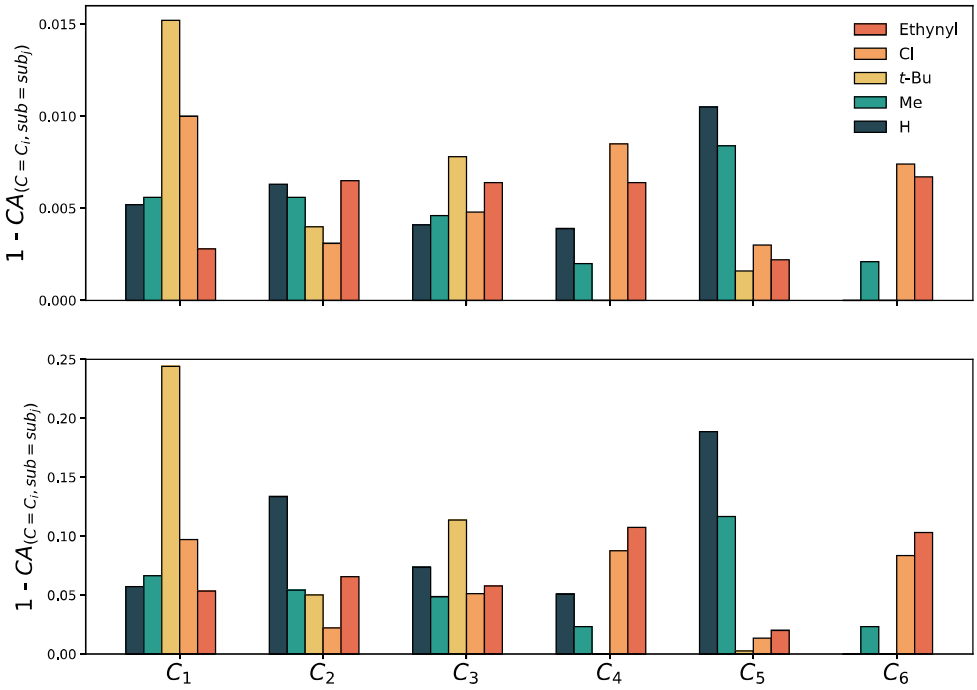
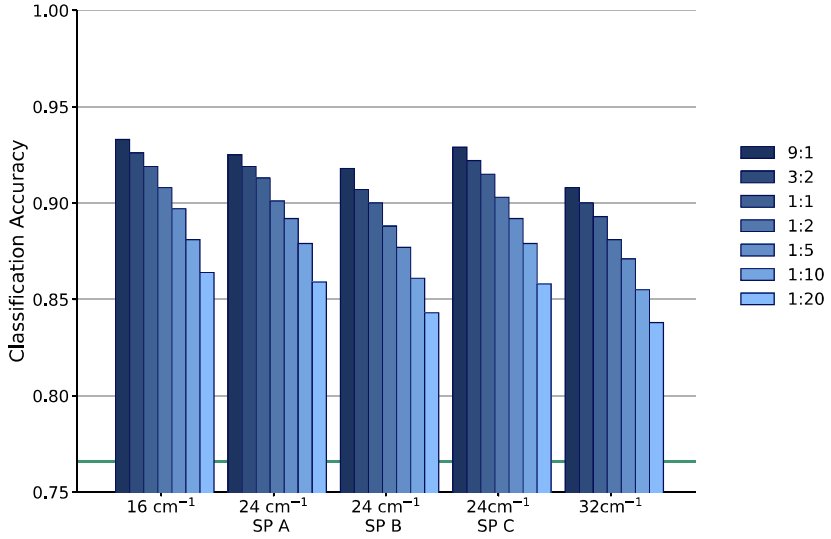


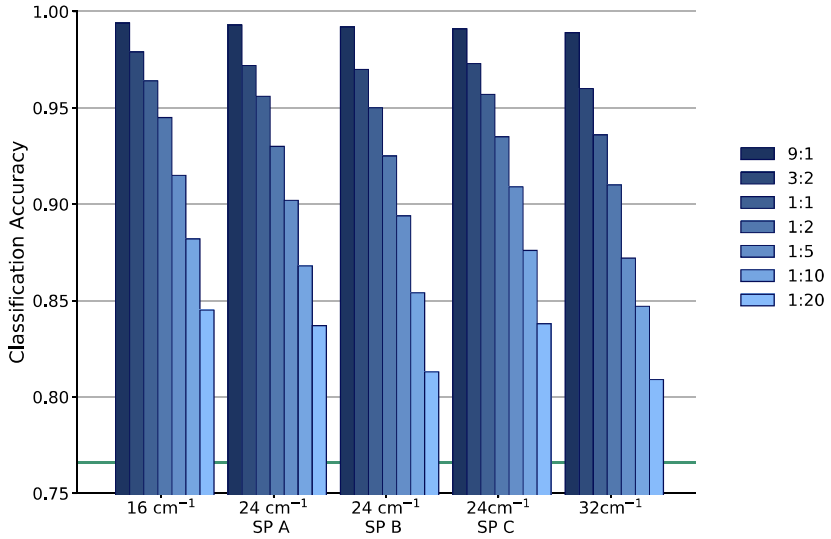
Fig. S3 Relative misclassification of the spectra for a certain substituent at each position 1-6 separately for feedforward neural network(top) and fandom forest(bottom).

E Influence of starting point on Classification Accuracy 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 S4, 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.



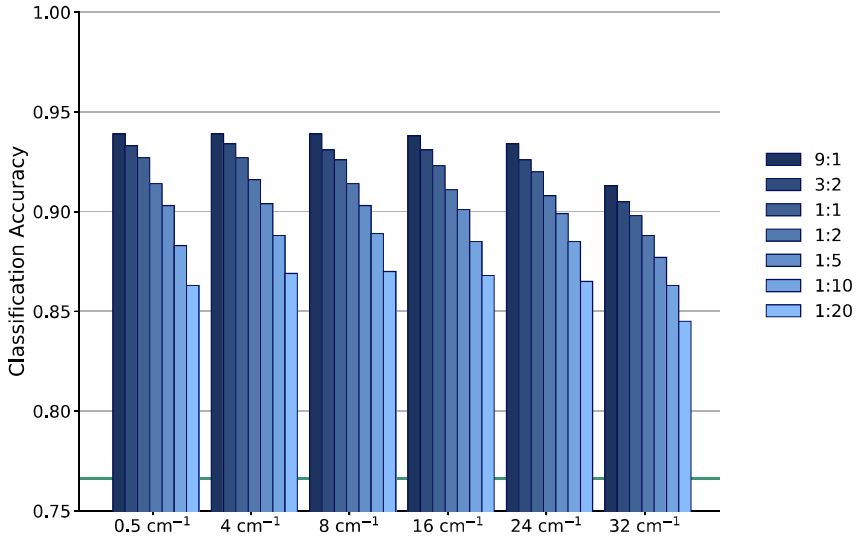
(a) Random forest



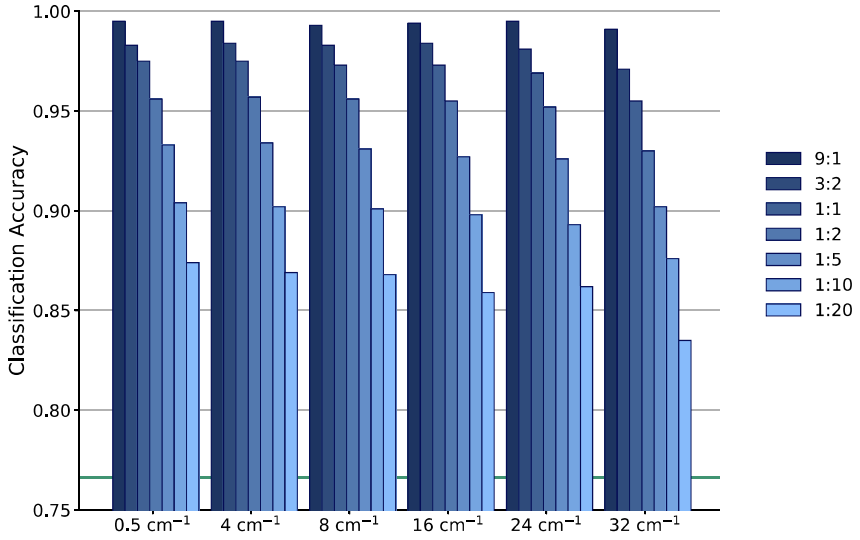
(b) Forward neural network

Fig. S4 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.

F Classification Accuracy for spectra with bandwidth of 15 cm^{-1}



(a) Random forest



(b) Forward neural network

Fig. S5 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.

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

In order to evaluate the stability 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 S6. We observe that the performance of LogReg, NB and, in particular, SVM is noticeably dependant on the level of theory, even when the a large majority of the data is provided for training.

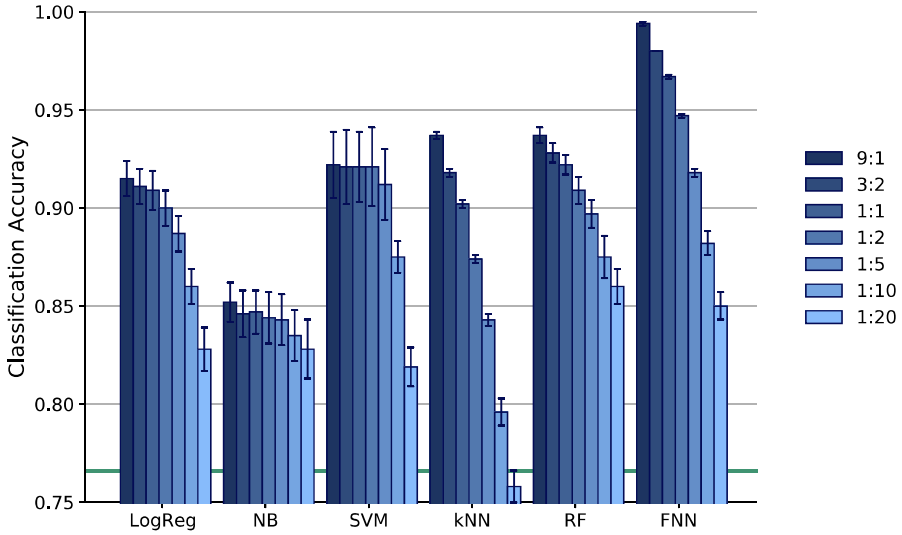
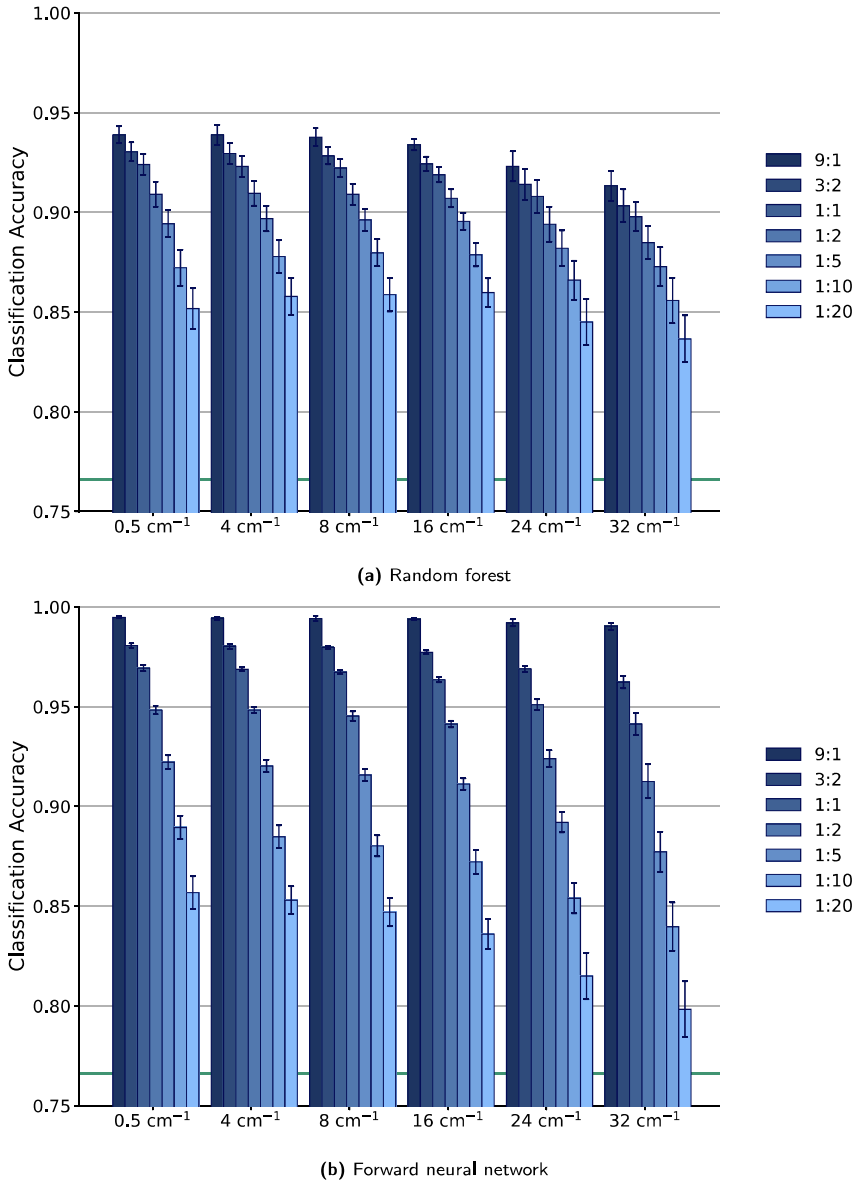


Fig. S6 Mean Classification accuracy of the spectra for the different ML models over all combinations of the B3LYP and B3PW91 functionals, with the 6-31G(d)6-31++G(d,p)/ 6-311++G(2d,2p) basis sets. The different data split ratios are coloured as described in the legend.

H 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 S1) 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 S7. 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 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

Fig. S7 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.

I Feature ranking for RF trained on various functional/basis set combinations

The question arises whether the similar performances discussed in section H and G are due to the robustness of the ML methods or the ML models themselves are identical. In this section, the workflow described in section 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 S8) 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.

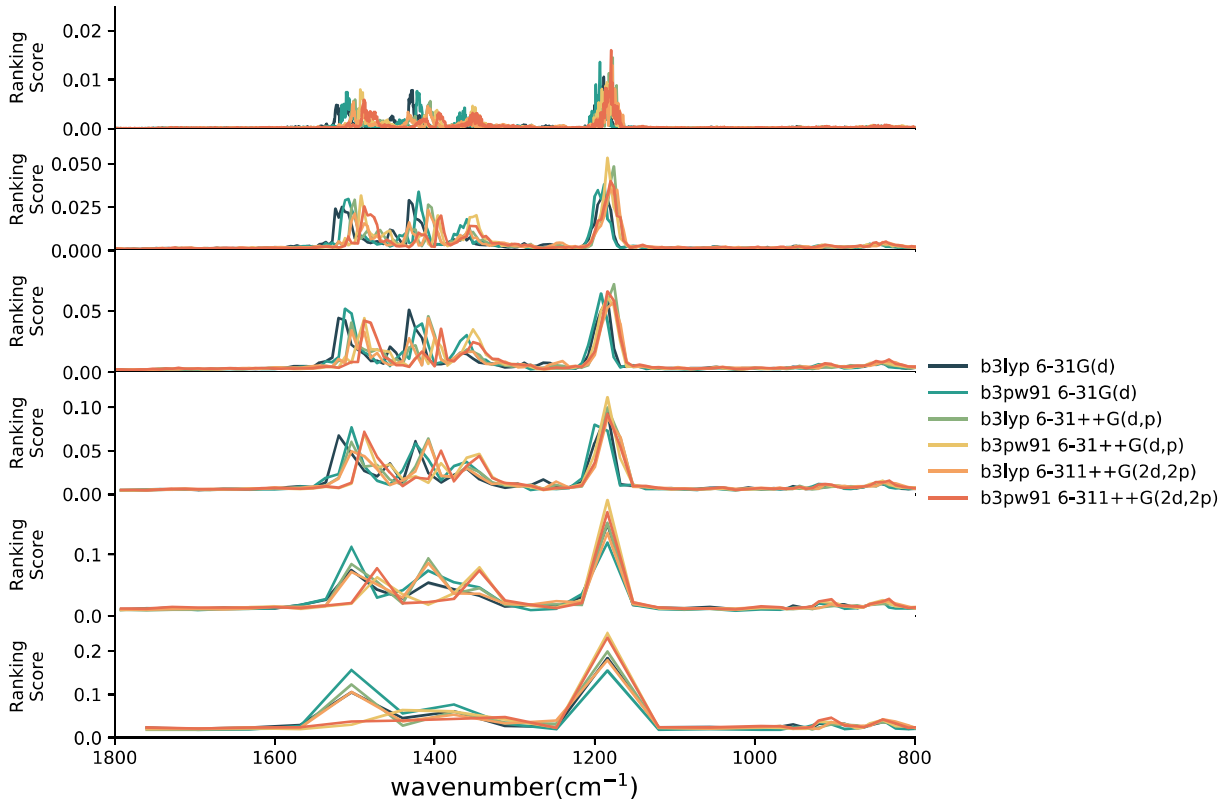


Fig. S8 Random forest ranking score of the spectral features for the prediction of the chirality of the compounds for the different sampling intervals and combinations of functional and basis set. From top to bottom the sampling interval equals 0.5, 4, 8, 16, 24, 32 cm^{-1}.

J Performance and structure of shallow decision trees trained on various functional/basis set

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 S9, 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.

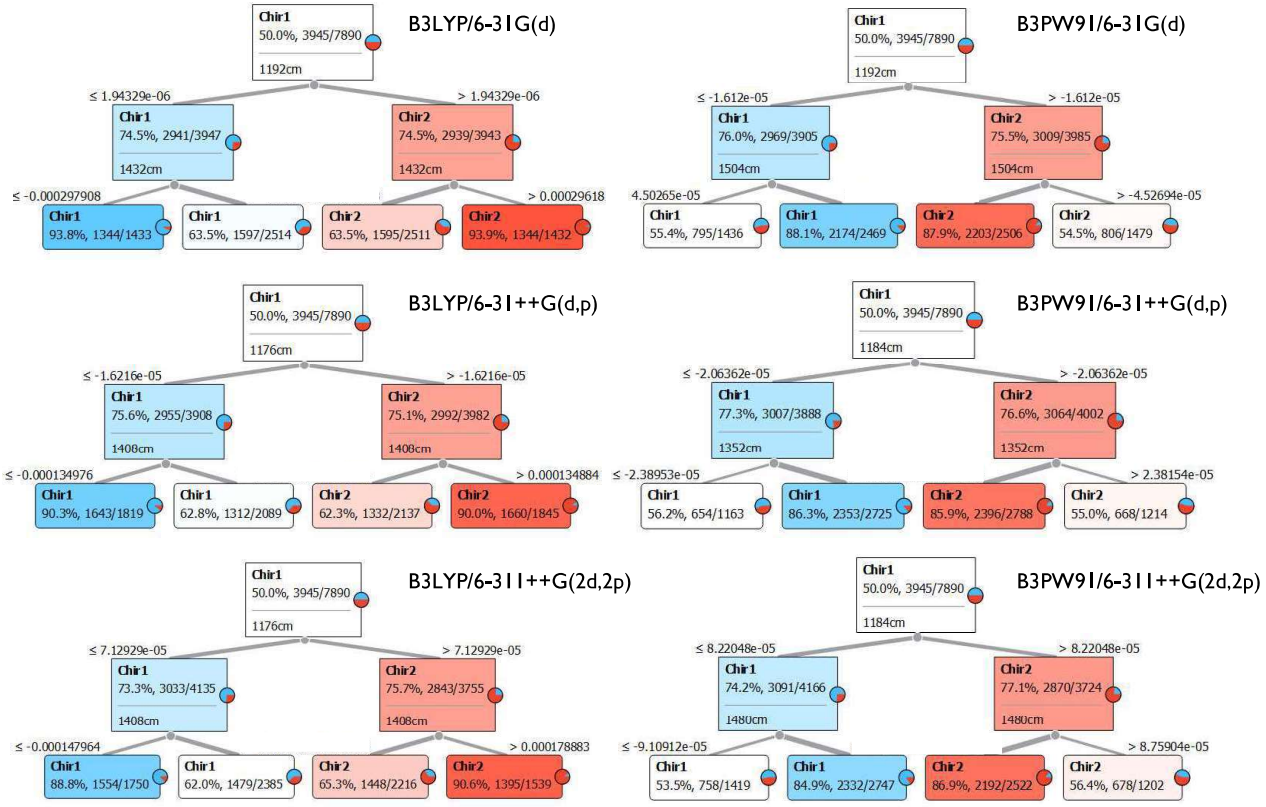


Fig. S9 Shallow decision trees trained on VCD spectra (SI 8 cm^{-1}) of different levels of theory as denoted in the figure. The nodes are coloured according to their purity, with a blue-white-red gradient, with the dominant chirality class present in each node denoted as 1 ((+)-α-pinene) or 2 ((-)-α-pinene). 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.