Supporting Information for:

Soil organic matter in its native state: unravelling the most complex biomaterial on Earth

Hussain Masoom[1], Denis Courtier-Murias[1], Hashim Farooq[1], Ronald Soong[1], Brian P. Kelleher[2], Chao Zhang[1], Werner E. Maas[3], Michael Fey[3], Rajeev Kumar[4], Martine Monette[4], Henry J. Stronks[4], Myrna J. Simpson[1], and André J. Simpson[1*]

1. Department of Chemistry, University of Toronto, Toronto, ON, Canada, M1C 1A4
2. School of Chemical Sciences, Dublin City University, Dublin, Ireland
3. Bruker BioSpin Corp., Billerica, Massachusetts, USA, 01821-3991
4. Bruker BioSpin Canada, Milton, ON, Canada, L9T 1Y6

Number of Pages: 25

Number of Figures: 4

Number of Tables: 2

S.1 Technological Advances of CMP-NMR

Comprehensive multiphase nuclear magnetic resonance (CMP-NMR) spectroscopy contains the following technology to study a continuum of physical states between crystalline solids and solution-like compounds:

- lock and susceptibility matching for ideal ^{1}H lineshape employed in solution state NMR
- magic angle spinning and magic angle pulsed field gradients used in HR-MAS NMR experiments
- high power circuitry to generate intense B_1 fields to accommodate for solid state NMR experimentation

In this manuscript a prototype MAS 4 mm ^{1}H-^{13}C-^{2}H CMP-NMR probe with a Magic Angle Gradient (~60G/cm) was employed. For this particular probe (3rd prototype built) when compared to a conventional dedicated (4mm ^{1}H-BB 2 channel solids probe) the sensitivity on carbon is reduced ~25%. The main sensitivity loss arises from in part the additional tuning and in part the interaction of the gradient with stator electronics. When compared to a dedicated ^{1}H-^{13}C-^{15}N HR-MAS probe sensitivity and line shape are comparable. The probe can deliver the full ^{1}H 100KHz decoupling required for uncompromised solids. Typical 90^{o} pulse are 2µs ^{1}H and 5µs ^{13}C. In addition to the probe, specific upgrades to spectrometers are required. On a standard solids system the addition of a gradient and lock board are required. For a standard solution state spectrometer amplifiers and pre-amplifiers need to be upgraded to both deliver and handle 200W per channel. When the CMP-NMR technology is combined with a suite of

spectral editing approaches liquids, gel, semi-solids and solids can be differentiated and studied *in-situ* in unaltered natural samples. Readers requiring more information and applications should refer to the original publication which includes a full description of the transmission line circuitry required to minimize the electronics to fit all components into the limited space of a narrow bore magnet.[1]

S.2 Detailed Experimental Parameters

NMR Experiment Parameters

1D ^{1}H experiments were performed with 1024 scans, a recycle delay of 5 x T_1, and 4096 time domain points. The solvent signal (water or DMSO) was suppressed using presaturation for water suppression. Spectra were apodized through multiplication with an exponential decay corresponding to 1 Hz line broadening in the transformed spectrum and a zero filling factor of 2. Diffusion edited experiments were performed with a bipolar pulse longitudinal encode-decode (BPLED) sequence. 2048 scans were collected using encoding/decoding gradients of 1.2 ms, 52.25 gauss/cm, sine shaped gradient pulse, a diffusion time of 180 ms, and 8192 time domain points. Spectra were apodized through multiplication with an exponential decay corresponding to 10 Hz line broadening in the transformed spectrum, and a zero filling factor of 2. In the ^{1}H T_2 filtered experiments, the T_2 delay was set to 600 µs and the pulse train was repeated 100 times resulting in a pulse train length of 120 ms. Inverse Diffusion Edited (IDE), Inverse T_2, and Recovering relaxation losses from Diffusion Editing (RADE) were created via subtraction from the appropriate controls as previously described.[1] For spectral editing, the

spectra were scaled until the spectra being subtracted was nulled leaving a difference spectrum containing positive peaks.[1]

^{1}H-^{13}C CP experiments were performed with a ramped (50%-100% linear ramp defined by 100 points) contact time of 2 ms, a recycle delay of 1 s, and 512 points in the time domain. The number of scans performed was 102400. Soil spectra were apodized through multiplication with an exponential decay corresponding to 30 Hz line broadening in the transformed spectrum, and a zero filling factor of 2. The summation spectrum of the model components was created by iterative fitting of individual components to create a spectrum most similar to that of the soil used (also see supporting information S.5). Note Total Suppression of Sidebands (TOSS) was attempted to supress the ^{13}C sidebands however the signal-to-noise losses were too large to be acceptable. Instead the sidebands for the carbonyl region were measured by integrating the first order side band downfield and reporting it as a % of the central transition. The side bands were 4.2% in the dry sample, 1.4% after swelling with water and 0.2% after swelling with DMSO. Seeing as the carbonyl signals themselves only change by ~20% on wetting with water the carbonyl sideband contribution to the aromatics is negligible (4.2 * 0.2) = 0.8%. However, to air on the side of caution changes in relative intensity are a reported as estimates rounded to the nearest 5% to account for variations in sidebands, baseline and phase properties.

All ^{1}H-^{13}C Heteronuclear Single Quantum Coherence (HSQC) NMR experiments were performed using a second prototype MAS 4 mm ^{1}H-^{13}C-^{19}F-^{2}H comprehensive multiphase NMR probe fitted with an actively shielded Z gradient. Spectra were collected in phase-sensitive mode using echo/anti-echo gradient selection, a $^{1}J^{1}H$-^{13}C (145 Hz), and a relaxation delay of 1 s.

1024 time domain points were collected for each of the 96 increments for all HSQC experiments. The F2 planes were multiplied by an exponential function corresponding to a 15 Hz line broadening, while the F1 dimension was processed using sine-squared functions with a $\pi/2$ phase shift and a zero-filling factor of 2. 7782 scans were collected for each increment for the soil sample while the number of scans for the standard compounds varied from 256 to 716 scans per increment. HSQC simulations were performed using ACD Labs Spec Manager (version 12.5. Parameters used for prediction including spectral resolution, sweep width, line shape, and base frequency were chosen to match a real data set as closely as possible. The predictions are based on HOSE code matches and incremental algorithms derived from a large database of assigned chemical shifts that represent the majority of published NMR data. Predictions using this approach have been shown to be accurate to <0.3ppm (^{1}H) and <3 ppm carbon (^{13}C) for over 90% of positions in any structure[2] and have been used to produce very accurate predictions for soil components.[3]

Integration

Spectral integration was performed using AMIX software and the regions were chosen based on the position of each of the specific regions for this soil (see Table S2 for a complete list). Overall integration percentages were calculated relative to the total integration value of the entire spectrum. Individual components were scaled appropriately to create the summation CP spectrum and the corresponding scaling factor was applied to each component for each integration region. The sum of all components for a given region was used to calculate relative percent contribution for each component.

Soil elemental analysis

Soil elemental composition was measured by first drying the soil in a 40^{o}C oven. Then, two different methods were employed to measure C and N, and Fe and Mn. For C and N, elemental analysis was performed on a Thermo Fisher Flash 2000 Elemental Analyzer externally calibrated with linear correlation of >0.9999 for carbon and nitrogen. Fe and Mn composition was measured using a Bruker S1 TRACER III-V X-Ray Fluorescence analyzer using a pre-installed soil calibration set.

S.3 Justification for the Components Chosen for the HSQC Model

To represent carbohydrates, xylan was used since it is found naturally in plant cell walls, bovine serum albumin was used to represent protein input, an alkaline lignin extract was used to represent lignin and a waxy lipid extract from pine needles to represent aliphatics in soil. Readers should note that the biopolymers chosen are not meant to represent the full diversity of components present in soil. For example, xylan will not represent the diverse range of carbohydrates including cellulose and hemi-cellulose. However, all carbohydrates are comprised of similar sugar sub-units. Therefore, these four components provide an indication as to where these major structural categories resonate in various NMR experiments.

S.4 Further Explanation of Figure 1

Each numbered box represents:

1) Aliphatic single bonds and amino acid side chains
2) Alpha proton on amino acids
3) Methoxy peak from lignin
4) CH_2 groups from carbohydrates
5) Anomeric protons from carbohydrates
6) Aromatic lignin and phenyl alanine
7) Aromatic inputs from amino acids tyrosine and tryptophan.

Readers should note that lignin does not contain strong aliphatic resonances (>3ppm) and the aliphatic region in the Aldrich lignin sample likely indicates some aliphatic species have been co-extracted along with the lignin. The red box in the biopolymer overlay highlights lignin syringyl units which are in higher abundance in the model lignin compared to the grass land soil used since the model lignin is derived from hardwood which in general tends to have a higher syringyl content.[4]

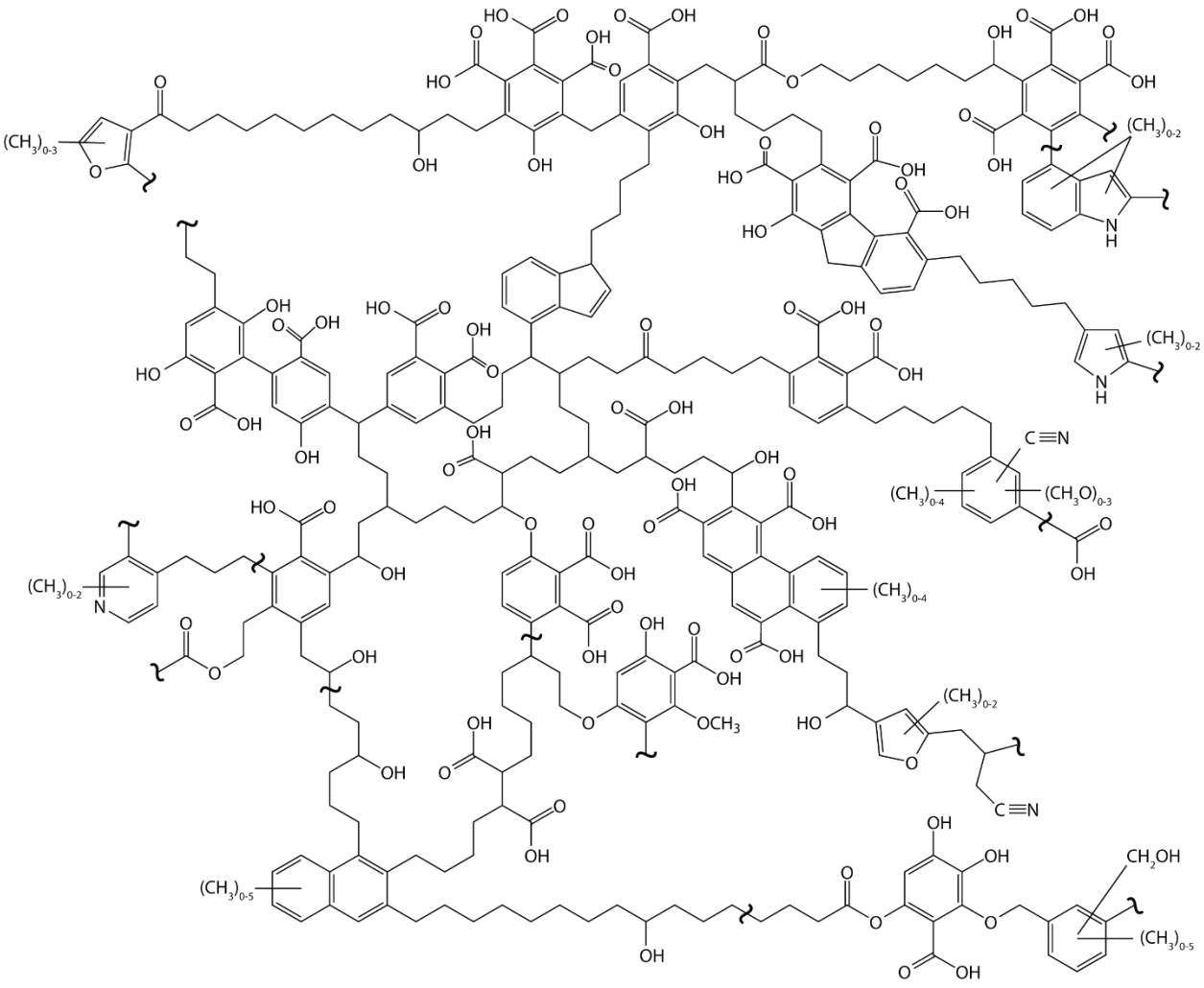


Figure S1. Representation of the cross linked humic model proposed by Schulten and Schnitzer[5] drawn using ChemDraw (Cambridgesoft, Version 12)

S.5 Quantification of the Biopolymer Model

With the finding that the four components mentioned (protein, aliphatics, lignin, and carbohydrates) dominate the NMR profile of soil organic matter (SOM), it was possible to determine the approximate relative contribution of each component. This was done by

performing a CP experiment on each of the components and then adding the resulting spectra to create the component CP as is seen in Figure 2A. This process is illustrated in Figure S2. Each component was iteratively weighted until the sum of the components best matched the soil spectrum. This was done using the algebraic module in AMIX (version 3.9.15) which utilizes the iterative Jacobi method for automated quantification.[6] Using these weightings as well as the chemical shift regions described in Table S1, it was possible to calculate the relative contribution of each component to each spectral region as well as each component's contribution to the entire spectrum. Table S2 summarizes these quantifications for the soil used in this study.

Figure S2 is an illustration of the process used to quantify and create the component CP model. First, each component's CP spectrum was collected. Then spectra were combined and then iteratively weighed until the sum model best matched the soil spectrum. These weightings we then used to assess how each model component (protein, carbohydrate etc.) contributed the main spectral regions (carbonyl, aromatic etc.).

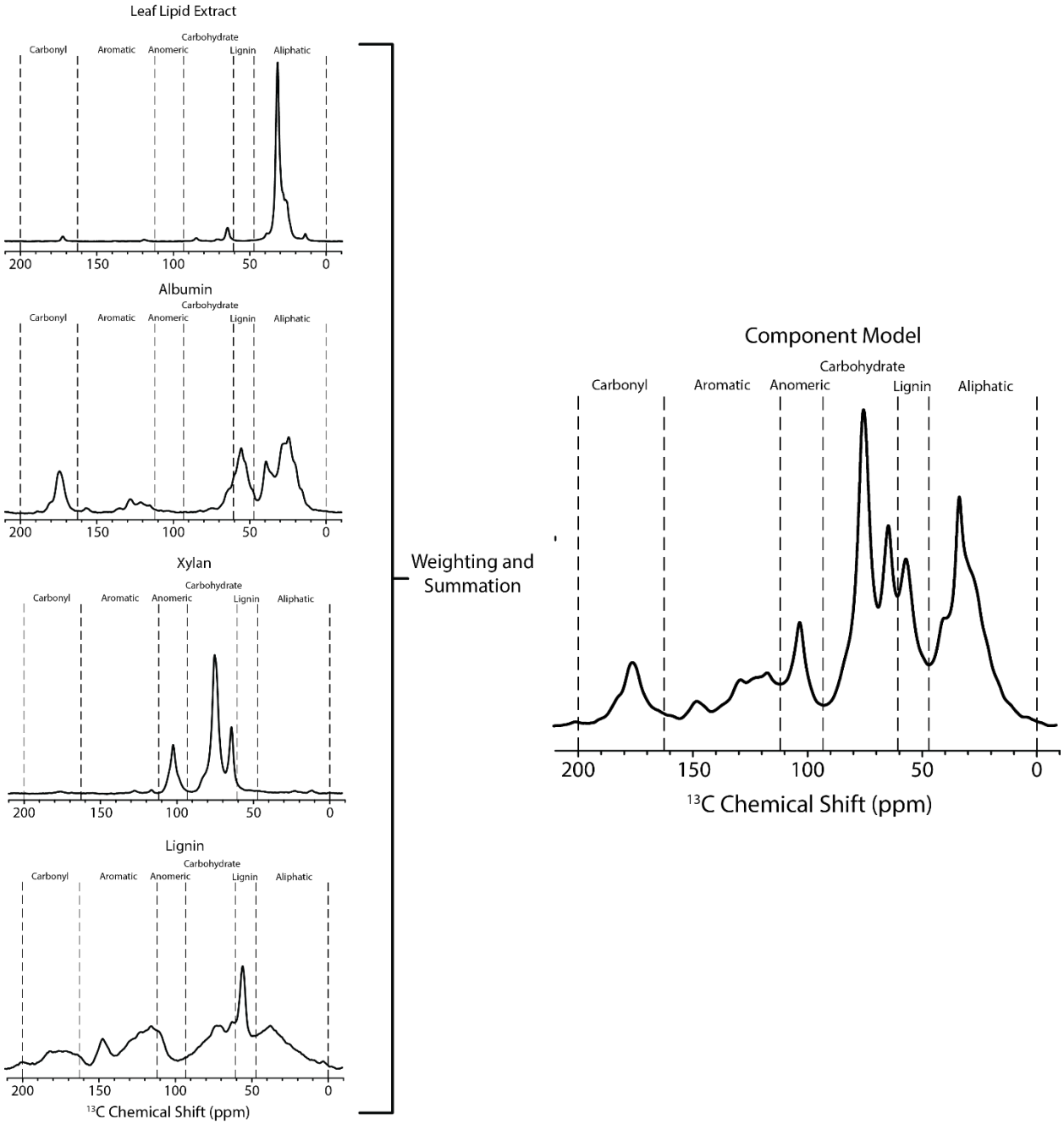


Figure S2. An illustration of the process used to quantify and create the component CP model.

Table S1: Chemical shift regions used for integration calculations

Region	Chemical Shift Range (ppm)	
Aliphatic	0	49.1
Lignin	49.1	61.7
Carbohydrate	61.7	95.2
Anomeric	95.2	114.1
Aromatic	114.1	167.2
Carbonyl	167.2	200

Table S2: The relative percent contribution of each model component to the specified spectral region in CP-MAS.

		Model Components			
		Protein	Aliphatic	Lignin	Carbohydrate
^{13}C Integration Regions	Aliphatic	46.9%	25.4%	23.7%	4.0%
	Lignin	44.0%	1.3%	43.8%	10.9%
	Carbohydrate	8.3%	1.9%	22.5%	67.3%

Anomeric	1.8%	0.2%	27.7%	70.3%
Aromatic	17.9%	0.7%	74.6%	6.8%
Carbonyl	46.1%	2.2%	47.3%	4.3%
Overall	**27.4%**	**8.6%**	**33.2%**	**30.7%**

It was found that the protein, lignin and carbohydrates all contributed roughly 30% to the signal intensity from the overlay spectrum. It is also interesting to note the significance of protein in the aliphatic, lignin and carbonyl regions. The large contribution from protein in soil organic matter has been addressed previously and has been shown to be from microbial biomass.[20]

S.6 Spectral Editing

Diffusion Based

The Inverse Diffusion Edited (IDE) spectra (Figure 3B) emphasizes molecules with unrestricted diffusion i.e. those truly in solution, and is created by the subtraction of the diffusion edited spectrum (contains larger components with restricted diffusion) from a reference spectrum (contains a profile of all components). Essentially the experiment select components that have free diffusion (i.e. the can physically move within the sample matrix) and represent dissolved species.[1]

Diffusion editing (DE) involves the coding of the spatial position of signals at the start of the experiment and then decoding the position at the end. The signal from molecules that diffuse i.e. change their physical position are not observed in the spectra resulting in a spectrum of large components or metabolites that show little to no diffusion. Essentially the DE spectrum (Figure 3C) is the opposite of the IDE spectrum and contains the signals of components that do not move position with the sample (i.e. components with restricted diffusion) these could include biopolymers, bounds species, etc.[1]

In a simple world, combined IDE and DE would describe all the components at the soil interface. However, due to the relatively long delays required for diffusion measurements some molecules may relax and not be observed at all. To fully account for these components an experiment termed Relaxation recovery Arising from Diffusion Editing (RADE) spectra can be used (Figure 3D).[1] The experiment recovers the components that otherwise may be missed due to their fast relaxation. This spectrum contains the most rigid component detectable by ^{1}H NMR, such as cell membranes, cell membranes, structural proteins etc.

Relaxation Based Editing

While diffusion based editing discriminates based on diffusion of the entire molecule, relaxation based editing discriminates based on individual localities within a molecule. Molecular entities that are macromolecular, or sorbed to a surface, exhibit fast relaxation whereas those with local dynamics (dissolved, in fast exchange) will exhibit slow relaxation.

However, if a larger molecule has a long flexible component, for example the lipid in a lipoprotein, the lipid will exhibit slow relaxation due to its high local dynamics.[1]

Figure 4B shows T_2 edited spectrum. This spectrum contains signals that exhibit long T_2 relaxation, a phenomena consistent with small dissolved species or flexible appendages of larger molecules. Figure 4C shows the inverse T_2 edited spectrum created by subtracting Figure 4B (mobile) components from a reference spectrum (Figure 4A all components). The result is a spectrum containing the most rigid components at the soil-water interface detectable by ^{1}H NMR.

S.7 Calculating contributions from different phases using filtered spectra

Fitting the contributions from the subcomponents (Figure 3B (free diffusion), Figure 3C (restricted diffusion) and Figure 3D (semi-solids) is possible but not obvious. First it is important to understand how spectrum 3D (the RADE spectrum (semi-solids)) is obtained. To create a ^{1}H RADE spectrum a reference spectrum is acquired (Figure 3A) and then another spectrum where the delays required for diffusion editing (needed later for Figure 3C) have been set to the appropriate length, however the gradient that encodes for diffusion is NOT turned on. As the diffusion gradient is off, there is no diffusion weighting whatsoever. The only difference between these two spectra (delays set to zero, and delays set to what is required for diffusion editing) is relaxation that can occur during the delays. The ^{1}H RADE spectrum (Figure 3C) is created by direct subtraction of these 2 spectra and highlights the fastest relaxing components that otherwise could be missed such as semi-solids. The total signal in the ^{1}H RADE (Figure 3C)

is easy to estimate by evaluating the % signal loss relative to the reference (Figure 3A) when the delays are introduced. Importantly the spectrum acquired with the delays set but the diffusion gradient off (i.e. the combined signal not lost through relaxation), actually represents the total signal from (Figure 3B (free diffusion) and Figure 3C (restricted diffusion) combined. The relative contribution of 3B and 3C is easy to obtain by estimating the relative weighting of each spectral envelope to this "sum" spectrum (delays set to what is required for diffusion editing but diffusion gradient off). As such estimates for the contributions from Figures 3B, 3C, and 3D can be performed providing approximations as to how much of the soil organic matter is in the freely diffusing, restricted diffusion, and semi-solid state relative the total observed by ^{1}H MAS (Figure 3A)

The contribution of each sub-component in Figure 4 (relaxation based editing) is luckily much less involved. In this case Figure 4A represents everything that can be observed by ^{1}H MAS. The contribution from 4B and 4C is simply obtained by weighted summing of Figure 4B and 4C to best match 4A and using the resulting ratio to estimate the contribution from each sub-spectrum.

S.8 Lipids are the Aliphatics Associated with the Soil-Water Interface

The triple peak pattern (~0.9pp, 1.3ppm and 2ppm) seen in the aliphatic region of Figures 4 and 5 is indicative of fatty acids/ lipids.[7-10] Figure S.3 shows a TOCSY in DMSO with the lipid resonances assigned. In Figure 4 and 5 in the main manuscript the peak residing at ~0.9ppm represents the terminal CH_3 groups of lipids. Since there is only one CH_3 group, it

makes sense that this peak is the smallest in intensity. The tallest peak (~1.3ppm) represents the long chain $(CH_2)_n$ groups of which there are many owing to its intensity. Of particular interest is the peak at ~2.0pm. While some of this peak is from acetic acid (a common degradation product in soil) a good portion of the intensity arises from CH_2 adjacent to the COOH in lipids. Note signals A and B are resolved and slightly shifted in DMSO (Figure S.3), whereas in D_2O they are less resolved and resonate close to 2ppm and partially overlap with acetic acid. Readers should note that no lipids were seen in the true solution phase (Figure 4B, free diffusion).

When considered together this indicates that the polar groups are sticking out of the water-soil interface hence showing rapid local dynamics at the polar end (Figure 5B). Meanwhile, their hydrophobic tails are interacting with, or partially buried under, the soil surface. This is an interesting finding and suggests that lipids could be a key component of soil likely contributing to properties such as swelling, soil aggregate structure and possibly representing a conduit between the polar interface and the more hydrophobic interior for organic contaminants.

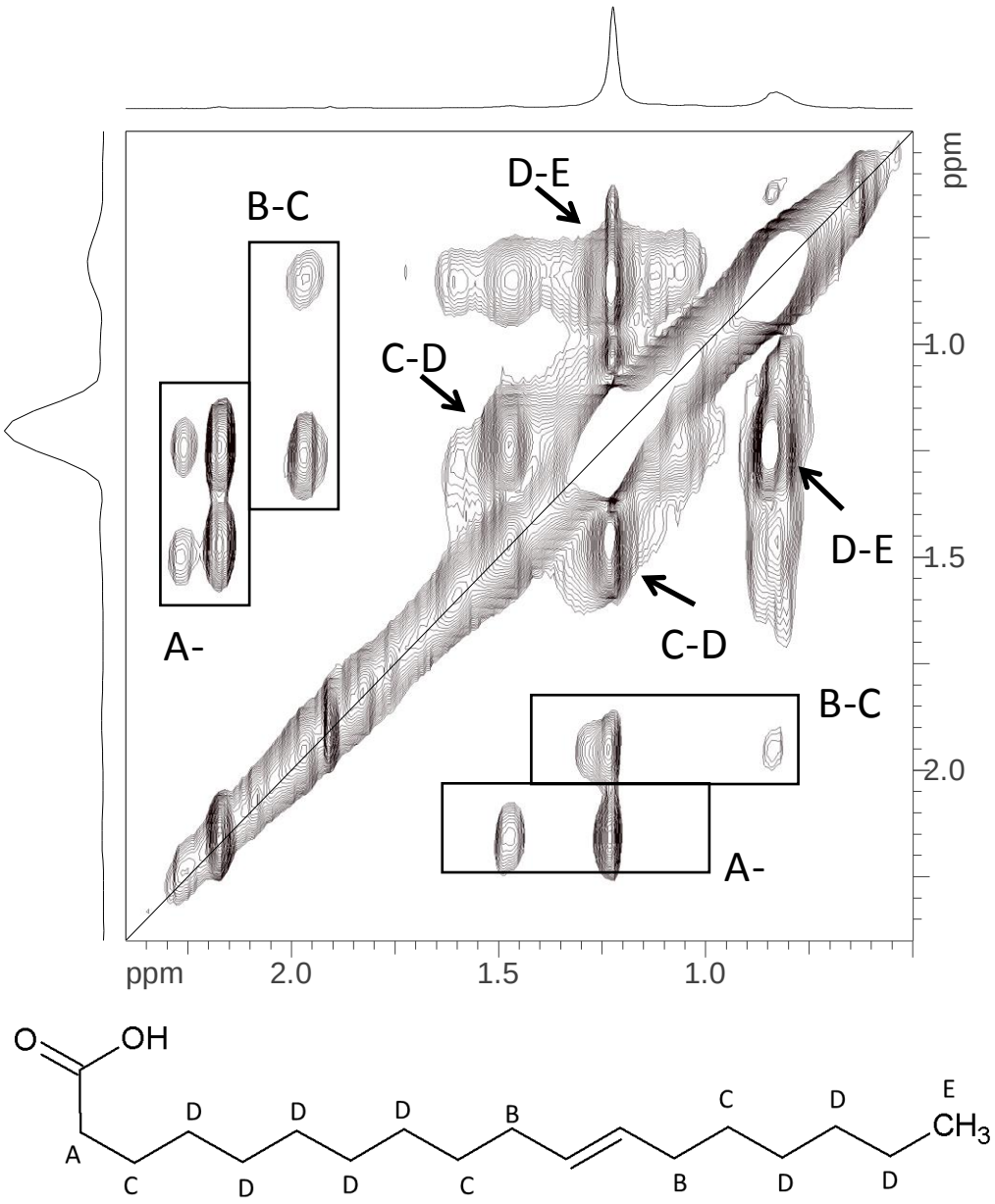
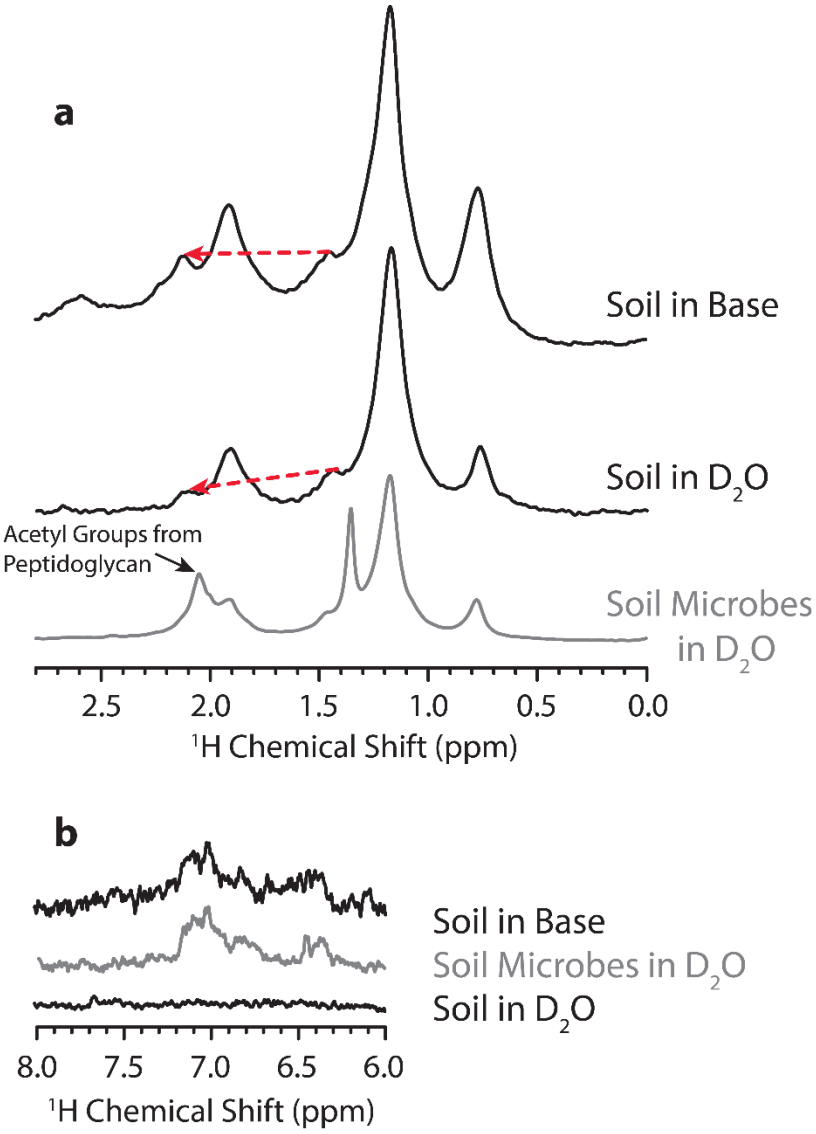


Figure S3. ^{1}H-^{1}H TOCSY spectrum of the soil swollen in DMSO. Major cross-peaks and couplings are consistent with aliphatic chains consistent with lipids/waxes and have been previously and extensively assigned in the literature.[7-10] Note the signals A and B are resolved and slightly shifted in DMSO, whereas in D_2O they are less resolved and resonate close to 2ppm and partially overlap with acetic acid.

S.9 Detailed Analysis on Microbe Associations in Soil

Figure S4



In Figure S4a, the acetyl group from peptidoglycan appears at 2.1ppm in the diffusion edited spectrum. The shoulder at 1.5ppm (start of red arrow) is dominated by lipids (units γ to a double bond and oxygen in an ester), while the 2.1ppm region is from the overlap of lipids and peptidoglycan. As the pH is raised this region appears to become more dominant and is difficult to assign to lipids alone, strongly suggesting additional contributions of peptidoglycan at the soil water interface. This indicates that microbial cell walls have more contact with the aqueous water phase as the pH is raised. This is confirmed with reference to the aromatic region. A T_2 filtered experiment is shown for the aromatic region as these experiments reduce baseline roll making spectral comparison easier.

There are strong similarities between the spectral profile of pure microbes (cultured from soil, run at neutral pH) and the profile of the alkaline soil itself in this region (Figure S2b). These signals arise from microbial protein/peptides and indicate at high pH, water can penetrate and swell the microbial cells. The data clearly shows that the isolated microbial cells on their own can be observed at neutral pH (see Figure S.4) but that they are not observed within the soil at neutral pH. This is strong evidence that at neutral pH water cannot reach the microbial cells and indicates microbes are largely removed from the soil-water interface.

S.10 Components of Soil Not Mentioned

While protein is a main component of soil, it was not included in this final diagram because previous work has demonstrated that of the vast majority the protein in soil originates from the microbes themselves.[11] This is consistent with studies that have shown free protein to be very labile and have short residence times within soil.[12, 13] This does not limit protein to only being within microbes, it certainly exists outside of the microbial cell wall at some level as their will be continual input and turnover from fresh organic deposition.

Inorganic species like minerals are not considered here, as the NMR evidence presented here does not provide direct information on the mineral component. This being said, minerals undoubtedly play a key role in determining how the soil is ordered. There have been studies suggesting that lignin binds strongly with minerals making minerals the possible nucleus of a soil globule where lignin surrounds it followed by carbohydrates, aliphatics and microbes.[14-16] Another area worth exploring is determining how organic components interact with one

another within soil. These missing links need future investigation and will be required to complete our understanding as to the molecular organization of soil as a whole.

S.11 Summary of Findings in this Manuscript

From the findings in this paper a number of conclusions can be drawn about SOM in general and its role in the formation of soil aggregates.

1) SOM in its native state appears to be dominated by four main structural categories, protein/peptides, aliphatics, carbohydrates and lignin, which reflect major soil inputs and is consistent with earlier work.[17, 18] It is likely these molecules exist at various stages of decay.
2) At the soil interface carbohydrate and polar groups of lipids tend to dominate. Lipid residues appear to orient with their polar group towards the water and hydrophobic tail buried. This suggests that lipids could be a key component in the soil contributing bulk negative charge and to properties of soil swelling, soil aggregate structure and possibly representing a conduit between the polar interface and more hydrophobic interior for organic contaminants. This has similarities with earlier micelle based concepts of organic matter.[19, 20]
3) Lignin residues are only accessible when a non-polar solvent is applied to the soil. Under aqueous conditions these domains are relatively dense and do not permit water penetration. This could be important in explaining why some hydrophobic contaminants accumulate in hydrophobic lignin domains in soil.[21]

4) Microbes appear relatively well removed from bulk water at natural pH. Microbes are slightly more exposed at high pH, but are fully detectable in DMSO. This suggests in part that the microbes are partially protected by the hydrophobic nature of SOM. Microbes may actively reside away from the aqueous-soil interface and possibly associate with mineral surfaces, preventing their erosion.

SUPPLEMENTARY REFERENCES

1. Courtier-Murias, D.; Farooq, H.; Masoom, H.; Botana, A.; Soong, R.; Longstaffe, J. G.; Simpson, M. J.; Maas, W. E.; Fey, M.; Andrew, B.; Struppe, J.; Hutchins, H.; Krishnamurthy, S.; Kumar, R.; Monette, M.; Stronks, H. J.; Hume, A.; Simpson, A. J.; Comprehensive multiphase NMR spectroscopy: Basic experimental approaches to differentiate phases in heterogeneous samples. *J. Magn. Reson.* **2012***, 217* 61-76.

2. Williams, A.; Lefebvre, B.; Sasaki, R. Putting ACD/NMR Predictors to the Test. **2004,** *2015,*

3. Simpson, A. J.; Lefebvre, B.; Moser, A.; Williams, A.; Larin, N.; Kvasha, M.; Kingery, W. L.; Kelleher, B.; Identifying residues in natural organic matter through spectral prediction and pattern matching of 2D NMR datasets. *Magn. Reson. Chem.* **2004***, 42* (1), 14-22.

4. Lupoi, J. S.; Smith, E. A.; Characterization of woody and herbaceous biomasses lignin composition with 1064 nm dispersive multichannel Raman spectroscopy. *Appl. Spectrosc.* **2012***, 66* (8), 903-910.

5. Schulten, H. R.; Schnitzer, M.; A State-Of-The-Art Structural Concept for Humic Substances. *Naturwissenschaften.* **1993***, 80* (1), 29-30.

6. Sleijpen, G. L. G.; Van Der Vorst, H. A.; A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.* **1996***, 17* (2), 401-425.

7. Simpson, A. J.; Kingery, W. L.; Hatcher, P. G.; The identification of plant derived structures in humic materials using three-dimensional NMR spectroscopy. *Environ. Sci. Technol.* **2003**, *37* (2), 337-342.

8. Deshmukh, A. P.; Simpson, A. J.; Hatcher, P. G.; Evidence for cross-linking in tomato cutin using HR-MAS NMR spectroscopy. *Phytochemistry.* **2003**, *64* (6), 1163-1170.

9. Deshmukh, A. P.; Simpson, A. J.; Hadad, C. M.; Hatcher, P. G.; Insights into the structure of cutin and cutan from Agave americana leaf cuticle using HRMAS NMR spectroscopy. *Org. Geochem.* **2005**, *36* (7), 1072-1085.

10. Simpson, A. J.; McNally, D. J.; Simpson, M. J.; NMR spectroscopy in environmental research: From molecular interactions to global processes. *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *58* (3-4), 97-175.

11. Simpson, A. J.; Simpson, M. J.; Smith, E.; Kelleher, B. P.; Microbially derived inputs to soil organic matter: Are current estimates too low? *Environ. Sci. Technol.* **2007**, *41* (23), 8070-8076.

12. Hill, P. W.; Farrell, M.; Jones, D. L.; Bigger may be better in soil N cycling: Does rapid acquisition of small L-peptides by soil microbes dominate fluxes of protein-derived N in soil? *Soil. Biol. Biochem.* **2012**, *48* 106-112.

13. Park, S. K.; Hettiarachchy, N. S.; Were, L.; Degradation behavior of soy protein-wheat gluten films in simulated soil conditions. *J. Agric. Food Chem.* **2000**, *48* (7), 3027-3031.

14. Heim, A.; Schmidt, M. W. I.; Lignin is preserved in the fine silt fraction of an arable Luvisol. *Org. Geochem.* **2007**, *38* (12), 2001-2011.

15. Heim, A.; Schmidt, M. W. I.; Lignin turnover in arable soil and grassland analysed with two different labelling approaches. *Eur. J. Soil Sci.* **2007**, *58* (3), 599-608.

16. Clemente, J. S.; Simpson, M. J.; Physical protection of lignin by organic matter and clay minerals from chemical oxidation. *Org. Geochem.* **2013**, *58* 1-12.

17. Burdon, J. Are the traditional concepts of the structures of humic substances realistic? *Soil Sci.* **2001**, *166* (11), 752-769.

18. Nelson, P. N.; Baldock, J. A.; Estimating the molecular composition of a diverse range of natural organic materials from solid-state 13C NMR and elemental analyses. *Biogeochemistry.* **2005**, *72* (1), 1-34.

19. Wershaw, R. L. Membrane-micelle model for humus in soils and sediments and its relation to humification. **1994**, *2410*

20. Wershaw, R. L. Molecular aggregation of humic substances. *Soil Sci.* **1999**, *164* (11), 803-813.

21. Longstaffe, J. G.; Courtier-Murias, D.; Soong, R.; Simpson, M. J.; Maas, W. E.; Fey, M.; Hutchins, H.; Krishnamurthy, S.; Struppe, J.; Alaee, M.; Kumar, R.; Monette, M.; Stronks, H. J.; Simpson, A. J.; In-Situ Molecular-Level Elucidation of Organofluorine Binding Sites in a Whole Peat Soil. *Environ. Sci. Technol.* **2012**, *46* (19), 10508-10513.