Thermal fluctuation and conformational effects on NMR parameters in β-O-4 lignin dimers from QM/MM and machine-learning approaches

Sonia Milena Aguilera-Segura, Dominik Dragún, Robin Gaumard, Francesco Di Renzo, Irina Malkin Ondík, Tzonka Mineva

ARTICLE

Thermal fluctuationσ and conformational effects on NMR parameters in β-O-4 lignin dimer from QM/MM and machine-learning approaches

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

DOI: 10.1039/x0xx00000x

[a]Sonia Milena Aguilera-Segura, [b]Dominik Dragún, [a]Robin Gaumard, [a]Francesco Di Renzo, [c]Irina Malkin Ondík ,[a]Tzonka Mineva*

Advanced solid-state and liquid-state nuclear magnetic resonance (NMR) approaches have enabled high throughput information about functional groups and types of bonding in a variety of lignin fragments from degradation processes and laboratory synthesis. The use of quantum chemical (QM) methods may provide detailed insight into the relations between NMR parameters and specific lignin conformations and their dynamics, whereas a rapid prediction of NMR properties could be achieved by combining QM with machine-learning (ML) approaches. In this study, we present the effect of conformations of β-O-4 linked lignin guaiacyl dimer on ^{13}C and ^{1}H chemical shifts while considering the thermal fluctuations of guaiacyl dimer in water, ethanol and acetonitrile, as well as their binary 75wt% aqueous solutions. Molecular dynamics and QM/MM simulations were used to describe the dynamics of guaiacyl dimer. The isotropic shielding of the majority of the carbon nuclei was found less sensitive toward a specific conformation than that of the hydrogen nuclei. The largest ^{1}H downfield shifts of 4-6 ppm were established in the hydroxy groups and the rings in the presence of organic solvent components. The Gradient Boosting Regressor model has been learned on 60% of the chemical environments in the dynamics trajectories with the NMR isotropic shielding (σ_{iso}), computed with the density-functional theory, for lignin atoms. It is established a high efficiency of this machine-learning model in predicting the remaining 40% σ_{iso} (^{13}C) and σ_{iso} (^{1}H) values.

Introduction

Characterization of molecular structures in lignin fractions takes an important part in the valorisation of lignin compounds - the promising environmentally sustainable feedstock of organic carbon[1]. Lignin is a three-dimensional, highly branched polyphenolic complex polymer and represents around 25-35% of the total dry weight of the wall substances in biomass[2]. Lignin monomeric units and their linkages vary with the source of lignin and the type of fractionation process. This makes their characterisation at the molecular level challenging. Despite this challenge, high throughput information about functional groups and types of bonding has been obtained with several spectroscopic techniques, such as infrared, ultraviolet-visible, Raman spectroscopy, and nuclear magnetic resonance (NMR)[1,3–5]. In particular, the higher resolution of NMR techniques has yielded a significant amount of information on lignin structural units and side-chain linkages[1,6]. In addition to the regular usage of ^{1}H and ^{13}C NMR chemical shifts, solid-state ^{13}C NMR and 2D heteronuclear single-quantum coherence (HSQC) NMR approaches have been applied with success to resolve structural properties in various lignin fractions[6–12].

To further decrease the ambiguities in NMR spectral assignments, artificial lignin models of dimers and small oligomers have been synthesized and analysed with NMR techniques[5,13,14]. Typically, the assignment of NMR signals in complex lignin polymers or smaller lignin fractions, arising from the degradation processes[5], is performed based on NMR chemical shifts in artificial lignin dimer models. The NMR signals in these synthetic models can however be broadened or even displace the chemical shift positions upon their interactions with solvent molecules used during the liquid NMR experiments [10,14]. The complexity of the lignin-solvent systems has been further increased by recent development in the use of mixed solvents in NMR analysis of whole swollen cell walls. The direct 2D NMR spectroscopy of solutions of dimethyl sulfoxide (DMSO) with swelling agents as different as tributylammonium, pyridinium, alkylimidazolium or hexamethylphosphoramide is becoming an established

[a] ICGM, Univ Montpellier, CNRS, ENSCM, Montpellier, France
[b] FIIT STU in Bratislava, Ilkovičova 2, 842 16 Bratislava, Slovakia
[c] MicroStep-MIS spol. s.r.o. Čavojského 1, 84104 Bratislava, Slovakia
*To whom correspondence should be addressed
Email: tzonka.mineva@enscm.fr

Electronic Supplementary Information (ESI) available: [Supporting Figures: $\sigma(^{13}C)$ rolling averaged values; $\sigma(^{13}C)$ and $\sigma(^{1}H)$ averaged shielding values with the standard deviation errors; supplementary ML-GBR vs. DFT isotropic shieldings. Supporting Table: ^{13}C chemical shifts in ring A and B; Cartesian coordinates and isotropic shielding values, used for ML-MBR are available in the ESI- files-2-6]. See DOI: 10.1039/x0xx00000x

analytical method [15–18]. The recently introduced techniques of characterization of the interaction of lignin with the cell wall environment by multidimensional ^{13}C solid-state NMR spectroscopy have opened a new field of investigation[19–21]. Despite the different methods of drying applied in these measurements, this technique opens untrodden ways to characterize the in-situ conformation of cell components in the presence of the organic-bearing vessel solutions[22,23]

Theoretical NMR chemical shifts in lignin monomers or small dimer models could constitute a basis for experimental NMR assignments. However, the complexity of lignin and its residues limited the theoretical works to only a few reports hitherto devoted to computations of NMR parameters. Very recently, a wide range of approximations of exchange-correlation functionals in density-functional theory (DFT) was examined for its ability to reproduce experimental ^{1}H and ^{13}C chemical shifts (δ) in optimized 5–5' lignin model dimers[24] and in sulphated monolignols[25]. Theoretical atomic charges and electron charge-density based parameters have been used as descriptors in an efficient neural-network-potential model developed to predict the experimental δ (^{13}C) in substituted phenol monomers [26]. We, therefore, find it of interest to study the effect of different conformations arising at room temperature and the thermal fluctuations of the lignin nuclei on the computed ^{1}H and ^{13}C chemical shifts. The inclusion of thermal fluctuations in the calculated NMR parameters, averaged over an ensemble of geometrical structures, can improve significantly the evaluation of NMR chemical shifts[27,28] and quadrupole coupling constant[29].

Considering the complexity of the lignin network, we have chosen here to focus on a guaiacyl dimer model of lignin with a β-O-4 linkage (G-b-G) shown in Figure 1A. This is the most common inter-unit linkage in lignin [30]. The basic structures of lignin are phenylpropanoid monolignols that consist of an aromatic ring and a 3-C side chain (Figure 1B). Differently substituted monolignol subunits are derived from coniferyl, *p*-coumaryl, and sinapyl alcohols[1,31]. The β-O-4 structures are flexible molecules that can adopt a large number of conformations, as established from both experimental and theoretical studies [32–34]. The β-O-4 linkage is relatively weak compared to other less common linkages, such as β-5, β-1, β-β', 5–5', and 4-O-5, more stable and consequently difficult to degrade. The first target of lignin degradation is indeed the β-O-4 linkage [4].

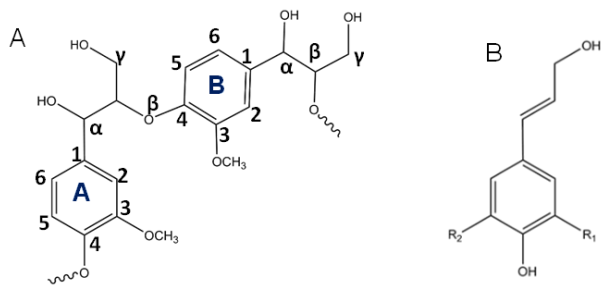

Figure 1. Chemical structures of (A) lignin dimer formed by two guaiacyl (G, G') monomers linked with a β-O-4 ether bond. (B) phenylpropanoid unit, from which are formed the three monolignols before polymerisation: p-coumaryl (R1=R2=H), coniferyl (R1=H, R2=OMe), and sinapyl (R1=R2=OMe) alcohols.

To capture the effect of those flexible β-O-4 structures on NMR parameters we considered five different solvents: pure water, ethanol (EtOH), acetonitrile (ACN), and 75wt% water-ethanol and water-acetonitrile mixtures. These solvents have been chosen because they were recently proposed for lignin fractionation with original properties in organosolv processes [35–37]. In particular, aqueous ethanol is among the most used alcoholic organic solvent in comparison to the other alcohols employed in the organosolv pulping of lignocellulosic biomass in biorefinery [37–41]. The organosolv lignin is recuperated by either removal of the organic solvent or dilution of organic solvent with water [39]. The isolated lignin structures vary significantly with pretreatment conditions in the organosolv pulping, notably with the type of solvent. Although, specific NMR fingerprints of each lignin unit are effectively used to analyze the resulting structures of the lignin fragments, knowledge about the impact of lignin unit conformational changes [14] are still scarce. In this study, we present detailed atomistic insight on the influence of guaiacyl conformations on the ^{13}C and ^{1}H NMR chemical shifts and propose a methodology for theoretical investigations using a combination of molecular dynamics, QM/MM and DFT calculations as well as predictions of NMR parameters based on Gradient Boosting Regressor (GBR) machine learning model (ML).

Methodology and computational details

The approach we employed combines molecular dynamics (MD) simulations for equilibration of the lignin dimers in the five solvents, followed by Born-Oppenheimer Molecular Dynamics (BOMD)/MD simulations within the QM/MM additive scheme to refine the descriptions of lignin dimer-solvent interactions. The BOMD/MD trajectories were further subjected to calculations of NMR shielding tensors at the DFT level, following the approach already applied successfully to study NMR parameters of flexible amphiphilic molecules in solvents and at air-solvent interfaces [42,43]. Trajectories and computed NMR isotropic shielding values were subsequently used to construct the descriptors and training sets for the machine-learning gradient-boosting regression (ML-GBR) model [44].

Molecular Dynamics simulations

All-atom Molecular Dynamics (MD) simulations of each system described in Table 1 were carried out using the GROMACS package [45] version 2016.3, along with the 4-sites Transferable Intermolecular Potential (TIP4P) for liquid water [46], the CHARMM36 (Chemistry at HARvard Macromolecular Mechanics) additive force field [47] for the organic solvent components, and the CHARMM-compatible force field for lignin [48]. Solvent structure for the organic solvents was available at the GROMACS molecule and liquid database[49]. For each simulation box, energy minimization was performed using the steepest descent algorithm until convergence to a tolerance of 100 kJ mol^{-1} nm^{-1}. After minimization, restrained simulations were performed for 200 ps at 298.15 K to allow solvent equilibration around the lignin dimer. Afterwards, 10-ns MD simulations were carried out with a frame-saving rate (for analysis) of 1 ps, to sample various conformations of lignin G-b-G dimer. Temperature and pressure coupling was handled using the leap-frog stochastic

dynamics integrator and the Parrinello-Rahman method, respectively. Initial velocities were generated from a Maxwell distribution at 298.15 K and the isothermal-isobaric (NPT) ensemble was considered for data collection. Neighbour searching and short-range nonbonded interactions were handled with the Verlet cut-off scheme. Electrostatics were treated with the Fast Smooth Particle-Mesh Ewald (SPME) method, with a Coulomb cut-off of 1.2 nm, a fourth-order interpolation and Fourier spacing of 0.12 nm. Van der Waals (vdW) interactions were treated using the Lennard - Jones potential with a cut-off distance of 1.2 nm.

Table 1. Configuration of simulated systems and equilibrium size of simulation boxes for the lignin dimer. Solvents studied include water, ethanol (EtOH), acetonitrile (ACN), and their 75 wt% aqueous mixtures.

Solvent system	Number of cosolvent molecules	Number of water molecules	Cubic box side length (nm)	Volume (nm^3)
Water	0	1448	3.54	44.55
EtOH -water	250	641	3.51	43.41
ACN-water	272	619	3.51	43.46
EtOH	399	0	3.4	39.48
ACN	456	0	3.45	41.01

Molecular dynamics simulations at QM/MM level

Born Oppenheimer Molecular Dynamics (BOMD) coupled to classical MD simulations (BOMD/MD) were carried out for 12 ps simulation time for each solvent composition. MD snapshots were extracted containing the lignin dimers, along with a 15 Å radius solvent drop. The lignin dimer was included in the QM (DFT) layer, whereas the solvent atoms were treated classically with the OPLS-AA force field [51]. In addition, the Onsager reaction field model is applied to represent the solvent as a continuum medium outside the explicitly treated solvent droplet at MM level using a BOMD/MD/PCM(Onsager) approach, implemented in deMon2k 6.0.2 developers version [52,53]. The dielectric constants for water, EtOH, ACN, EtOH-water and ACN-water were set up to 78.0, 38.0, 25.0, 35.0 and 45.0, [54,55] respectively. BOMD/MD/PCM(Onsager) simulations were carried out at 300 K in the canonical ensemble using a Nosé-Hoover chain of 5 thermostats with frequencies of 400 cm^{-1}. The integration time-step was set to 1 fs. The linear and the angular momenta of the whole lignin-explicit solvent system were conserved with a threshold of 10^{-8}, and therefore the rotational and translational degrees of freedom of the whole QM/MM system were kept frozen to avoid spurious translation or rotation in the space. These simulations were carried out with deMon2k.6.0.2 developer's version. PBE [56] exchange-correlation functional with double-ζ quality wave functions (DZVP)[57] were used. Automatically generated auxiliary functions up to l = 2 (labelled as GEN-A2*) were used for fitting the density with the GGA functionals, thus decreasing the computational time with comparable accuracy to density calculations from the molecular orbitals [58]. An empirical dispersion-like term (D2 approximation[59]) was added to the DFT energy and energy gradients. This level of approximation in combination with PBE functional has already been found to be a good compromise between accuracy and computational time of NMR properties in organic and bio-organic compounds, [29,42,43,52] and solids [60].

DFT calculations of NMR parameters

The last 7.4 ps out of the 12 ps BOMD/QM/PCM trajectories was subjected to statistical analysis with frames extracted at intervals of 8 fs for the DFT calculations of the shielding tensor of lignin atoms. The shielding tensors were computed with the gauge-independent atomic orbitals (GIAO) scheme, as implemented [61] in the deMon2k program, using triple-ζ bases [57], GEN-A2* auxiliary functions and PBE functional for the lignin dimer, treated at DFT level. The ^{13}C and ^{1}H isotropic shielding (σ_{iso}) values were averaged over the 925 frames. Larger sets of geometrical frames (see below) were considered in the training and validation data sets in the ML-GBR model. These data sets were also augmented by the computed with the same approach σ_{iso} (^{17}O) values. Note that in the ML-GBR model only the isotropic shielding values were treated. The chemical shift is obtained as the difference between the DFT isotropic shielding and the reference isotropic shielding $\left(\boldsymbol{\sigma}_{\boldsymbol{iso}}^{\boldsymbol{ref}}\right)$ in tetramethylsilane (TMS) that was computed at the same level of theory. We obtained $\boldsymbol{\sigma}_{\boldsymbol{iso}}^{\boldsymbol{ref}}$ (^{13}C) = 162.72 ppm and $\boldsymbol{\sigma}_{\boldsymbol{iso}}^{\boldsymbol{ref}}$ (^{1}H) = 31.10 ppm.

Machine-learning Gradient Boosting Regressor (ML-GBR) model

We have used the gradient-boosting regressor method, which enables us to create strong learning trees from poorly learning trees

[44]. This approach utilizes boosting so that the trees are created sequentially as opposed to random forests where the trees are generated in parallel. Each new tree is created with an effort to reduce the prediction error learning from the errors of the previous tree. The goal of the method is to achieve the lowest possible error while keeping the predicted values as accurate as possible.

Two data sets in CSV format were prepared using the Cartesian coordinates of the selected trajectories, subjected to NMR calculations. The first dataset contains the Cartesian coordinates (x, y and z) of the G-b-G, the calculated σ_{iso}, the name of the chemical element and the name of the solvent (provided in the Electronic Supplementary Information (ESI) 2-6 files). The second CSV file contained 3x3 tensors and the name of the corresponding solvent-lignin dimer model. We have used the Dscribe package [62] to convert our data to Smooth Overlap of Atomic Positions (SOAP) descriptor vectors. Individual structures were represented as Atoms class objects from the ASE package [63] with the use of 3x3 tensors. We have begun by creating a DScribe.SOAP object, for which the parameters such as the number of basis functions, range, level l and a list of all elements in our data were set. Subsequently, the DScribe.SOAP.create[62] function was used to create a SOAP vector for each atom. The complete dataset was split into a training and test set in a ratio of 6:4. The sizes of the whole data sets (training + testing) are provided for every individual combination as the number of chemical environments in the "Results and discussion" section. Individual training times and predictions are reported in the same section for each combination.

We have used the Anaconda distribution for Python 3.8.5, utilizing the scikit-learn program package [44,64] with the GBR model, where the *random_state hyperparameter* was set to 0 and the rest of the hyperparameters were set to the default values.

Results and discussion

Dynamic (time-averaged) ^{13}C chemical shift and conformational effects

Our previous studies[37,65] on the evolution of lignin dimer conformations along the equilibrium dynamics in several solvents revealed that the β–O-4 linked guaiacyl dimer has the smallest solvent accessible surface area and the most stacked conformation in water. This effect has been attributed to lignin-water hydrophobic interactions favouring the phenolic G ring-ring interactions, instead of lignin-water interactions. The increase of the organic co-solvent components decreased the hydrophobic effect of water and resulted in open T-shaped conformations. The averaged conformations in the five solvents, obtained from the averaged coordinates during the 12 ps trajectories, are presented in Figure 2. The general shape of these conformations, established along the 10 ns MD simulations, did not change during the significantly shorter 12 ps BOMD/MD/PCM trajectories.

The DFT time-averaged ^{13}C chemical shifts ($<\sigma_{iso}(^{13}C)>$) can capture the vibrational degrees of freedom due to the thermal fluctuations of the lignin nuclei and the effect of solvent-induced conformational changes when averaged over a large number of structures along the BOMD/MD trajectories [42,43,66,67]. An expectation of the convergence of the ^{13}C isotropic shielding, examined from the rolling averaged with a step of 1, reported in ESI, Figures S1-S3, indicates that the σ_{iso} values converged reasonably well for most of the carbons after averaging over 600-700 geometrical frames in the five solvents with a standard deviation error (STDE) of the averaged σ_{iso} ≤ 3.2%. The averaged σ_{iso} (^{13}C) with STDEs for each of the twenty carbons in the G-b-G dimer models are plotted in Figures S4-S8.

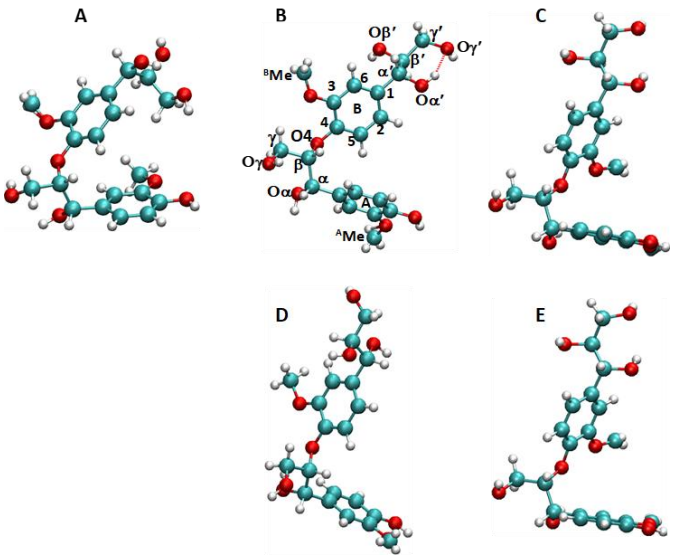

Figure 2. Averaged structures from the 12 ps BOMD/MD/PCM simulations of β-O-4 linked G-G lignin dimer (G-b-G) in (A) water, (B) ethanol, (C) acetonitrile, (D) water-ethanol and (E) water-acetonitrile solvents. In (B) the atomic labelling, as referred to throughout the "Results and discussion" section, is provided.

Computed $<\delta(^{13}C)>$ in Table 2 are the values additionally averaged over the chemically equivalent carbons in A and B rings (see Fig. 1), whereas $<\delta(^{13}C)>$ in the ring A and the ring B, separately, are presented in Table S1. As follows from Table S1, there is a non-negligible difference in the shielding of carbons in each ring. Moreover, the split of $<\delta(^{13}C)>$ in rings A and B varies with the type of the solvent component(s). Experimentally, a difference of $\delta(^{13}C)$ of carbons belonging to different rings has been established as well[2]. For example, $\delta(^{13}C)$ in rings A and B in the β-O-4 linked G-G dimer (compound number 248 in ref. 2) amounts to ~6 ppm in deuterated chloroform ($CDCl_3$) solvent, decreases slightly in acetone and further reduces to 2.5-3 ppm in DMSO. This shows that the local structural arrangement varies in the rings, an effect well captured by our averaged NMR calculations, including the thermal fluctuations of the lignin nuclei in addition to the dimer conformations

ARTICLE

Table 2. Dynamic (time-averaged) ^{13}C chemical shifts in ppm in water, ethanol (EtOH), acetonitrile (ACN), ethanol-water (EtOH-water) and acetonitrile-water (ACN-water) solvents. Calculated values are compared with literature results (exp).

Carbon atom	Water	EtOH	ACN	EtOH-Water	ACN-Water	exp
C2	108.1	108.6	105.0	101.7	103.1	106 -114[a]
C5	104.7	120.1	113.7	109.1	104.9	114–117[a]
C6	109.7	114.1	109.0	115.6	114.2	117–123[a]
C3	155.6	151.1	150.8	150.0	154.6	140–155[a]
C4	155.2	129.2	146.3	151.6	154.3	140–155[a]
C1	146.9	133.5	133.5	130.0	141.2	127–140[a]
Cα	72.1	72.6	82.0	76.9	78.6	67–78[a]
Cβ	75.3	86.2	77.5	81.3	77.6	78–90[a]
Cγ	59.0	68.1	61.1	64.5	59.5	59.8[b] ; 60.2[c]
CMe	56.1	54.4	56.1	53.2	57.9	54–57.5[a]

[a]from refs. [1,2,5,13,68]; [b]from ref. [13]; c from ref. [2]

The last column in Table 2 reports the intervals of known experimental $\delta(^{13}C)$ values in β-O-4 linkages between G-units [1,2,5,13,68] collected from solid-state or liquid NMR measurements. Our calculated chemical shifts fall in the same ranges, despite the different solvents considered here. The few outliers, observed in Table 2, are shifted by a maximum of 10 ppm from the known experimental intervals. This is the case of C5 and C6 atoms in all the solvents, and C4 in ethanol. In the other solvents, the chemical shifts of the carbons involved in the β-O-4 linkage, e.g. C4 and Cβ, fall in the experimental range of data with very good agreement between the experimental and computed $<\delta(^{13}C\beta)>$, independently of the solvent. The variations of $<\delta(^{13}C)>$ according to a particular solvent, allows therefore distinguishing carbon sites more sensitive to specific conformations.

A more detailed analysis of the solvent effect on $<\delta(^{13}C)>$ is provided below, using $<\delta(^{13}C)>$ in G-b-G conformation in water solvent as a reference. As follows from the results in Table 2 and Table S1, a clear tendency of displacing chemical shifts, if organic co-solvent components are present, is obtained for C1, Cα, Cβ and Cγ sites. The chemical shift of C1 decreases by ~5 to ~15 ppm in the pure organic and organic-water mixed solvents. This is most probably related to the conformational changes from the stacked (in water) to the T-shaped like conformers, whose population dominates in the solvents with organic components. On the contrary, a tendency to deshield the carbon nuclei (an increase of $<\delta(^{13}C)>$) is found for Cα, Cβ and Cγ in the pure organic solvents, which might be attributed to a concomitant effect of conformational changes and H-bonds arising between (co)solvent molecules and OH groups (see below) attached to these carbons. Indeed, our previous analysis of the MD trajectories revealed well-organized and structured first solvation shells around O-Cα and O-Cγ sites in all the solvents [65]. The previously studied radial distribution functions of N(ACN) – O-CMe in the G-b-G dimer revealed acetonitrile coordination to the oxygen atoms that are bound to methyl carbons[42,65]. The chemical shift of the methyl carbon appears, however, to be only little affected (by a maximum of 3 ppm) by the presence of EtOH co-solvent, and almost not affected by ACN co-solvent.

Dynamic (time-averaged) ^{1}H chemical shifts, internal H-bonds and conformation dynamics

The time-averaged calculations of the ^{1}H chemical shift, $<\delta(^1H)>$, are presented in Table 3. All the averaged ^{1}H isotropic shieldings are in the range 22-28 ppm, with STD errors of 0.7-0.8 ppm of the averaged σ_{iso} per H site, reported in Figures S4 - S8. The chemical shifts of methyl (Me) hydrogen atoms in Table 3 are the averaged value over the three H atoms in the Me group. In Figure 3, the

chemical shifts of all the individual hydrogens are reported, thus illustrating visually the effect of conformations arising in the considered solvents. The computed $<\delta(^1H)>$ spread in the interval 3.8 – 12.1 ppm, which generally fall in the range of the experimental values between 0 - 12 ppm[1, 49, 50, 54], knowing that a precise spectral assignment of ^1H suffers from the signal overlaps [70].

Table 3. Dynamic (time-averaged) ^1H chemical shifts in ppm in water, ethanol (EtOH), acetonitrile (ACN), ethanol-water (EtOH-water) and acetonitrile-water (ACN-water) solvents. For the atomic labels and numbers see Figure 2B. Calculated values are compared with literature results (Exp).

Hydrogen atom	water	EtOH	ACN	EtOH-water	ACN-water	Exp.[a]
			Ring A			
H_{Me}	3.8	4.7	4.5	4.3	5.7	3.8
$H_{O\alpha}$	6.3	3.1	5.3	6.2	6.3	-
$H_{O\gamma}$	4.4	2.7	5.4	2.3	2.6	-
H_{C2}	4.9	11.2	9.8	10.8	9.6	-
H_{C5}	8.5	8.1	8.2	8.4	8.1	-
H_{C6}	8.6	9.8	8.2	8.4	7.0	-
			Ring B			
H_{Me}	5.2	6.0	5.4	6.0	6.4	3.9
$H_{O\alpha'}$	4.0	11.2	4.5	4.1	7.8	-
$H_{O\beta'}$	4.5	5.4	5.9	4.1	3.0	-
$H_{O\gamma'}$	4.1	3.0	1.3	4.5	2.6	-
H_{C2}	6.1	10.2	10.4	11.7	11.1	-
H_{C5}	5.8	12.1	10.3	9.5	9.4	6.9
H_{C6}	5.1	11.1	9.2	10.3	10.2	6.7
$H_{C\alpha}$	5.1	9.6	7.0	7.3	6.4	4.9
$H_{C\alpha'}$	4.9	8.8	7.3	5.4	8.7	
$H_{C\beta}$	5.4	10.8	8.9	7.8	5.9	4.1
$H_{C\beta'}$	2.9	9.6	5.1	6.9	7.3	
$H_{C\gamma}$	4.5; 4.6	6.7; 7.4	6.7; 5.6	6.3; 6.2	5.9; 5.9	3.5; 3.7
$H_{C\gamma'}$	3.4; 2.6	4.4; 4.3	4.3; 4.0	3.9; 4.6	2.7; 2.0	

[a] from ref. [2]

Following the results in Table 3 and Figure 3, the $<\delta(^1H)>$ values of hydrogens in the same chemical group, but in two different rings, may vary even by 5-6 ppm, according to a specific conformation and H-site. The largest splitting in $<\delta(^1H)>$ is found in the conformers in the organic solvents, either pure or mixed with water. In particular, an important deshielding in the organic solvents is found for aromatic hydrogens in ring B. In pure ethanol, $<\delta(^1H)>$ increases up to 11.2 (H_{C2}-ringA), 10.2 (H_{C2}-ringB) and 12.1 ppm (H_{C5}-ring A). The addition of water to ethanol even increases the chemical shift of H_{C2} by 1.5 ppm, but deshields H_{C5} by 1.6 ppm in ring B. In water, these values are closer to the experimental ones, recently reported to be 6.94 (H_{C2}), 6.65 (H_{C5}) and 6.75 (H_{C6}) ppm in G units in milled wood lignin of Chinese quince fruit [68]. Similar $\delta(^1H)$ values (between 6.7-7.1 ppm) were assigned to the ring hydrogens in guaiacyl from NMR measurements in DMSO [69] and acetone solvents [2]. Another spike in the computed ^1H NMR is $<\delta(^1H)>$ in the hydroxyl group in ring B ($H_{O\alpha'}$) in ethylene solvent. Focusing on this hydrogen, we noticed its involvement in an internal hydrogen bond (HB) with the neighbour hydroxy group oxygen, $O\gamma'$, (see atomic labelling in Fig. 2B). We note that the geometrical H-bond analysis was carried out by setting cut-offs of 3 Å for OH…O distances and 145° for the O –H…O angles. This analysis revealed the formation of an internal HB only in ethanol solvent, preserved along the 12 ps QM/MM dynamics. Other internal OH…O distances < 3 Å have been identified in the G-b-G dimer in the five solvents. However, their angles are < 130°.

ARTICLE

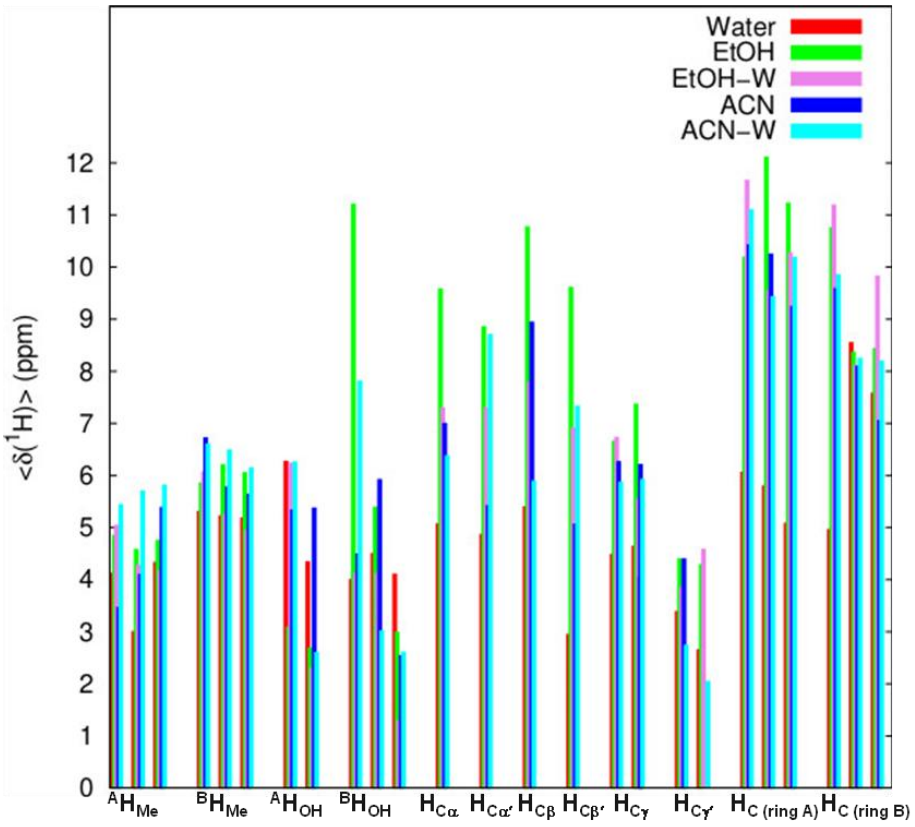

Figure 3 Visualization of the averaged DFT ^{1}H chemical shifts of all hydrogens in G-b-G model in water, ethanol (EtOH), aqueous 75wt% ethanol (EtOH-Water), acetonitrile (ACN), and 75wt% acetonitrile-water (ACN-Water) solvents.

The evolution of the $H_{O\alpha'}\ldots O_{\gamma}'$ distance in the five solvents, along the snapshots considered for the NMR shielding calculations, is presented in Figure 4. It demonstrates indeed that in EtOH, the $H_{O\alpha'}\ldots O\gamma'$ distance predominantly oscillates around 2 Å. Its average value is 2.08 ± 0.36 Å. The addition of water to EtOH changes the G-b-G conformer by breaking this internal H-bond leading to an averaged $H_{O\alpha'}\ldots O\gamma'$ distance of 4.61 ± 0.58 Å. In water, ACN and ACN-water solvents, the averaged $H_{O\alpha'}\ldots O_{\gamma}'$ separation amounts to 2. 23 ± 0.50 Å, 4.16 ± 0.67 Å and 2.54 ± 0.45 Å, respectively. The two relatively short distances in water and ACN-water mixture are associated with averaged $O_{\alpha'} - H_{O\alpha'}\ldots O_{\gamma'}$ angles < 145°, equal, respectively, to 124.4 ± 24.2° and 108.1 ± 27.2°. The only internal HB is therefore established in ethanol. This internal HB decreases the electronic density around $H_{O\alpha}$ and limits its movements during dynamics, which indeed is expected to favour a downfield shift or deshielding (higher chemical shift) in $H_{O\alpha}$.

ARTICLE

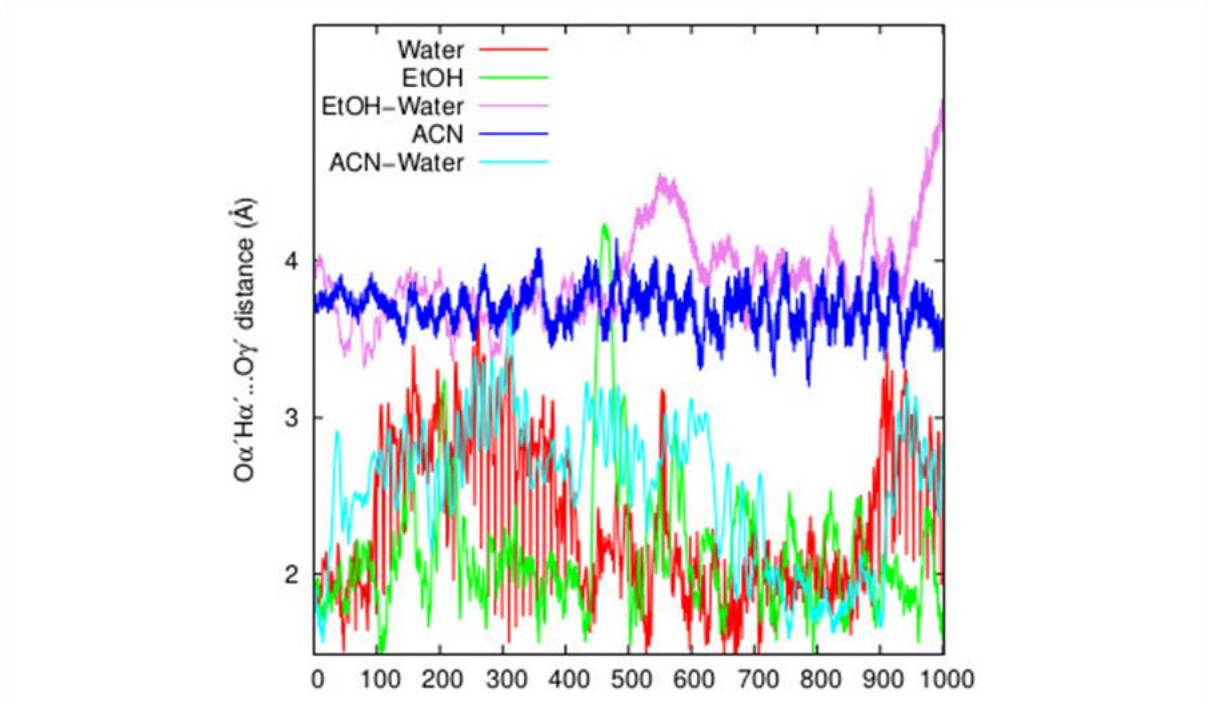

Figure 4. Evolution of the Oα′ –$H_{Oα′}$…Oγ′ distance in the G-b-G dimer model in water, ethanol (EtOH), aqueous 75wt% ethanol (EtOH-Water), acetonitrile (ACN), and 75wt% acetonitrile–water (ACN-Water).

In the theoretical models, the shielding/deshielding of H-atoms in hydroxy groups might be also affected by the thermal fluctuations of the G-b-G conformer in the different solvents, because of HBs formation with external solvent molecules. A closer examination of the G-b-G dynamics in the five solvents evidenced the formations of external HBs only between the hydroxy groups and the solvent molecules; however, their occurrence depends on the specific solvent composition. In Table 4, the averaged H-bond distances between hydroxy group H-sites and the solvent molecules are collected.

Table 4. Averaged bond lengths and STD errors in Å of all the hydrogen bonds formed between H_O site in the hydroxy groups of the lignin dimer and the solvent molecules. The bond lengths are averaged over the trajectory frames, used to compute the ^{1}H chemical shifts.

H atom	Water	EtOH	ACN	EtOH-Water	ACN-Water
$H_{0α}$	4.28 ±1 .43	1.87 ± 0.38	1.88 ±0.19	1.84* ± 0.21	1.77** ± 0.16
$H_{0γ}$	1.83 ± 0.26	-	1.98 ± 0.30	1.65** ± 0.13	1.69* ± 0.16
$H_{0α′}$	-	-	1.92 ± 0.30	1.77* ± 0.19	2.02** ± 0.43
$H_{0β′}$	-	1.80 ± 0.21	-	1.83* ± 0.31	-
$H_{0γ′}$	1.68 ± 0.17	1.87 ± 0.38	-	1.94** ± 0.45	1.67** ± 0.15

*external HB with O in C_2H_5OH or with N in $CNCH_3$; **external HB with a water molecule

Interestingly, $H_{Oα′}$, for which a strong deshielding is found (see above) in ethanol, is not involved in an external HBs with this solvent. This, therefore, confirms that the main factor contributing to the downfield shift of $H_{Oα′}$, is indeed the internal Oα′ –H_Oα′…Oγ′ HB. In the aqueous organic solvents, HBs with both solvent components exists, as follows from Table 4. This is in agreement

with our previous analysis from the classical MD simulations [65]. Water forms HBs only with HOγ,γ′, whereas it is more attracted by the hydroxy groups if mixed with the organic solvents. The latter finding is fully in line with the hydrophobic character of lignin. Indeed, in pure water solvent, water molecules prefer to interact between them than with the lignin dimer. In the mixtures of organic solvents with water, the HB network of water is strongly disrupted (especially in ACN-water), which favours the water-hydroxy group interactions. The formation of various external HBs certainly influences the amount of the computed ^{1}H isotropic shielding and respectively, the chemical shift, because of the constrained movements of lignin hydroxy groups involved in H-bonds with solvent components.

This is not the case for the other H nuclei ($H_{C\alpha,\alpha'}$, $H_{C\beta,\beta'}$ and H_{C2}, H_{C5}, H_{C6}), undergoing notable downfield shifts by more than 4 ppm in the organic solvents, or their mixtures with water. An observation of the structures revealed that solvent molecules are not in proximity within 3 Å to these H atoms. Therefore, contributions of external HBs between solvent molecules and the ring hydrogens in the G-b-G model to their downfield shifts can be certainly ruled out. The downfield shift induced by the organic solvents appeared to be therefore related to a specific conformational geometry and its dynamics in a particular solvent. These structural analyses concerning the time-averaged ^{1}H and ^{13}C NMR shieldings demonstrate the higher sensitivity of the H nuclei than the C nuclei toward a specific G-b-G conformation and its dynamics.

Machine learning Gradient boosted regressor model applied to ^{13}C and ^{1}H isotropic shielding.

The BOMD/MD/PCM trajectory frames, subjected to NMR isotropic shielding calculations, were further used to examine the performance of the gradient boosted regressor ML model for prediction of NMR isotropic shielding. A successful ML model would therefore speed up the processes for time-averaged computations of NMR parameters, including the thermal nuclei fluctuations while training the ML model to predict σ_{iso} on a reduced number of geometrical frames for various lignin isomers. These predicted σ_{iso} values for each atomic site of interest can be consequently averaged to give the final result, thus capturing the effect of local structural changes due to different conformations, thermal fluctuations of the lignin nuclei and their interactions with the environment.

ML models, based on Gaussian Processes Regression, Neural Network and Kernel Ridge Regression approaches have been already found successful for the prediction of isotropic shielding in molecular crystals [71], crystalline and amorphous silica and aluminosilicates [72,73], as well as in proteins[74] and isolated lignin monomers [26]. Several types of descriptors have been tested, among them SOAP, used by us in the present study. The SOAP descriptors have been established to provide a good relationship between the local atomic structural environment and the isotropic shieldings. To the best of our knowledge, ML applications to predict σ_{iso} in flexible structures has been carried out on proteins and, more generally, in the domain of life science, using experimental datasets and attempting to achieve an automated assignment of the signals [75].

The challenge in our study is to examine the applicability of an ML model for large-size and heterogeneous systems, composed of solvent molecules (studied at MM level) and the G-b-G lignin dimer, described at QM level. Because the NMR properties are computed at QM level only, the training dataset consisted of the G-b-G dimer atomic positions along the dynamics and the corresponding isotropic shielding for every nucleus (H, C, or O) in lignin model. These data-sets are available in the ESI files.

The correlations between the σ_{iso} values predicted with ML–GBR model and those, computed with DFT, are plotted in Figure 5 for σ_{iso} (^{13}C) and in Figure 6 for σ_{iso} (^{1}H). The first 60% of the trajectory frames were used to train the ML-GBR model, while the remaining 40% were set as test cases, as described in the section "Computational details". In the training and test sets, all carbon, hydrogen, and oxygen atoms, respectively, were considered without differentiating between a specific atomic site, but differentiating among the solvents. We, therefore, ensure a larger number of chemical environments, necessary for the good training of the GBR algorithm, because our aim here is to examine the capability of the ML – GBR model with SOAP descriptors in predicting correctly the computed DFT values for each atomic chemical environment.

ARTICLE

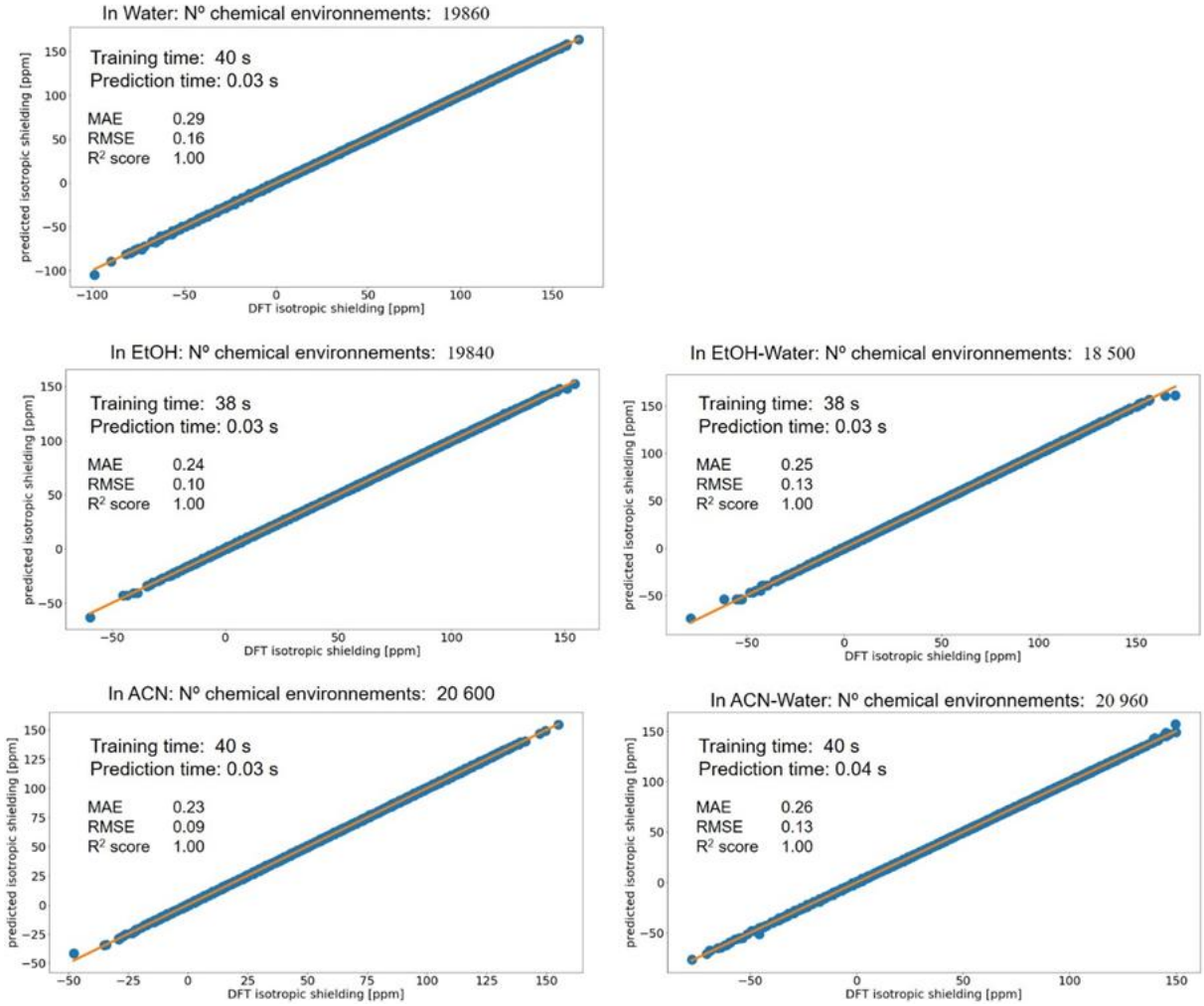

Figure 5. Correlation between he predicted with ML – GBR model and the computed with DFT σ_{iso} (^{13}C) values of G-b-G lignin dimer in water (top); ethanol (middle left), aqueous 75wt% ethanol (middle right) - acetonitrile (down left), and 75wt% acetonitrile –water (down right). The number of chemical environments, training and prediction time, mean averaged error (MAE), and root mean square errors (RMSE) are reported for each solvent.

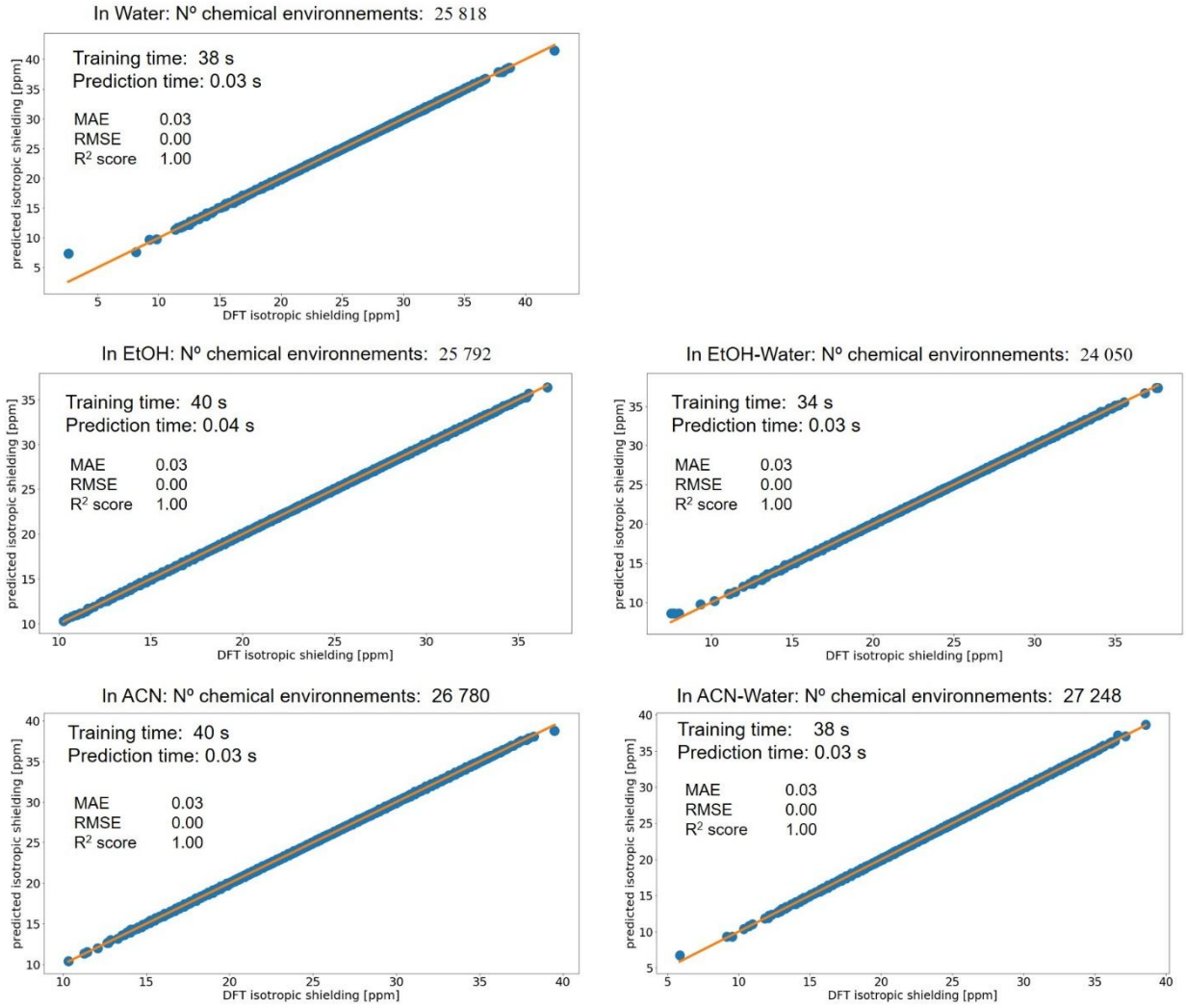

Figure 6. Correlation between the predicted with ML – GBR model and the DFT computed σ_{iso} (^{1}H) values of G-b-G lignin dimer in water (top); ethanol (middle left), aqueous 75wt% ethanol (middle right), acetonitrile (down left), and 75wt% acetonitrile –water (down right). The number of chemical environments, training and prediction time, mean averaged error (MAE), and root mean square errors (RMSE) are reported for each solvent.

The correlations between the ML – GBR and the DFT results in Figures 5 and 6 revealed a R^2 coefficient of 1 for both ^{1}H and ^{13}C NMR shieldings. There is not an outlier among the carbons. The mean averaged error (MAE), reported in Figures 5 and 6, are in the interval 0.23-0.29 ppm for carbons and 0.03 ppm for hydrogens. The only outlier is established among σ_{iso} (^{1}H) in water solvent, which is $^{A}H_{C5}$ (in ring A) in the 686-th trajectory frame of G-b-G dimer in water. For this hydrogen site, DFT σ_{iso}(^{1}H) equals to 2.61 ppm, whereas the predicted value is 7.39 ppm. This very strong deshielding in the 686th frame is also an outlier among the DFT computed isotropic shielding values, as follows from the distribution of the ^{1}H isotropic shieldings presented in Figure 7 (A). The σ_{iso} (^{13}C) distribution is provided in Figure 7(B) and we can clearly distinguish the three peaks corresponding to the three distinct intervals of carbons in Table 2. In addition to the very good correlation between the predicted and computed σ(^{1}H) and σ(^{13}C) in Figures 5 and 6, we note the very short training and prediction time of a maximum of 40 s and 0.04 s, respectively. The combination of SOAP with the GBR model is therefore found to be a promising approach for isotropic chemical shift prediction in flexible organic molecules in solvents.

ARTICLE

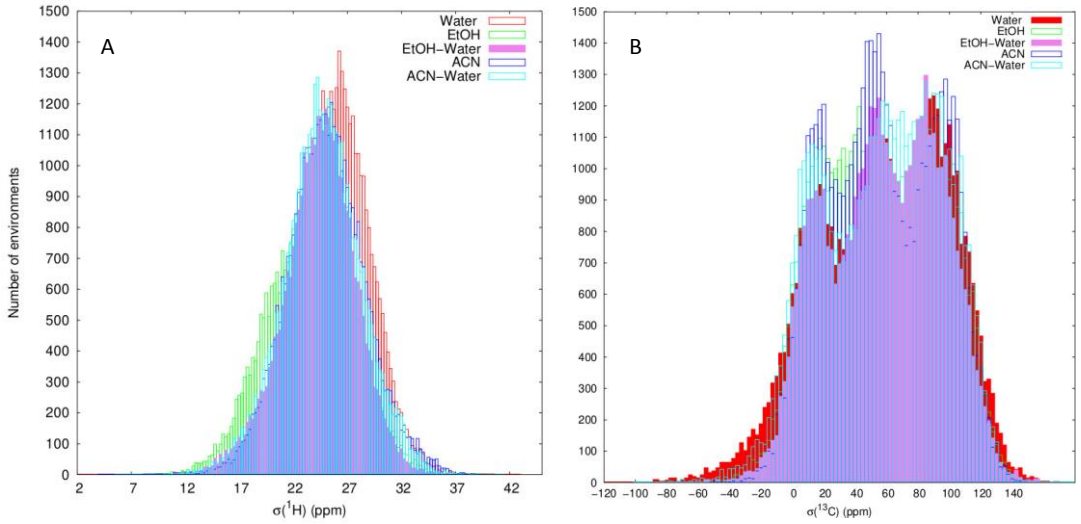

Figure 7. Distribution of the full set of ^{1}H (A) and ^{13}C (B) isotropic shieldings of G-b-G lignin model in water, ethanol (EtOH), ethanol-water (EtOH-water), acetonitrile (ACN), and acetonitrile-water (ACN-Water) solvent. The histograms are obtained with an interval of 0.25 ppm for ^{1}H and 2.5 ppm for ^{13}C in the frequency count.

Training the model over a sufficiently large set of chemical environments is however crucial. In the case of σ_{iso} (^{17}O), the number of O-environments for the same trajectories, that we have considered for ^{1}H and ^{13}C calculations, is significantly smaller, ranging from 8928 in EtOH to 9432 in the mixture of acetonitrile and water (see Figure S9). The correlation between predicted and DFT σ_{iso} (^{17}O), reported in Figure S9 is somewhat less satisfactory, giving MAE of 0.5 – 0.6 ppm and several outliers. Interestingly, in the pure ethanol and acetonitrile solvents, there are no outliers among the predicted ML-GBR values. Details about the outliers are given in the ESI section, below the Figure S9. Taking into account all the lignin elements in the five solvents together allowed to increase the number of chemical environments to 274 340 (see Figure S10) and additionally increase the heterogeneity of data in the training set. Two σ_{iso} values of ^{17}O and one of ^{13}C outliers have been found in this example, thus evidencing the necessity to take into consideration not only a large number of the chemical environments but also their degree of heterogeneities. In the latter case, the three O, C and H nuclei have too diverse shielding values and more outliers have been identified.

Conclusion

In this work we propose a multiscale methodology, combining classical and quantum/classical dynamic simulations and subsequent DFT calculations of dynamic (time-averaged) NMR parameters of β-O-4 linked guaiacyl dimer, which enables us to take into consideration the conformational changes and thermal fluctuations of lignin nuclei in different solvents relevant for organosolv lignocellulose pulping. Moreover, this allows us to better analyze and understand the factors playing a predominant role in NMR shielding constants concerning lignin conformations and their dynamics.

The majority of time-averaged ^{13}C chemical shifts fit very well within the experimentally reported NMR data intervals, independently whether the NMR measurements were carried out in different solvents (typically DMSO and acetone solvents) or using solid-state NMR techniques. It, therefore, follows that those carbons are practically unaffected by the lignin interaction with solvents or by the specific lignin conformation. However, a clear tendency of displacing $<\delta(^{13}C)>$ in the presence of ethanol and acetonitrile co-solvent components compared to the experimental data is obtained for C1, Cα, Cβ and Cγ sites. This result has been attributed to the effect of conformational changes from the stacked (in water) to the T-shaped G-b-G conformer that dominates in the solvents with organic components.

Hydrogens experience significantly larger sensitivity toward conformational changes in the solvents. Only one stable internal H-bond ($O\alpha' - H_{O\alpha'}...O\gamma'$) was identified in the G-b-G conformer in ethanol solvent, which explains the very strong downfield shift of $<\delta(^{1}H_{O\alpha'})>$. Interestingly, also H atoms linked to ring carbons and not

directly interacting with solvent molecules have downfield shifts of ~4-5 ppm in ethanol, acetonitrile and in their 75 wt% aqueous mixtures.

The examination of SOAP descriptors with ML-GBR model demonstrated that NMR parameters can be effectively learned and predicted in only a few tenths of seconds. This opens the possibility to build NMR datasets of the rich manifold of lignin, among other flexible organic molecules in solvents, including different monomers, dimers or larger oligomers and their interlinkage.

Author Contributions

S.M.A.S. prepared the models, carried out MD and QM/MM/PCM simulations. R.G and T.M. performed NMR calculations and prepared the data sets. D.D. and I.M.O. adapted and carried out GBR training and prediction calculations and the ML-GBR analysis. F.D. aided in interpreting the results and editing the manuscript. T.M. conceived and supervised the project, and took the lead in writing the manuscript with the inputs from all authors. All authors discussed the results and contributed in the revision of the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

This work was partially funded by an ENSCM PhD grant in the framework of the SINCHEM Joint Doctorate program under the Erasmus Mundus Action 1 Programme (FPA 2013-0037). Access to the HPC resources of CCRT/CINES/IDRIS was granted under the allocation A0070807369 and A0090807369 by GENCI.

1 Y. C. Y. Lu, Y. C. Y. Lu, H. Q. Hu, F. J. Xie, X. Y. Wei and X. Fan, *J. Spectrosc.*, , DOI:10.1155/2017/8951658.
2 S. Ralph, J. Ralph, L. Landucci, U. F. Service and J. Ralph, .
3 J. L. Wen, S. L. Sun, B. L. Xue and R. C. Sun, *Materials (Basel).*, 2013, **6**, 359–391.
4 B. Jiang, Y. Zhang, T. Guo, H. Zhao and Y. Jin, *Polymers (Basel).*, , DOI:10.3390/polym10070736.
5 F. Yue, F. Lu, S. Ralph and J. Ralph, *Biomacromolecules*, 2016, **17**, 1909–1920.
6 J. F. K. Ewellyn A Capanema 1 , Mikhail Y Balakshin, *J Agric Food Chem*, 2004, **52**, 1850–60.
7 M. Y. Balakshin, E. A. Capanema, C. L. Chen and H. S. Gracz, *J. Agric. Food Chem.*, 2003, **51**, 6116–6127.
8 J. Ralph and L. Landucci, *Lignin and Lignans*, 2010, 137–243.
9 L. Zhang, G. Gellerstedt, J. Ralph and F. Lu, *J. Wood Chem. Technol.*, 2006, **26**, 65–79.
10 Y. Tobimatsu, T. Takano, T. Umezawa and J. Ralph, *Solution-state multidimensional NMR of lignins: approaches and applications*, 2019.
11 J. Ralph, C. Lapierre, F. Lu, J. M. Marita, G. Pilate, J. Van Doorsselaere, W. Boeŕjan and L. Jouanin, *J. Agric. Food Chem.*, 2001, **49**, 86–91.
12 J. Ralph, C. Lapierre, J. M. Marita, H. Kim, F. Lu, R. D. Hatfield, S. Ralph, C. Chapple, R. Franke, M. R. Hemm, J. Van Doorsselaere, R. R. Sederoff, D. M. O'Malley, J. T. Scott, J. J. MacKay, N. Yahiaoui, A. M. Boudet, M. Pean, G. Pilate, L. Jouanin and W. Boerjan, *Phytochemistry*, 2001, **57**, 993–1003.
13 T. Kishimoto, Y. Uraki and M. Ubukata, *Org. Biomol. Chem.*, 2008, **6**, 2982–2987.
14 Y. Tokunaga, T. Nagata, K. Kondo, M. Katahira and T. Watanabe, *Holzforschung*, 2021, **75**, 379–389.
15 C. G. Yoo, Y. Pu, M. Li and A. J. Ragauskas, *ChemSusChem*, 2016, **9**, 1090–1095.
16 K. Cheng, H. Sorek, H. Zimmermann, D. E. Wemmer and M. Pauly, *Anal. Chem.*, 2013, **85**, 3213–3221.
17 S. D. Mansfield, H. Kim, F. Lu and J. Ralph, *Nat. Protoc.*, 2012, **7**, 1579–1589.
18 F. Lu and J. Ralph, *Plant J.*, 2003, **35**, 535–544.
19 R. Dupree, T. J. Simmons, J. C. Mortimer, D. Patel, D. Iuga, S. P. Brown and P. Dupree, *Biochemistry*, 2015, **54**, 2335–2345.
20 X. Kang, A. Kirui, M. C. Dickwella Widanage, F. Mentink-Vigier, D. J. Cosgrove and T. Wang, *Nat. Commun.*, 2019, **10**, 1–9.
21 O. M. Terrett, J. J. Lyczakowski, L. Yu, D. Iuga, W. T. Franks, S. P. Brown, R. Dupree and P. Dupree, *Nat. Commun.*, 2019, **10**, 1–11.
22 T. Wang, Y. B. Park, M. A. Caporini, M. Rosay, L. Zhong, D. J. Cosgrove and M. Hong, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 16444–16449.
23 T. Wang, P. Phyo and M. Hong, *Solid State Nucl. Magn. Reson.*, 2016, **78**, 56–63.
24 T. T. Nguyen, P. Q. Le, J. Helminen and J. Sipilä, *J. Mol. Struct.*, 2021, **1226**, 129300.
25 A. L. C. O. H. O. Ls, 2021, **55**, 41–54.
26 M. Jalali-Heravi, S. Masoum and P. Shahbazikhah, *J. Magn. Reson.*, 2004, **171**, 176–185.
27 M. Dračínský, J. Vícha, K. Bártová and P. Hodgkinson, *ChemPhysChem*, 2020, **21**, 2075–2083.
28 R. Pohl, O. Socha, P. Slavíček, M. Šála, P. Hodgkinson and M. Dračínský, *Faraday Discuss.*, 2018, **212**, 331–344.
29 T. Mineva, P. Gaveau, A. Galarneau, D. Massiot and B. Alonso, *J. Phys. Chem. C*, 2011, **115**, 19293–19302.
30 S. Guadix-Montero and M. Sankar, *Top. Catal.*, 2018, **61**, 183–198.
31 J. Rencoret, A. Gutiérrez, L. Nieto, J. Jiménez-Barbero, C. B. Faulds, H. Kim, J. Ralph, Á. T. Martínez and J. C. del Río, *Plant Physiol.*, 2011, **155**, 667–682.
32 S. Besombes and K. Mazeau, *Biopolymers*, 2004, **73**, 301–315.
33 S. Besombes, J. P. Utille, K. Mazeau, D. Robert and F. R. Taravel, *Magn. Reson. Chem.*, 2004, **42**, 337–347.
34 R. Stomberg, S. Li, K. Lundquist and B. Albinsson, *Acta Crystallogr. Sect. C Cryst. Struct. Commun.*, 1998, **54**, 1929–1934.
35 J. Ralph, J. Peng, F. Lu, R. D. Hatfield and R. F. Helm, *J. Agric. Food Chem.*, 1999, **47**, 2991–2996.
36 X. Gong, Y. Meng, J. Zhu, X. Wang, J. Lu, Y. Cheng, Y. Tao and H. Wang, *Ind. Crops Prod.*, 2021, **166**, 113471.
37 S. M. Aguilera-Segura, J. Bossu, S. Corn, P. Trens, T.

Mineva, N. Le Moigne and F. Di Renzo, *Macromol. Symp.*, 2019, **386**, 1–10.
38 M. Wu, J. Pang, X. Zhang and R. Sun, *Int. J. Polym. Sci.*, , DOI:10.1155/2014/194726.
39 X. Meng, S. Bhagia, Y. Wang, Y. Zhou, Y. Pu, J. R. Dunlap, L. Shuai, A. J. Ragauskas and C. G. Yoo, *Ind. Crops Prod.*, , DOI:10.1016/j.indcrop.2020.112144.
40 N. Hanis, A. Latif, N. Brosse, I. Ziegler-devin, L. Chrusiel, R. Hashim, M. H. Hussin, N. Hanis, A. Latif, N. Brosse, I. Ziegler-devin, L. Chrusiel, R. Hashim, N. Hanis, A. Latif, N. Brosse, I. Ziegler-devin and L. Chrusiel, , DOI:10.15376/biores.17.1.469-491.
41 R. Rinaldi, R. T. Woodward, P. Ferrini and H. J. E. Rivera, *J. Braz. Chem. Soc.*, 2019, **30**, 479–491.
42 T. Mineva, Y. Tsoneva, R. Kevorkyants and A. Goursot, *Can. J. Chem.*, 2013, **91**, 529–537.
43 Y. Tsoneva, A. Tadjer and T. Mineva, *Int. J. Quantum Chem.*, 2016, **116**, 1419–1426.
44 T. Hastie, R. Tibshirani and J. Friedman, 2009, 1–51.
45 Van Der Spoel, D., et al., *GROMACS Fast, flexible, Free. J. Comput. Chem. 2005. 26(16) p. 1701-1718.*
46 T. J. and J. D. M. Dick, in *Annual Reports in Computational Chemistry*, ed. David Spellmeyer, Elsevier, 2005.
47 A. D. MacKerell, E. P. Raman and O. Guvench, *J. Phys. Chem. B*, 2010, **114**, 12981–12994.
48 Petridis, L. and Smith J. C., *J Comput Chem*, 2009, **30**, 457–467.
49 D. van der Spoel, P. J. van Maaren and C. Caleman, *Bioinformatics*, 2012, **28**, 752–753.
50 D. Marx and J. Hutter, *Ab initio molecular dynamics: Theory and implementation*, 2000, vol. 1.
51 W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.
52 A. De La Lande, A. Alvarez-Ibarra, K. Hasnaoui, F. Cailliez, X. Wu, T. Mineva, J. Cuny, P. Calaminici, L. López-Sosa, G. Geudtner, I. Navizet, C. G. Iriepa, D. R. Salahub and A. M. Köster, *Molecules*, , DOI:10.3390/molecules24091653.
53 M. C. (2018). A.M. Koster, G. Geudtner, A. Alvarez-Ibarra, P. Calaminici, M.E. Casida, J. Carmona-Espindola, V.D. Dominguez, R. Flores-Moreno, G.U. Gamboa, A. Goursot, T. Heine, A. Ipatov, A. de la Lande, F. Janetzko, J.M. del Campo, D. Mejia-Rodriguez, J. U. Reveles, 2018.
54 R. M. Shirke, A. Chaudhari, N. M. More and P. B. Patil, *J. Chem. Eng. Data*, 2000, **45**, 917–919.
55 L. G. Gagliardi, C. B. Castells, C. Ràfols, M. Rosés and E. Bosch, *J. Chem. Eng. Data*, 2007, **52**, 1103–1107.
56 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1997, **78**, 1396.
57 N. Godbout, D. R. Salahub, J. Andzelm and E. Wimmer, *Can. J. Chem.*, 1992, **70**, 560–571.
58 A. M. Köster, J. U. Reveles and J. M. Del Campo, *J. Chem. Phys.*, 2004, **121**, 3417–3424.
59 A. Goursot, T. Mineva, R. Kevorkyants and D. Talbi, *J. Chem. Theory Comput.*, 2007, **3**, 755–763.
60 S. E. Ashbrook and D. McKay, *Chem. Commun.*, 2016, **52**, 7186–7204.
61 B. Zuniga-Gutierrez, G. Geudtner and A. M. Kster, *J. Chem. Phys.*, 2011, **134**, 1–11.
62 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.
63 ASE package, https://wiki.fysik.dtu.dk/ase/.
64 L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt and G. Varoquaux, 2013, 1–15.
65 S. M. Aguilera-Segura, F. Di Renzo and T. Mineva, *Langmuir*, 2020, **36**, 14403–14416.
66 N. Özcan, J. Mareš, D. Sundholm and J. Vaara, *Phys. Chem. Chem. Phys.*, 2014, **16**, 22309–22320.
67 K. Aidas, A. Møgelhøj, H. Kjær, C. B. Nielsen, K. V. Mikkelsen, K. Ruud, O. Christiansen and J. Kongsted, *J. Phys. Chem. A*, 2007, **111**, 4199–4210.
68 Z. . Wang, W.-Y.; Qin, J.-H. . Liu, H.-M.; Wang, X.-D.; Gao and G.-Y. Qin, *Molecules*, 2021, **26**, 398.
69 Q. Wang, S. Liu, G. Yang and J. Chen, *Energy and Fuels*, 2014, **28**, 3167–3171.
70 Y. Pu, B. Hallac and A. J. Ragauskas, *Aqueous Pretreat. Plant Biomass Biol. Chem. Convers. to Fuels Chem.*, 2013, 369–390.
71 F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, *Nat. Commun.*, 2018, **9**, 1–10.
72 Z. Chaker, M. Salanne, J. M. Delaye and T. Charpentier, *Phys. Chem. Chem. Phys.*, 2019, **21**, 21709–21725.
73 J. Cuny, Y. Xie, C. J. Pickard and A. A. Hassanali, *J. Chem. Theory Comput.*, 2016, **12**, 765–773.
74 X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov and Z. Chen, *Angew. Chemie - Int. Ed.*, 2020, **59**, 10297–10300.
75 D. Chen, Z. Wang, D. Guo, V. Orekhov and X. Qu, *Chem. - A Eur. J.*, 2020, **26**, 10391–10401.

