SUPPORTING INFORMATION

Determining the Regiochemistry and Relative Stereochemistry of Small and Drug-like Molecules Using an Alignment Algorithm for Infrared Spectra

Lennard Böselt,[†] Reinhard Dötzer,[‡] Sandra Steiner,[‡] Michaela Stritzinger,[‡] Susanne Salzmann,[¶] and Sereina Riniker[*,†]

[†]*Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland*

[‡]*Competence Center Analytics, BASF SE, Carl-Bosch-Strasse 38, 67056 Ludwigshafen, Germany*

[¶]*Digitalization of R&D, BASF SE, Carl-Bosch-Strasse 38, 67056 Ludwigshafen, Germany*

E-mail: *sriniker@ethz.ch

Contents

1 Additional Computational Details

Frequency calculations were performed using the Gaussian modeling software package[1] (compounds **1**-**4**, **13**, **14**, **16**-**18**: B3LYP/G**6-31-D3, compound **15**: B3LYP/G*6-31-D3, compounds **5**-**12**, **19**-**38**: BP86/cc-pVTZ-D3,[2–4] ultrafine grid, very tight conversion criteria). The choice of the functional/basis set combination is based on the following considerations: Pople double zeta basis sets are relatively cheap even for larger compounds. It is therefore possible to sample a higher number of conformers, which is crucial to obtain a good estimate of the conformational ensemble of flexible molecules and thus meaningful results. BP86/cc-pVTZ is relatively expensive, which makes the study of larger and flexible compounds cumbersome. However, this combination of functional and basis set is known to give good results in frequency calculations,[5–7] since it benefits from a favourable error cancellation. Thus, σ_2 (the "error" for the frequency shifting) can be chosen to be lower, providing better results. In general, large basis sets are to be preferred. However, for flexible molecules smaller basis sets may be used to avoid restricting the conformational search.

The following values were used for μ, σ_1, and σ_2 in the IRSA scoring function (Eq. (1) in the main text):

- $\sigma_1 = 1$ for compounds **1**-**4**, and $\sigma_1 = 0.2$ for compounds **5**-**38**.
- Values for μ and σ_2 were based on Ref. 6, specifically $\mu = (1.011 + 0.961)/2 = 0.986$ and $\sigma_2 = 1.011 - \mu = 0.025$ for B3LYP and the pople basis sets, and $\mu = (1.0325 + 0.9941)/2 = 1.0133$ and $\sigma_2 = 1.0325 - \mu = 0.0192$ for BP86/cc-pVTZ-D3.

μ is thus computed as the mean of the optimal anharmonic shifting values of fundamental (0.9941 for BP86/cc-pVTZ) and combination (1.0325 for BP86/cc-pVTZ) vibrations, and σ_2 is the span of these two values.[5–7] These values can be adjusted depending on the system studied.

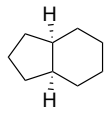
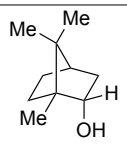
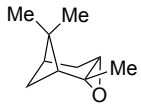
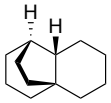
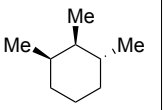
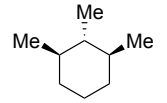
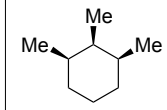
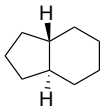
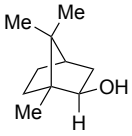
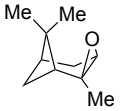
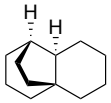
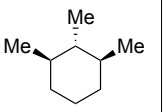
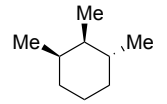
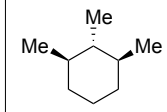
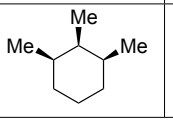
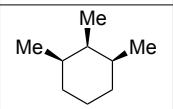
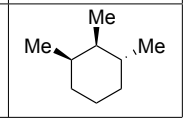
2 Number of Conformers

Table S1: Number of conformers per compound/isomer, which were considered in the computation of the final IR spectrum (those with negative frequencies were excluded, conformers with an energy difference of 10^{-4} kJ mol^{-1} were considered to be duplicates). The number in parentheses indicates the number of conformers with an energy below 10 kJ mol^{-1} for compounds **8** to **15**.

Compound	isomer$_0$	isomer$_1$	isomer$_2$	isomer$_3$	isomer$_4$	isomer$_5$	isomer$_6$	isomer$_7$
1	1							
2	1							
3	1							
4	1							
5	1	1						
6	3	3						
7	1	1						
8	2 (1)	4 (1)						
9	4 (1)							
10	4 (1)							
11	4 (1)							
12	283 (18)	210 (6)						
13	174 (5)	130 (8)	156 (4)	163 (4)				
14	34 (8)	31 (5)	34 (5)	21 (6)				
15	190 (21)	195 (11)	195 (4)	196 (9)	195 (10)	191 (1)	196 (4)	196 (6)
16	1							
17	1							
18	1							
19	1							
20	1							
21	1							
22	1							
23	1							
24	1							
25	1							
26	1							
27	13							
28	12							
29	13							
30	6							
31	4							
32	3							
33	1							
34	1							
35	1							
36	2							
37	2							
38	2							

3 Definition of Isomers of Compounds 5–15

Table S2: Diastereoisomers of compounds **5**–**11**.

Compound	5	6	7	8	9	10	11
$isomer_0$							
$isomer_1$							
$isomer_2$							

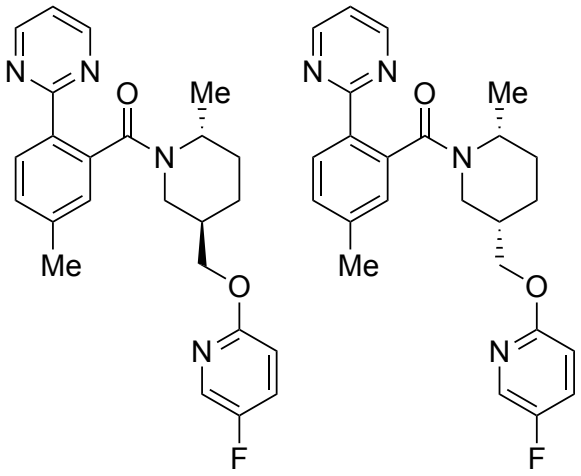


Figure S1: Filorexant (**12**): $isomer_0$ (left) and $isomer_1$ (right).

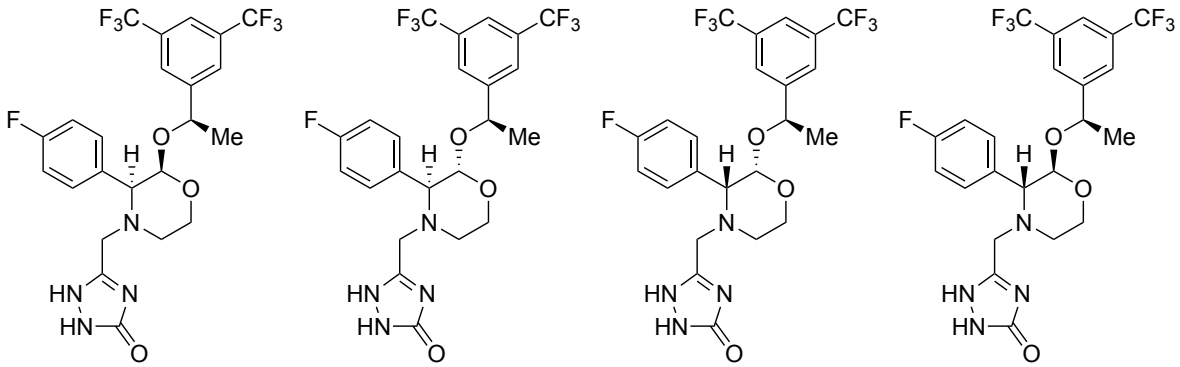


Figure S2: Aprepitant (**13**): from left to right $isomer_0$ to $isomer_3$.

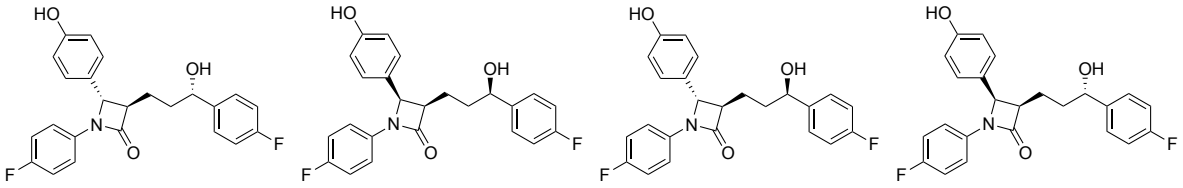


Figure S3: Ezetimibe (**14**): from left to right isomer$_0$ to isomer$_3$.

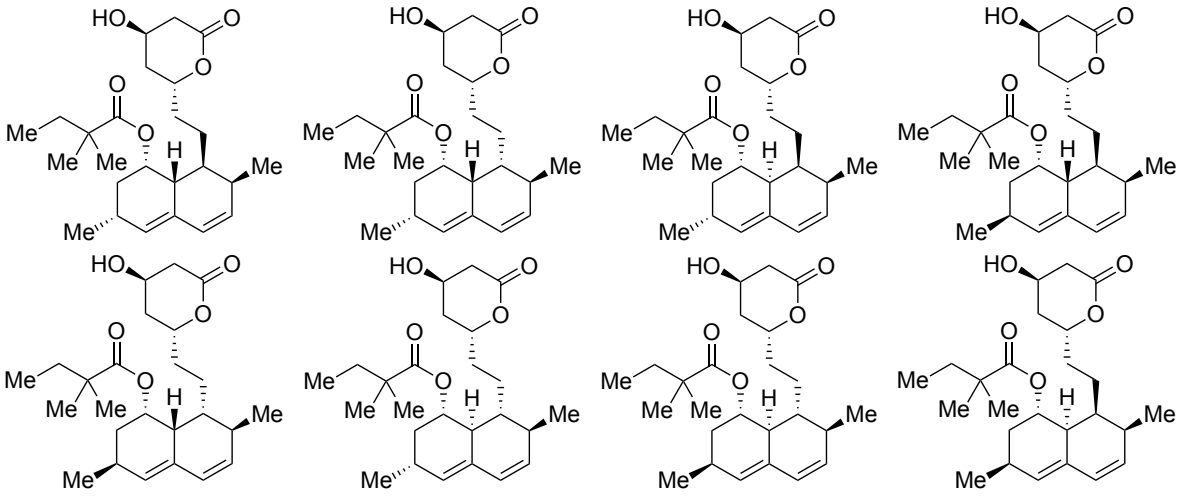


Figure S4: Simvastatin (**15**): (top) from left to right isomer$_0$ to isomer$_3$, (bottom) from left to right isomer$_4$ to isomer$_7$.

4 Pearson Correlation Coefficients and Alignment Scores for Compounds 5–15

Table S3: IRSA Pearson correlation coefficient r_p for compounds **5**–**15**. Isomer$_0$ corresponds to the correct isomer, whereas isomer$_1$ to isomer$_5$ are the other diastereoisomers.

Compound	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
isomer$_0$	0.16	0.88	0.94	0.35	0.79	0.76	0.93	0.79	0.90	0.91	0.83
isomer$_1$	0.16	0.58	0.67	0.24	0.83	0.66	0.46	0.77	0.94	0.70	0.79
isomer$_2$					0.63	0.56	0.75		0.93	0.62	0.70
isomer$_3$									0.85	0.12	0.56
isomer$_4$											0.62
isomer$_5$											0.53
isomer$_6$											0.65
isomer$_7$											0.69

Table S4: IRSA alignment score s for compounds **5**–**15**. Isomer$_0$ corresponds to the correct isomer, whereas isomer$_1$ to isomer$_5$ are the other diastereoisomers.

Compound	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
isomer$_0$	0.61	0.49	0.51	0.67	0.37	0.25	0.47	0.29	0.43	0.18	0.23
isomer$_1$	0.24	0.30	0.50	0.25	0.33	0.23	0.21	0.26	0.38	0.18	0.18
isomer$_2$					0.17	0.12	0.35		0.36	0.19	0.14
isomer$_3$									0.32	0.16	0.15
isomer$_4$											0.15
isomer$_5$											0.10
isomer$_6$											0.16
isomer$_7$											0.09

5 Spectral Data for Compounds 1–4

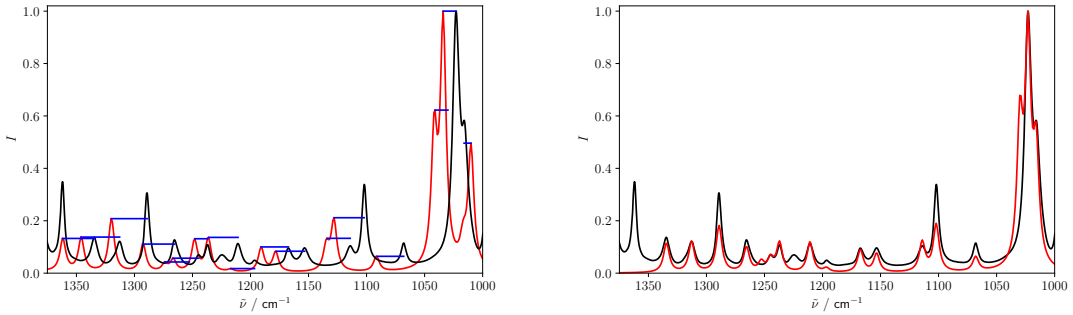


Figure S5: (Left): Experimental (black) and theoretical (red) IR spectra of fenchone (**1**). The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectrum was calculated with B3LYP/G**6-31-D3.

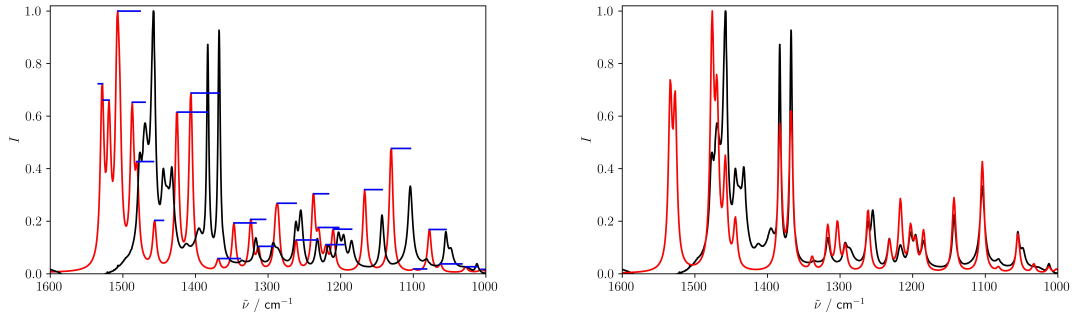


Figure S6: (Left): Experimental (black) and theoretical (red) IR spectra of β-pinene (**2**). The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectrum was calculated with B3LYP/G**6-31-D3.

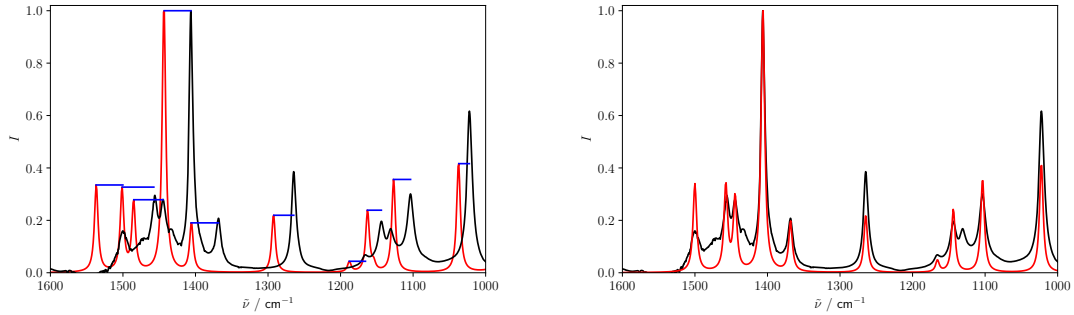


Figure S7: (Left): Experimental (black) and theoretical (red) IR spectra of propylene oxide (**3**). The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectrum was calculated with B3LYP/G**6-31-D3.

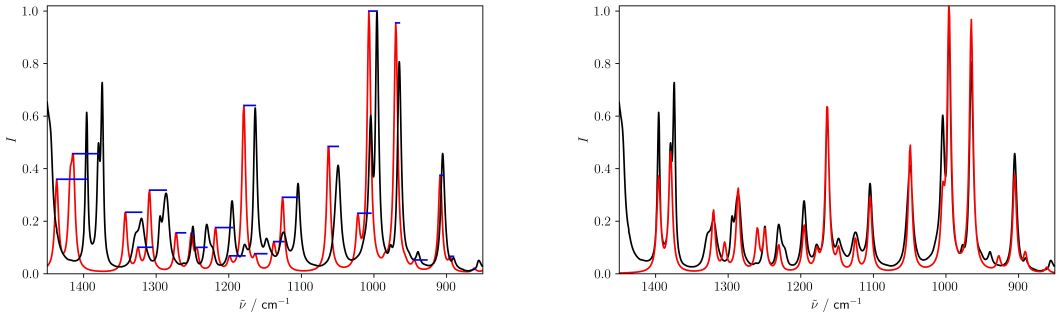


Figure S8: (Left): Experimental (black) and theoretical (red) IR spectra of camphor quinone (**4**). The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectrum was calculated with B3LYP/G**6-31-D3.

6 Spectral Data for Compounds 5–15

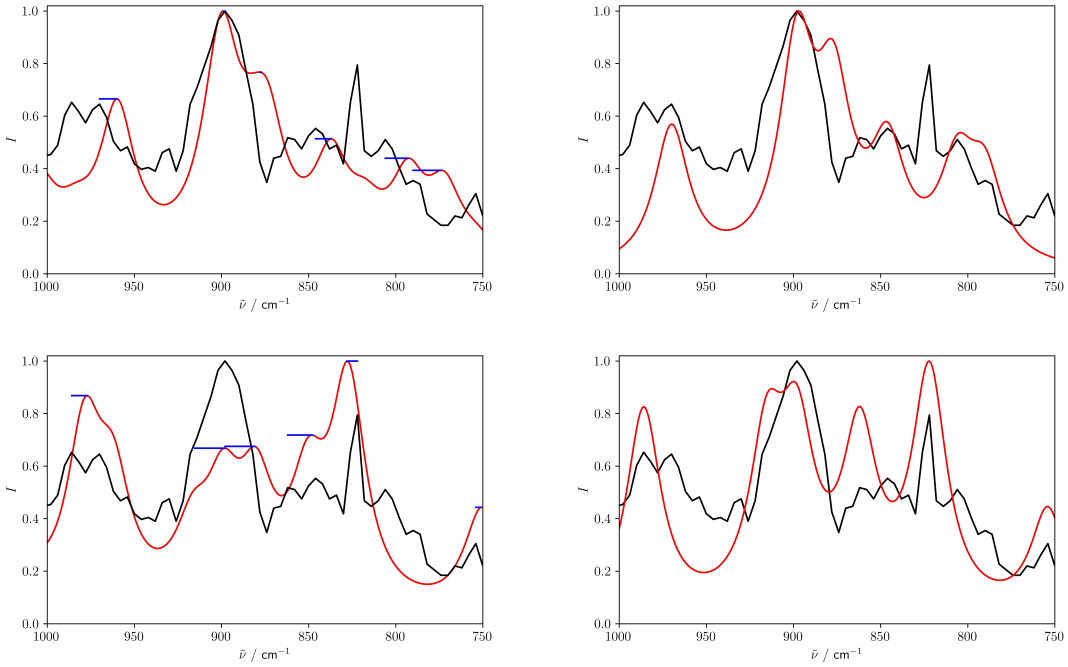


Figure S9: Isomer$_0$ (top) and isomer$_1$ (bottom) of bicyclo[4.3.0]nonane (**5**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase and taken from NIST SRD 35 database.[9] The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

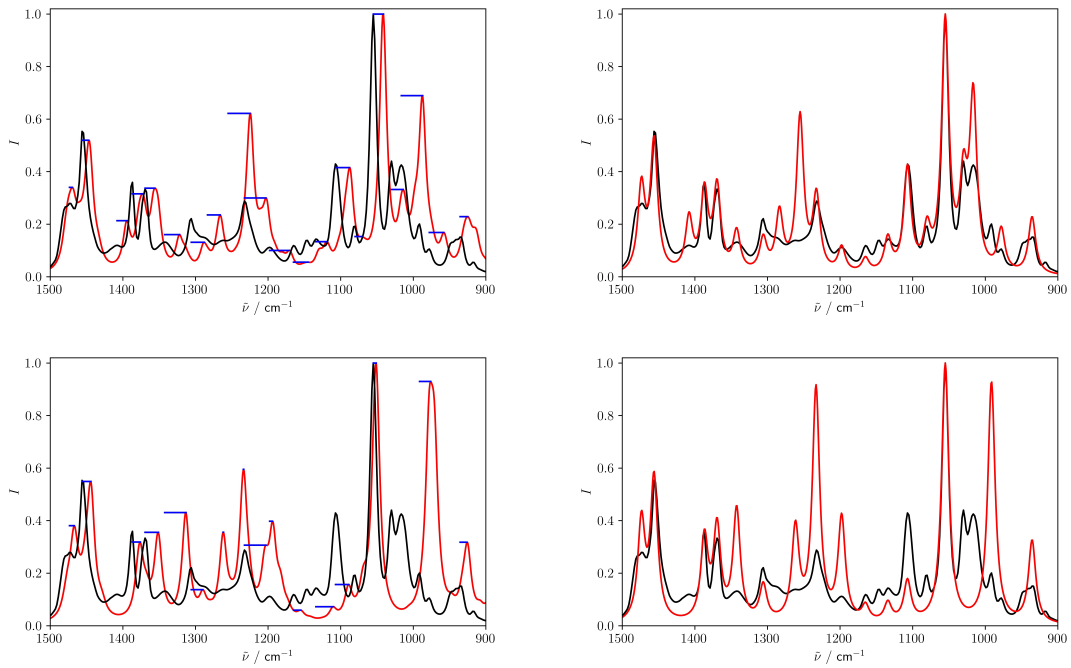


Figure S10: Isomer$_0$ (top) and isomer$_1$ (bottom) of borneol (**6**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

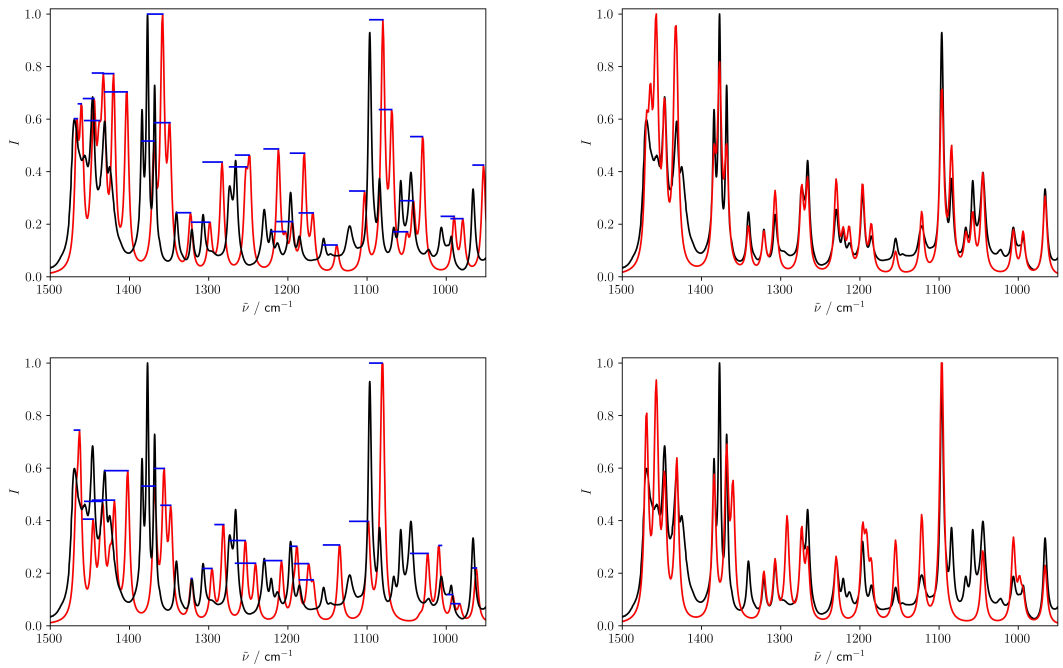


Figure S11: Isomer$_0$ (top) and isomer$_1$ (bottom) of pinene oxide (**7**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

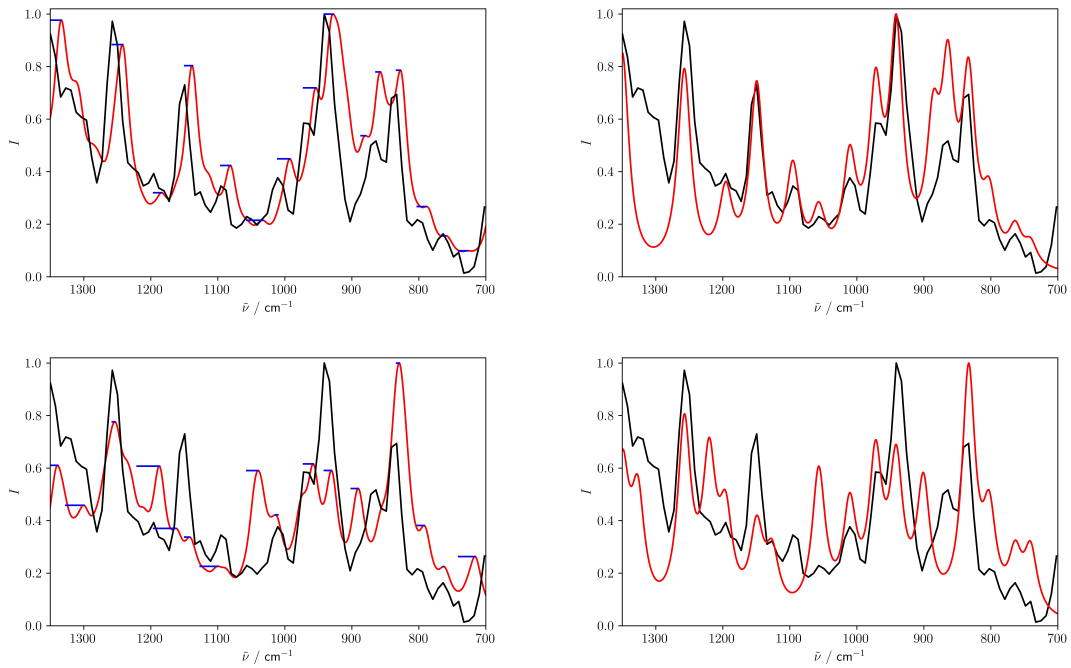


Figure S12: Isomer$_0$ (top) and isomer$_1$ (bottom) of bridged decalin (**8**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase (GC-IR). The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

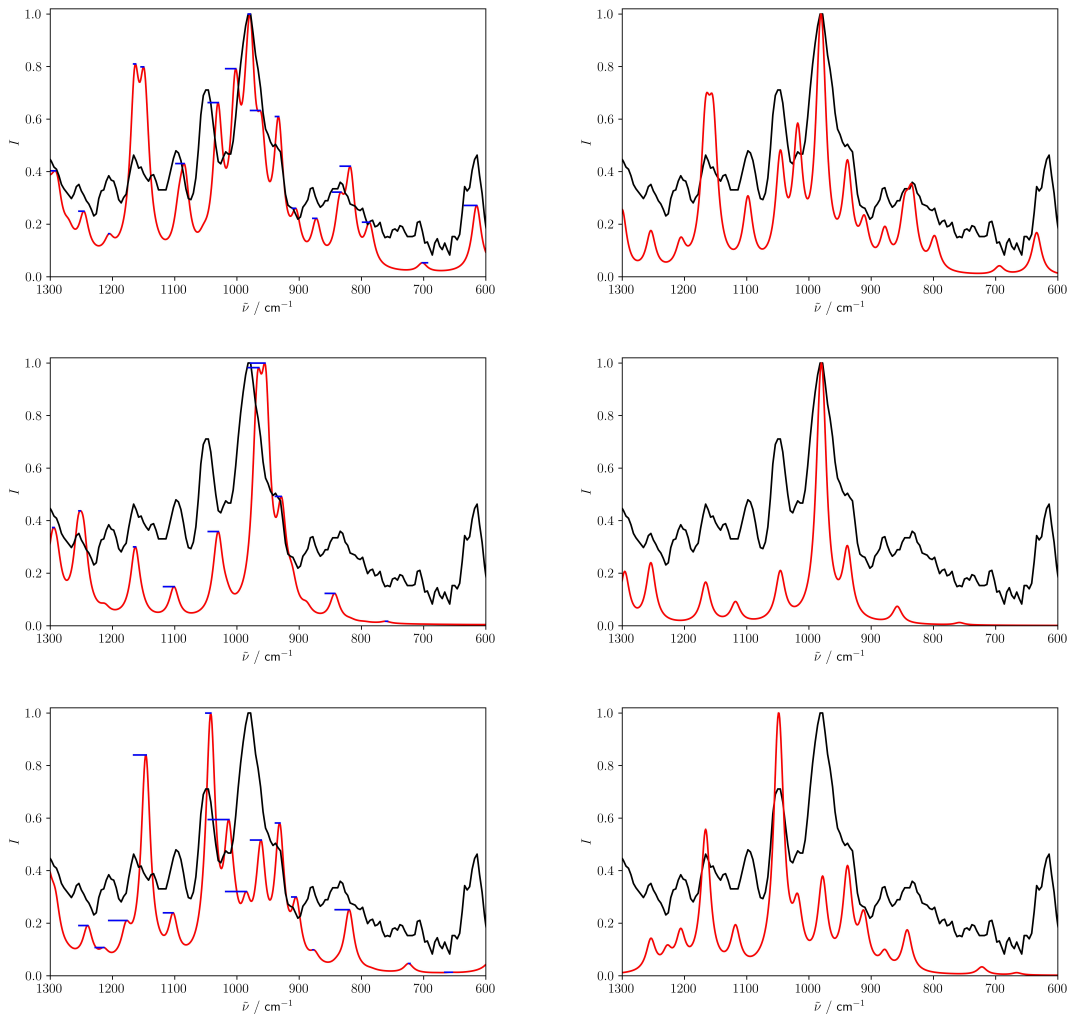


Figure S13: Isomer$_0$ (top), isomer$_1$ (middle) and isomer$_2$ (bottom) of 1,2,3-trimethylcyclohexane (**9**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase and taken from NIST SRD 35 database.[9] The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

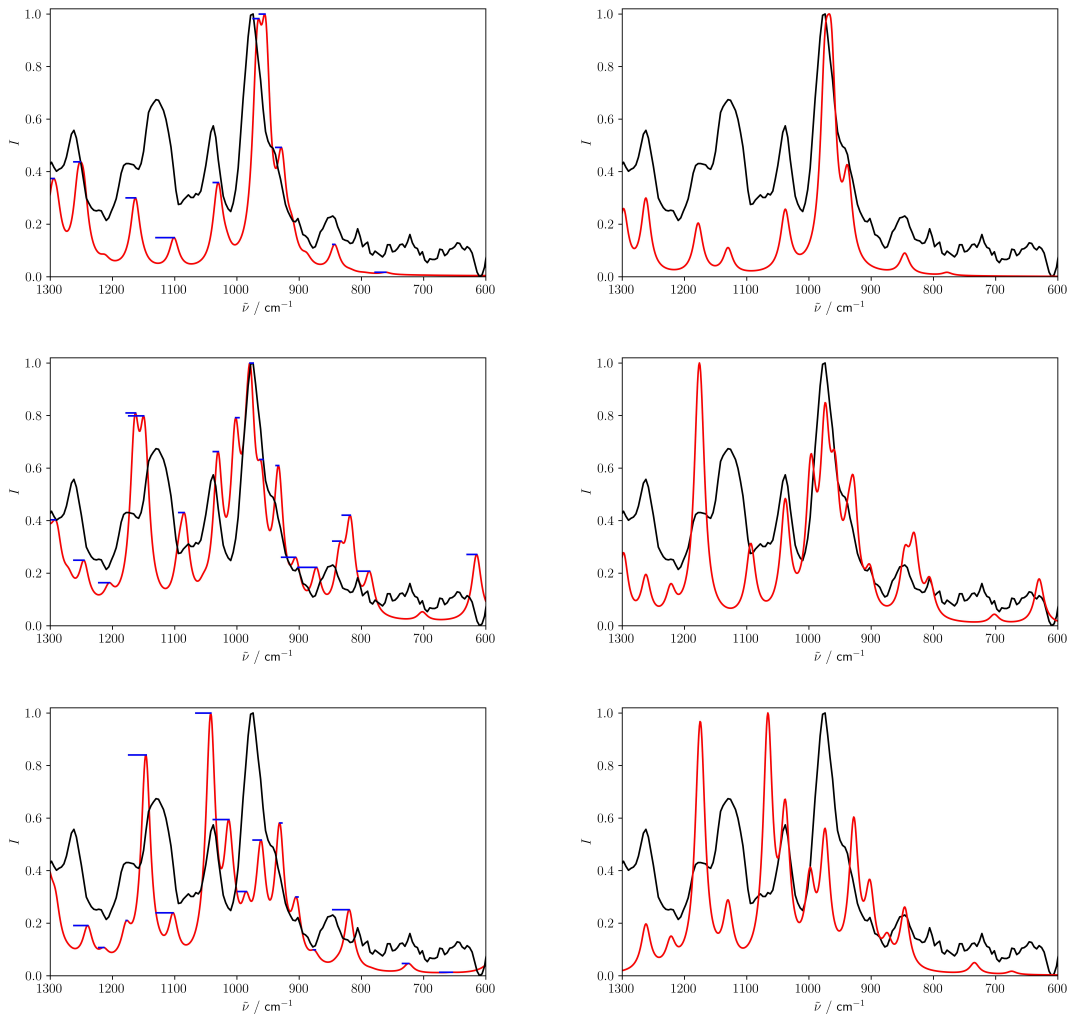


Figure S14: Isomer$_0$ (top), isomer$_1$ (middle) and isomer$_2$ (bottom) of 1,2,3-trimethylcyclohexane (**10**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase and taken from NIST SRD 35 database.[9] The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

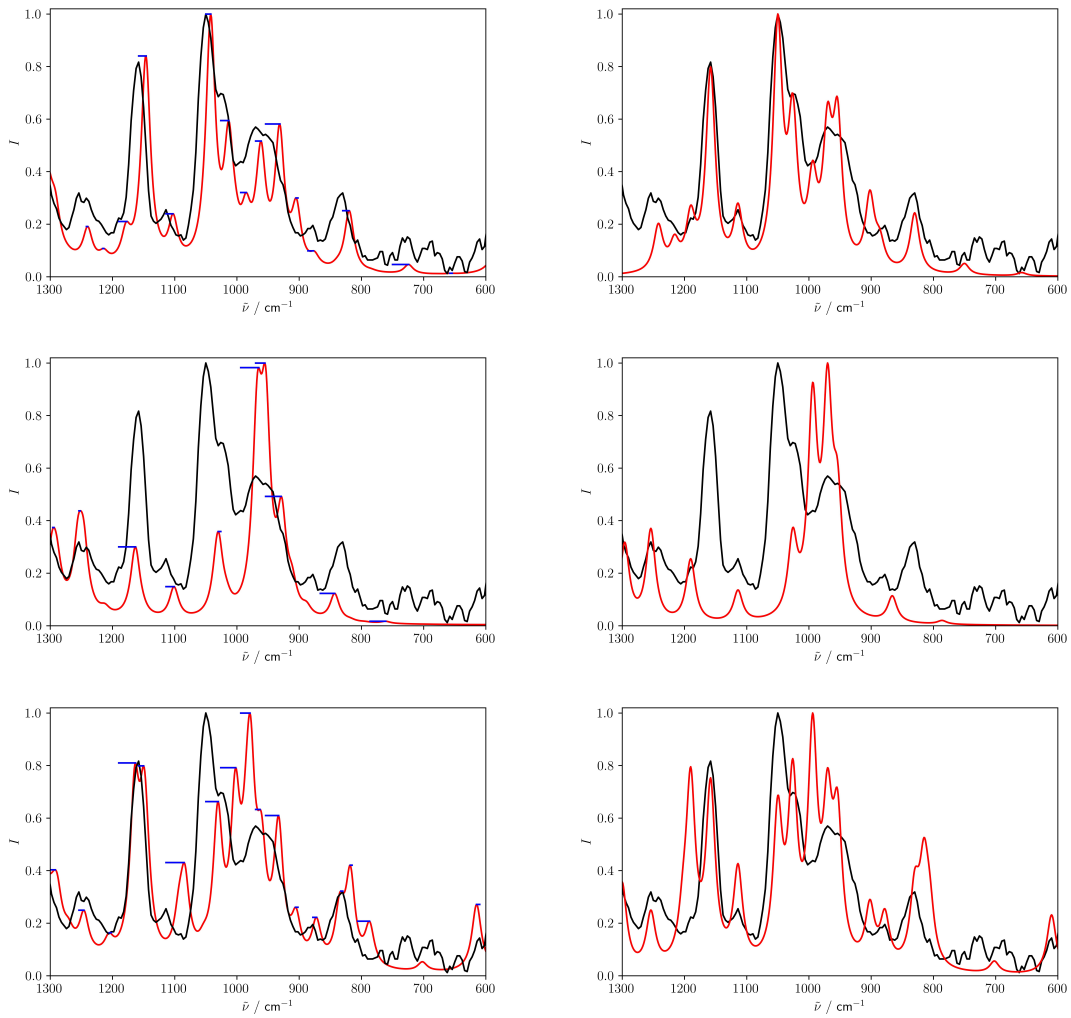


Figure S15: Isomer$_0$ (top), isomer$_1$ (middle) and isomer$_2$ (bottom) of 1,2,3-trimethylcyclohexane (**11**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase and taken from NIST SRD 35 database.[9] The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

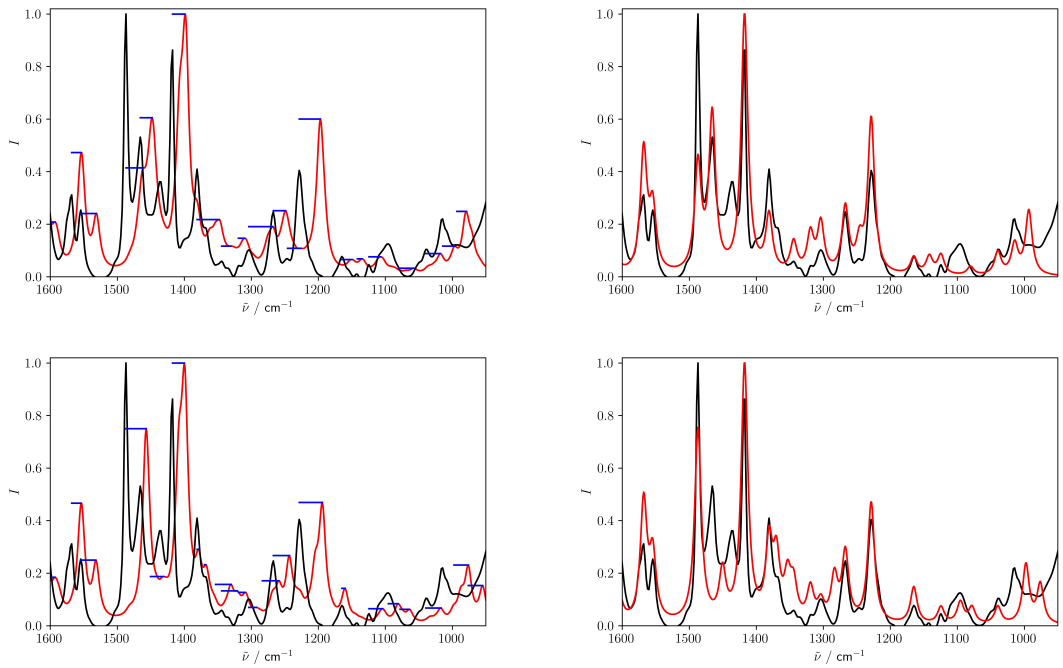


Figure S16: Isomer$_0$ (top) and isomer$_1$ (bottom) of filorexant (**12**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 10. The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

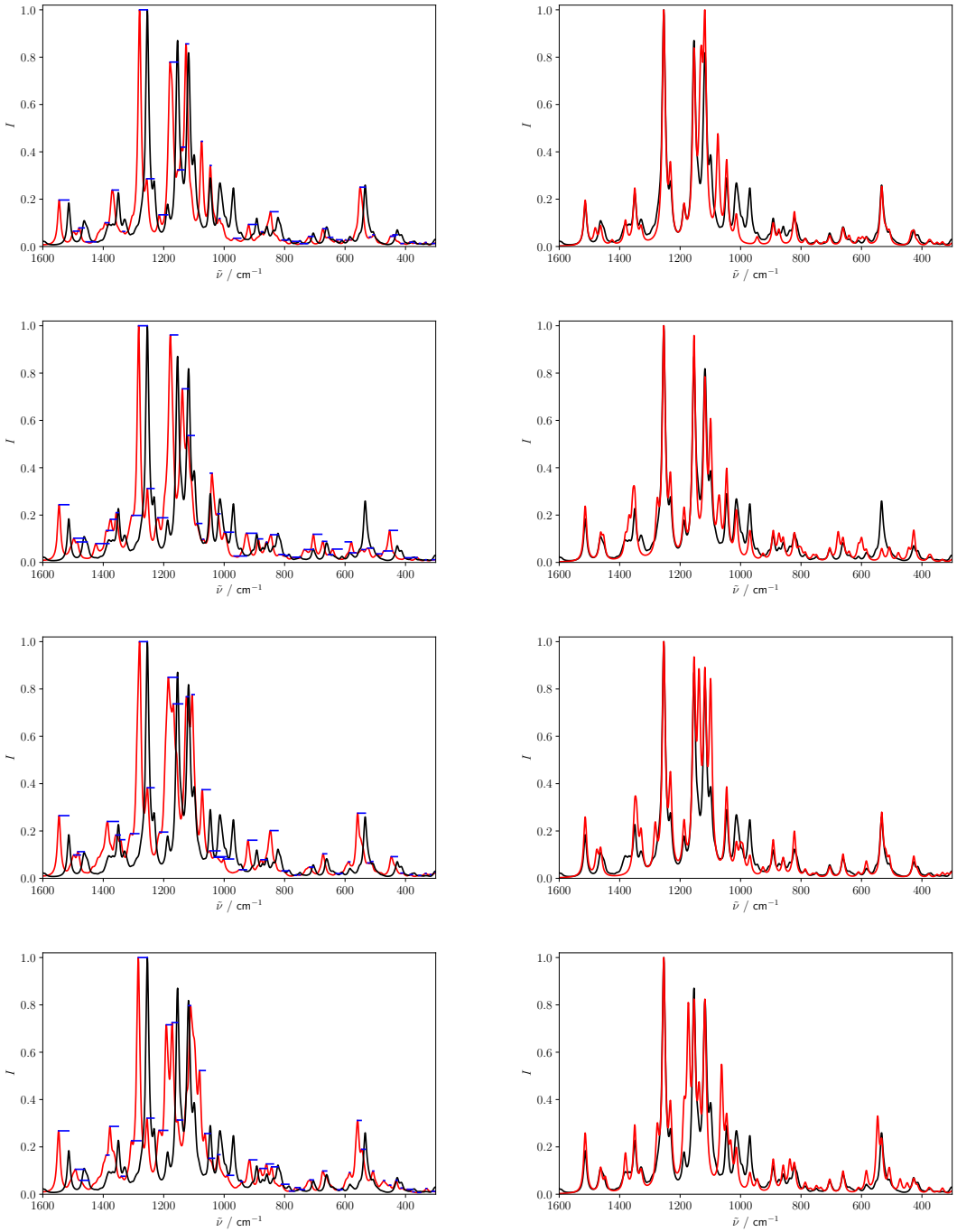


Figure S17: Isomer$_0$ (first row), isomer$_1$ (second row), isomer$_2$ (third row) and isomer$_3$ (fourth row) of aprepitant (**13**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 10. The theoretical spectra were calculated with B3LYP/G**6-31-D3.

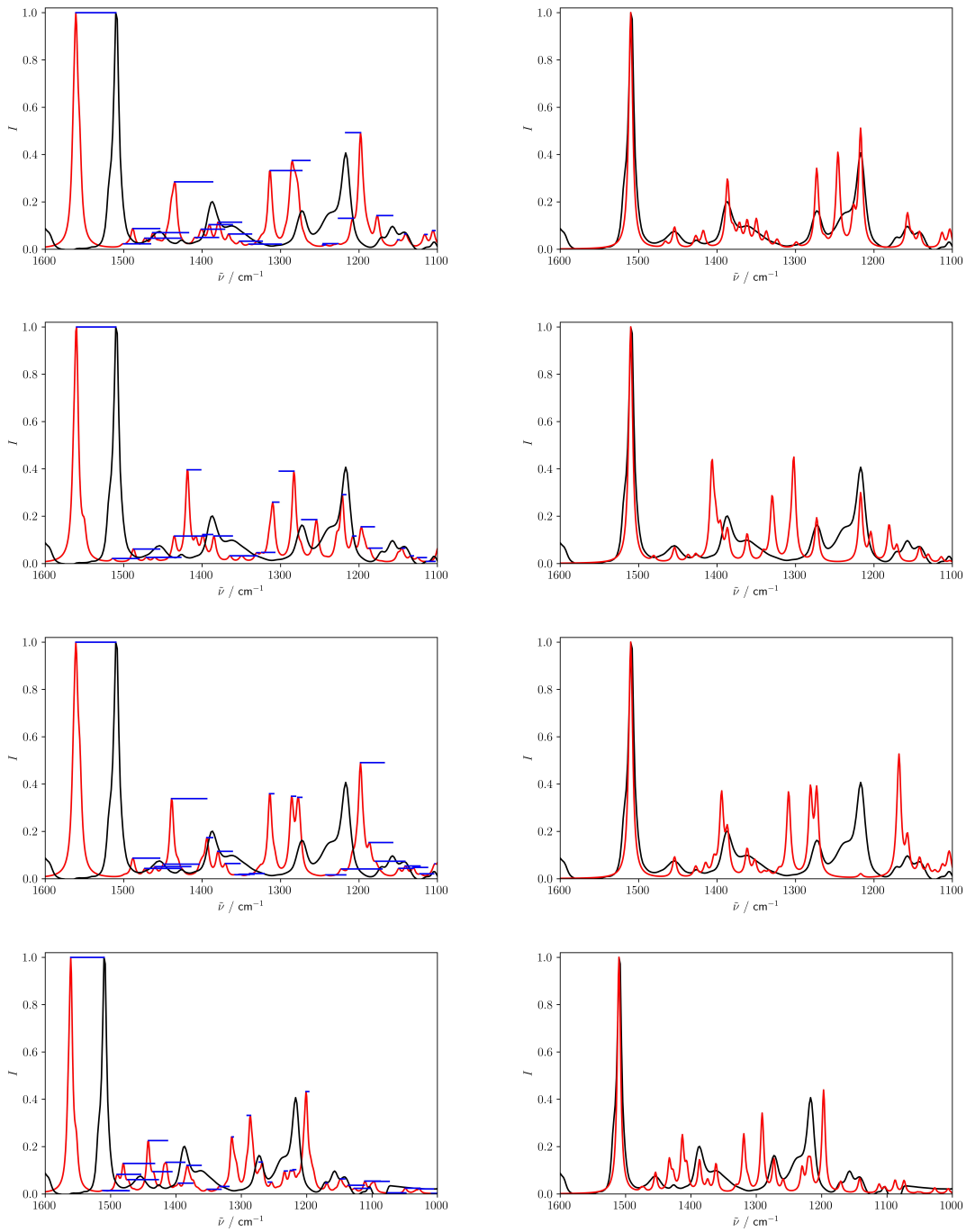


Figure S18: Isomer$_0$ (first row), isomer$_1$ (second row), isomer$_2$ (third row) and isomer$_3$ (fourth row) of ezetimibe (**14**). (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 10. The theoretical spectra were calculated with B3LYP/G**6-31-D3.

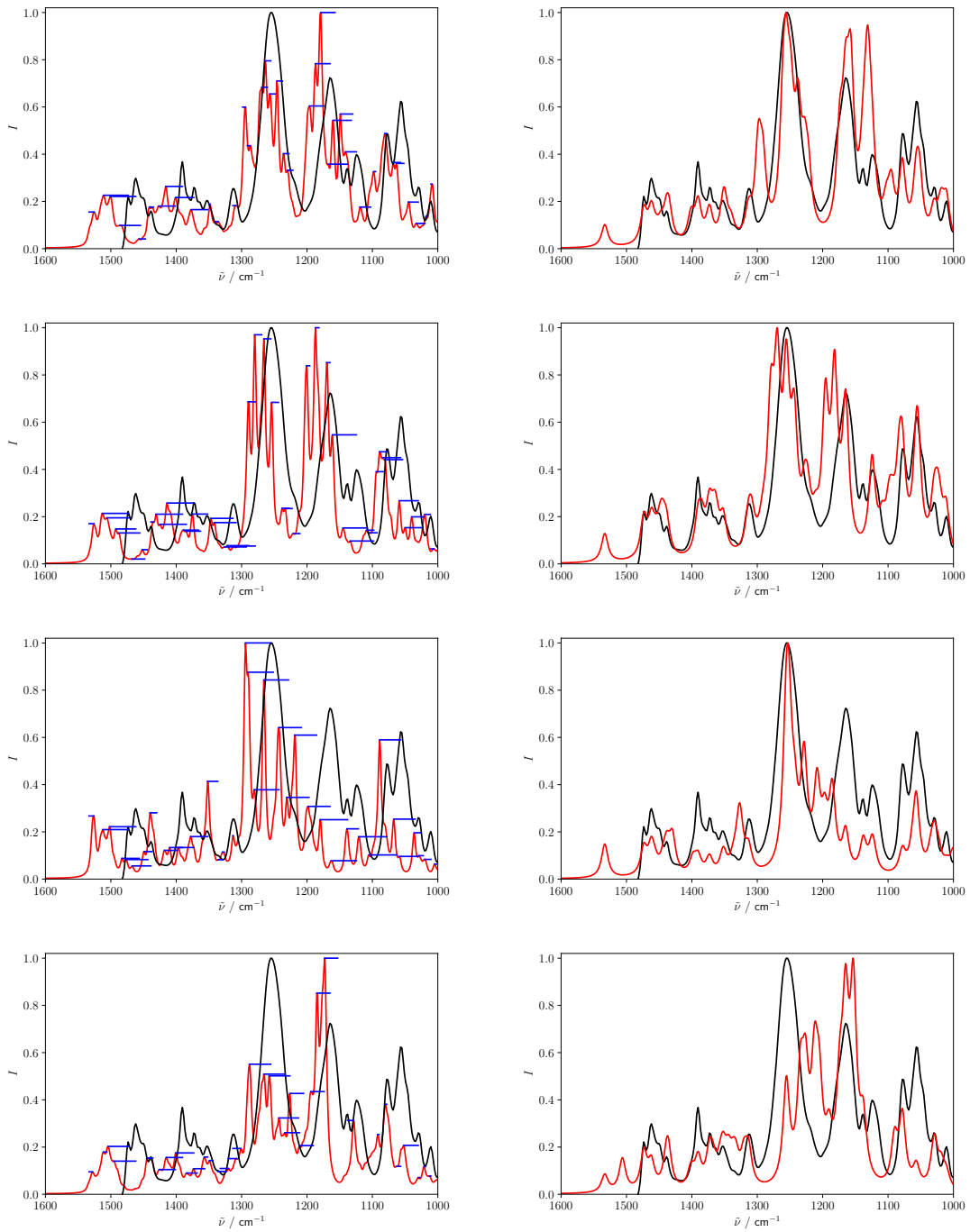


Figure S19: Simvastatin (**15**): isomer$_0$ (first row), isomer$_1$ (second row), isomer$_2$ (third row), and isomer$_3$. (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 10.

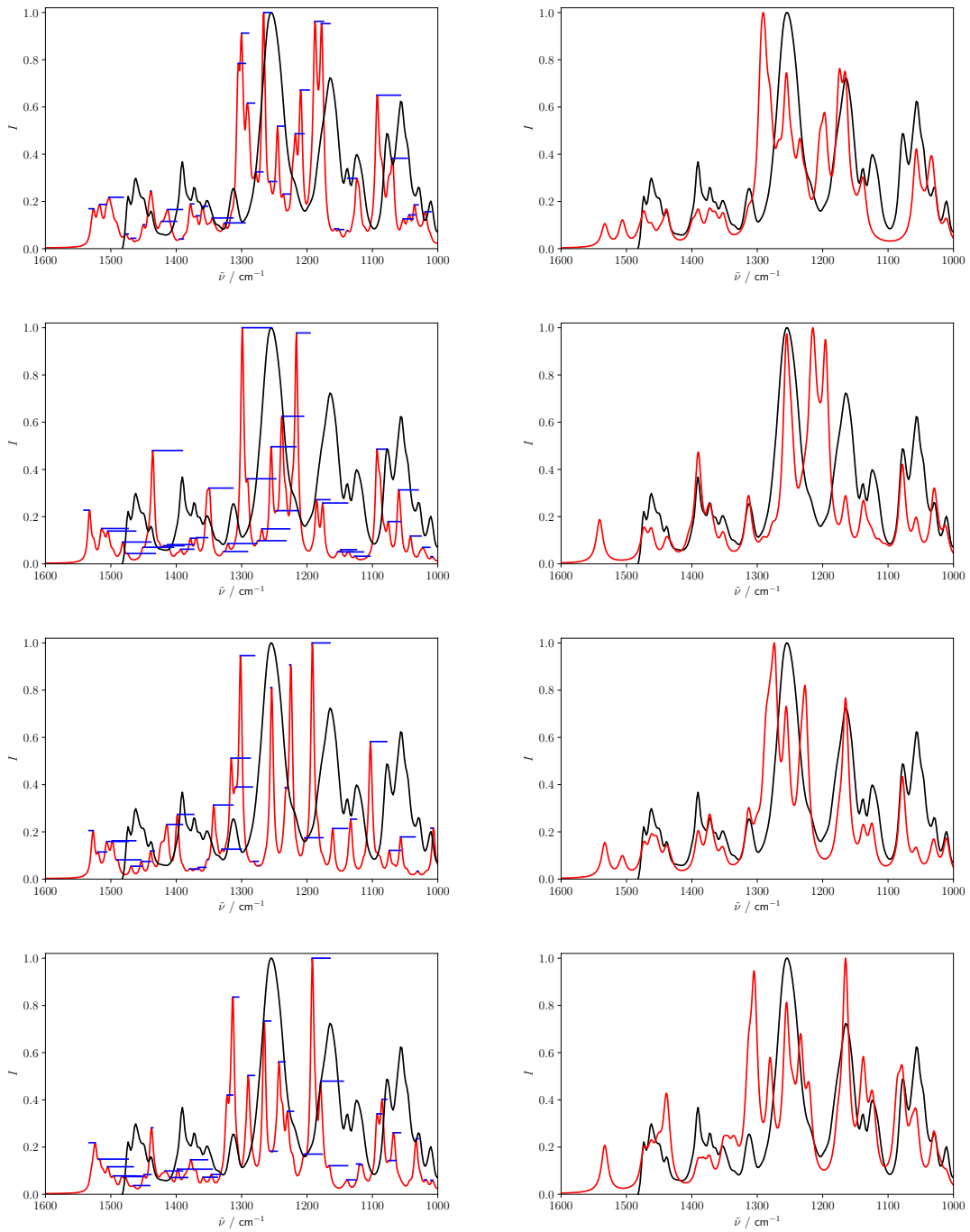


Figure S20: Simvastatin (**15**): isomer$_4$ (first row), isomer$_5$ (second row), isomer$_6$ (third row), and isomer$_7$. (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in solution and taken from Ref. 10. The theoretical spectra were calculated with B3LYP/G*6-31-D3.

7 Spectral Data of Compounds 16 and 17

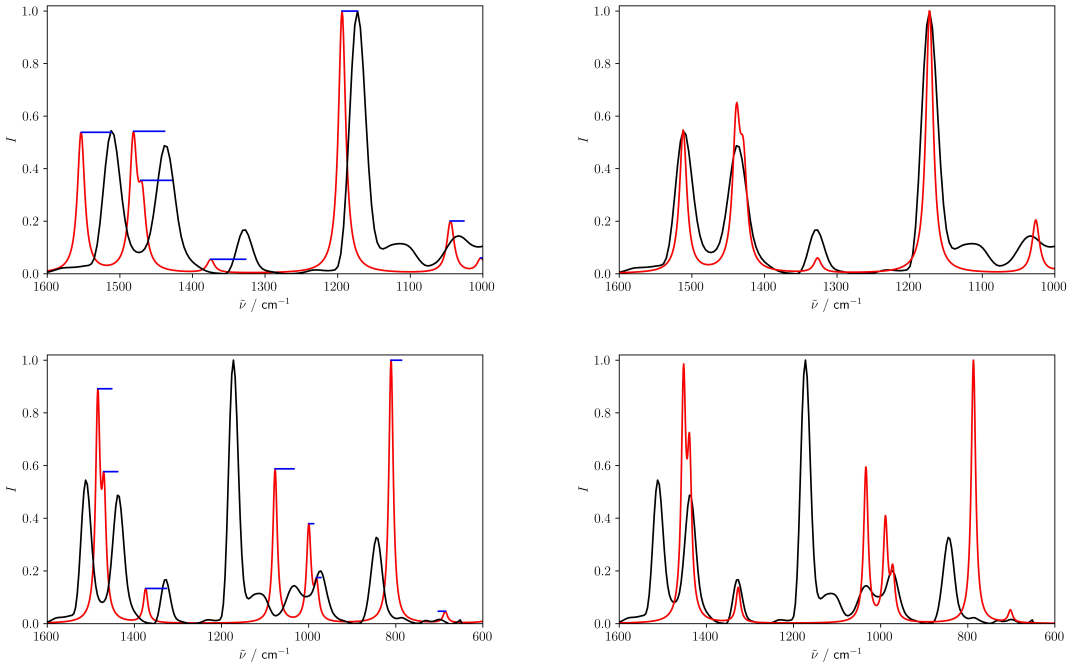


Figure S21: Z-isomer (top) and E-isomer (bottom) of compound **16**. (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase (GC-IR).

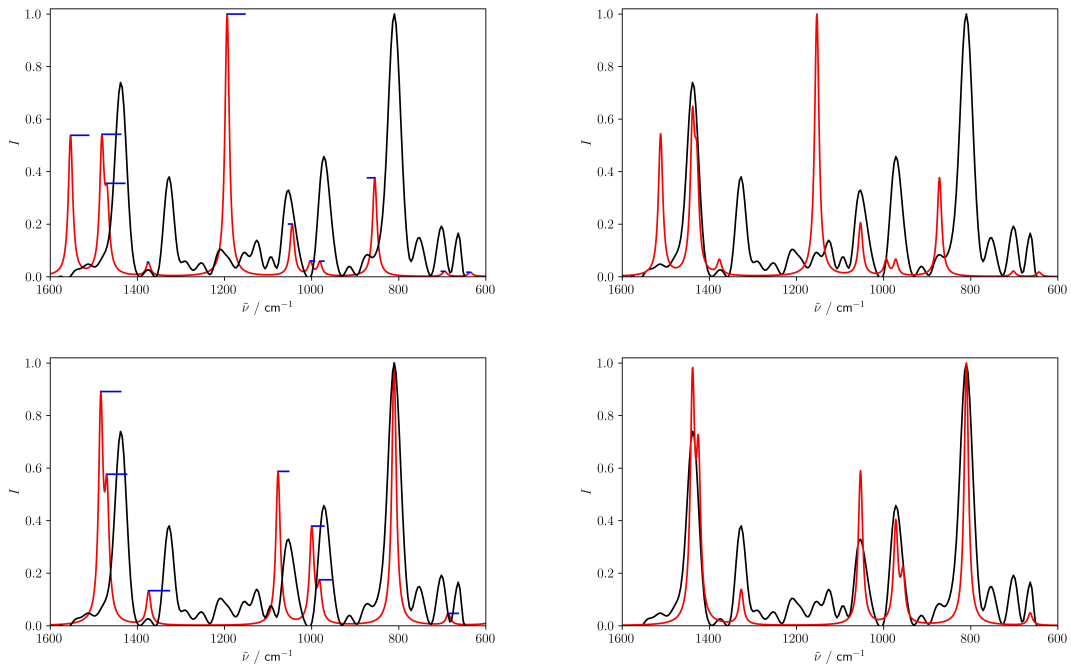


Figure S22: E-isomer (top) and Z-isomer (bottom) of compound **17**. (Left): Unaligned experimental (black) and theoretical (red) IR spectra. The blue lines correspond to the peak assignment made by the algorithm. (Right): Overlap between the experimental (black) and theoretical (red) IR spectra after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectrum was measured in the gas phase (GC-IR).

8 Spectral Data of Compounds 18–26

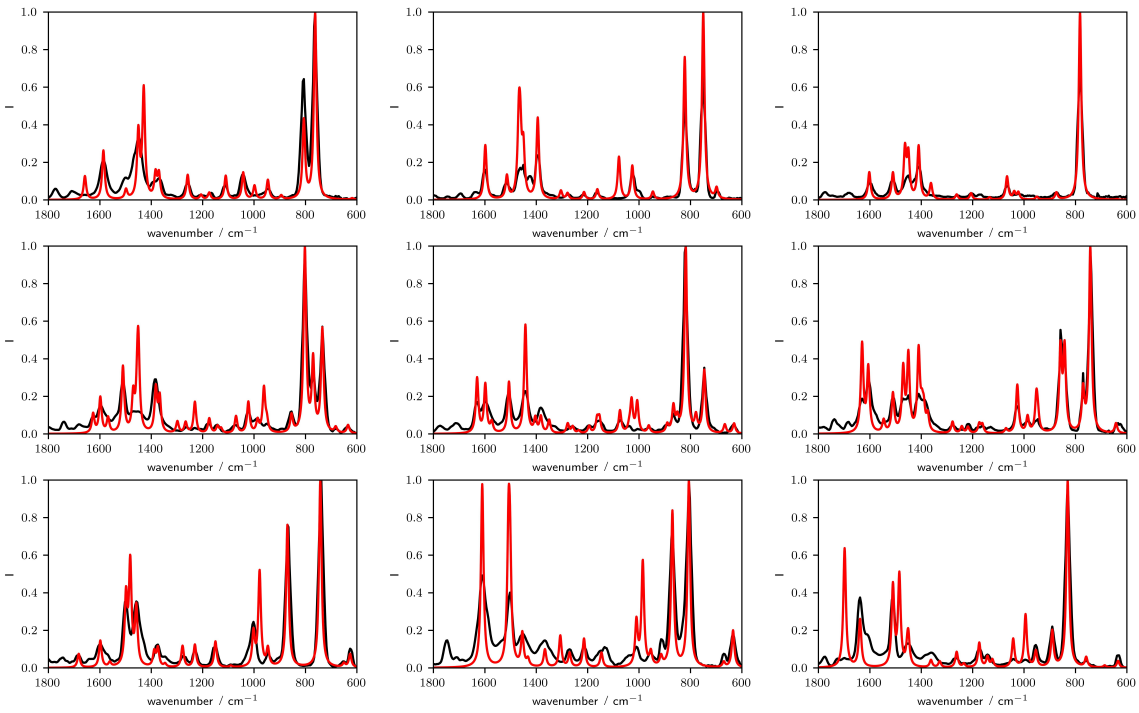


Figure S23: Top row: compounds **18** (left), **19** (mid) and **20** (right). Middle row: compounds **21** (left), **22** (mid) and **23** (right). Bottom row: compounds **24** (left), **25** (mid) and **26** (right). Overlap between the experimental (black) and theoretical (red) IR spectra of the correct isomer (isomer$_0$) after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectra were measured in the gas phase and taken from NIST SRD 35 database.[9] The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

9 Spectral Data of Compounds 27–38

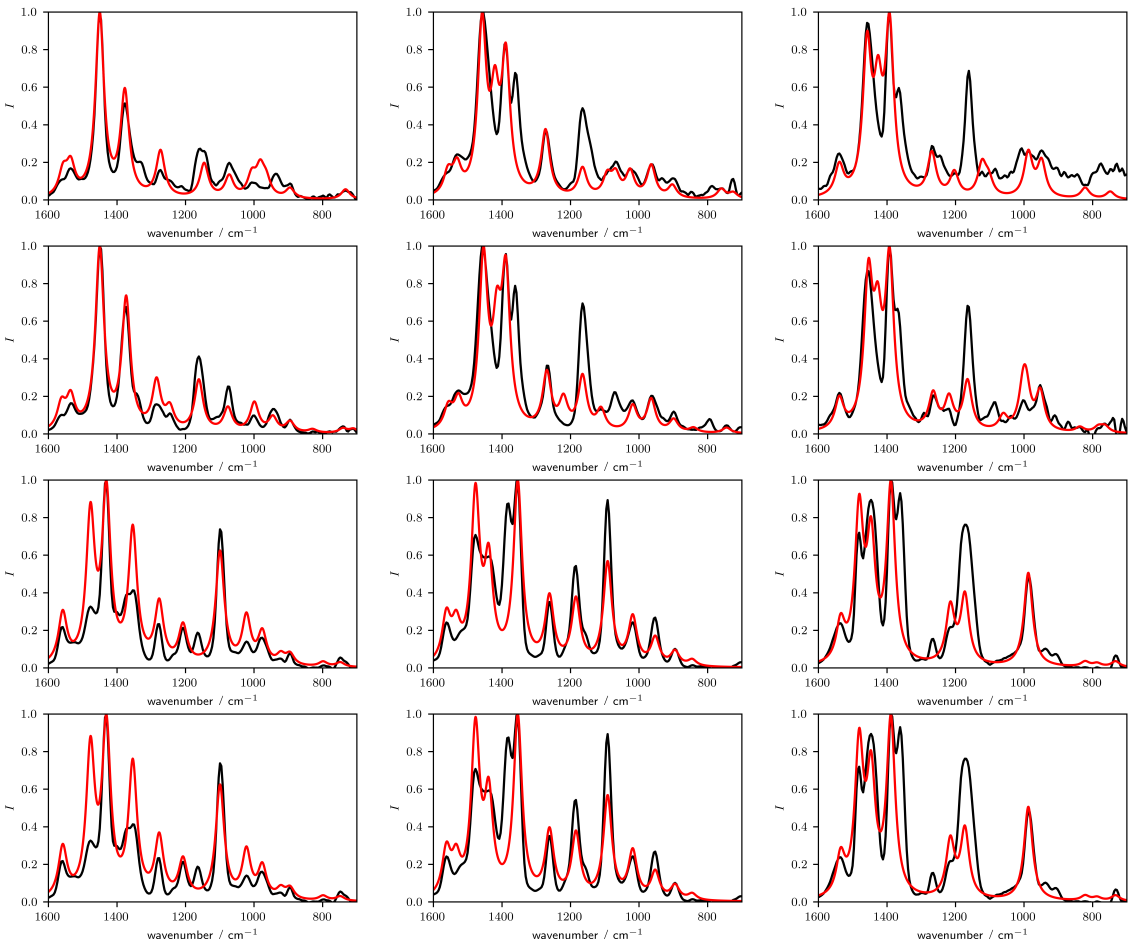


Figure S24: Top row: compounds **27** (left), **28** (mid) and **29** (right). Second row: compounds **30** (left), **31** (mid) and **32** (right). Third row: compounds **33** (left), **34** (mid) and **35** (right). Bottom row: compounds **36** (left), **37** (mid) and **38** (right). Overlap between the experimental (black) and theoretical (red) IR spectra of the correct isomer (isomer$_0$) after shifting the theoretical spectrum based on the IRSA matching. The experimental IR spectra were measured in the gas phase (GC-IR). The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

10 Combined Analysis of IR and VCD Spectra for Borneol

If both experimental IR and VCD spectra are available the alignment and analysis can be performed in a combined manner using the scoring function,

$$s_{e,t}(\tilde{\nu}_e, \tilde{\nu}_t; \sigma_1, \sigma_2, \mu) = e^{\frac{1}{2}\frac{(\min(\frac{I_{e,\mathrm{VCD}}}{I_{t,\mathrm{VCD}}}, \frac{I_{t,\mathrm{VCD}}}{I_{e,\mathrm{VCD}}})-1)^2}{\sigma_1^2}} \cdot e^{-\frac{1}{2}\frac{(\min(\frac{I_e}{I_t}, \frac{I_t}{I_e})-1)^2}{\sigma_1^2}} \cdot e^{-\frac{1}{2}\frac{(\frac{\tilde{\nu}_e}{\tilde{\nu}_t}-\mu)^2}{\sigma_2^2}} . \quad (1)$$

The output of the alignment program is shown for the correct and incorrect diastereomer of borneol (**6**) as example in Figures S25 and S26.

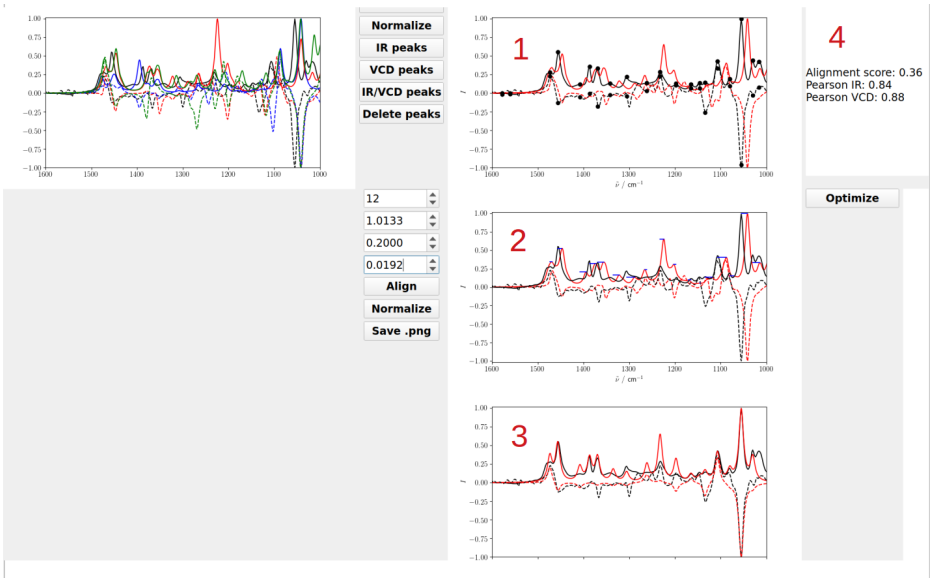


Figure S25: Combined IR/VCD analysis of (+)-borneol (**6**). (1): Boltzmann weighted theoretical IR (red solid line) and VCD spectra (red dashed line), as well as the experimental IR (black solid line) and VCD (black dashed line) spectra. The black dots mark the experimental peaks selected automatically in the IR spectrum, and the corresponding peaks in the VCD spectrum. (2): The matched peak assignments performed by the algorithm. (3): The aligned spectra. (4): The alignment score, and the Pearson correlation coefficient of the VCD spectra and the IR spectra are in good agreement. Note that also the absolute stereochemistry can directly be determined here, since the correlation of the VCD spectra is strongly positive. The experimental IR spectrum was measured in solution and taken from Ref. 8. The theoretical spectra were calculated with BP86/cc-pVTZ-D3.

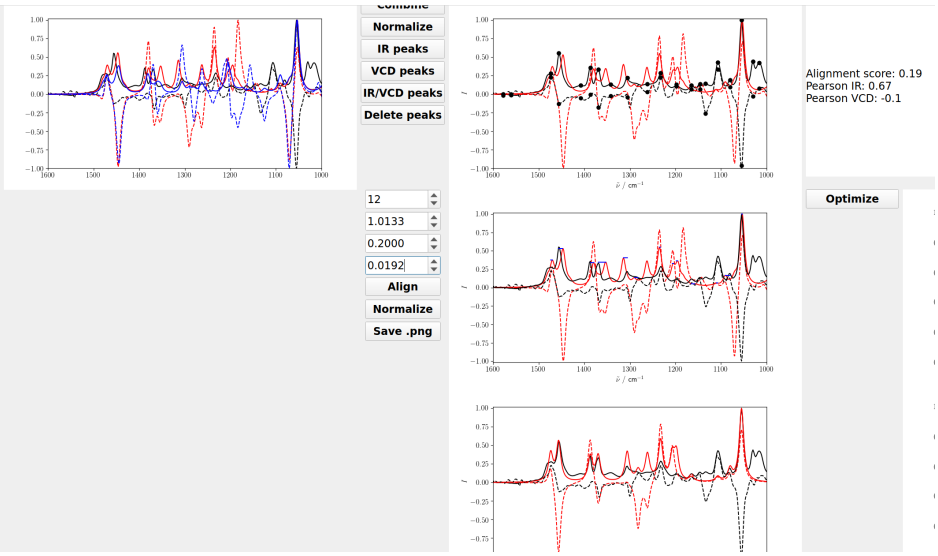


Figure S26: Same analysis performed with the other diastereomer.

References

(1) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery,; Jr.,; A., J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 09. Gaussian, Inc.: Wallingford, CT, 2004.

(2) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange With Density Functional Approximations. *J. Chem. Phys.* **1996**, *105*, 9982–9985.

(3) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724–728.

(4) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, S. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-d) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(5) Sinha, P.; Boesch, S. E.; Gu, C.; Wilson, A. K. Harmonic Vibrational Frequencies: Scaling Factors for HF, B3LYP, and MP2 Methods in Combination with Correlation Consistent Basis Sets. *J. Phys. Chem. A.* **2004**, *108*, 92113–92117.

(6) Merrick, J. P.; Moran, D.; Radom, L. An Evaluation of Harmonic Vibrational Frequency Scale Factors. *J. Phys. Chem. A.* **2007**, *111*, 11683–11700.

(7) Kesharwin, M. K.; Brauer, B.; Martin, J. M. L. Frequency and Zero-Point Vibrational Energy Scale Factors for Double-Hybrid Density Functionals (and Other Selected Methods): Can Anharmonic Force Fields Be Avoided? *J. Phys. Chem. A.* **2015**, *119*, 1701–1714.

(8) Böselt, L.; Sidler, D.; Kittelman, T.; Stohner, J.; Zindel, D.; Wagner, T.; Riniker, S. Determination of Absolute Stereochemistry of Flexible Molecules Using A Vibrational Circular Dichroism Spectra Alignment Algorithm. *J. Chem. Inf. Model.* **2019**, *59*, 1826–1838.

(9) Clifton, C.; Gallagher, J.; Shamin, A.; Stein, S. NIST Standard Reference Database 35. 2007.

(10) Sherer, E. C.; Lee, C. H.; Shpungin, J.; Cuff, J. F.; Da, C.; Ball, R.; Bach, R.; Crespo, A.; Gong, X.; Welch, C. J. Systematic Approach to Conformational Sampling for Assigning Absolute Configuration Using Vibrational Circular Dichroism. *J. Med. Chem.* **2014**, *57*, 477–494.