PCCP

ROYAL SOCIETY OF CHEMISTRY

ARTICLE TYPE

DOI:10.1039/c9cp05682c

ELECTRONIC SUPPLEMENTARY INFORMATION (ESI):

Quantitative Structural Determination of Monosaccharides in Solution from Raman and Raman Optical Activity Spectra

Vladimír Palivec,[a‡] Vladimír Kopecký Jr.,[b] Pavel Jungwirth,[a] Petr Bouř,[a] Jakub Kaminský,[ac*] and Hector Martinez-Seara[a*]

Contents

[a] *Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Flemingovo náměstí 542/2, CZ166 10, Prague 6, Czech Republic; E-mail: jakub.kaminsky@uochb.cas.cz; hseara@gmail.com*

[b] *Institute of Physics, Faculty of Mathematics and Physics, Charles University, Ke Karlovu 5, CZ121 16, Prague 2, Czech Republic*

[c] *Gilead Sciences & IOCB Research Center*

‡ PhD. Student of Faculty of Sciences, Charles University in Prague

1 Simulation Approaches

Tables 1, 2, and 3 contain all used simulation protocols. The first line describes solvation protocol, the second line describes the QM level of theory used (DFT functional and basis set), and the third line describes optimization procedure which was used to refine the MD snapshots. Final Raman/ROA spectra were obtained using geometry which was obtained by given optimization procedure at the QM level of theory mentioned with a given protocol. All simulation spectra were generated using 100 snapshots.

Table 1 Calculation protocols used in this work — Part 1/3

Protocol	Description
`CPCM`	• CPCM continuum solvation • B3LYP/6-311++G** • 10 opt steps with HR
`MM`	• Water molecules closer than 12 Å as MM water molecules • B3LYP/6-311++G** • 10 opt steps in Cartesian coordinates
`MM_w`	• Water molecules closer than 12 Å as MM water molecules • B3LYP/6-311++G** • opt=water molecules frozen, sugar fully optimized, optimization in Cartesian coordinates
`hybrid`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • 10 opt steps
`hybrid_w`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • all water molecules frozen, saccharide optimized until gradient criteria reached
`hybrid_fo`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • optimized until gradient criteria reached
`opt_normal_modes`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • 50 opt steps with modes <300 cm^{-1} frozen [1]
`hybrid(X)`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • X opt steps

HR = harmonic restraints
CPCM = conductor-like polarizable continuum model

Table 2 Calculation protocols used in this work — Part 2/3

Protocol	Description
`dft1(hybrid)`	• Ater molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • 10 opt steps
`dft2`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • M06L/6-311++G** • 10 opt steps
`dft3`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • CAM-B3LYP/6-311++G** • 10 opt steps
`dft4`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • ωB97XD/6-311++G** • 10 opt steps
`bs1`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-31G* • 10 opt steps
`bs2(hybrid)`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/6-311++G** • 10 opt steps
`bs3`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • B3LYP/aug-cc-pvtz • 10 opt steps
`bs4`	• Water molecules closer than 3 Å as MM water molecules • CPCM continuum solvation • geometry and force field calculations B3LYP/6-31G* • Raman/ROA tensors B3LYP/rDPS [2] • 10 opt steps

HR = harmonic restraints
CPCM = conductor-like polarizable continuum model
rDPS basis set = 3-21++G with semidiffuse p(0.2) functions on hydrogens

Table 3 Calculation protocols used in this work — Part 3/3

Protocol	Description
cpcm_free	• CPCM • B3LYP/6-311++G** • full optimization without any restraints
cpcm_HR	• CPCM • B3LYP/6-311++G** • 10 opt steps with HR
hybrid_QM	• Water molecules closer than 3 Å as QM water molecules • CPCM • B3LYP/6-311++G**(water molecules:6-311G++**) • 10 opt steps
hybrid_QM_fo	• Water molecules closer than 3 Å as QM water molecules • CPCM • B3LYP/6-311++G**(water molecules:6-311++G**) • optimized until gradient criteria reached

HR = harmonic restraints
CPCM = conductor-like polarizable continuum model

In the hybrid_QM/hybrid_QM_fo simulation protocols (Table 3) the signal originating from water molecules was removed by setting the polarizability derivatives of all water molecules to zero by in house developed script[3]. When harmonic restraints were deployed we used Cartesian harmonic restraints with a force constant 10^{-6} Hartree·Å^{-2} on every atom.

2 Results of the CPCM/MM/hybrid/hybrid_w/hybrid_fo Approaches

Figures 1, 2, 3, 4, 5, and 6 show the performance of CPCM, MM, hybrid, hybrid_w, and hybrid_fo approaches (see Table 1) on the calculation of Raman/ROA spectra of all studied monosaccharides. The resulting spectra can be found in the middle, while scaling function with optimized parameters a, b, c, and d is to be found above. Below the spectrum we display the differences between experimental and simulation data. Note that the scaling function is the same for each pair of Raman and ROA spectra, but they are different between simulation protocols.

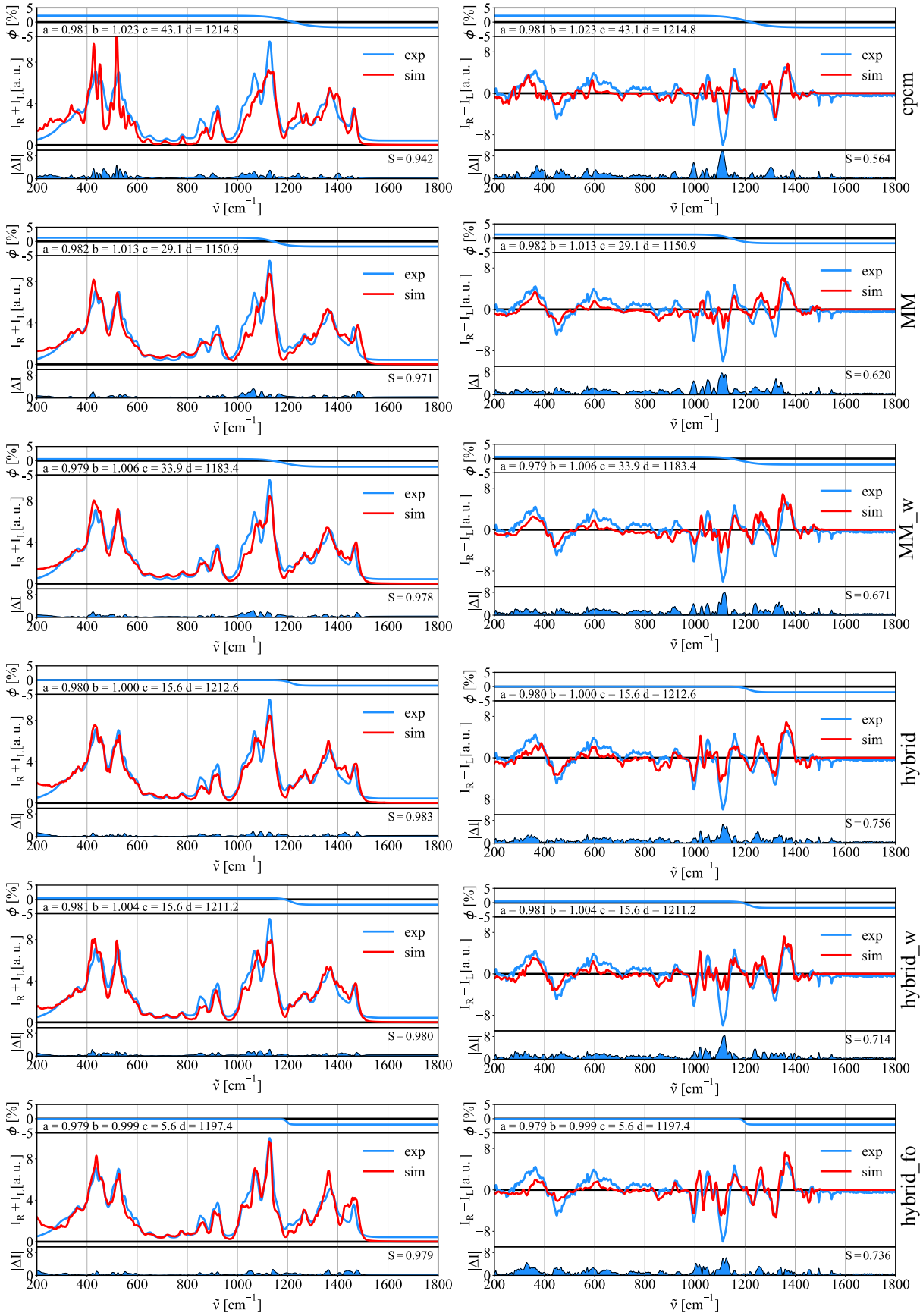

Fig. 1 Simulated and experimental Raman (left) and ROA (right) spectra of **D-glucose** obtained by CPCM, MM, MM_w, hybrid, hybrid_w, and hybrid_fo simulation approach. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between the experimental and the model spectrum are shown below the spectrum, together with the overlap integral S.

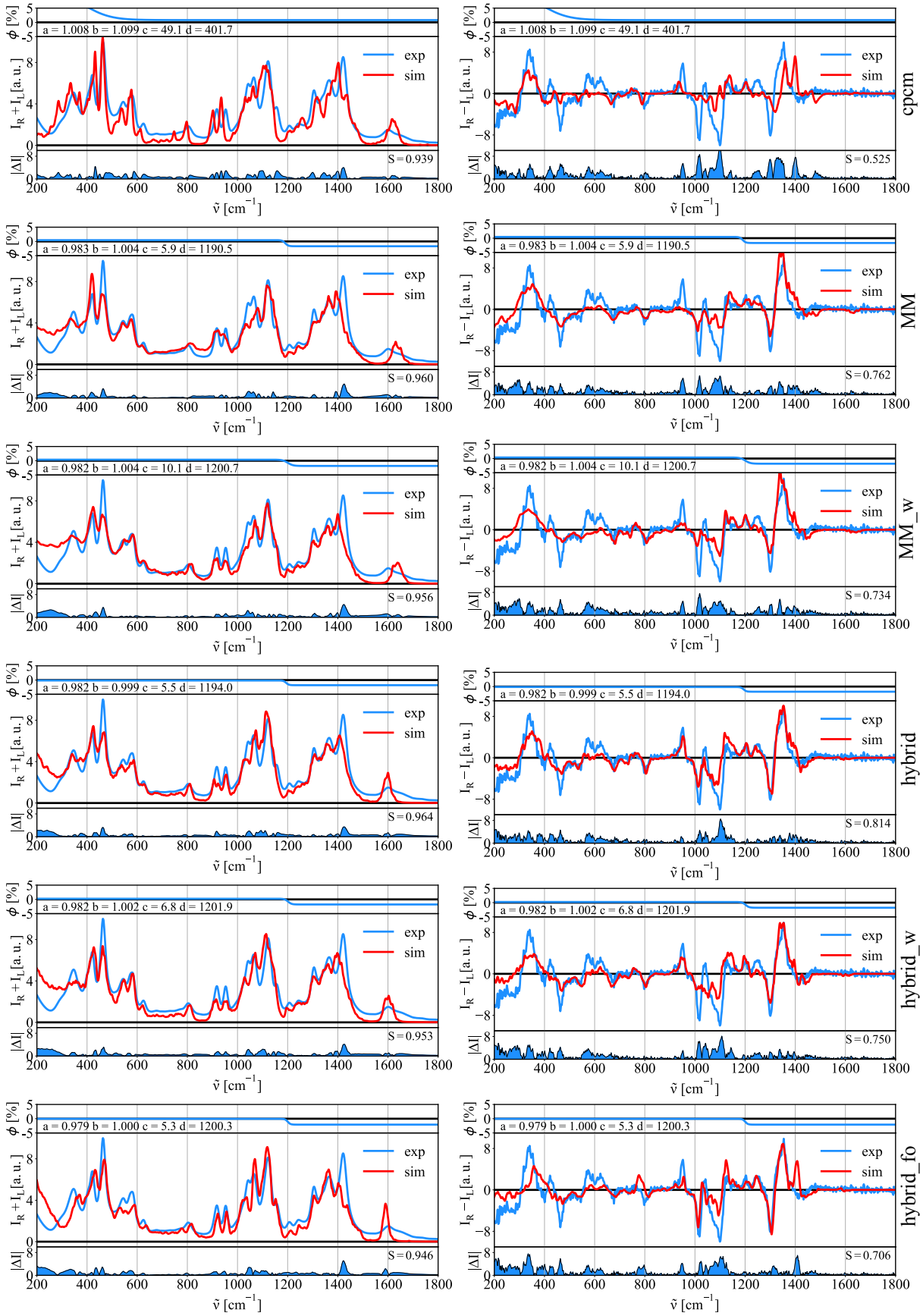

Fig. 2 Simulated and experimental Raman (left) and ROA (right) spectra of **D-glucuronic acid** obtained by CPCM, MM, MM_w, hybrid, hybrid_w, and hybrid_fo simulation approach. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between the experimental and the model spectrum are shown below the spectrum, together with overlap integral S.

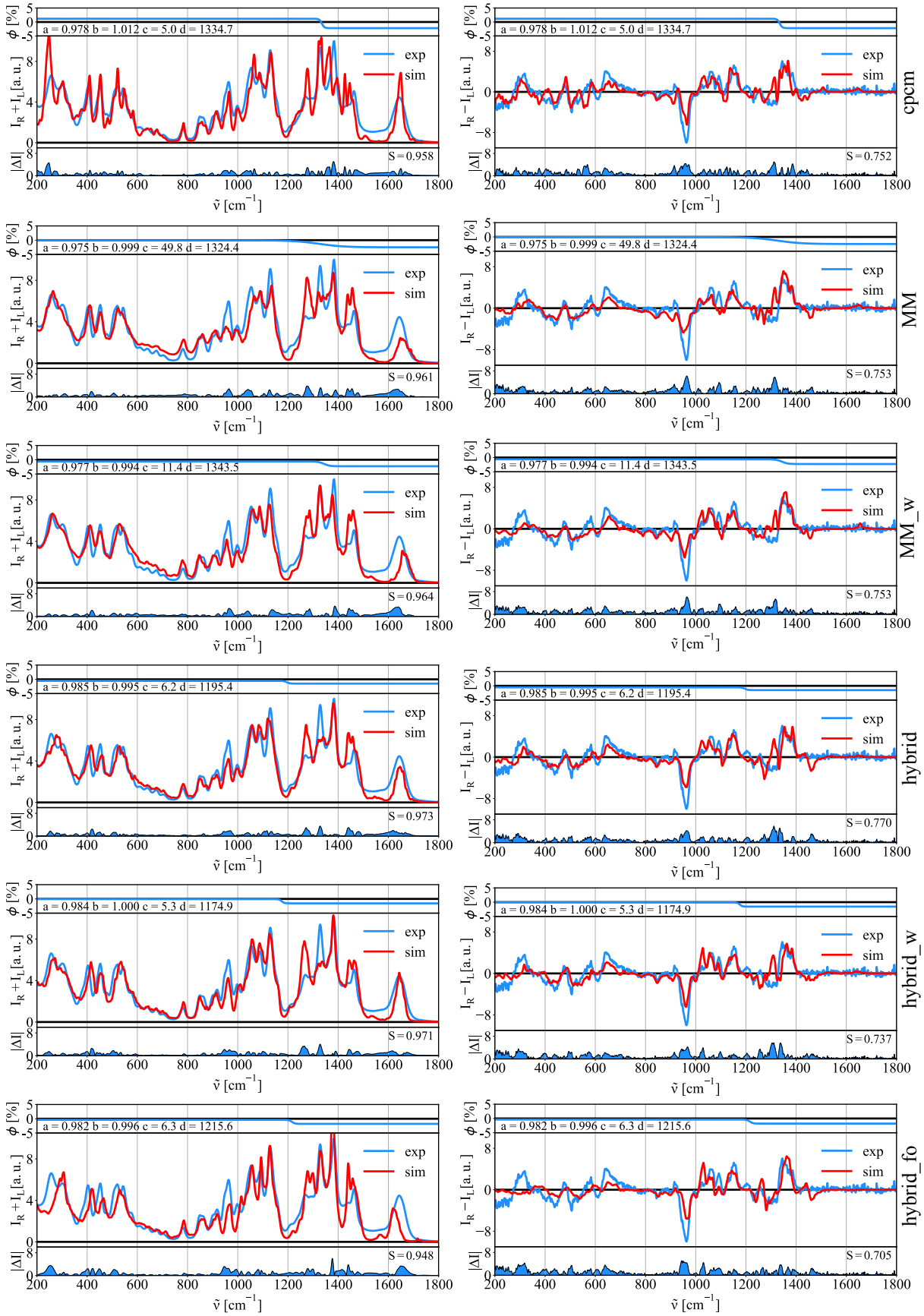

Fig. 3 Simulated and experimental Raman (left) and ROA (right) spectra of ***N*-acetyl-D-glucosamine** obtained by CPCM, MM, MM_w, hybrid, hybrid_w, and hybrid_fo simulation approach. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between experimental and model spectrum are shown below the spectrum, together with the overlap integral S.

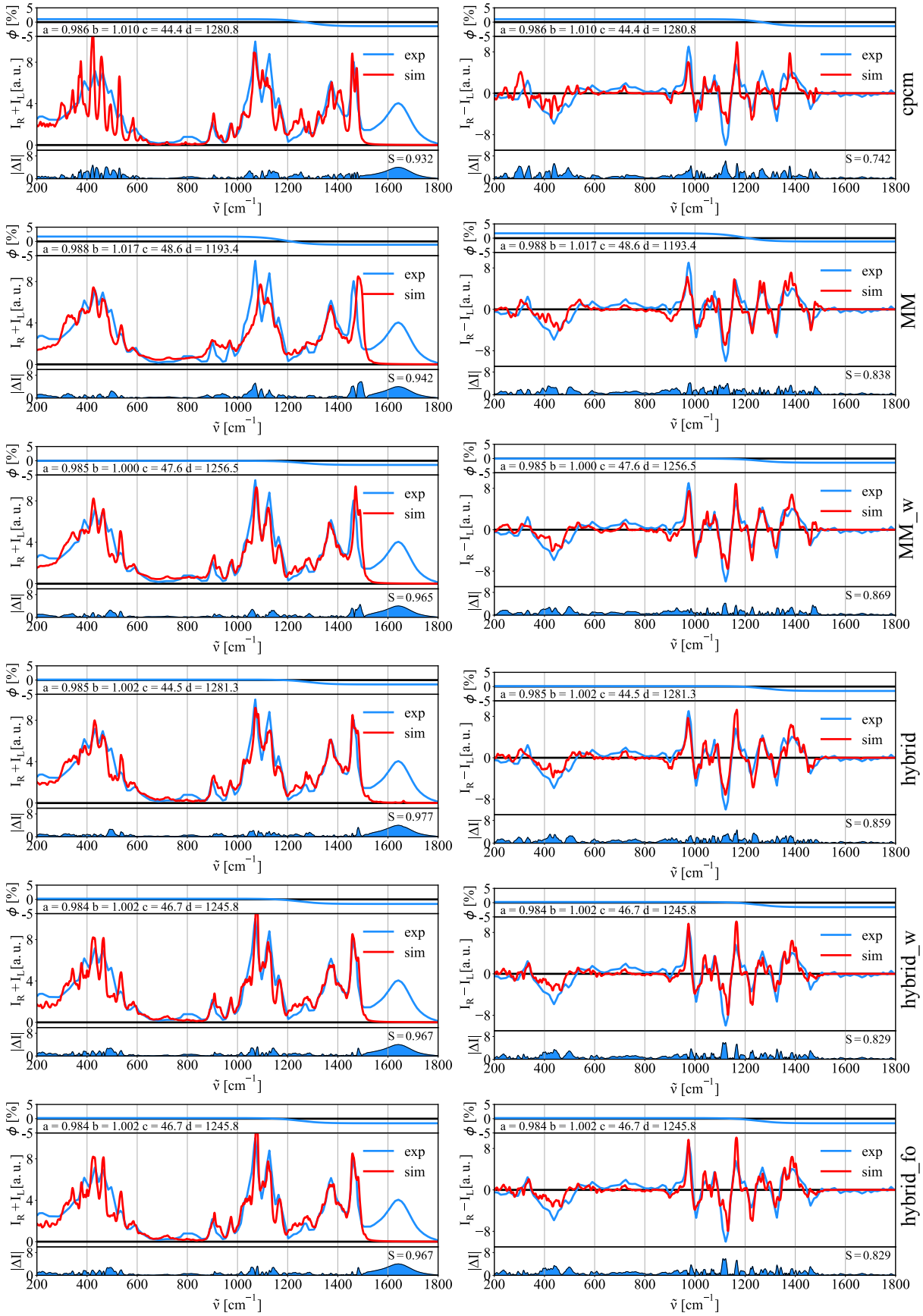

Fig. 4 Simulated and experimental Raman (left) and ROA (right) spectra of **methyl β-D-glucopyranoside** obtained by CPCM, MM, MM_w, hybrid, hybrid_w, and hybrid_fo simulation approach. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between the experimental and the model spectrum are shown below the spectrum, together with the overlap integral S.

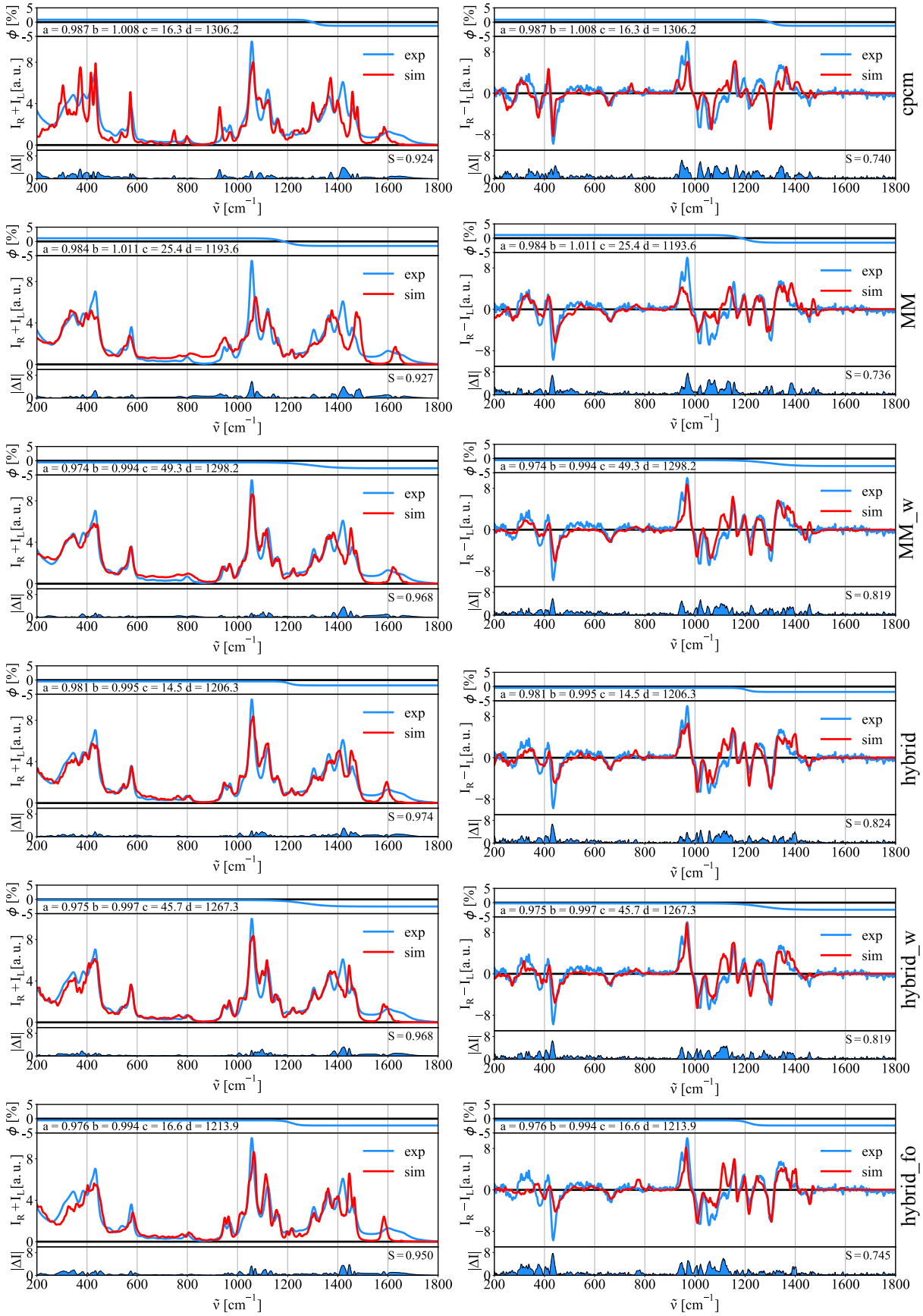

Fig. 5 Simulated and experimental Raman (left) and ROA (right) spectra of **methyl β-D-glucuronide** obtained by CPCM, MM, MM_w, hybrid, hybrid_w, and hybrid_fo simulation approach. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between the experimental and the model spectrum are shown below the spectrum, together with the overlap integral S.

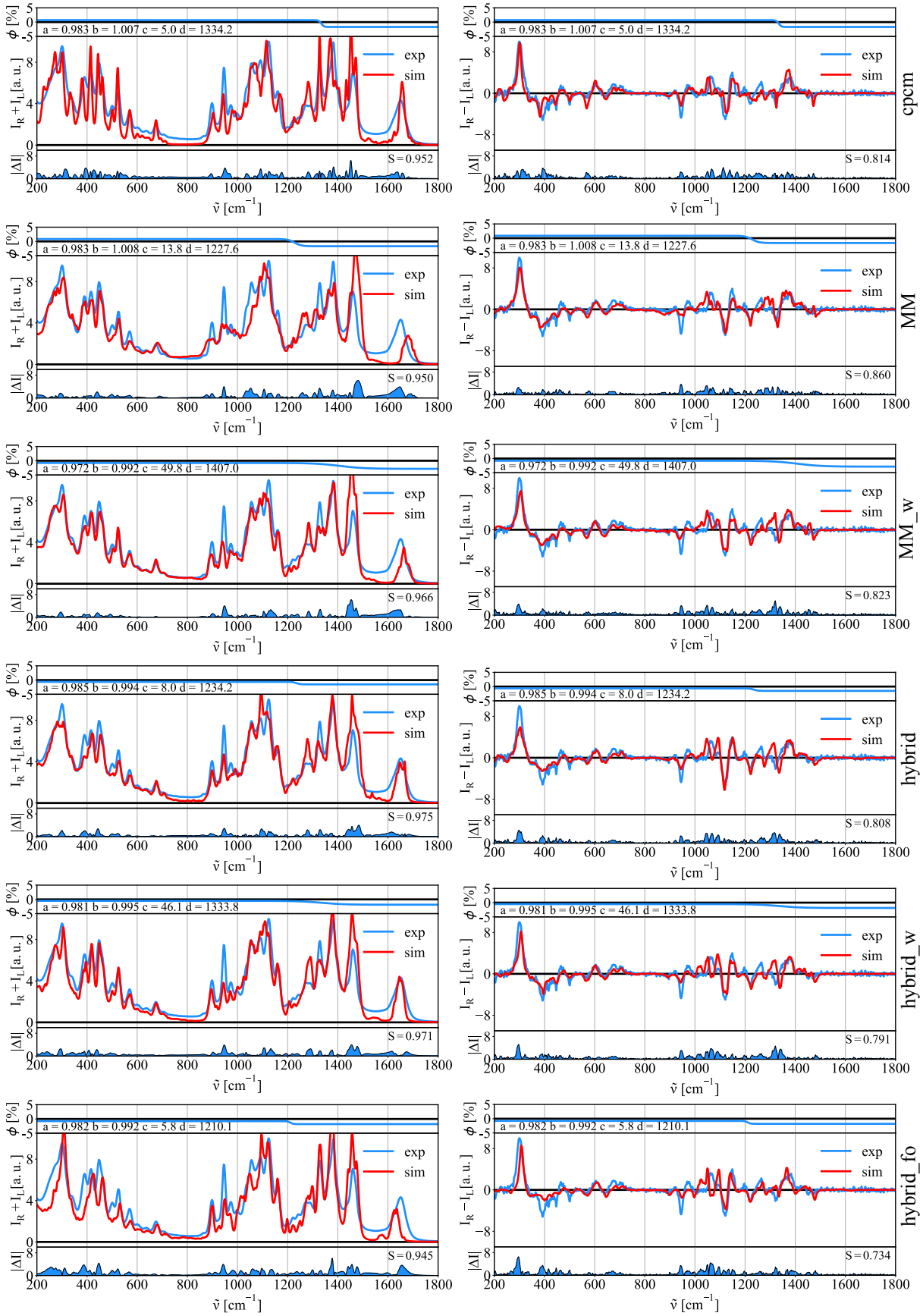

Fig. 6 Simulated and experimental Raman (left) and ROA (right) spectra of **methyl** β-***N*-acetyl-D-glucosaminide** obtained by CPCM, MM, MM_w, hybrid, hybrid_w, and hybrid_fo simulation approach. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between the experimental and the model spectrum are shown below the spectrum, together with the overlap integral S.

3 Raman/ROA Spectra of Individual Anomers

In all figures, where the spectra of glucose, glucuronic acid, and N-acetyl glucosamine are shown, a single spectrum is always compared to experimental data. However, these spectra are formed as a linear combination of spectra of individual anomers. Figure 7 shows spectra of glucose (top)/ glucuronic acid(middle)/ N-acetyl glucosamine(bottom), where on the left, we see calculated spectra of α/β anomer and on the right comparison of α/β weighted average to experimental data.

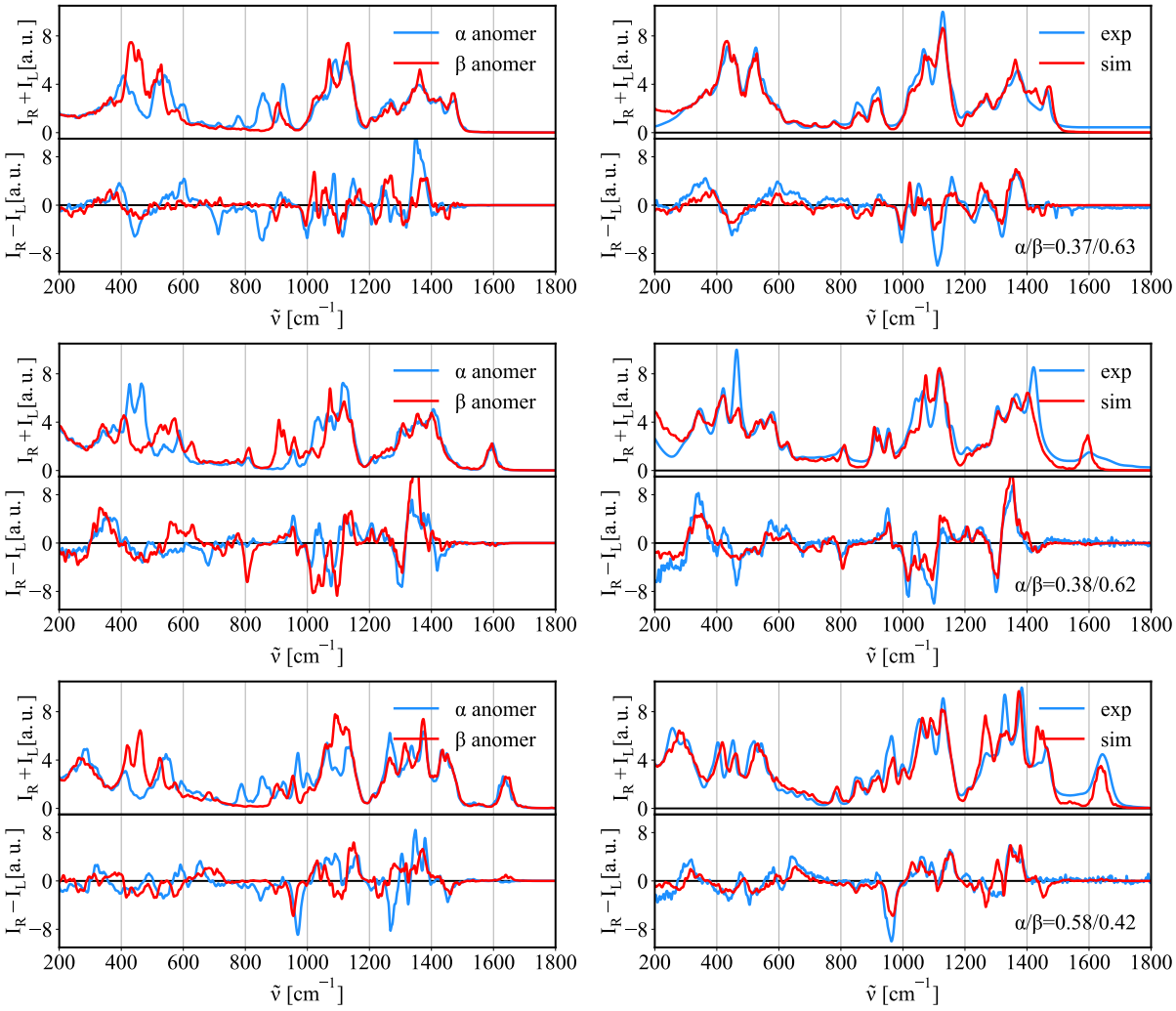

Fig. 7 Calculated Raman/ROA spectra of glucose(top)/glucuronic acid(middle)/N-acetyl glucosamine(bottom) α/β anomers, their weighted average(sim, glucose - 0.63, glucuronic acid - 0.62, N-acetyl glucosamine - 0.42) and comparison to exp data(exp). All spectra were calculated using `hybrid` approach.

It can be seen that it is essential to include α/β anomerism when calculating Raman/ROA spectra as spectra (especially ROA spectra) of both anomers differ quite a lot. Only by averaging them we obtain a good agreement with experimental data.

4 Performance of Different Simulation Protocols

In this section, we explore the use of normal modes optimization, DFT functionals, basis sets, and usage of QM described water molecules.

4.1 Performance of Normal Modes Optimization

Figure 8 compares the performance of `opt_normal_modes/hybrid` simulation protocols (see Table 1) on calculation of Raman/ROA spectra of methyl β-D-glucuronide and methyl β-*N*-acetyl-D-glucosaminide. The figure shows that both used optimization protocols results into similar outcomes, with exception in ROA spectrum of methyl β-D-glucuronide, where frequencies near 1000 cm^{-1} are poorly described.

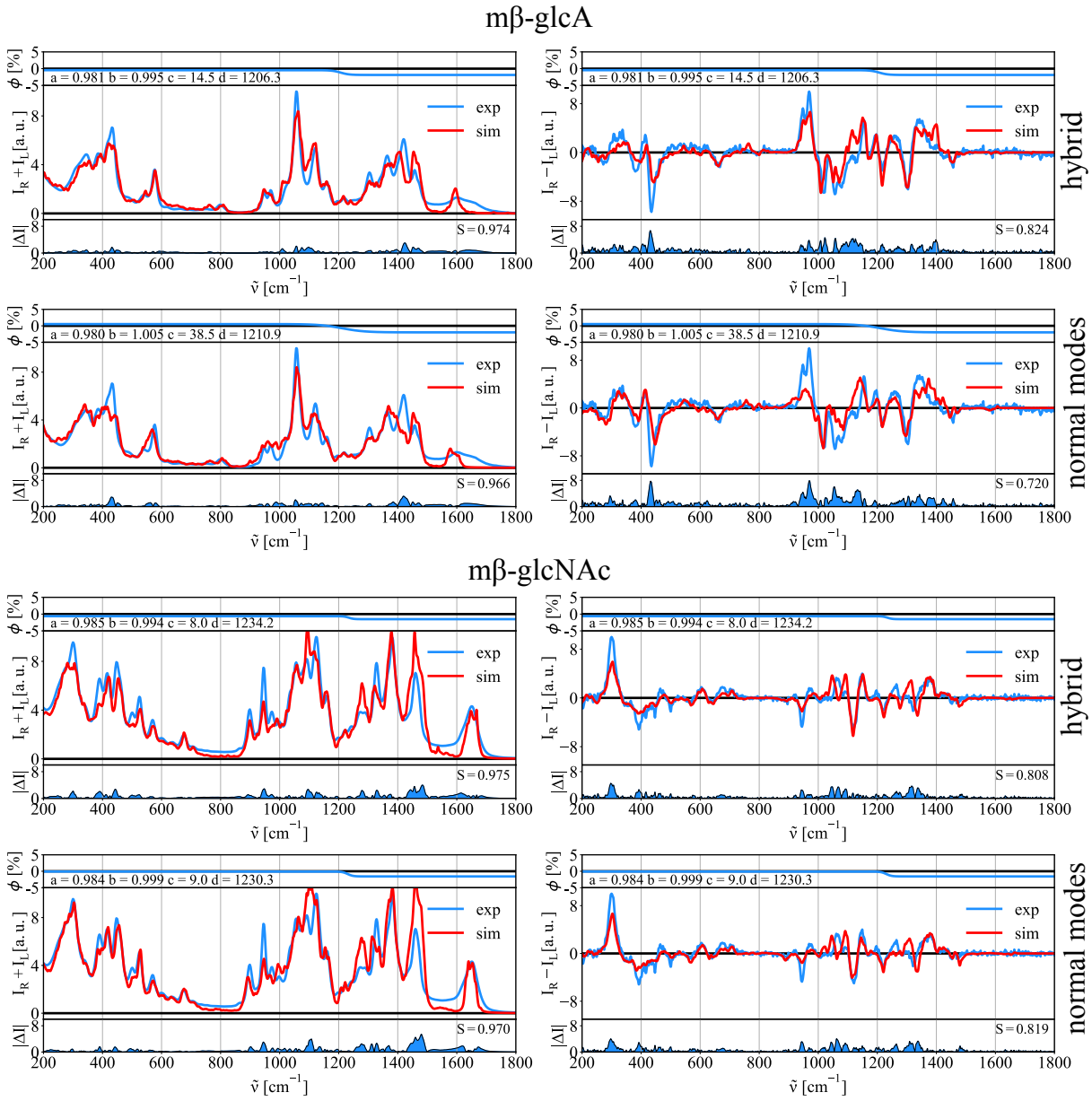

Fig. 8 Performance of different optimization methods (`opt_normal_modes/hybrid`) as described in Table 1 on reproduction of the Raman/ROA spectra of methyl β-D-glucuronide and methyl β-*N*-acetyl-D-glucosaminide.

4.2 Dependence on the Number of Optimization Steps

Figure 6 in the main text shows convergence of calculated Raman and ROA spectra using the `hybrid` solvation approach with a number of optimization steps. Here, we show 2 examples - glcA and mβ-glcA, how their spectra evolve with number of optimization

steps (Figure 9 and Figure 10). We see that spectra quickly gain its converged shape at 5 optimizations steps and it further deteriorates after 10-30 optimization steps.

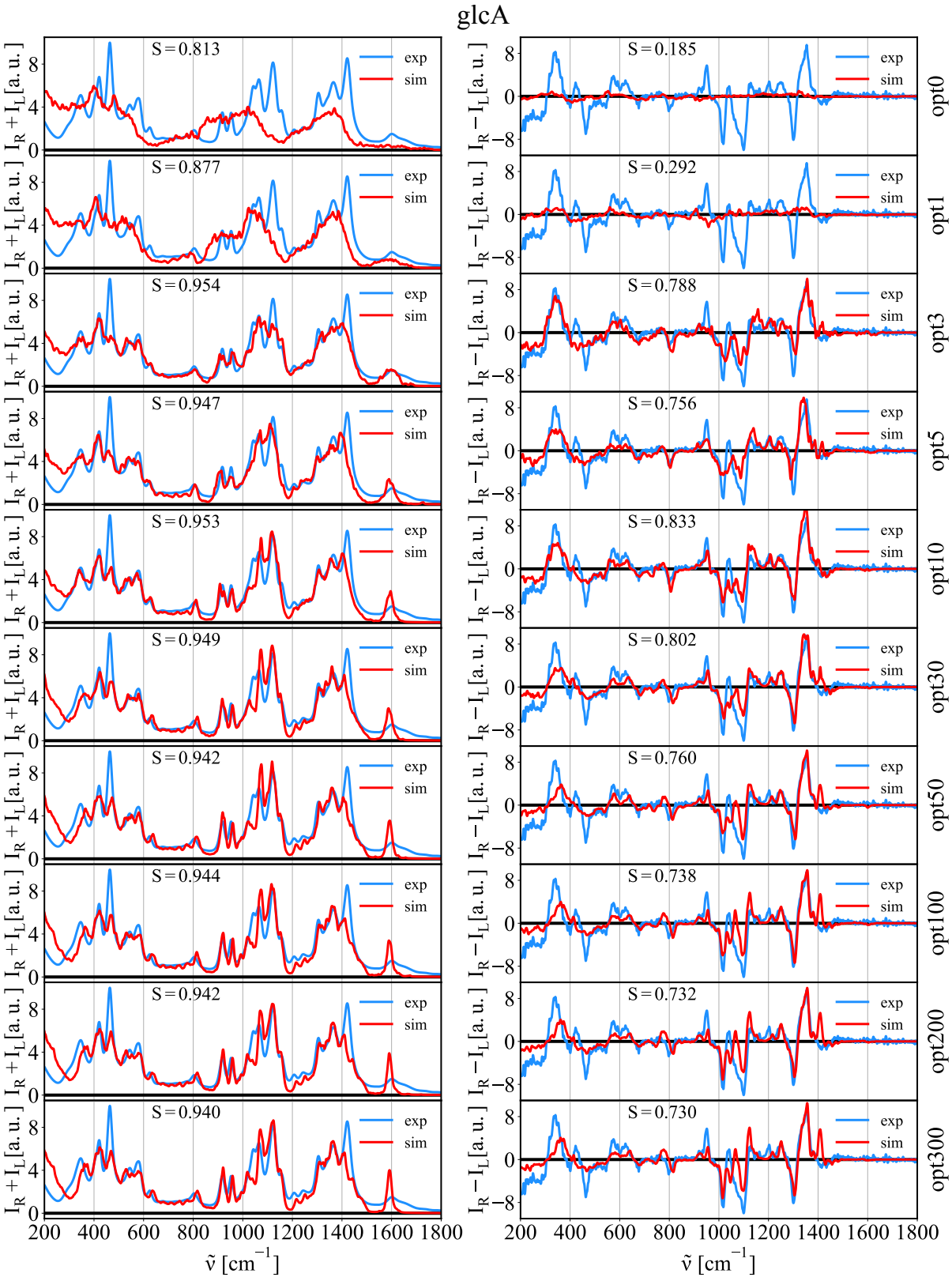

Fig. 9 Raman and ROA spectra of methyl β-D-glucuronide and obtained by `hybrid(X)` simulation approach using 0/1/3/5/10/30/50/100/200/300(full) optimization steps. 0 steps means that raw MD structures were used.

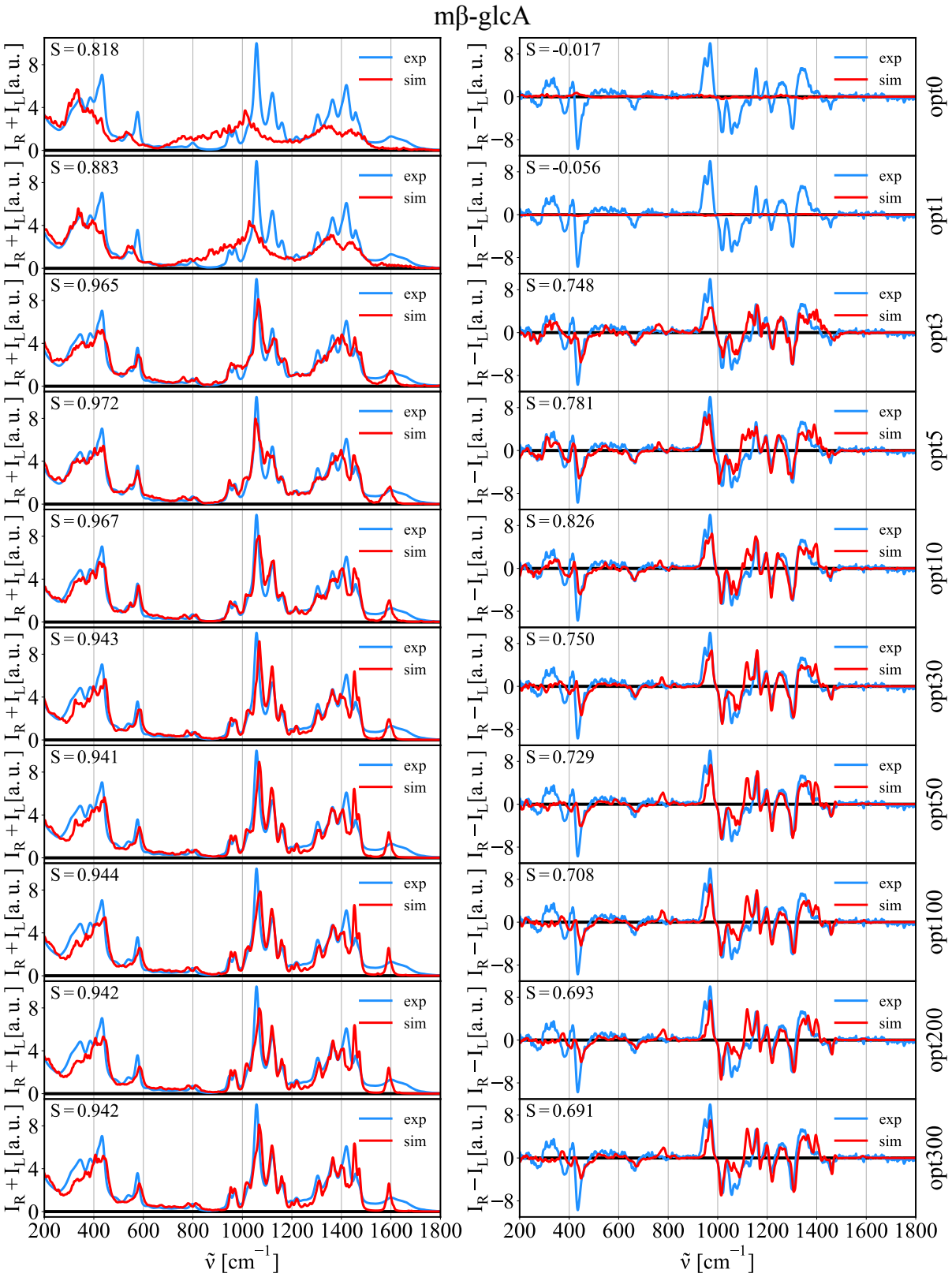

Fig. 10 Raman and ROA spectra of D-glucuronic acid obtained by `hybrid(X)` simulation approach using 0/1/3/5/10/30/50/100/200/300(full) optimization steps. 0 steps means that raw MD structures were used.

4.3 Dependence on the Basis Set

The 6-31G* basis set is too small to provide meaningful description of Raman/ROA properties, while aug-cc-pvtz provides similar spectra as 6-311++G** basis set. Therefore, in Figures 11 and 12 we provide comparison of rDPS and 6-311++G** basis sets. The preparation a calculation details can be found in the Table 1.

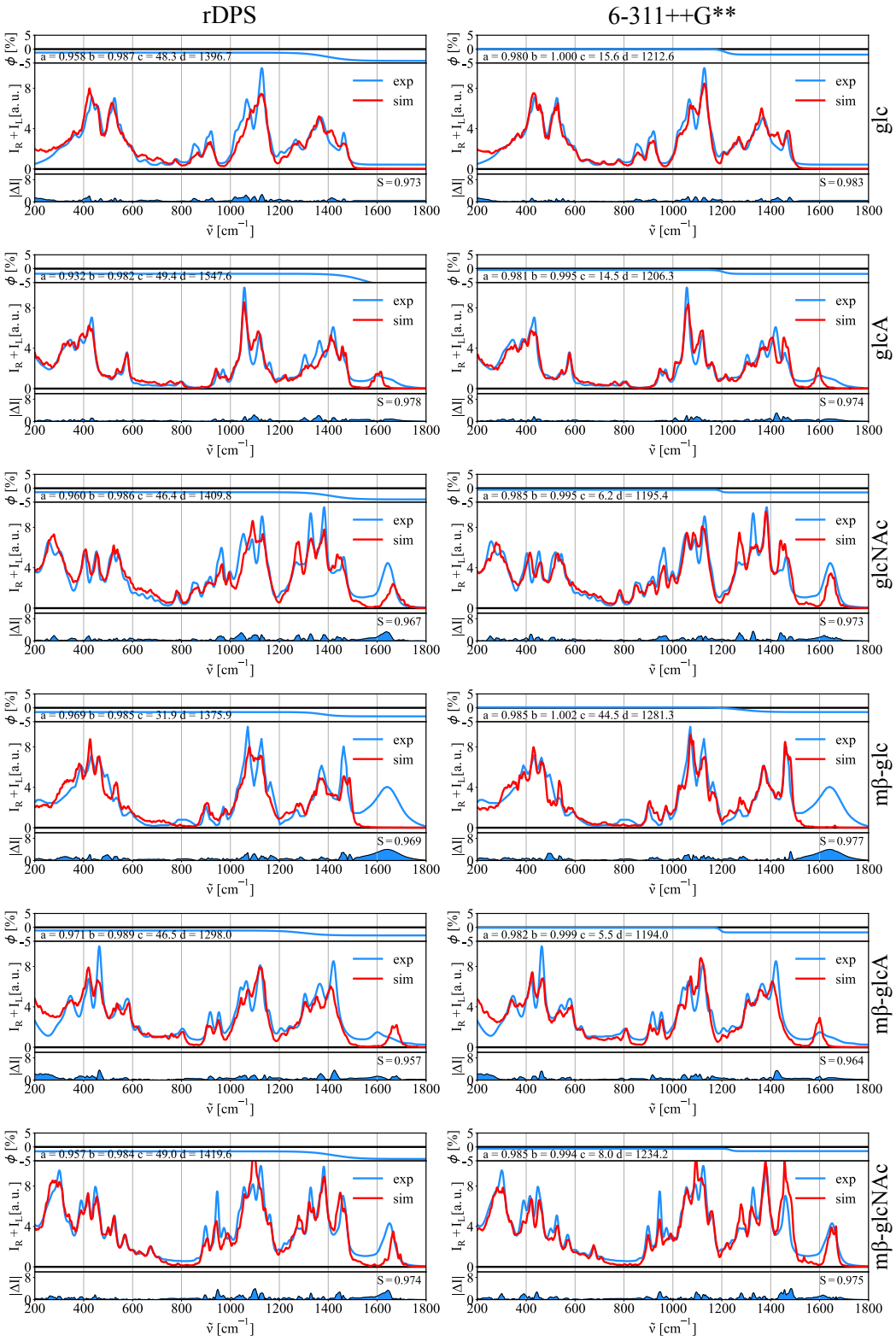

Fig. 11 Part 1/2. Comparison of Raman spectra obtained by using either rDPS(left) or 6-311++G**(right) basis sets. ROA spectra are shown in the Figure 12. The details of calculations can be found in the Table 1.

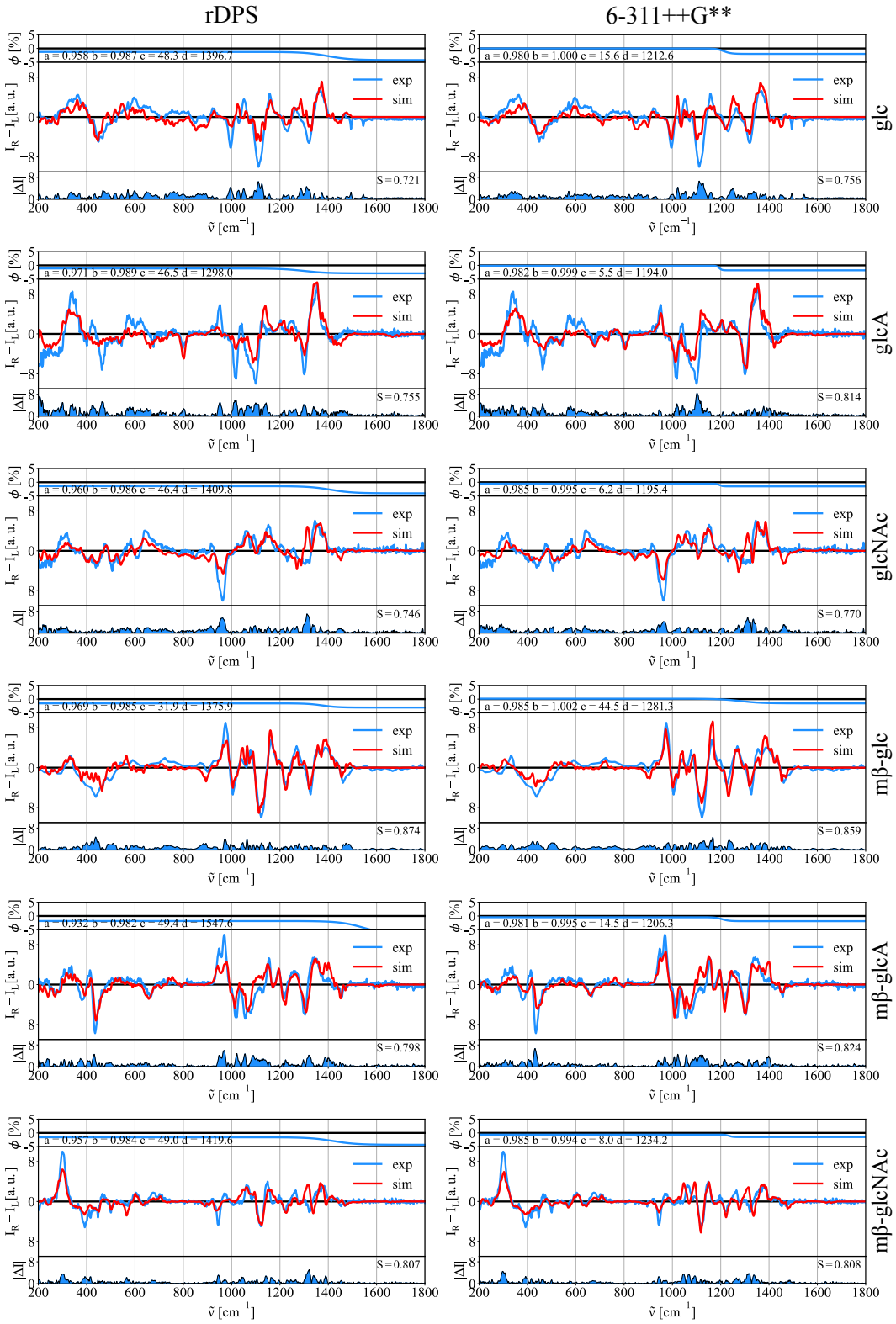

Fig. 12 Part 2/2. Comparison of ROA spectra obtained by using either rDPS(left) or 6-311++G**(right) basis sets. Raman spectra are shown in the Figure 11. The details of calculations can be found in the Table 1.

Overall, spectra obtained by using the rDPS basis set agree quite well with the ones obtained by using larger 6-311++G** basis set. Nevertheless, there are several semi-reproduced spectral features, e.g. ROA spectra of glucuronic acid and methyl β-glucuronide, which might be due to the charged nature of the molecules. Moreover, using the rDPS basis set leads to the scaling function converging to different values, see Figure 13 and Table 4.

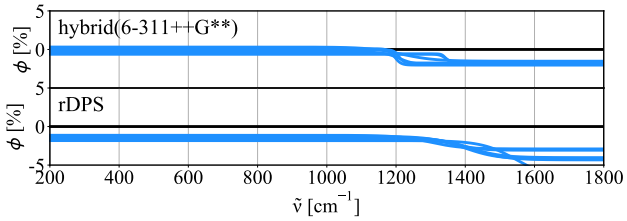

Fig. 13 Scaling functions produced when using either 6-311++G** basis set or rDPS basis set (1 per molecule, i.e. 6 per approach). The average scaling function parameters are in Table 4

Table 4 The average scaling function parameters while using 6-311++G**(`bs2` protocol) or rDPS(`bs4` protocol) basis sets. The scaling functions can be found in the Figure 13.

$\phi(\tilde{\nu}^{sim},a,b,c,d)$	a	b	c	d
`hybrid`(6-311++G**)	0.983	0.997	15.7	1220.1
`rDPS`(rDPS)	0.958	0.986	44.5	1405.5

4.4 Performance of Various DFT Functionals

Since all DFT functionals tested here behave similarly we omit these spectra.

4.5 Performance of Unrestrained Optimization When Using CPCM

Figure 14 shows beneficial effect of using harmonic restraints when using only continuum solvation (`CPCM`).

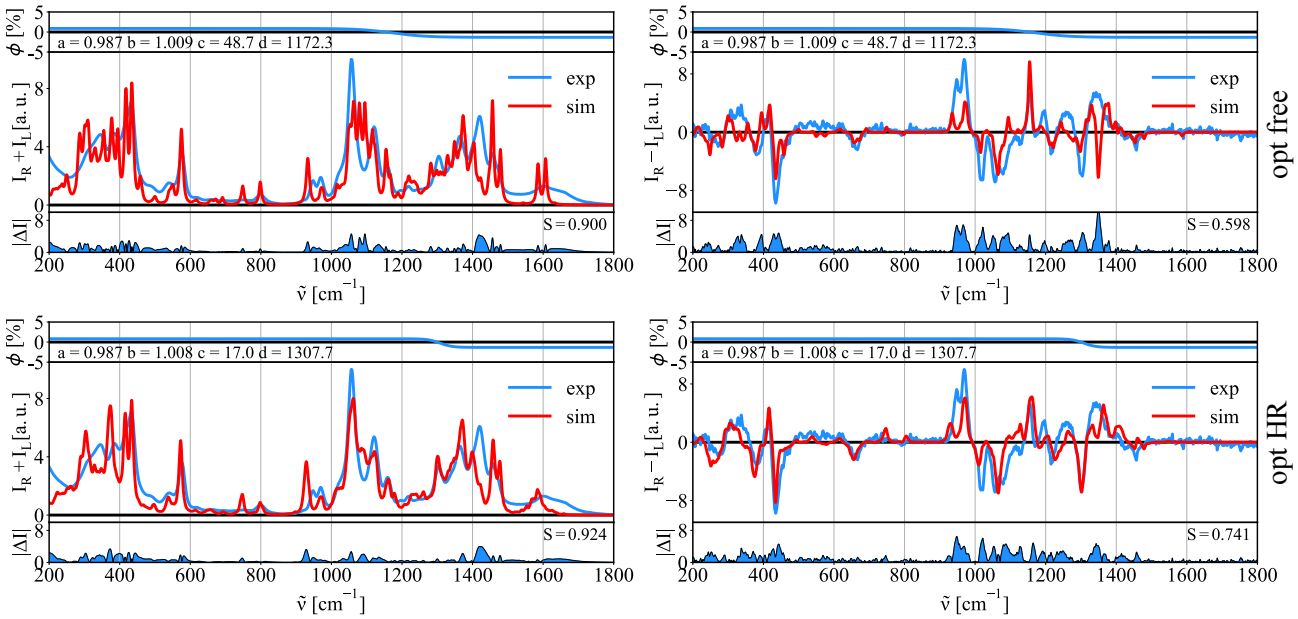

Fig. 14 Raman and ROA spectra of methyl β-D-glucuronide calculated after performing an unrestrained optimization (`opt_free`), or after using 10 optimization steps with harmonic restraints (`opt_HR`) when using continuum solvation protocols as `CPCM`, see Table 3.

4.6 Full optimization of glucose using QM water molecules

Because spectra of glucose were reproduced relatively poorly using `hybrid` approach, we decided to carry out the full optimization calculations (using both QM or MM water molecules). Figure 15 shows calculated spectra of glucose using `hybrid` and `hybrid_fo`, which compares calculations using MM water molecules, where in case of hybrid, the optimization is run only for 10 steps, while in `hybrid_fo`, the optimization is run until gradient criteria are reached. Moreover, the Figure also compares the same setup, but using QM water molecules(`hybrid_QM` and `hybrid_QM_fo`).

Using `hybrid_fo` approach with MM water molecules gives slightly worse results as compared to 10 steps optimization using `hybrid` approach. Using QM described water molecules with `hybrid_QM` 10 steps optimization lead to worse description of the high

frequency region (>1000 cm^{-1}), however, the low frequency region of ROA spectrum was described better. Using full optimization with QM water molecules (`hybrid_QM_fo`) results into the improvement of the high frequency region as compared to `hybrid_QM` approach, while the low frequencies are still reproduced.

Overall, the data suggests that higher frequency region (>1000 cm^{-1}) was better reproduced using full optimization, regardless the theory that was used for description of water molecules. In the lower frequency region the description of water molecules mattered and QM description was superior to MM. Note that using QM water molecules is expensive as in the relative cpu times for glucose molecule the time needed for a single calculation was: `hybrid`:`hybrid_fo`:`hybrid_QM`:`hybrid_QM_fo`=1:8:25:70.

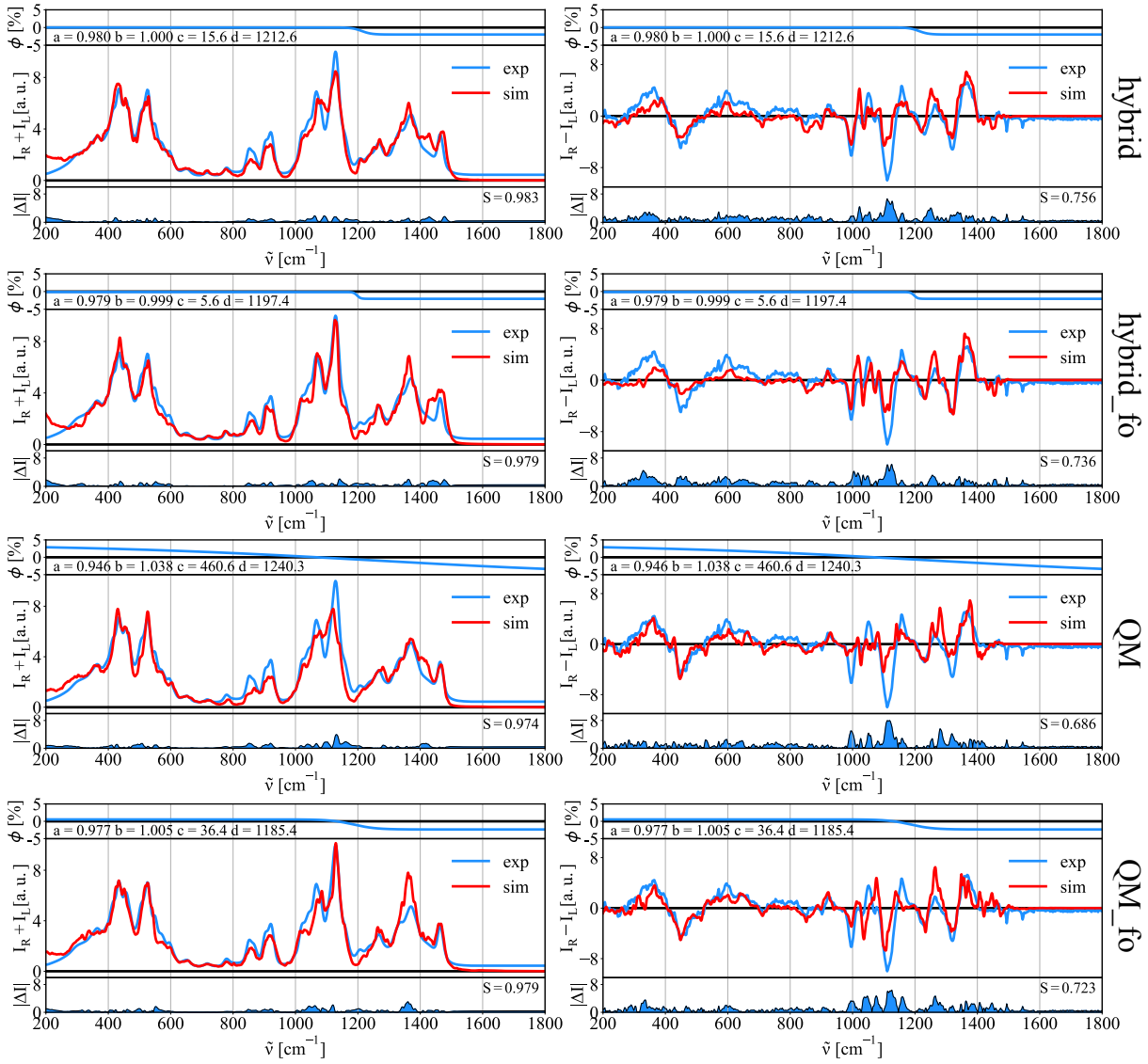

Fig. 15 Simulated and experimental Raman (left) and ROA (right) spectra of glucose obtained by `hybrid`, `hybrid_fo`, `hybrid_QM`, and `hybrid_QM_fo` simulation approaches. Vibrational frequency scaling function ϕ with optimized parameters is provided with each of the simulation (above spectrum). Absolute differences between the experimental and the model spectrum are shown below the spectrum, together with the overlap integral S.

5 Overestimation of the Vibrational Wavenumbers When Using Hybrid DFT Functionals

Calculated vibrational wavenumbers, $\tilde{\nu}^*_{ij}$, are usually overestimated depending on the used method. The Hartree-Fock approximation describes poorly the potential energy surface as the molecule is distorted from its equilibrium value[4]. The problem lies in the very core of the Hartree-Fock method as it forces the double occupation of the orbitals, which does not correctly reproduce energetics of non-equilibrium states. This leads to the overestimation of the vibrational wavenumbers. Since hybrid DFT functionals employ part of the exact Hartree-Fock exchange energy, they also exhibit this feature but to a lesser degree.

To partially correct these theoretical inaccuracy, an empirical redshift of the spectra produced by *ab initio* calculations is usually

performed by using a scaling factor on the obtained vibrational wavenumbers[5–8]. Numerous studies focus on distinct types of molecules aiming to evaluate scaling factors for a given DFT functional and basis set combination[5,6]. Moreover, scaling factors as a function on the wavenumber of a given vibration have also been introduced[5,6].

The usual strategy is that high energy vibrations (high wavenumber) requires larger redshift scaling, e.g., B3LYP/6-311++G** $\sim 0.97\tilde{\nu}$, while low energy vibrations (low wavenumber) almost do not need to be scaled, e.g., B3LYP/6-311++G** $\sim 1.01\tilde{\nu}$[5,6]. Unfortunately, there are neither clear rules defining both scaling regions, nor agreement on the transition between the regions. Therefore, we used an empirically fitted scaling function $\phi(\tilde{\nu}^*_{ij})$ to correct calculated wavenumbers in this work.

5.1 Example of How Scaling Function Works

The scaling function transforms vibrational frequencies in a similar manner as widely used simple 1 number scaling factors. Figure 16 shows calculated spectra of methyl β-glucuronide without any scaling (top), with use of widely used one factor scaling (0.98 in this case), and with use of our proposed scaling function.

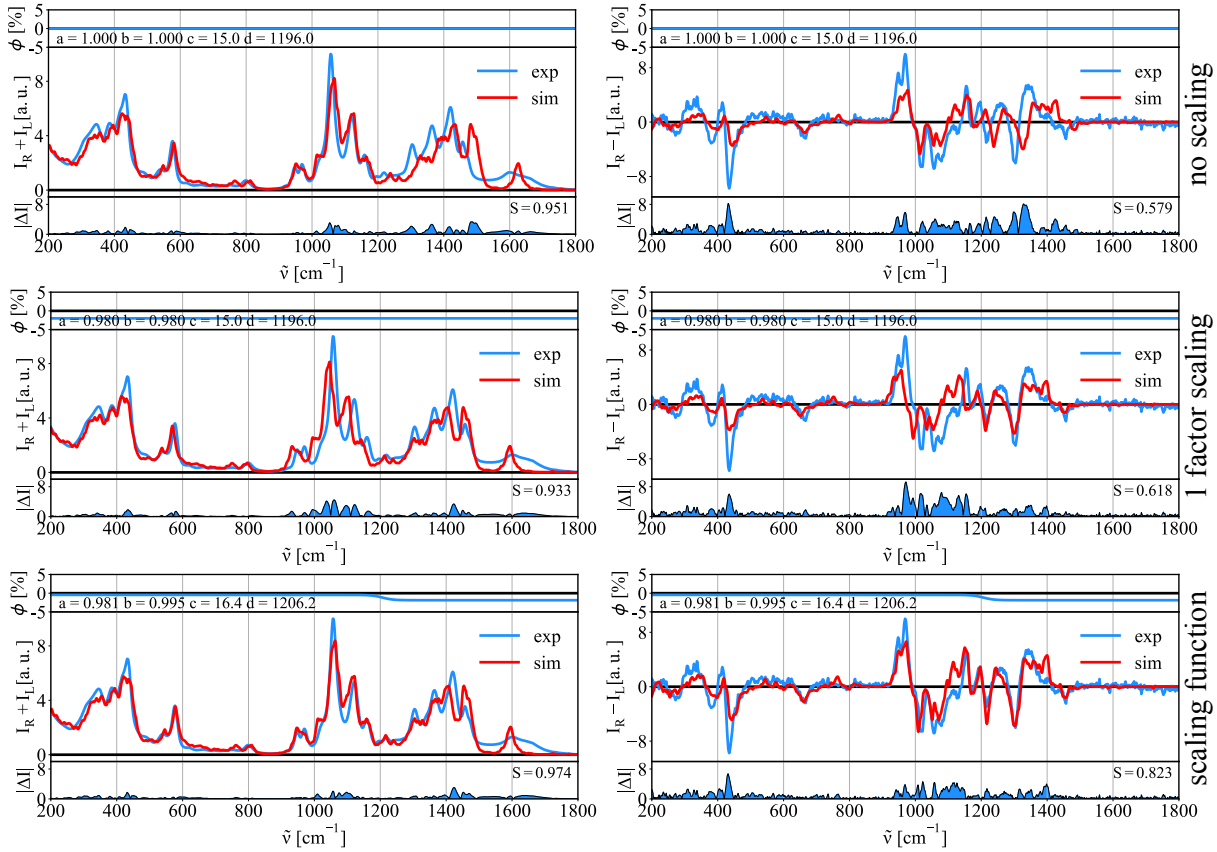

Fig. 16 Raman and ROA spectra of methyl β-glucuronide calculated by hybrid approach, where calculated frequencies were not scaled (top), were scaled as found in the literature by single scaling factor 0.98(middle), and scaled by our adaptive scaling function(bottom). In each graph the scaling function is found at the top, simulation and experimental data in the middle, and differences in experiment and simulation intensities, together with error function and overlap integral at the bottom.

We see that when we do not scale the spectra is shifted after 1200 cm^{-1} and it does not correspond to experimental data. When we use one scaling factor as widely used in the literature (0.98) we can see that >1200 cm^{-1} region now corresponds to the experimental data, however, the low frequency region is badly adjusted. Only when properly separating the scaling factors for low/high frequency regions we can obtain good agreement with experimental data. This can be easily seen on the value of the overlap integral S for unscaled/scaled(0.98)/scaling function-Raman: 0.951, 0.933, and 0.974; ROA: 0.579, 0.618, and 0.823. The obtained values points to the strength of using proposed scaling function.

Finally, we would like to point out that all of our approaches converge to similar shape scaling functions, regardless the molecule or optimization procedure (Figure 17). The average values of scaling functions are reported in the Table 5. From our experience, the average optimal scaling function for the `hybrid` protocol would be $\phi(\tilde{\nu}^{sim},a,b,c,d) = \phi(\tilde{\nu}^{sim},0.982,1.00,15,1210)$.

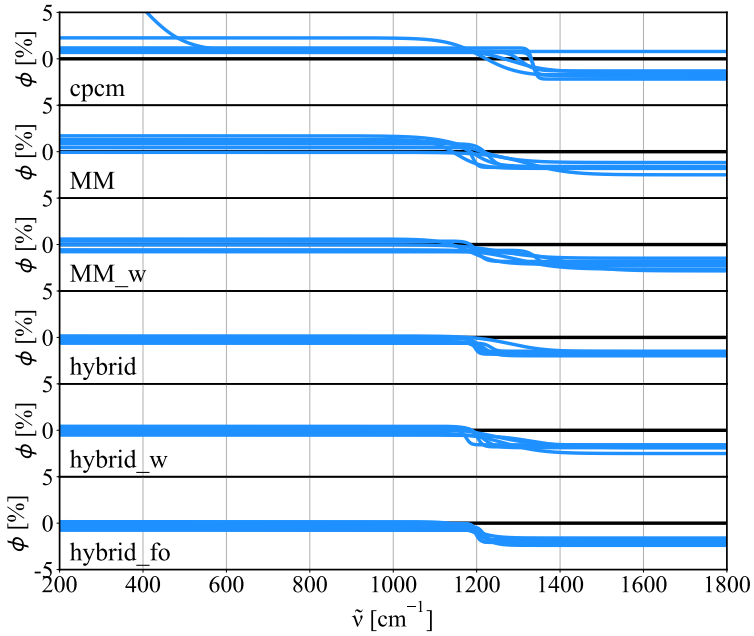

Fig. 17 Scaling functions produced by different optimization approaches (1 per molecule, i.e. 6 per approach)

Table 5 Optimal scaling function parameters.

$\phi(\tilde{\nu}^{sim},a,b,c,d)$	a	b	c	d
cpcm[a]	0.983	1.012	22.8	1294.9
MM	0.983	1.009	28.8	1213.3
MM_w	0.982	0.998	33.7	1281.2
hybrid	0.983	0.997	15.7	1220.1
hybrid_w	0.981	1.000	27.7	1239.1
hybrid_fo	0.980	0.997	10.8	1203.5

[a] glcA values excluded

6 Convergence with Number of Frames

Figure 18 shows the convergence of calculated Raman/ROA of spectra with the number of frames (glcNAc, hybrid approach).

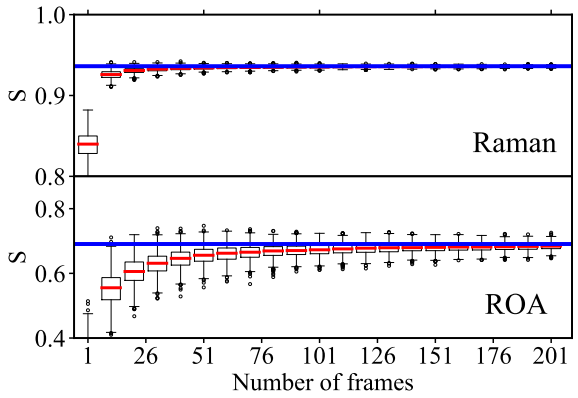

Fig. 18 Box plot of the convergence of the error(overlap integral) of simulation spectra from the experimental Raman/ROA spectra of glcNAc as a function of the number of frames used in the construction of the simulation spectra. In blue, the value of overlap integral obtained when using all 500 frames.

At first, we uniformly sampleed 500 snaphots from MD simulation. Then, we calculated the corresponding average Raman/ROA spectra using the `hybrid` simulation approach together with the corresponding scaling function $\phi(\tilde{\nu}^{sim}, a, b, c, d) = \phi(\tilde{\nu}^{sim}, 0.982, 1.00, 15, 1210)$. Using the same scaling function and random draws of different sizes using the same 500 spectra, we estimated the convergence of the method as a function of the number of frames used. For each number of frames, we made 1000 randomized draws. Each draw was compared to experimental data utilizing the overlap integral. Figure 18 shows that Raman spectrum converges when using $\sim$30 snapshots, while ROA spectrum converges much more slowly.

References

1 P. Bouř and T. A. Keiderling, *J. Chem. Phys.*, 2002, **117**, 4126–4132.

2 G. Zuber and W. Hug, *The Journal of Physical Chemistry A*, 2004, **108**, 2108–2118.

3 K. H. Hopmann, K. Ruud, M. Pecul, A. Kudelski, M. Dračínský and P. Bouř, *J. Phys. Chem. B*, 2011, **115**, 4128–4137.

4 G. E. Scuseria, C. A. Jiménez-Hoyos, T. M. Henderson, K. Samanta and J. K. Ellis, *J. Chem. Phys.*, 2011, **135**, 124108.

5 M. K. Kesharwani, B. Brauer and J. M. L. Martin, *J. Phys. Chem. A*, 2015, **119**, 1701–1714.

6 M. L. Laury, M. J. Carlson and A. K. Wilson, *J. Comput. Chem.*, 2012, **33**, 2380–2387.

7 P. Sinha, S. E. Boesch, C. Gu, R. A. Wheeler and A. K. Wilson, *J. Phys. Chem. A*, 2004, **108**, 9213–9217.

8 M. Jeffrey P., D. Moran and L. Radom, *J. Phys. Chem. B*, 2007, **111**, 11683–11700.