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

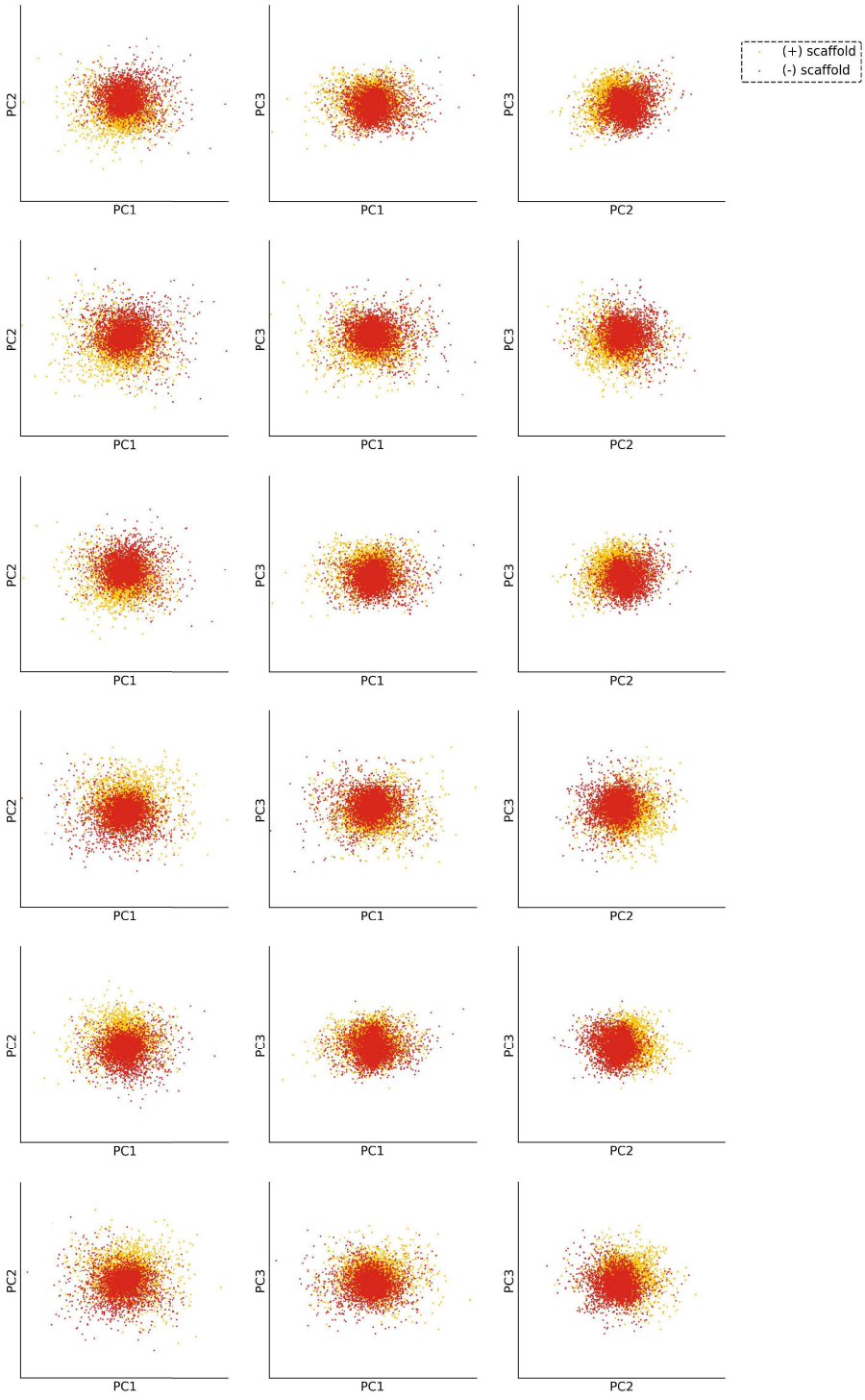


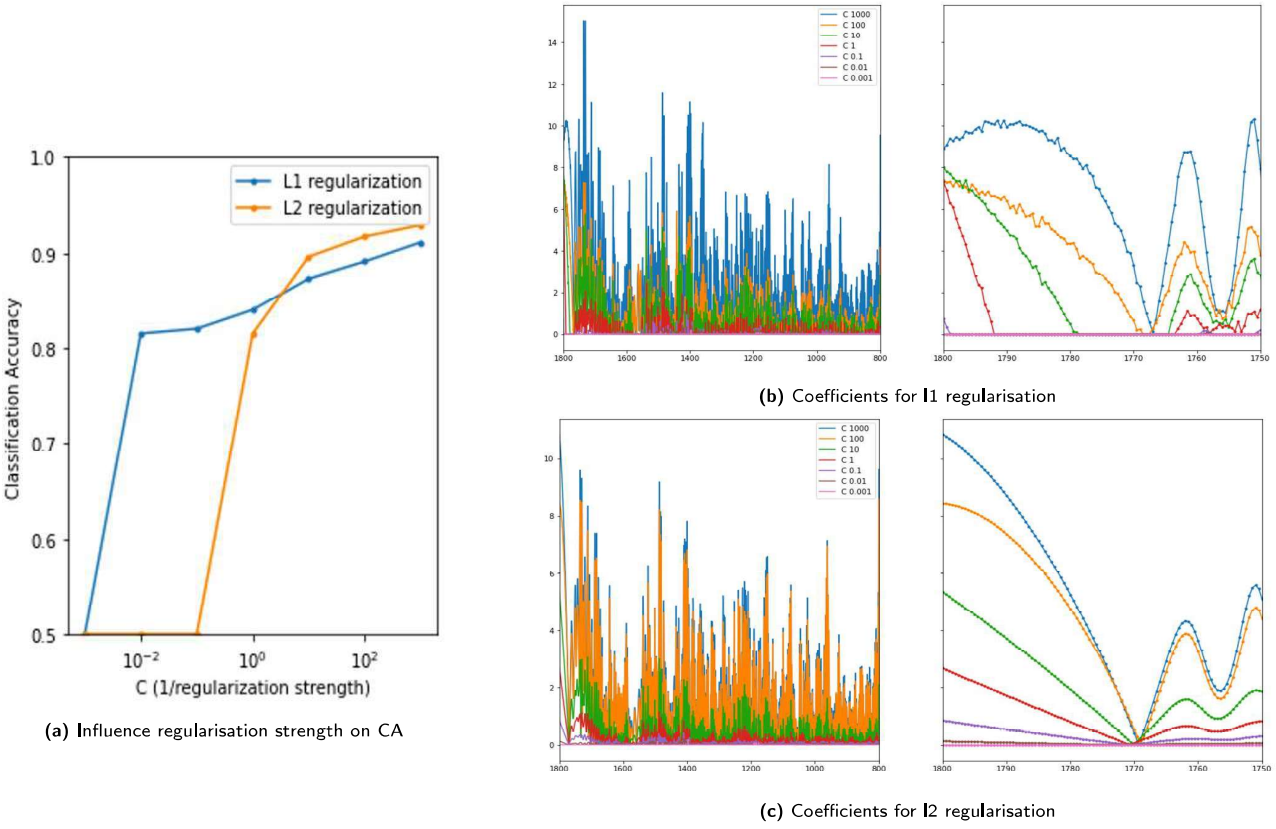
Fig. S1 Comparison of the enantiomers' PCA transformed spectra, from top to bottom B3LYP/6-31G(d), B3PW91/6-31G(d), B3LYP/6-31++G(d,p), B3PW91/6-31++G(d,p), B3LYP/6-311++G(2d,2p), B3PW91/6-311++G(2d,2p).

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.

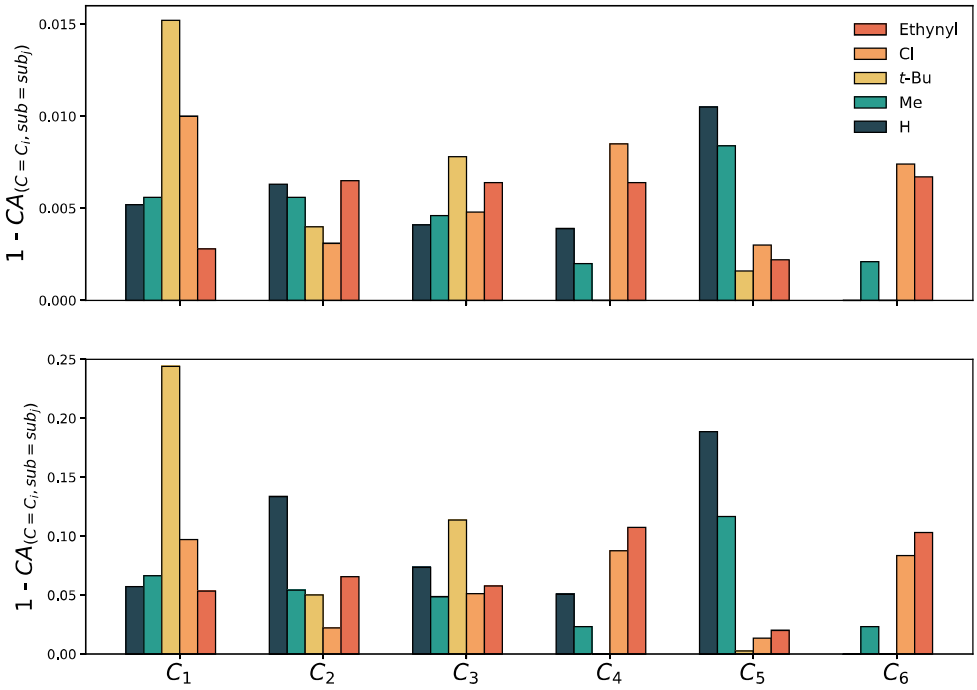
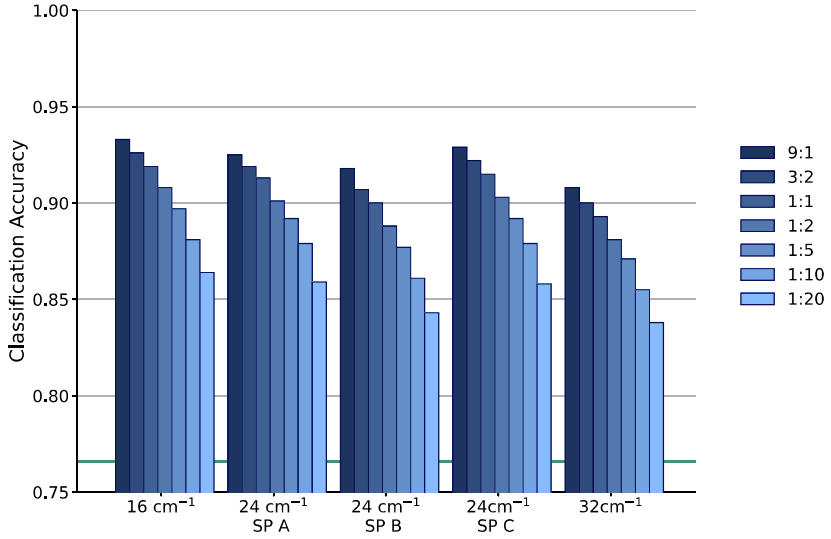


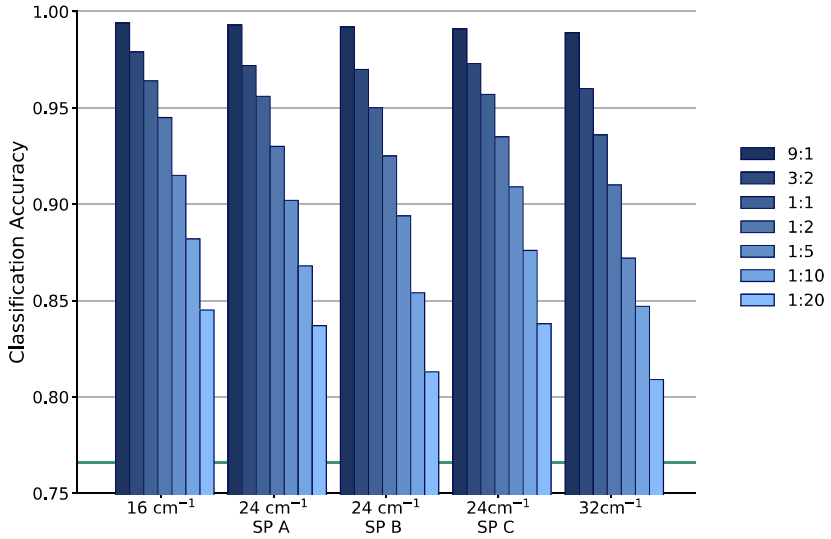
Fig. S3 Relative misclassification of the spectra for a certain substituent at each position 1-6 separately for feedforward neural network(top) and fandom forest(bottom).

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.



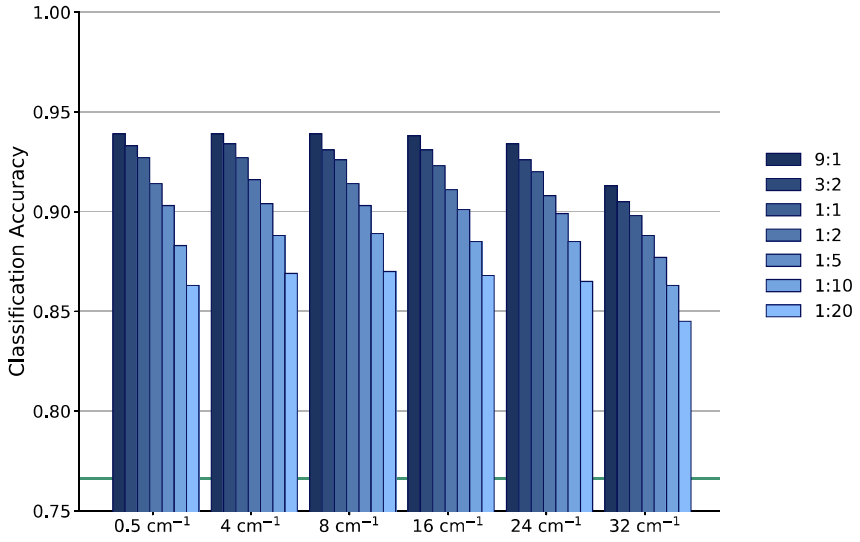
(a) Random forest



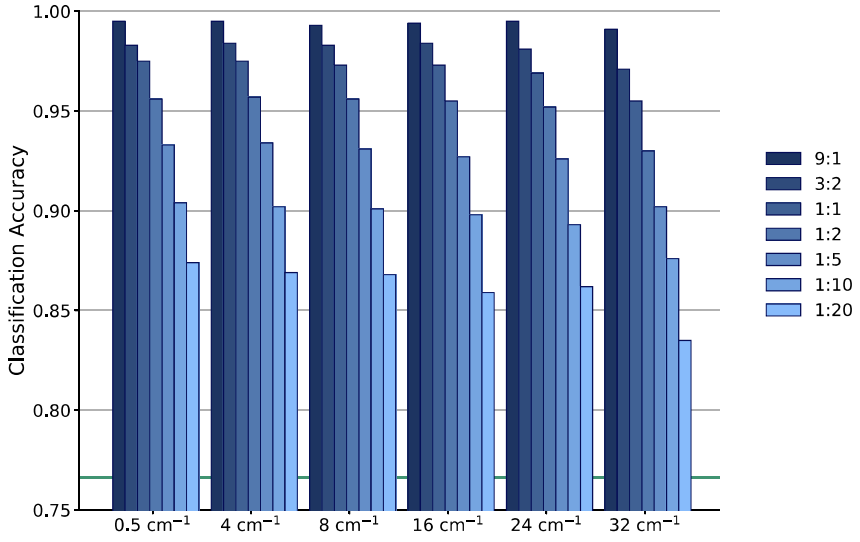
(b) Forward neural network

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.

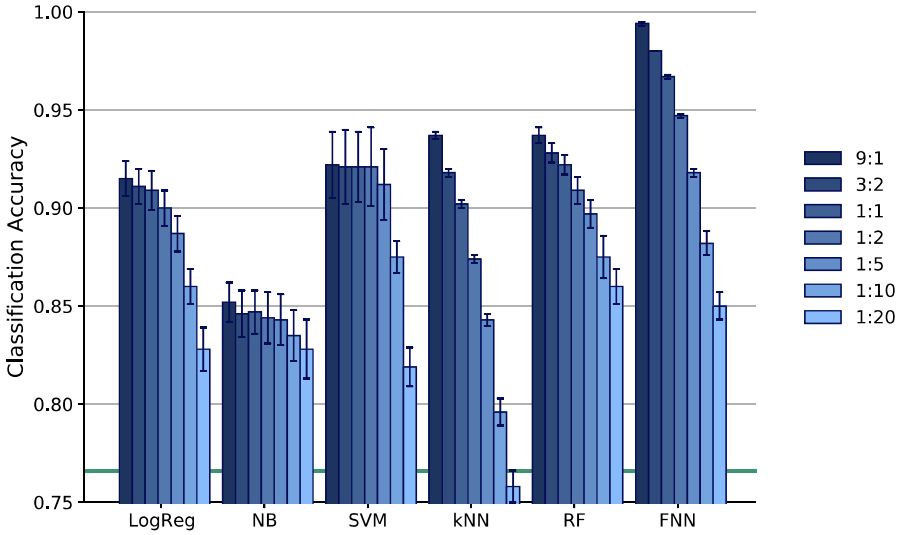
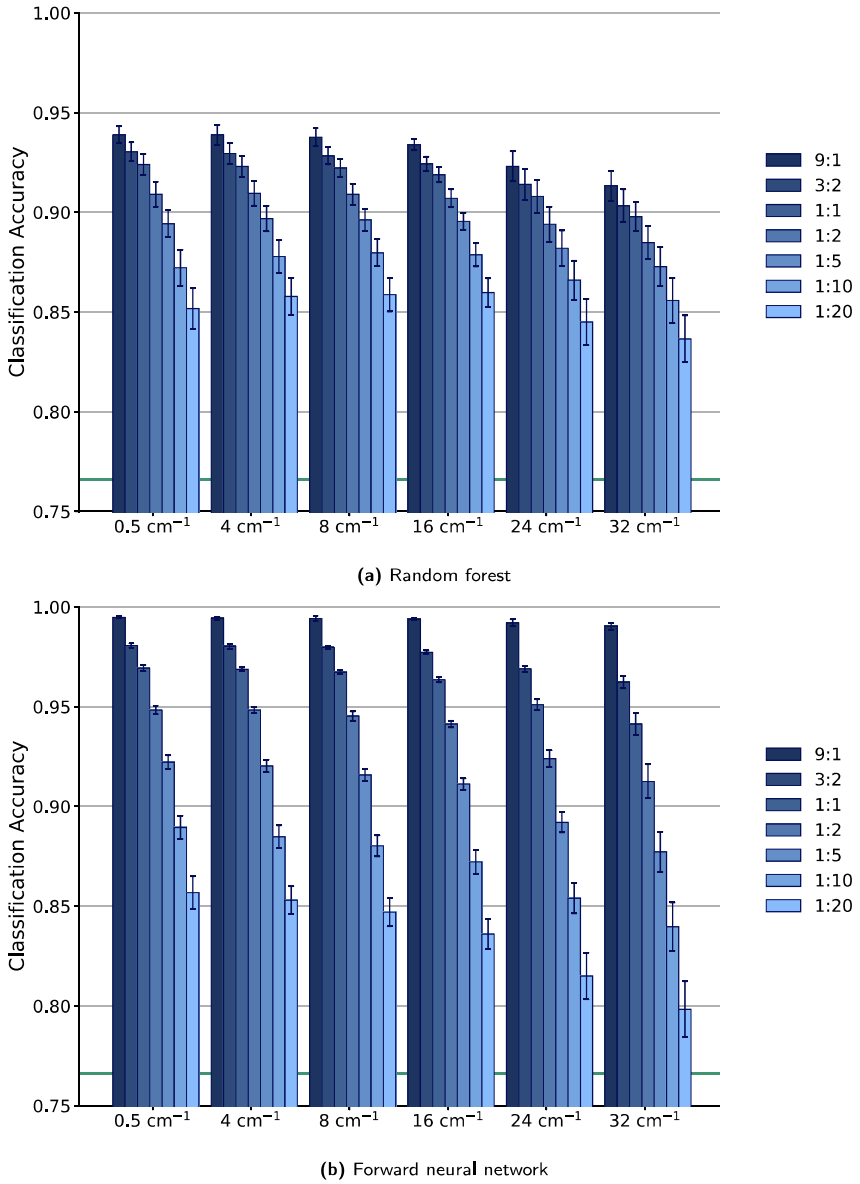


Fig. S6 Mean Classification accuracy of the spectra for the different ML models over all combinations of the B3LYP and B3PW91 functionals, with the 6-31G(d)6-31++G(d,p)/ 6-311++G(2d,2p) basis sets. The different data split ratios are coloured as described in the legend.

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.

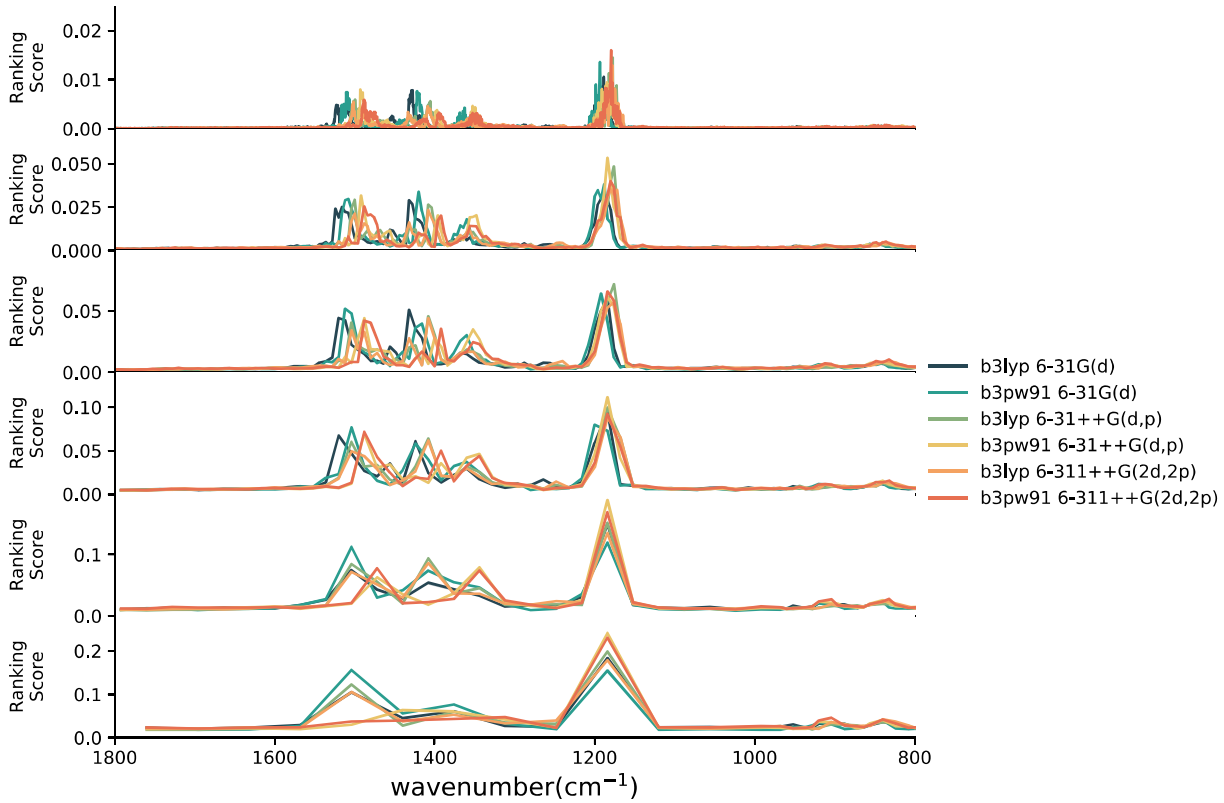


Fig. S8 Random forest ranking score of the spectral features for the prediction of the chirality of the compounds for the different sampling intervals and combinations of functional and basis set. From top to bottom the sampling interval equals 0.5, 4, 8, 16, 24, 32 cm^{-1}.

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.

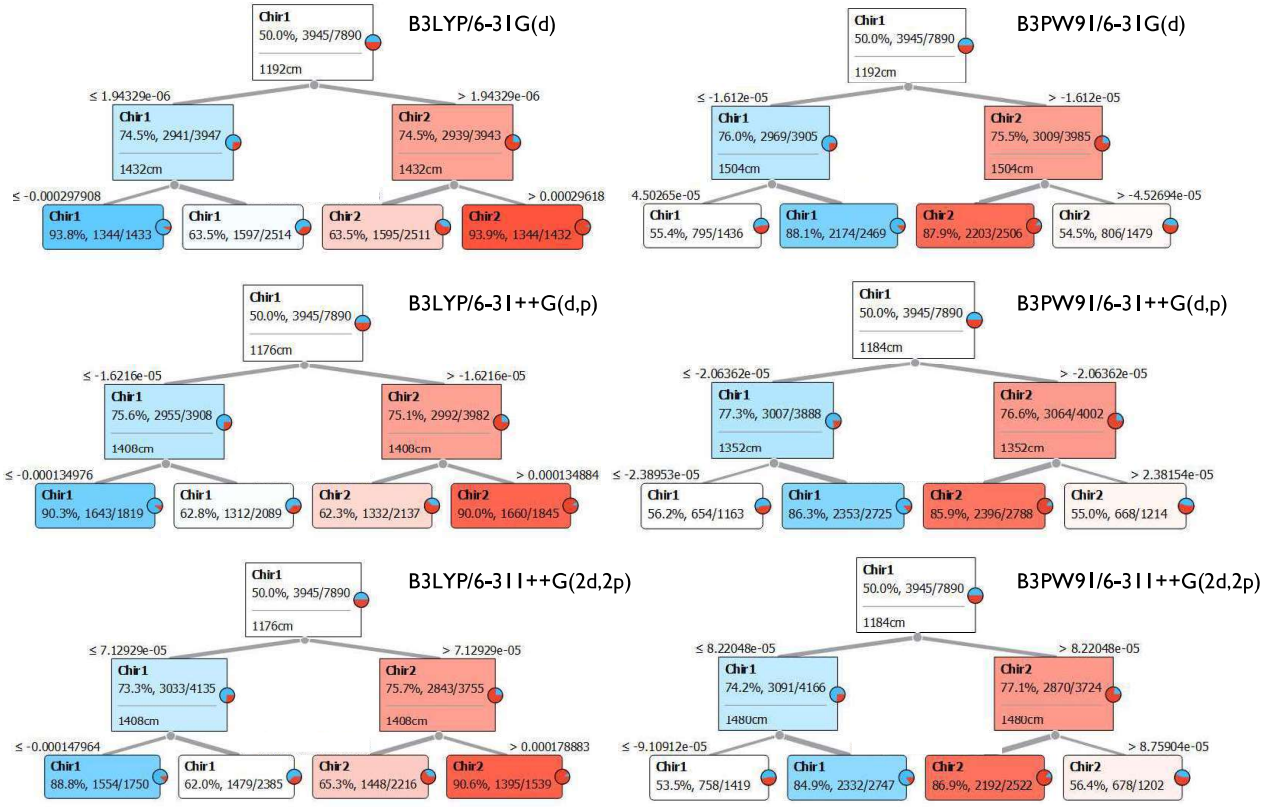


Fig. S9 Shallow decision trees trained on VCD spectra (SI 8 cm^{-1}) of different levels of theory as denoted in the figure. The nodes are coloured according to their purity, with a blue-white-red gradient, with the dominant chirality class present in each node denoted as 1 ((+)-α-pinene) or 2 ((-)-α-pinene).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