



OPEN

# Soil as an extended composite phenotype of the microbial metagenome

Andrew L. Neal<sup>1</sup>✉, Aurélie Bacq-Labreuil<sup>2,4</sup>, Xiaoxian Zhang<sup>3</sup>, Ian M. Clark<sup>3</sup>, Kevin Coleman<sup>3</sup>, Sacha J. Mooney<sup>2</sup>, Karl Ritz<sup>2</sup> & John W. Crawford<sup>3,5</sup>

**We use a unique set of terrestrial experiments to demonstrate how soil management practises result in emergence of distinct associations between physical structure and biological functions. These associations have a significant effect on the flux, resilience and efficiency of nutrient delivery to plants (including water). Physical structure, determining the air–water balance in soil as well as transport rates, is influenced by nutrient and physical interventions. Contrasting emergent soil structures exert selective pressures upon the microbiome metagenome. These selective pressures are associated with the quality of organic carbon inputs, the prevalence of anaerobic microsites and delivery of nutrients to microorganisms attached to soil surfaces. This variety results in distinctive gene assemblages characterising each state. The nature of the interactions provide evidence that soil behaves as an extended composite phenotype of the resident microbiome, responsive to the input and turnover of plant-derived organic carbon. We provide new evidence supporting the theory that soil-microbe systems are self-organising states with organic carbon acting as a critical determining parameter. This perspective leads us to propose carbon flux, rather than soil organic carbon content as the critical factor in soil systems, and we present evidence to support this view.**

Soil—the basis of terrestrial life on Earth—continues to defy our comprehensive understanding despite the evident catastrophic consequences of mismanagement, such as the North American “Dust Bowl” of the 1930s which was exacerbated by poor stewardship of agricultural soils<sup>1</sup>. Faced with the multiplicity of processes which constitute soil, scientific reductionism has led to studies which have advanced our knowledge of soil’s biological, chemical or physical components predominantly in isolation. However, soil—in common with many biological phenomena—is more appropriately considered a hierarchical assemblage of interacting processes, stabilized and actively maintained at different timescales<sup>2</sup>: soil is processual and not comprehensible based on single-discipline experimentation. Tisdall and Oades’ pioneering conceptual model<sup>3</sup> linking microbial activity to soil structural development advanced the importance of interaction between biotic and abiotic phenomena in the process of generating soil structural complexity (topology and connectivity).

Soil organic matter (SOM) is the fundamental causative agent generating structural complexity, as it acts to bind mineral particles and colloids together. Plant and animal residues are processed by microbes before joining the SOM pool<sup>4,5</sup>: this step is an important facet of both the Tisdall and Oades model, and its subsequent extension<sup>6</sup>. SOM may take the form of microbial polysaccharidic and proteinaceous exudates as well as cell debris and is chemically structurally diverse<sup>4</sup>; in effect, SOM is a continuum of progressively more extensively oxidized compounds<sup>7</sup>. Much of this SOM is associated with pores of 30–100  $\mu\text{m}$  diameter<sup>8</sup>, scales comparable to the 12–13  $\mu\text{m}$  distances observed in soil between microbial cells<sup>9</sup>. As a result, the effect of microbial processes—metabolism, extracellular degradation of compounds, polymer secretion and cell lysis—on soil structure is particularly evident at the scales  $< 50 \mu\text{m}$  responsible for regulating convective and diffusive flow rates, as well as the balance of air and water at any given matric potential<sup>6</sup>. These hierarchical processes exhibit characteristic properties of self-organizing and emergent systems<sup>10,11</sup>.

<sup>1</sup>Department of Sustainable Agriculture Sciences, Rothamsted Research, North Wyke EX20 2SB, UK. <sup>2</sup>Division of Agriculture and Environmental Science, School of Biosciences, The University of Nottingham, Sutton Bonington Campus, Sutton Bonington LE12 5RD, UK. <sup>3</sup>Department of Sustainable Agriculture Sciences, Rothamsted Research, Harpenden AL5 2JQ, UK. <sup>4</sup>Present address: Greenback, 4 rue de l’église, 27440 Lisors, France. <sup>5</sup>Present address: Adam Smith Business School, University of Glasgow, Glasgow G12 8QQ, UK. ✉email: andy.neal@rothamsted.ac.uk

Such experiential and theoretical approaches are formulating a new understanding of how microbial activity controls soil structure—in effect, how soil should be viewed as an expression of biological process. They also provide evidence supporting a view of soil as a product of genes, manifest through the combined effects of multiple organism phenotypes: in essence, an *extended composite phenotype* (Phillips<sup>12</sup>, after Dawkins<sup>13</sup>). The identifying features of this phenomenon are a strong influence of at least one organism upon the form or structure of a soil environment—termed a *process-form relationship*; demonstrable *synchrony* between the activity of influencing organisms and form development; *selective pressure* arising from form development acting, in Dawkins' strict sense upon alleles<sup>13,14</sup> and in Phillips' broader concept upon soil organisms<sup>12</sup>; which results in *positive feedback* where selective pressure favours alleles (or organisms) associated with the process-form state, manifest as the influence of microbial turnover of SOM upon soil structural development, discussed above.

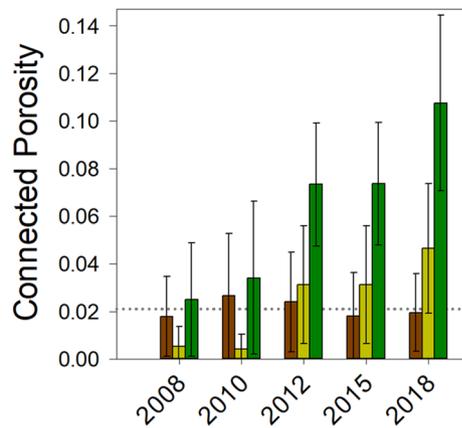
There is compelling evidence implicating plant-derived organic carbon inputs in the soil extended composite phenotype<sup>15,16</sup>. However, complete description of such a phenotype requires, in turn, a well-developed understanding of the consequences of evolving soil structure for the genetic manifestation of on-going microbial processes—such feedback is necessary for emergence of organisation, observable at the whole-system level in complex biological, chemical and physical systems. Currently, few studies present comprehensive description of the influence of soil structure upon microbial processes, and those that do, typically address only the association of metabolically defined bacterial groups with soil aggregate or particle size, rather than soil structure per se (see Lensi et al.<sup>17</sup> and Chotte et al.<sup>18</sup>). The principal influence of soil structural complexity is predicted to be on diffusion processes dictating the microenvironments surrounding surface-associated cells<sup>19</sup>. Observation of anaerobic regions of soil aggregates associated with denitrification processes<sup>20</sup>, and the influence of anaerobic microsites in ostensibly oxygen-rich soils upon microbial respiration and carbon compound oxidation rates<sup>21</sup> provide indirect evidence for such metabolic constraints arising from soil structure. However, this view of soil as an extended composite phenotype requires two specific conditions to be met. The first we term the *Process-Form Condition*, where the biological structures and functions that emerge from interactions between individual genotypes and their microenvironments should result in soil structural changes beyond the scale of individual cells. The second we term the *Allelic Response Condition*, where the process-form interaction should be reflected in significant modification at the level of individual alleles in soil microbiomes (i.e., fundamental changes in gene abundance patterns and whole metabolic pathways) such that alleles that correspond with specific processes are preferentially selected for—extending beyond short-term quantitative changes in specific gene expression profiles.

In this paper, we integrate biological and physical data relating to dynamics of the soil system with mathematical modelling to explore these conditions. This approach is used to interpret results from a unique long-term field-experiment within the context of the proposed view of soil as an extended composite phenotype: linking organic carbon inputs to soil with emergence of key soil structural properties; and describing the gene-level microbiome responses to contrasting emergent soil structural complexity arising from long-term carbon input regimes. The experiment uses the Highfield Ley-Arable Experiment at Rothamsted Research, Harpenden, U.K.

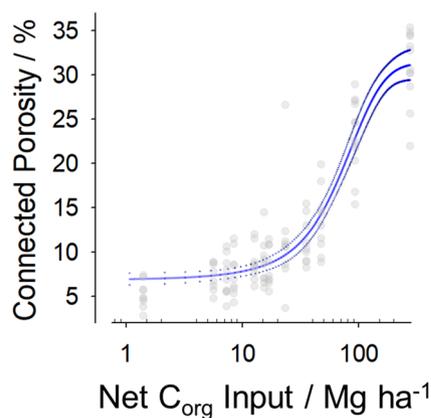
## Results

**Process-form relationships in soil are expressed through fine-scale connected porosity.** We first investigated the influence of added organic carbon ( $C_{org}$ ) on the development of soil structural complexity, testing the hypothesis that greater inputs of  $C_{org}$  to soil are associated with development of improved soil structure; assessed as the degree of connectedness between pores (connected porosity). The bare fallowed soil used as a starting point for these experiments experienced a demonstrable decline in  $C_{org}$ <sup>22</sup> and microbial abundance<sup>23</sup> over forty-nine years of continuous management. Estimates of  $C_{org}$  are approximately 3 g C kg<sup>-124</sup> and the soil has a significantly reduced total porosity compared to mixed grass sward soils<sup>25</sup>. To assess the influence of newly imposed managements, we followed structure development for a decade (2008–2018) in these degraded soils following conversion to either arable or mixed grass sward. The resulting estimates of soil structure demonstrated clear differences in the development of connected porosity between soil managed continuously as bare fallow and the converted soils (Fig. 1). Testing treatment effects upon connected porosity (square-root transformed to stabilize variances) between 2012 and 2018 inclusive, by analysis of covariance employing time as a covariate, indicated no significant heterogeneity of slopes ( $F_{2,75}=0.537$ ,  $p=0.587$ ) due to land management. The resulting equal slopes model indicated a significant effect of management upon connected porosity development ( $F_{2,72}=26.2$ ,  $p<0.001$ ). Post hoc pair-wise comparisons indicated that connected porosity generated in grassland soil (mean<sub>adjusted</sub>, 0.079) was significantly greater than in either continuous bare fallow (mean<sub>adjusted</sub>, 0.010) or converted arable (mean<sub>adjusted</sub>, 0.025) soils (smallest difference,  $t=4.79$ ,  $p<0.001$ ). A significant difference between connected porosity generated in arable and bare fallow soils was also apparent ( $t=2.30$ ,  $p=0.024$ ). No significant differences were detected when the complete 2008–2018 dataset was included, suggesting that significant differences only become apparent after a period of at least 5 years *post* conversion. This represents *prima facie* evidence for an extended composite phenotype, since there is clear synchrony between the establishment and continued growth of plants and development of connected porosity, representing a process-form relationship.

To test the potential role of  $C_{org}$  in the observed relationship, we compared connected porosity development with soil  $C_{org}$  inputs. We modelled changes in  $C_{org}$  content (see “Methods” for details) in bare fallow soil from inception of the experiment in 1949 up to the year before conversion, 2007. From 2008,  $C_{org}$  changes were modelled for each year, based upon continued bare fallow management or management as either winter wheat based arable or mixed grass sward. Continuously managed arable and grassland soils were also modelled in the same manner but using their respective starting dates. The extremes of the cumulative  $C_{org}$  input-connected porosity relationship (Fig. 2) are derived from consistently managed soils; bare fallow being associated with the lowest net  $C_{org}$  input and connected porosity, and arable and grassland soils being associated with the second highest and



**Figure 1.** Grassland soils generate connected pore space more rapidly than arable soils. Degraded soil (managed as bare fallow since 1959) developed greater connected micro-porosity following conversion in 2008 to grassland than bare fallow soil converted to arable. The mean and standard error of the mean of connected porosity measured in soil aggregates collected from soil managed continuously as bare fallow (brown), soil converted to arable management (dark yellow) and soil converted to grassland (green) over the ten years following conversion are shown. The dotted line marks the mean connected porosity of continuously managed bare fallow soil over the entire ten-year period.



**Figure 2.** Soil process-form relationships reflect biotic organic carbon inputs and turnover. The connected pore space in degraded soils converted after 49 years of bare fallow management to either arable or grassland increases in association with the net input of organic carbon ( $C_{org}$ ). Soils managed continuously as either arable (for 66 years) or grassland (for > 175 years) which have each accumulated over  $100 \text{ Mg ha}^{-1}$  of  $C_{org}$  over their history follow this trend. The relationship is described by an asymptotic function; the resulting fit (solid line) is shown, together with the upper and lower 95% confidence intervals of the fit (dotted lines).  $R^2 = 0.85$ .

highest  $C_{org}$  input and connected porosity, respectively. Modelled  $C_{org}$  inputs and measured connected porosity for soils converted to both arable and grassland for each of the ten years between 2008 and 2018 is distributed between the consistently managed bare fallow and arable soils data. We assumed that  $C_{org}$  in soils managed as grassland since 1838 represented the maximum which could be stored; this and the Akaike information criterion was used to guide selection of a sigmoidal function. There was a clear non-linear relationship between  $C_{org}$  inputs to soil and connected porosity, with all converted and continuously managed soils following the same trend (Fig. 2). This establishes that process-form relationships can be explained in terms associated with biotic  $C_{org}$  inputs and turnover in soil.

**Contrasting long-term soil management results in different process-form states.** Although the converted soils provided compelling evidence for structural development contemporaneous with the establishment and development of plant populations (albeit that arable soils are subject to external processes such as tillage and fertilization), the relatively short time span did not allow comparison of maximal differences in structural development, or evaluation of the potential for any resulting selective pressures to influence the representation of specific genes within the soils. Soils which had been under continuous management, at the time of sampling, as bare fallow for 52 years, arable for 62 years and mixed grass swards for over 170 years presented an

	Porosity/%	Permeability/mm <sup>2</sup>	Connectivity/ $\mu\text{m}^{-3}$	Pore surface density/ $\mu\text{m}^2 \mu\text{m}^{-3}$	$d_{\text{crit}}/\mu\text{m}$	Pore neck size/ $\mu\text{m}$
Grassland ( $n = 14$ )	$31.1 \pm 1.18$	$1.13 \pm 0.310$	$-0.206 \pm 0.025$	$0.088 \pm 0.003$	$9.74 \pm 0.37$	$11.19 \pm 0.34$
Arable ( $n = 14$ )	$23.4 \pm 1.22$	$0.62 \pm 0.154$	$-0.236 \pm 0.033$	$0.092 \pm 0.004$	$7.17 \pm 0.26$	$8.79 \pm 0.48$
Bare fallow ( $n = 9$ )	$15.0 \pm 2.21$	$0.55 \pm 0.339$	$-0.018 \pm 0.080$	$0.059 \pm 0.010$	$3.10 \pm 0.76$	$4.72 \pm 0.95$

**Table 1.** Topology-related parameters derived from binary images generated from X-ray computed tomography of aggregates from Highfield soils. The mean and standard error of each parameter is shown.

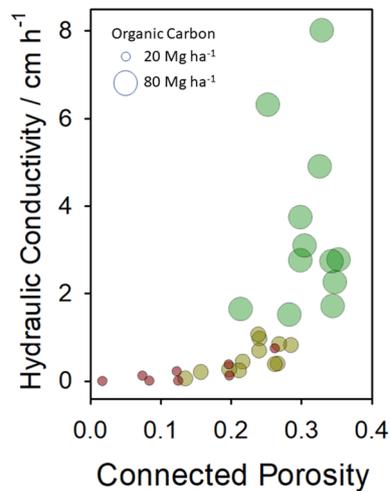
opportunity to test the hypothesis that established process-form relationships result in selection of organisms or genes, the fitness of which is suited to each particular soil biotope.

We have already determined that the continuously managed grassland and arable soils have significantly greater total porosity, a wider range of pore sizes and greater pore connectivity than continuously managed bare fallow soil<sup>25</sup>. Here we extend these findings, generating detailed information concerning pore topology and connectivity since they exert a strong influence upon diffusive flow in porous materials<sup>26</sup>, and modelling the effect of the observed pore networks upon diffusion processes within the soils. Euler connectivity functions [ $\chi(d)/V$ , see “Methods” section for calculation] for each soil are shown in Supplementary Fig. 1. For connected pores  $\chi(d)/V < 0$ , the value is positive for unconnected pores.  $\chi(d)/V = 0$  represents a critical threshold diameter ( $d_{\text{crit}}$ ) describing the maximum pore throat size of connected pores controlling hydraulic conductivity<sup>27</sup>. Mean estimates ( $\pm$  standard error) of  $d_{\text{crit}}$  were  $9.7 \pm 0.37 \mu\text{m}$  for grassland soils,  $7.2 \pm 0.26 \mu\text{m}$  for arable soils, and  $3.1 \pm 0.76 \mu\text{m}$  for bare fallowed soils. There was a significant effect of soil management upon  $d_{\text{crit}}$  ( $F_{2,6} = 42.3$ ,  $p < 0.001$ ) and each mean was significantly different from all others (smallest difference, grassland versus arable,  $Q = 4.99$ ,  $p = 0.029$ ). Topology-related parameters derived from X-ray computed tomography of aggregates (Table 1) showed a consistent trend of greater parameter estimates in grassland soils than degraded bare fallow soils. We chose porosity and  $d_{\text{crit}}$  as measures of pore topology since their implications are readily defined and they are of direct relevance to cells within the soil matrix because of their influence upon advective and diffusional processes. Porosity measurements from X-ray CT were used to derive diffusion coefficients for solutes within saturated soil aggregates, relative to unconstrained solute diffusion in water ( $D/D_0$ , see “Methods” for details). For grassland soils, mean  $D/D_0$  was determined at  $0.399 \pm 0.014$ ,  $0.285 \pm 0.009$  for arable and  $0.161 \pm 0.001$  for bare fallow. These estimates were significantly different ( $F_{2,70} = 106.4$ ,  $p < 0.001$ ). Normalised diffusion coefficients for each treatment were all significantly different from each other ( $p < 0.001$  for all comparisons).

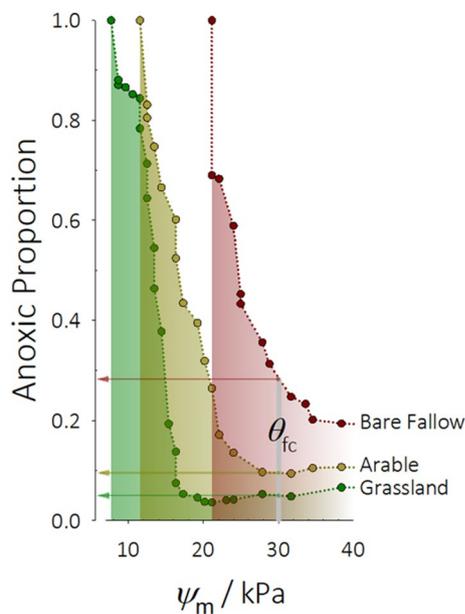
Direct measurement of pore network topology and modelling of the consequences for diffusion demonstrate that different long-term soil management results in quantitatively different process-form states. We also used estimates of pore network topology to model the hydrodynamic behaviour of the pore networks under saturated conditions, estimating hydraulic conductivity as a function of connected porosity. This measures the dynamical state of the pore space and the maximum potential flow rate at which resources can move through the networks—effectively the capacity for flux within the soil pore space. Figure 3 shows the combined effects of soil  $C_{\text{org}}$  and connected porosity upon the predicted hydraulic conductivity of soils. Combined direct measurements and modelling indicate a power law relationship between connected porosity and conductivity and that  $C_{\text{org}}$  is associated with these changes. Regions of this relationship correspond to the process-form states of continuously managed bare fallow, arable and grassland soils. The fraction of anoxic volume characterising each process-form state was estimated using a multi-phase lattice-Boltzmann approach<sup>28,29</sup>, described in detail in the “Methods” section. The results (Fig. 4) indicate that the predicted anoxic fraction is significantly lower in grassland soil, compared with arable and bare fallowed soils, the latter is predicted to have the highest fraction of anoxic volume at all matric potentials (moisture contents).

Microorganisms in each land management-associated process-form state are likely to experience markedly different hydraulic environments, particularly in degraded bare fallow soils where reduced delivery of dissolved nutrients and  $\text{O}_2$  is predicted compared to grassland soils. This is a direct result of constraints placed upon diffusive flow by reduced connected porosity and  $d_{\text{crit}}$  resulting from different biotic  $C_{\text{org}}$  inputs and turnover. These constraints are likely to exert significant selective pressures in soil microbiomes which should be reflected in changes in the assemblages of organisms or alleles in the different process-form states. To test this hypothesis, we generated shotgun metagenome sequence datasets from nucleic acids extracted directly from the different soils. These were analysed to determine whether any observed differences in phylogenetic community assemblages or in gene abundance were directly attributable to the differences in porosity or  $d_{\text{crit}}$  described above.

**Process-form states in soil do not exert selective pressure at the organismal level.** Chao-1 lower bound estimates of Prokaryote OTU richness ( $S_{\text{Chao1}}$ ) for each land management ranged from 562 to 578 (mean 570) for grassland, 530–547 (mean 540) for arable, and 482–542 (mean 513) for bare fallow soils. There was a significant effect of soil treatment upon  $S_{\text{Chao1}}$  ( $F_{2,6} = 7.6$ ,  $p = 0.023$ ), the difference between grassland and bare fallow mean richness was significant ( $Q = 5.49$ ,  $p = 0.019$ ). There was no significant difference between arable and grassland or arable and bare fallowed soils. Grassland soils also exhibited the largest Fungal OTU richness, range 35–44 (mean 39) compared to either arable (range 19–27, mean 24) or bare fallowed (range 17–27, mean 23) soils. There was again a significant treatment effect upon  $S_{\text{Chao1}}$  ( $F_{2,6} = 11.8$ ,  $p = 0.008$ ) and pairwise comparison indicated grassland was significantly more rich in fungal OTUs than either arable or bare fallowed soils (smallest difference,  $Q = 5.68$ ,  $p = 0.017$ ), but there was no difference between arable and bare fallowed soils. Weighted UniFrac distance-based comparison of  $\beta$ -diversity (Fig. 5) indicated significant effects of

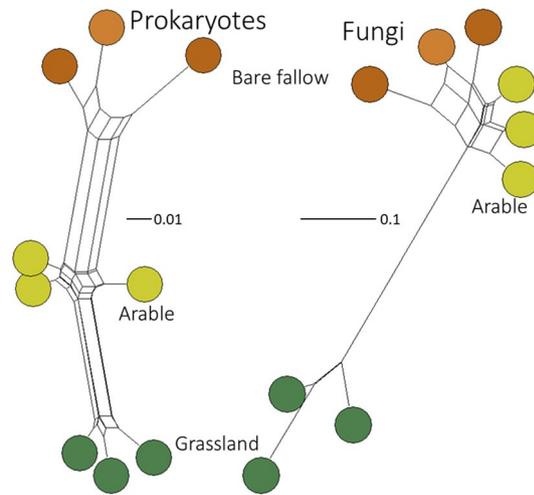


**Figure 3.** Contrasting long-term soil management results in quantitatively different process-form states. Soils are described by a combination of the connectivity of pore space, established from X-ray connected porosity (CT) and modelled hydraulic conductivity—a measure of capacity, representing the maximum potential movement of resources through pore networks to organisms. Grassland soils (green data points) are characterized as having high pore connectivity and hydraulic conductivity and are associated with the greatest stocks of  $C_{org}$ . In contrast, degraded bare fallow soils (brown data points) are associated with extremely limited connected porosity and hydraulic conductivity and the lowest stocks of  $C_{org}$ . Arable soil (dark yellow) is intermediate between these two extremes. Data point size is proportional to  $C_{org}$  ( $Mg\ ha^{-1}$ ) in each soil, the extremes of which are shown in the key.



**Figure 4.** Process-form states control anoxia within soil. Low- $C_{org}$ , low-connected porosity soil contains much larger volumes of anoxic microsites than high- $C_{org}$ , high-connected porosity soil. Across a range of matric potential ( $\psi_m$ ), the predicted volume of anoxic sites is consistently larger in degraded bare fallowed soil than arable or grassland. At field capacity ( $\theta_c$ ), approximately 30% of degraded soil is anoxic, falling to 5% in grassland soil. At 21 kPa degraded soil is completely anoxic while the volume remains between 4–5% in grassland soil. In arable soil 10% of the soil volume is predicted to be anoxic at  $\theta_c$ —double that in grassland.

soil management upon both prokaryotic (PERMANOVA,  $pseudo-F_{2,6} = 15.5$ ,  $p_{perm} < 0.0001$ ) and fungal ( $pseudo-F_{2,6} = 19.0$ ,  $p_{perm} = 0.0032$ ) community structures. Prokaryote communities were significantly different between all three treatments (smallest difference,  $pseudo-t = 2.9$ ,  $p_{MC} < 0.0001$ ) but fungal communities in arable and bare fallowed soils did not differ ( $pseudo-t = 1.7$ ,  $p_{MC} = 0.111$ ); both were significantly different from the grassland community (smallest difference,  $pseudo-t = 5.0$ ,  $p_{MC} = 0.0015$ ). Inspection of individual fungal OTU abundance



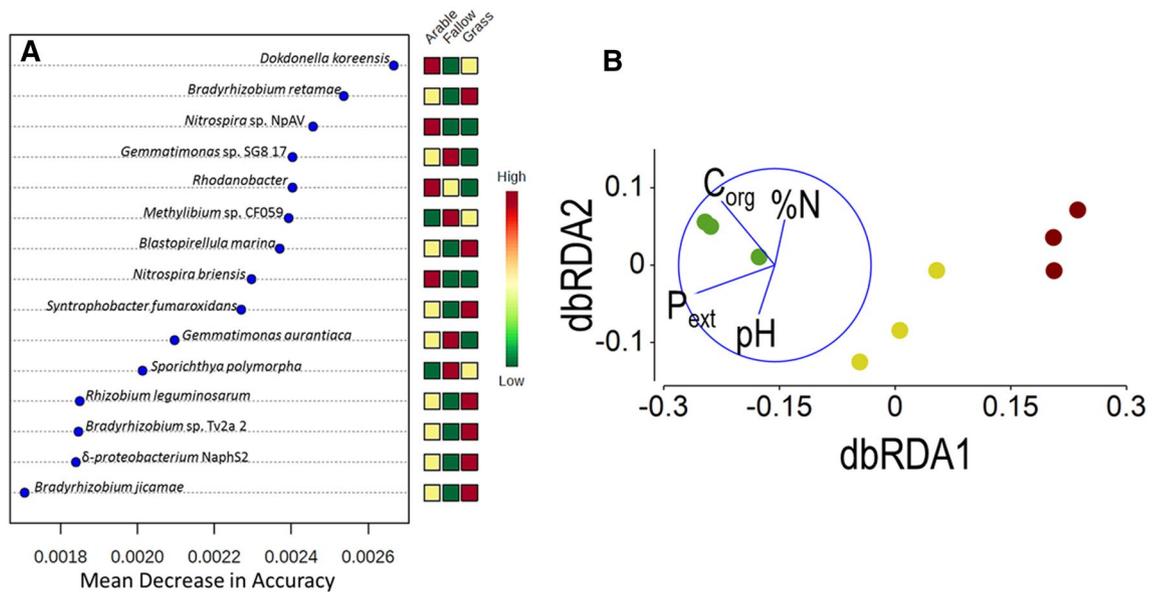
**Figure 5.** Grassland, arable and bare fallowed soil microbial community  $\beta$ -diversity. Neighbour-Net networks of prokaryotic and fungal community profiles from the three soil managements based on weighted UniFrac distance. Prokaryotic community assemblages were significantly phylogenetically different between all three managements; for fungi there was no difference between arable and bare fallow soil assemblages, which were both significantly different from grassland assemblages.

indicated that this was due to several OTUs, including *Rhizophagus irregularis* (formerly *Glomus intraradices*) and other Glomeromycetes, Agaricomycetidae, Onygenales, Eurotiomycetidae, Aspergillaceae and Atheliaceae, being abundant in grassland soils but not detected in either arable or bare fallowed soils: arbuscular mycorrhizal *R. irregularis*, for example, had a mean abundance in grassland soils of  $3.5 \times 10^5$ , but was not detected in the other soils. This large, qualitative, difference between the soils is consistent with the effect of soil tillage<sup>30,31</sup> upon fungal communities and cannot be interpreted as a response to selective pressures arising from different process-form states. Since prokaryotes appeared to be less sensitive to the effects of tillage than fungi, the effects of soil management upon prokaryotic communities were studied in detail.

Random Forest machine learning classification of biomarkers (Fig. 6A) indicated prokaryotic communities in grassland soils were characterized by Rhizobiaceae including *Bradyrhizobium* spp. and *Rhizobium leguminosarum* as well as the planctomycete *Blastopirellula*. At the other extreme, taxa characteristic of degraded, low input bare fallow soils were *Gemmatimonas* spp., an organism related to the aromatic compound degrader *Methylibium* and the actinomycete *Sporichthya*. The influence of nitrogenous fertilization was evident in the organisms identified as characteristic of arable soils; nitrite-oxidizing *Nitrospira* spp. were particularly characteristic of these soils together with the denitrifying *Rhodanobacter* and *Dokdonella koreensis*<sup>32</sup>.

16S rRNA gene-conditional phylogenetic diversity based upon placement of exact sequence variants for each treatment was compared using Kantorovich-Rubinstein (KR) distance metrics. PERMANOVA identified a significant effect of treatment ( $pseudo-F_{2,6} = 17.9$ ,  $p_{perm} < 0.0001$ ) and all post hoc comparisons were significantly different (smallest difference: bare fallow vs. arable,  $pseudo-t = 3.2$ ,  $p_{MC} = 0.0018$ ) consistent with the weighted UniFrac approach described above. Principal coordinates analysis (PCoA) was used to present an unconstrained view of differences in 16S rRNA gene-conditional microbiome assemblages (Supplementary Fig. 2) using KR distance. The first two principal coordinates separated treatments clearly, the ordination accounting for 89% of total variation across the two axes. Distance-based linear modelling (distLM) was used to describe the relationship between the 16S rRNA gene-conditional phylogenetic community structure and edaphic variables shown in Table 2. All combinations of variables were considered: the most parsimonious model, identified using Bayesian information criterion (BIC), was a combination of the chemical factors pH, soil  $C_{org}$ , soil organic nitrogen (%N) and NaOH-EDTA extractable phosphorus ( $P_{ext}$ ). Distance-based redundancy analysis (dbRDA) indicated the model accounted for 84% of total variation on the two axes (Fig. 6B). Separation of treatments on dbRDA axis 1 was associated most highly with  $P_{ext}$  ( $r = -0.81$ ; marginal test,  $pseudo-F = 7.4$ ,  $p_{perm} = 0.013$ ) and  $C_{org}$  ( $r = -0.53$ ; marginal test,  $pseudo-F = 12.2$ ,  $p_{perm} = 0.0035$ ), both greatest in grassland soils and least in bare fallowed soils. The second axis was most highly associated with  $C_{org}$  ( $r = -0.88$ ) and %N ( $r = 0.41$ ; marginal test,  $pseudo-F = 11.3$ ,  $p_{perm} = 0.004$ ). Using these four chemical parameters to model the distribution, addition of neither porosity (sequential test,  $pseudo-F = 0.7$ ,  $p_{perm} = 0.565$ ) nor  $d_{crit}$  (sequential test,  $pseudo-F = 0.5$ ,  $p_{perm} = 0.691$ ) accounted for a significant amount of additional variation.

Although there are clear qualitative and phylogenetic differences between prokaryote assemblages associated with each soil, distLM suggests that these differences are adequately described by chemical edaphic parameters. They are therefore unlikely to be due to selective pressure arising from the respective process-form states directly. Instead, assemblage differences are likely to reflect organism traits: for example, *Gemmatimonadetes* are common in soil and show adaptation to low soil moisture<sup>33</sup>, so identification of *Gemmatimonas* as characteristic of bare fallow soil is likely to reflect the fact that direct isolation experienced by these soils renders them much drier than the other soils; nitrogenous fertilization of arable soils is reflected by the organisms identified as characteristic of



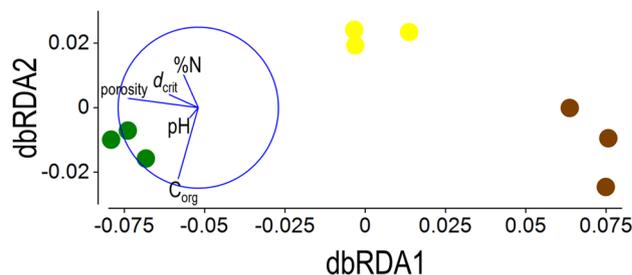
**Figure 6.** Taxonomy- and phylogeny-based community responses to land management. **(A)** Predictive modelling using a supervised Random Forest algorithm identified the fifteen OTUs that were most discriminatory between the different soils, based upon the mean decrease in model accuracy of a leave-one-out cross-validation procedure. **(B)** Management-conditional distance-based redundancy analysis (dbRDA) of chemical edaphic factors and 16S rRNA gene-based assessment of microbiomes associated with the Highfield Ley-Arable experiment using Kantorovich–Rubinstein phylogenetic distances calculated from placement of homologous metagenome reads on the 16S rRNA gene reference phylogenetic tree. Data points represent individual replicate plots of grassland (green), arable (yellow) and bare fallow (brown) soils. Environmental factors (pH, NaOH-EDTA extractable P [ $P_{ext}$ ],  $C_{org}$  and % organic N [%N]) were determined by distance-based linear modelling as the most parsimonious combination of variables to model the multivariate data and are represented as vectors, increasing in the direction of the vector: vector length indicates the degree of partial correlation of each environmental variable with the dbRDA axes. The circle has an arbitrary origin and radius of  $r = 1$ . dbRDA axis 1 accounted for 83.95% of variation accounted for by the model (74.71% of total variation) and dbRDA axis 2 accounted for 9.96% of variation accounted for by the model (8.87% of total variation).  $R^2 = 0.8899$ . The corresponding unconstrained PCoA ordination is shown in Supplementary Fig. 2. See text for a detailed description of the analysis.

	pH ( $H_2O$ ) <sup>a</sup> /-log(g[H <sup>+</sup> ]L <sup>-1</sup> )	Organic carbon <sup>a</sup> /mg g <sup>-1</sup> soil	Total nitrogen <sup>a</sup> /μg g <sup>-1</sup> soil	NaOH-EDTA extractable phosphorus <sup>b</sup> /μg g <sup>-1</sup> soil
Grassland	6.2 ± 0.13	3.72 ± 0.44	340 ± 39.0	661.7 ± 31.3
Arable	5.8 ± 0.11	1.85 ± 0.06	190 ± 5.08	517.0 ± 12.6
Bare fallow	5.3 ± 0.19	1.07 ± 0.10	110 ± 6.71	235.0 ± 3.8

**Table 2.** Summary of chemical data of Highfield Ley-Arable experiment soils. The mean and standard error of the mean are shown ( $n = 3$ ). <sup>a</sup>Gregory et al.<sup>22</sup>. <sup>b</sup>Neal et al.<sup>68</sup>.

these soils to be either nitrite-oxidisers or denitrifiers; and while identification of Rhizobiaceae as characteristic of the mixed swards of grassland soils suggests association with legumes—and therefore possibly responsive to selection pressure exerted by the plant population, *Bradyrhizobium* spp. in these soils lack genes and gene clusters for symbiosis and nitrogen fixation<sup>34</sup>.

**Process-form states in soil exert selective pressure at the level of alleles.** A total of 1,197 KEGG orthologs were identified as having significantly different abundance between the soils (selected orthologs are presented in detail in Supplementary Figs. 3–7). We adopted a similar approach to analysing the effects of soil management upon microbiome genetic variation, determined by binning reads to KEGG orthologs, as for the effect upon community assemblage, described above. Multivariate ortholog analysis was based on Hellinger distance, calculated from square root-transformed ortholog abundance. PERMANOVA identified a significant effect of land use upon ortholog assemblage ( $pseudo-F_{2,6} = 26.8$ ,  $p_{perm} < 0.0001$ ) and all post hoc comparisons were significantly different (smallest difference: arable vs. bare fallow,  $pseudo-t = 3.6$ ,  $p_{MC} = 0.0006$ ). PCoA separated each land use, the first two axes accounting for 91% of total variation (Supplementary Fig. 8). The most parsimonious model identified by distLM and BIC included a combination of both chemical and physical edaphic variables; namely pH,  $C_{org}$ , %N, porosity and  $d_{crit}$ . dbRDA (Fig. 7) showed clear separation between



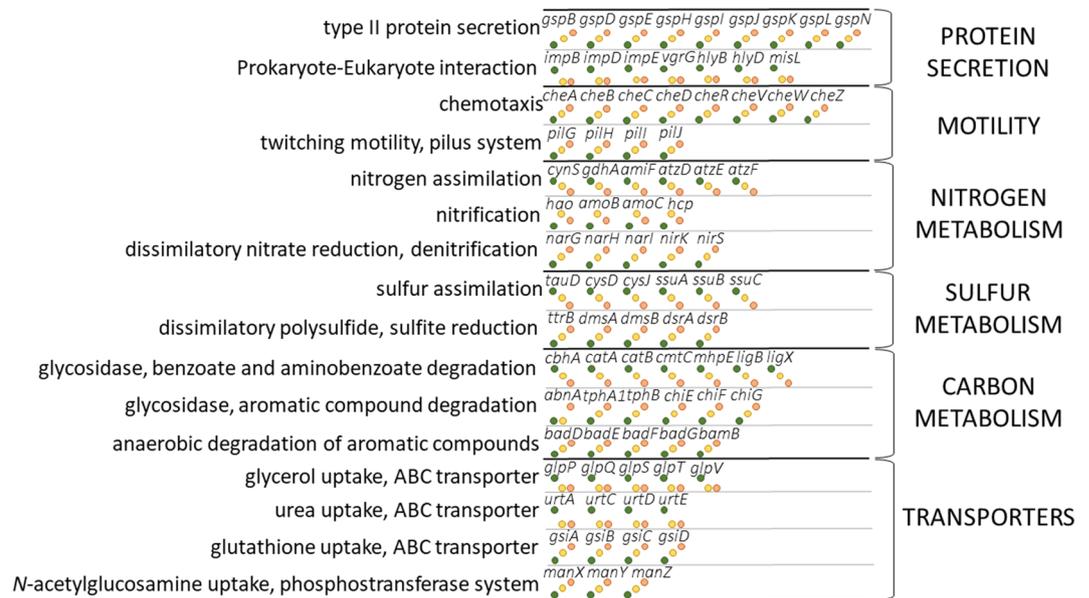
**Figure 7.** Function-based community responses to process-form states. Management-conditional dbRDA of chemical and physical edaphic factors and function-based assessment of genes associated with the Highfield Ley-Arable experiment. Square root transformed KEGG ortholog abundances were used to calculate Hellinger distances between the nine samples. Data points represent individual replicate plots of grassland (green), arable (yellow) and bare fallow (brown) soils. Environmental factors (pH,  $C_{org}$ , % organic N [%N], porosity and  $d_{crit}$ ) were selected by distLM as the most parsimonious combination of variables to model the multivariate data and are represented as vectors, increasing in the direction of the vector: **vector length indicates the degree of partial correlation of each environmental variable with the dbRDA axes.** The circle has an arbitrary origin and radius of  $r=1$ . **dbRDA axis 1 accounted for 88.40% of variation** accounted for by the model (83.21% of total variation) and dbRDA axis 2 accounted for 7.17% of variation accounted for by the model (6.75% of total variation).  $R^2=0.9414$ . The corresponding PCoA ordination is shown in Supplementary Fig. 8. See text for a detailed description of the analysis.

the treatments on dbRDA axis 1. The edaphic variables associated most highly with this axis were both physical parameters; porosity ( $r=-0.87$ ; marginal test,  $pseudo-F=24.7$ ,  $p_{perm}=0.0009$ ) and  $d_{crit}$  ( $r=-0.36$ ; marginal test,  $pseudo-F=15.2$ ,  $p_{perm}=0.0019$ ). Both variables were greatest in grassland soil and least in bare fallowed soil. The treatments showed little separation on the second axis. Edaphic variables associated most highly with this second axis were chemical,  $C_{org}$  ( $r=-0.88$ ; marginal test,  $pseudo-F=17.3$ ,  $p_{perm}=0.0002$ ), %N ( $r=0.47$ ; marginal test,  $pseudo-F=0.42$ ,  $p_{perm}=0.633$ ) and pH ( $r=0.41$ ; marginal test,  $pseudo-F=13.9$ ,  $p_{perm}=0.0038$ ). Using these edaphic parameters to model the distribution, addition of  $P_{ext}$  (sequential test,  $pseudo-F=0.97$ ,  $p_{perm}=0.484$ ) did not account for a significant amount of additional variation.

This analysis presents clear evidence for a large and direct influence of process-form state upon gene assemblages. The topological parameters shown to be so influential upon gene distributions are both fundamental properties of pore networks which we have shown to be sensitive to biotic  $C_{org}$  inputs and turnover in soil (Figs. 2, 3). We have also demonstrated that they exert a dominant influence upon hydrodynamic conductance of the pore network and the potential for anaerobic sites across a range of soil matric potential (Figs. 3, 4). However, the analysis cannot demonstrate preferential selection of genes dependent upon their fitness within each process-form state. To test whether the differences in gene abundance could be due to selection pressures arising from different plant inputs and emergent soil structural properties, we characterised the genes shown to be sensitive to the different land managements. Consideration of changes in individual gene abundance indicated clear shifts in both cellular behaviour and metabolic potential (Fig. 8). For cell behaviour, there were a number of genes associated with protein secretion, among them *impB*, *impD*, *impE* and *vgrB* associated with bacterial type VI protein secretion systems, *hylB* and *hylD* associated with type I protein secretion systems and the autotransporter gene *misL*: all typify bacterial-bacterial and bacterial-Eukaryotic interactions and were more abundant in grassland soils. Genes associated with type II protein secretion systems (T2SS) were more abundant in arable and bare fallow soils, suggesting a greater reliance upon exoenzymes in these soils. Consistent with this latter observation, several genes coding for exoenzymes were more abundant in these soils, including *abnA* (glucosyl hydrolase [GH] family 43 endo-arabinanase), *chiE* (GH family 18 chitinase) and *chiF* and *chiG* (both GH family 19 chitinases) associated with carbohydrate metabolism, and *dmsA* and *dmsB* (dimethyl sulfoxide reductase) associated with sulfur metabolism. Genes coding for chemotaxis and twitching motility were also more abundant in arable, and particularly bare fallowed soils compared to grassland soil.

The increase in abundance of *dmsAB* was also part of a general trend of an increase in genes associated with dissimilatory anaerobic metabolism of nitrogen and sulfur in arable and bare fallowed soils combined with reductions in genes associated with assimilatory pathways. Nitrification-associated genes were most abundant in arable soils, and genes associated with dissimilatory reduction of nitrate and sulfate most abundant in bare fallowed soils. There was also an increase in genes associated with anaerobic degradation of aromatic compounds in arable and bare fallowed soil. Transport pathways also differed between land use with genes associated with ATP-binding cassette (ABC) transporter pathways of glycerol and urea being most abundant in grassland soil and least abundant in bare fallowed soil, while genes associated with the ABC transport pathway for glutathione and the *N*-acetylglucosamine phosphotransferase pathway exhibited the opposite trend.

These genetic shifts were related to nutrient status and, saliently, changes in soil structure (in this case pore topology) controlling gaseous and nutrient diffusion. They present direct evidence for genetic selection of genes and pathways based upon fitness under the different process-form states. The increase in genes associated with less efficient anaerobic processes in arable and bare fallowed soil can be considered a response to reduced diffusion of  $O_2$  in these progressively more poorly connected pore networks. Other responses, such as the increase



**Figure 8.** Schematic representation of the relative abundance of genes for which significant differences between process-form states was determined. The central column indicates the general trend in relative abundance for genes grouped according to specific functions. Grassland gene abundance is represented as green points, arable gene abundance as yellow points and bare fallow gene abundance as brown points; each specific function is described in the left-hand column; specific functions are organized into higher-level KEGG ontologies, shown in the right-hand column. Absolute abundances for each gene and associated significance ( $p$ ) and positive false discovery rate ( $q$ ) of the difference in abundance between the three treatments are shown in Supplementary Figs. 3–7.

in gene abundance for chemotaxis and protein secretion, may also be responses to reduced diffusion of soluble nutrients, and hence a requirement to search out nutrients, or avoidance of anaerobic niches within the soil. Microbial community structure is often considered as a balance of cooperative behaviours between individuals, mediated by “public goods” or soluble nutrients arising from leaky processes (nutrients which are lost through the outer membrane or released by cell lysis) or the activity of exoenzymes<sup>35,36</sup>. Producers of public goods support populations of “cheaters” which exploit goods without contributing to them. In well-mixed systems, cheaters maintain a competitive advantage over producers, but this advantage is lost in structured environments where diffusive constraints are manifest<sup>37</sup>. In this context, the increase of T2SS and arabinanase and chitinase exoenzyme genes in arable and bare fallowed soils may be a response to both qualitative changes to organic inputs, and reduced delivery of soluble nutrients by advective flow and diffusion to cheaters, and thus an increase in abundance of producer organisms. Additionally, the reduced diffusive processes predicted for arable and particularly bare fallowed soil may result in an increased efficiency of exoenzymes since reduced diffusion allows for a greater accumulation of product near producer organisms<sup>38</sup>. Thus, production of exoenzymes, and cell motility as a searching or avoidance behaviour provide adaptations in response to spatially constrained circumstances arising from reduced pore connectivity as a result of reduced  $C_{org}$  inputs in arable and bare fallowed soils.

## Discussion

We have presented data consistent with the conditions which should be met if soil is an extended composite microbial phenotype: the emergent physical states are organised at several orders of magnitude above the scale of individual microbes; different physical states are associated with different genetic states at the level of individual alleles rather than organisms. The data also demonstrates that both physical and biotic states of the system can be manipulated by nutritional interventions, particularly relating to the flux of energy through the system.

The data do not prove the associations are causal, however comparison with simpler systems that are amenable to deeper theoretical analysis and direct manipulation provides additional evidence of a causal feedback between allelic abundance, process and form. Specifically, the soil-microbe system exhibits behaviour seen in physical systems that display spontaneous (and endogenously driven) emergence of large-scale self-organisation as a result of such feedback. In a subset of such systems, there is a critical point at which the state of the system changes discontinuously (a phase transition) with continuous change in one of the system parameters. The rate of change of the system with respect to that parameter is characterised by a power law close to the phase transition, reflecting the emergence of coherence across a wide range of scales (often referred to as fractal scaling). In Fig. 3 we observe such behaviour between connected porosity and hydraulic conductivity. The state of the soil system changes from one with a disconnected pore space to one with a connected pore space where  $C_{org}$  (energy) flux is a critical parameter. There is a power law relation between conductivity and porosity, consistent with the emergence of large-scale spatial coherence in soil structure at a critical value of  $C_{org}$  flux. In this sense, soil displays many of the properties of self-organising systems<sup>39</sup>. Results presented here provide further evidence for a

causal feedback between allelic abundance, process and form. We have previously posited a mechanism for this in soil and shown how soils with and without plants are capable of spontaneously generating emergent structures at important scales<sup>6</sup> compared with sterile soils, which do not. This interpretation predicts that soils which are more self-organising will be more metabolically active in any given situation than a soil where the interaction between biological process and form is weak or non-existent. We see that, after a minimum of 52 years, each soil in our study is a different expression of its multiple biotic components; a phenomenon termed an extended composite phenotype<sup>12</sup>.

With plants present, such as land managed as long-term mixed grass sward, the extended composite phenotype has an increased capacity to store water and soluble nutrients, a property which may confer a degree of resilience to the soil–plant–microbe system during periods of low rainfall or nutritional inputs. Independent analysis of these same soils has demonstrated greater water storage capacity in the grassland soil<sup>40</sup>. In addition, the more extensive and more connected pore network selects for assimilatory, and against dissimilatory, processes by permitting greater flux of O<sub>2</sub> through the system: it thus improves the efficiency of metabolic processes and C<sub>org</sub> conversion into biomass while reducing potential losses of plant nutrients arising from leaching or emission to the atmosphere. Thus, the extended phenotype interacts with plants to increase the flux, resilience and efficiency of nutrient transport to plants (including water).

The finding that soil under grassland management has significantly higher capacity, efficiency and resilience compared with arable or bare fallowed management is associated with greater C<sub>org</sub> inputs and turnover. Furthermore, the rate of recovery of degraded soil is also linked to stocks and flows of C<sub>org</sub> (Fig. 1). Our experiments cannot distinguish between C<sub>org</sub> flux or storage as the dominant mechanism supporting improved soil function. However, interpreting results in terms of soil remodelling through self-organizing processes, we predict that the biophysical state of soil and rate of change of that state will both be related to cumulative metabolic activity. Our data are consistent with recovery rate being limited by cumulative soil metabolism: soil C<sub>org</sub> content acts as a diagnostic for this. This raises the important questions of what limits soil metabolism and incorporation of C<sub>org</sub> in soil<sup>41</sup>, and how it can be manipulated in each context to maximise the rate of soil recovery. We know both anaerobic niches and physical dislocation of microbes from resources result from low pore connectivity, and both significantly limit microbial metabolism. We also know soil recovery is associated with more voluminous and better-connected pore space and significantly lower levels of anaerobic respiration. We speculate that the rate-limiting factor in recovery of degraded soil is the process of microbially-mediated micro-structure remodelling, and that this is soil texture dependent<sup>25</sup>. Sandy-textured soil would be less able to recover compared to soils with higher fractions of silt and clay, where remodelling fine-scale structure is inherently more feasible due to a greater proportion of “raw materials” to enable such fine-scale architecture to be manifest. It is also likely to be dependent on the quality and quantity of organic inputs to soil, especially in relation to the latent energy contained in them. This is apparent in our data, though we are not able to distinguish the relative importance of each.

Tillage is known to contribute to decreases in soil C<sub>org</sub>, and the most effective recovery rate and highest metabolizing end-state in our data was achieved with management under grassland without tillage. Tillage has the effect of significantly changing the distribution of microenvironments in soil through increased aeration and exposure of previously physically protected prey organisms and soil C<sub>org</sub>. This results in the immediate release of physical and chemical constraints on metabolism and therefore to loss of soil C<sub>org</sub>. More importantly, rearrangement of microenvironments—i.e. within and between soil macro- and micro-aggregates—will have the effect of “re-setting” microbial remodelling of soil microarchitecture, slowing down establishment of connected pore space and longer-term cumulative metabolism.

This new interpretation of the role of nutritional and physical management of soil is a step towards a more general theory of soil. Such a theory is needed as a framework upon which to synthesize data and knowledge on biological, chemical and physical properties of soil that are typically studied in isolation. Theory leading to quantitative prediction is also essential in seeking synergistic interventions that recognise the interplay between capacity, efficiency and resilience of soil, and to avoid the unintended consequences of land management that are directing us towards systemic collapse of productive land and an amenable climate.

## Methods

**Soils.** We analysed soil from plots of the Rothamsted Highfield Ley-Arable field experiment (00:21:48° W, 51:48:18° N) set on soil that has been under permanent grass since at least 1838. The soil is a silty clay loam (25% clay: 62% silt: 13% sand) (Chromic Luvisol according to FAO criteria). In 2008, plots of severely degraded soil managed as bare fallow by regular tillage to remove any plants since 1959, were converted to arable and grassland managements. Arable soil was placed under continuous wheat rotation (winter wheat, *Triticum aestivum* L., currently cv. “Hereward” seed coated with Redigo Deter combination insecticide/fungicide treatment, Bayer CropScience) receiving ammonium nitrate fertilization to provide approximately 220 kg-N ha<sup>-1</sup> annum<sup>-1</sup>, and additional 250 kg-K ha<sup>-1</sup> and 65 kg-P ha<sup>-1</sup> every three years, and grassland plots were maintained as a managed sward of mixed grasses and forbs. Plots were sampled annually, and soil was air-dried and sieved (< 2 mm) before being archived. To follow the development of soil structure in these soils, soil aggregates from continuous bare fallow, bare fallow converted to arable and bare fallow converted to grassland were selected for the years 2008, 2010, 2012, 2015 and 2018. In addition, we also sampled plots which had been managed consistently as bare fallow for fifty-two years, arable for sixty-two years or mixed grass swards since 1838. Physical and biological data has already been reported for these consistently managed soils (Table 1). Over these periods, the bare fallowed soils have become depleted in more labile organic carbon and enriched in persistent organic carbon<sup>42</sup> and soil organic carbon has been reduced to a greater extent than in arable soil. There has also been an observable progressive shift, from grassland to arable and bare fallowed soil, in the distribution of organic carbon between different pools in the three soil managements, particularly a relative decline in discrete organic particles independ-

ent of stable soil aggregates, and a corresponding increase in the proportion of organic particles encapsulated in stable aggregates<sup>23</sup>. Confirmation of this apparent shift in soil structure has been provided by high-resolution X-ray Computed Tomography<sup>25</sup>.

**X-ray computed tomography and image analysis.** We generated X-ray Computed Tomography (CT) images at 1.5  $\mu\text{m}$  resolution and scales relevant to microbes ( $10^0$ – $10^2$   $\mu\text{m}$ ), requiring imaging of 0.7–2.0 mm diameter soil aggregates. The connectivity of pores within networks was assessed from binary images derived from X-ray CT using Minkowski functions<sup>27</sup>, basic geometric measures defined for binary structures. Aggregates were selected at random from soil collected from each plot of the Highfield experiment. Each was scanned using a Phoenix Nanotom system (GE Measurement and Control solution, Wunstorf, Germany) operated at 90 kV and a current of 65  $\mu\text{A}$ . Initial image analysis was performed using Image-J. Images were threshold-adjusted using the bin bi-level approach of Vogel et al.<sup>27</sup> using QuantIm version 4.01 (<https://www.quantim.ufz.de/>). Porosity and mean pore neck size were estimated directly from the threshold-adjusted binary images and Minkowski functions including Euler number— $\chi(d)$ , pore size distribution, pore connectivity and surface area density were determined according to Vogel et al.<sup>27</sup>.  $\chi(d)$  is a well-defined characteristic related to pore space topology and shown to be critical to hydraulic properties<sup>43</sup>. In three dimensions,  $\chi(d)$  is defined as the number of isolated pores (of diameter,  $d$ ) minus the number of redundant connections within the pore space, plus the number of enclosed pores<sup>44</sup>. Using this approach, we estimated Euler number density— $\chi(d)/V$ , where  $V$  represents the image volume—of the pore network of aggregates from each continuously managed Highfield soil.

**Calculation of diffusion in soil pore networks.** The hierarchical soil structures revealed in X-ray CT images indicate that gaseous  $\text{O}_2$  in the atmosphere moves into soil primarily through its inter-aggregate pores and is then dissolved in water prior to moving into the aggregates largely by molecular diffusion. Since gaseous  $\text{O}_2$  diffuses up to  $10^3$ -fold more quickly than  $\text{O}_2$  dissolved in water, microbial community activity is thus constrained mainly by  $\text{O}_2$  diffusion within aggregates. The ability of aggregates to conduct dissolved  $\text{O}_2$  and other soluble substrates depends on the intra-aggregate pore geometry, and we quantified it with effective diffusion coefficients calculated directly by mimicking solute movement through the pore geometry using numerical simulations. The movement of solutes, including  $\text{O}_2$  and substrates, within the pore geometry is assumed to be diffusion dominated. For the images illustrated in Supplementary Fig. 9, the temporal change in solute concentration inside any pore voxel can be calculated using the finite volume approach, as follows:

$$\begin{aligned} \frac{c_o^{t+\delta t} - c_o^t}{\delta t} &= q_w + q_e + q_s + q_n + q_u + q_d, \\ q_w &= \begin{cases} D(c_w^{t+\delta t} - c_o^{t+\delta t}) & \text{if voxel } w \text{ is pore} \\ 0, & \text{if voxel } w \text{ is solid} \end{cases}, \quad q_e = \begin{cases} D(c_e^{t+\delta t} - c_o^{t+\delta t}) & \text{if voxel } e \text{ is pore} \\ 0, & \text{if voxel } e \text{ is solid} \end{cases} \\ q_s &= \begin{cases} D(c_s^{t+\delta t} - c_o^{t+\delta t}) & \text{if voxel } s \text{ is pore} \\ 0, & \text{if voxel } s \text{ is solid} \end{cases}, \quad q_n = \begin{cases} D(c_n^{t+\delta t} - c_o^{t+\delta t}) & \text{if voxel } n \text{ is pore} \\ 0, & \text{if voxel } n \text{ is solid} \end{cases} \\ q_d &= \begin{cases} D(c_d^{t+\delta t} - c_o^{t+\delta t}) & \text{if voxel } d \text{ is pore} \\ 0, & \text{if voxel } d \text{ is solid} \end{cases}, \quad q_u = \begin{cases} D(c_u^{t+\delta t} - c_o^{t+\delta t}) & \text{if voxel } u \text{ is pore} \\ 0, & \text{if voxel } u \text{ is solid} \end{cases} \end{aligned} \quad (1)$$

where  $c$  is concentration,  $q$  is diffusive flux,  $D$  is molecular diffusion of the solute in liquid water, superscripts  $t$  and  $t + \delta t$  represent time,  $\delta t$  is a time increment, subscript  $o$  represents the pore voxel being calculated, and subscripts  $w$ ,  $e$ ,  $s$ ,  $n$ ,  $u$  and  $d$  represents the face-to-face neighbours of voxel  $o$  on the west, east, south, north, top and bottom sides respectively. Applying Eq. (1) to all pore voxels leads to linear systems which was solved by the bi-conjugate gradient stabilized method<sup>45</sup>.

**Calculation of diffusion coefficients.** To calculate the effective diffusion coefficient of each aggregate, we applied a constant concentration  $C_1$  on the top and a constant concentration  $C_0$  on the bottom of the image, and then simulated solute diffusion to steady state. The diffusive flux in the three directions in each pore voxel was calculated by Eq. 1. Taking the vertical direction as the  $z$  direction for the image, the effective diffusion coefficient of the image was calculated as follows:

$$D_{\text{eff}} = \frac{L_z \sum_{i=1}^N q_z(x_i)}{N(C_1 - C_0)}, \quad (2)$$

where  $D_{\text{eff}}$  is the effective diffusion coefficient,  $N$  is the total number of pore voxels in the simulated images,  $q_z(x_i)$  is the vertical diffusive flux in pore voxel centred at location  $x_i$ ,  $L_z$  is the height of the image as shown in Supplementary Fig. 9. To address the impact of change in pore geometry due to management on the ability of the aggregate to diffuse solute, in result analysis we normalized the effective diffusion coefficient  $D_{\text{eff}}$  of all solutes by their associated molecular diffusion coefficient in non-constrained water,  $D$ .

**Modelling of oxygen diffusion and anoxia.** The impact of soil structure on  $\text{O}_2$  diffusion and its subsequent consumption by microbes under various saturations was studied using pore-scale simulations. We first calculated the spatial distribution and connectedness of different pores and then determined water distributions in pores under different matric potentials ( $\psi_m$ ). We assumed the soil was initially saturated and then applied a negative pressure  $p$  at the bottom to drain water. We assumed the soil was essentially hydrophilic in that only pores whose associated capillary pressure  $p_c$ , calculated by  $p_c = \sigma/r$  with  $\sigma$  being water–air surface tension, is

less than  $p$  and that they form clusters which stretch from the top to the bottom of the structure can be drained. Supplementary Fig. 10A shows an example illustrating water distribution in the structure calculated using the method described above when the saturation is 55%.

Once the water distribution was determined for a given  $\psi_m$ , we treated the water–air interfaces inside the structure as a boundary at which gaseous  $O_2$  dissolves and then moves toward the solid–water interface to be reduced by microbial reactions. The partial pressure of gaseous  $O_2$  in the simulated structure was assumed to be constant. Movement of dissolved  $O_2$  in the liquid water was simulated using the following diffusion–reaction equation:

$$\begin{aligned}\frac{\partial c}{\partial t} &= \nabla D \nabla c - s, \\ c|_{\Gamma_{aw}} &= c_s,\end{aligned}\quad (3)$$

where  $c$  is the concentration of dissolved  $O_2$ ,  $D$  is the molecular diffusion coefficient of  $O_2$  in water,  $\Gamma_{aw}$  is the air–water interface,  $s$  is microbial consumption,  $c_s$  is the saturated dissolved  $O_2$  concentration at the water–air interface calculated from Henry’s law,  $c_s = p_o/H$  in which  $H$  is the Henry constant and  $p_o$  is the partial pressure of the gaseous  $O_2$  inside the structure. Microbial consumption was assumed to occur in water-filled voxels adjacent to the water–solid wall and described by the following Monod kinetic equation:

$$s = m_c k_0 \frac{[C]}{k_C + [C]} \frac{c}{k_O + c}, \quad (4)$$

where  $m_c$  is microbial biomass,  $k_0$  is kinetic parameter,  $[C]$  is the concentration of dissolved carbon. Since we are interested in impact of soil structure on development of anaerobic sites, we simulated  $O_2$  diffusion and reduction to steady state. In all simulations, we normalized Eqs. (3) and (4) as follows

$$\begin{aligned}\frac{\partial c'}{\partial t'} &= \nabla D' \nabla c' - s', \\ c'|_{\Gamma_{aw}} &= 1, \\ s' &= k' \frac{c}{k'_O + c},\end{aligned}$$

where  $t' = t/T_0$ ,  $D' = DT_0/L^2$ ,  $c' = c/c_s$  and  $k' = m_c k_0 T_0 [C]/(k_C + [C])$  in which  $L$  is the side length of the voxels and  $T_0$  is a characteristic chosen to make  $D' = 1$  in our simulations.

The above equation was solved by a finite volume method with each water-filled voxel being the element used to calculate the mass balance. In all simulations, water was assumed to be initially free of  $O_2$  and we simulated the system to steady state. As the development of anaerobic areas was a balance between the ability of soil to diffuse dissolved  $O_2$  and the microbial consumption rate, to elucidate that the relative anaerobicity of soils under the same  $\psi_m$  is the consequence of their structures and does not change with microbial activity rate, we simulated two scenarios: a fast microbial decomposition ( $k' = 1 \times 10^{-2}$ ) and a slow microbial decomposition ( $k' = 1 \times 10^{-4}$ ). For each scenario, once the system was deemed to have reached a steady state, we sampled sites and where the concentration of dimension-less dissolved  $O_2$  was less than 20% assumed them to be anaerobic<sup>46</sup>. Supplementary Fig. 10B shows an illustrative example of the location of anaerobic areas simulated by the above method in which soil particles were made transparent. We repeated the procedure to achieve different water distributions calculated by varying  $\psi_m$  and then calculated the proportional change in the volumetric anaerobic sites with the  $\psi_m$  for both the fast and slow microbial reactions. The results are shown in Fig. 4 for soil samples taken from all treatments.

**Modelling of organic carbon dynamics in soil.** We used RothC-26.3<sup>47</sup> to model the turnover of soil  $C_{org}$  in the experimental soils, accounting for the effects of soil type, plant cover and historical temperature and moisture content on organic carbon turnover processes. We used the same inputs of  $C_{org}$  to the soil as those used by Johnston et al.<sup>48</sup>. To obtain the starting soil  $C_{org}$  of 63.6 Mg-C ha<sup>-1</sup>, input to the soil from plant debris, roots, and root exudates was 2.7 Mg-C ha<sup>-1</sup>, with inert organic matter (IOM) being 3.0 Mg-C ha<sup>-1</sup>. The incoming  $C_{org}$  from plant residues were assumed to have decomposable plant material (DPM) and resistant plant material (RPM) in the proportion 0.59 and 0.41, respectively; these are the default proportions for arable cropping and managed grassland. For the first 12 years after the experiment started, grass was grazed by sheep before management changed to a grass/clover sward, harvested once or twice a year for conservation. To reflect this, the modelled grass management received inputs of 5 Mg-C ha<sup>-1</sup> annum<sup>-1</sup> between 1949 and 1960, or 4 Mg-C ha<sup>-1</sup> annum<sup>-1</sup> between 1961 and 2016. Arable management received a carbon input of 1.4 Mg-C ha<sup>-1</sup> annum<sup>-1</sup> and bare fallow management received no inputs of carbon to the soil.

**DNA extraction and metagenome sequencing.** Soil was collected from triplicate plots for each management to a depth of 10 cm using a 3-cm diameter corer. The top 2-cm of soil containing root mats and other plant detritus was discarded. Ten cores per plot were pooled and thoroughly mixed whilst sieving through a 2-mm mesh; samples were then frozen at  $-80$  °C. All implements were cleaned with 70% ethanol between sampling/sieving soil from each plot. Soil community DNA was extracted from a minimum of 2 g soil using the MoBio PowerSoil DNA isolation kit (Mo Bio Laboratories, Inc. Carlsbad, CA, USA) with three replicates for each soil treatment. When necessary, extracts from individual replicates were pooled to provide sufficient

material for sequencing. 10 µg of high-quality DNA was provided for sequencing for each of the nine continuous management plots. Shotgun metagenomic sequencing of DNA was provided by Illumina (Great Abington, UK) using a HiSeq 2000 sequencing platform, generating 150-base, paired-end reads. The generated sequences were limited to a minimum quality score of 25 and a minimum read length of 70-bases using Trimmomatic<sup>49</sup>. After filtering to remove substandard sequences, the average metagenome size for each soil was  $4.96 \times 10^8$  reads for grassland,  $2.86 \times 10^8$  for arable and  $2.88 \times 10^8$  for bare fallow soils. Since differences in library sizes were less than tenfold, we did not employ rarefaction before comparing the datasets<sup>50</sup>.

**Bioinformatical analysis of metagenome sequences.** To assess general abundance of taxa and genes in metagenomes, we mapped individual metagenomic sequences to the RefSeq non-redundant (NR) protein database held at NCBI (downloaded August 22nd, 2018) using DIAMOND version 0.8.27<sup>51</sup> in BLASTX mode using a bitscore cut-off of 55. For each sequence, only the match with the highest bitscore was considered. Sequences not matching the NR database were considered currently unclassified. MEGAN Ultimate version 6.10.2<sup>52</sup> was used to associate metagenome sequences with both taxa and Kyoto Encyclopaedia of Genes and Genomes (KEGG) functional orthologs and modules<sup>53</sup>. For taxa, MEGAN was used to establish Prokaryotic and Fungal community assemblages and calculate weighted UniFrac distances<sup>54</sup> between assemblages associated with each land management. To identify the most diagnostic microorganisms characterising communities of each soil, we used supervised Random Forests (RF), a classification algorithm approach based upon a collection of unpruned decision trees<sup>55</sup>, each built using a bootstrap sample of training data using a randomly selected subset of OTUs. The RF classifier was built by growing 5,000 classification trees. The prediction performance and confusion matrices were determined using out-of-bag cross-validation. The percent mean decrease in accuracy of the importance matrix was used to select taxa that were most predictive of each microbiome assemblage. RF was employed as implemented in MicrobiomeAnalyst<sup>56</sup>. In addition, bacterial communities were also compared based upon the abundance and phylogenetic relatedness of metagenome reads homologous to the bacterial 16S rRNA gene. A 16S rRNA profile hidden Markov model (pHMM) was generated based upon an alignment of the set of 4,528 reference sequences associated with paprica<sup>57</sup>, built December 2017. Metagenome reads with homology to the 16S rRNA pHMM were identified using hmmsearch<sup>58</sup> with a  $1 \times 10^{-5}$  Expect-value (*E*) cut-off and assigned to branches of the fixed maximum likelihood 16S rRNA phylogenetic tree using a phylogenetic placement algorithm, pplacer version 1.1.alpha10<sup>59</sup>. To assess 16S rRNA gene-based  $\beta$ -diversity in the different soils, Kantorovich-Rubinstein (KR) phylogenetic distance metrics<sup>60</sup> were calculated from phylogenetic placements of metagenome reads using the guppy kr binary (part of the pplacer suite), treating each query as a point mass concentrated on the highest-weight placement. The advantage of the KR distance metric is that it compares gene assemblage distributions on a phylogenetic tree (of 16S rRNA or other genes), in units of nucleotide substitutions per site, and is therefore a biologically meaningful approach to comparing communities.

From all of the reads binned to a KEGG orthologous group, we selected those associated with carbohydrate metabolism (ko09101) [including glycolysis/gluconeogenesis (ko00010), citrate cycle (ko00020), pentose phosphate pathway (ko00030), pentose and glucuronate interconversions (ko00040), fructose and mannose metabolism (ko00051), galactose metabolism (ko00052), ascorbate and aldarate metabolism (ko00053), starch and sucrose metabolism (ko00500), amino sugar and nucleotide sugar metabolism (ko00520), pyruvate metabolism (ko00620), glyoxylate and dicarboxylate metabolism (ko00630), propanoate metabolism (ko00640), butanoate metabolism (ko00650), C5-branched dibasic acid metabolism (ko00660), inositol phosphate metabolism (ko00562)], methane metabolism (ko00680), carbon fixation pathways in prokaryotes (ko00720), nitrogen metabolism (ko00910), sulfur metabolism (ko00920), xenobiotics biodegradation and metabolism (ko09111) [including benzoate degradation (ko00362), aminobenzoate degradation (ko00627), fluorobenzoate degradation (ko00364), chloroalkane and chloroalkene degradation (ko00625), chlorocyclohexane and chlorobenzene degradation (ko00361), toluene degradation (ko00623), xylene degradation (ko00622), nitrotoluene degradation (ko00633), ethylbenzene degradation (ko00642), styrene degradation (ko00643), atrazine degradation (ko00791), caprolactam degradation (ko00930), dioxin degradation (ko00621), naphthalene degradation (ko00626), polycyclic aromatic hydrocarbon degradation (ko00624), furfural degradation (ko00365), steroid degradation (ko00984), metabolism of xenobiotics by cytochrome P450 (ko00980) and drug metabolism—other enzymes (ko00983)], enzyme families (ko09112), membrane transport (ko09131) [including transporters (ko02000), ABC transporters (ko02010), phosphotransferase systems (ko02060), bacterial secretion systems (ko03070) and secretion systems (ko02044)], two-component systems (ko02020 and 02022), biofilm formation—*Vibrio cholerae* (ko05111), —*Pseudomonas aeruginosa* (ko02025), —*Escherichia coli* (ko02026), bacterial chemotaxis (ko02030), bacterial motility proteins (ko02035), and flagellar assembly (ko02040) for detailed study of abundance differences between the soils. Where necessary, KEGG orthologs were associated with higher-order functions by mapping to the KEGG BRITe functional hierarchy classification. In total, 8,857 KEGG functional orthologs were identified. To identify genes for which a significant difference in abundance between the treatments was observed we used DESeq2<sup>61</sup> which employs a negative binomial generalized linear model to generate maximum-likelihood estimates for the  $\log_2$ -fold change between conditions associated with each gene. Bayesian shrinkage, based upon a zero-centred normal distribution as a prior, reduces the  $\log_2$ -fold change towards zero for genes with low mean counts or a high dispersion in their count distribution. The resulting shrunken fold-changes are used in tests of significance using Wald's test. DESeq2 has been shown to be particularly sensitive to differences in gene abundance on small datasets<sup>50</sup> such as those in this study. Before analysis, 3,930 low abundance genes were removed (minimum mean count of 20) as well as 986 genes with the lowest coefficients of variation. Differential abundance of the remaining 3,940 genes was tested for significance employing  $\alpha = 0.05$  and a Benjamini–Hochberg false discovery rate (*q*) of 0.1 to control type I error rate in the face of multiple comparisons. DESeq2 was employed as implemented in MicrobiomeAnalyst.

**Statistical analysis.** One-factor analysis of variance was employed to test the effect of soil treatment upon  $d_{crit}$  and modelled diffusion coefficients arising from X-ray CT, and phylogenetic diversity estimates of  $\alpha$ -diversity arising from metagenomic analysis. Where a significant treatment effect was observed, post hoc pair-wise comparisons were performed using Tukey's HSD test (Q) employing the Copenhaver and Holland multiple comparisons procedure. These tests were performed in PAST version 3.25<sup>62</sup>. One-factor analysis of covariance (ANCOVA) was used to test for treatment effects upon the formation of connected porosity in degraded soil following conversion to either arable or grassland, using time post conversion as the covariate. The assumption of homogeneity of slopes was tested first, before ANCOVA was used to test for treatment effects using an equal slopes model. Post hoc Holm–Šidák multiple pair-wise comparisons were used to establish whether differences in adjusted mean connected porosity between treatments were significant. ANCOVA was performed in Sigma-Plot for Windows version 14.0 (Systat Software Inc., San Jose, CA, USA).

For metagenome-associated multivariate data, we first compared prokaryotic and fungal communities by calculating unrooted phylogenetic Neighbour-Nets<sup>63</sup> using weighted UniFrac distances and compared the 16S rRNA-conditional bacterial assemblages using KR distances. KR distance-based analyses were performed after testing for heteroscedasticity using the PERMDISP test<sup>64</sup>. Hypothesis testing was based upon permutational multivariate analysis of variance<sup>65</sup> (PERMANOVA) and post hoc pair-wise tests. Differences between treatment were visualized using Principal Coordinates Analysis using KR distance. To identify associations between chemical and physical edaphic factors and any treatment effects, distance-based linear modelling<sup>66</sup> was used to identify the best combination of edaphic factors to model the multivariate data; the resulting model was visualized using distance-based redundancy analysis. All multivariate tests were performed in PRIMER PERMANOVA + version 7.0.13 (PRIMER-e, Auckland, New Zealand) and probabilities were based upon 99,999 permutations (denoted  $p_{perm}$ ). For post hoc pair-wise comparisons, since the number of observations was insufficient to allow a reasonable number of permutations, Monte Carlo probabilities (denoted  $p_{MC}$ ) were calculated based upon an asymptotic permutation distribution<sup>67</sup>.

## Data availability

Data, code, and materials described in this research are available together with extensive chemical, climate and treatment data and history on the *e*-RA database, maintained by Rothamsted Research.

Received: 24 October 2019; Accepted: 5 June 2020

Published online: 30 June 2020

## References

- Cook, B. I., Miller, R. L. & Seager, R. Amplification of the North American “Dust Bowl” drought through human-induced land degradation. *Proc. Nat. Acad. Sci. USA*. **106**, 4997–5001 (2009).
- Dupré, J. & Nicholson, D. J. A manifesto for a processual philosophy of biology. In *Everything Flows: Towards a Processual Philosophy of Biology* (eds Nicholson, D. J. & Dupré, J.) 3–45 (Oxford University Press, Oxford, 2018).
- Tisdall, J. M. & Oades, J. M. Organic matter and water-soluble aggregates in soils. *Eur. J. Soil Sci.* **33**, 141–163 (1982).
- Kallenbach, C., Frey, S. & Grandy, A. Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. *Nat. Commun.* **7**, 13630. <https://doi.org/10.1038/ncomms13630> (2016).
- Barré, P. *et al.* Microbial and plant-derived compounds both contribute to persistent soil organic carbon in temperate soils. *Bio-geochemistry* **140**, 81–92 (2018).
- Crawford, J. W. *et al.* Microbial diversity affects self-organization of the soil-microbe system with consequences for function. *J. R. Soc. Interface*. **9**, 1302–1310 (2012).
- Lehmann, J. & Kleber, M. The contentious nature of soil organic matter. *Nature* **528**, 60–68 (2015).
- Kravchenko, A. N. *et al.* Microbial spatial footprint as a driver of soil carbon stabilization. *Nat. Commun.* **10**, 3121. <https://doi.org/10.1038/s41467-019-11057-4> (2019).
- Raynaud, X. & Nunan, N. Spatial ecology of bacteria at the microscale in soil. *PLoS ONE* **9**, e87217. <https://doi.org/10.1371/journal.pone.0087217> (2014).
- Addiscott, T. M. Soil mineralization: An emergent process?. *Geoderma* **160**, 31–35 (2010).
- Addiscott, T. M. Entropy, non-linearity and hierarchy in ecosystems. *Geoderma* **160**, 57–63 (2010).
- Phillips, J. D. Soils as extended composite phenotypes. *Geoderma* **149**, 142–151 (2009).
- Dawkins, R. *The Extended Phenotype. The Gene as the Unit of Selection* (Oxford University Press, Oxford, 1982).
- Dawkins, R. Extended Phenotype—but not too extended. A reply to Laland, Turner & Jablonka. *Biol. Philos.* **19**, 377–396 (2004).
- ter Horst, C. P. & Zee, P. C. Eco-evolutionary dynamics in plant–soil feedbacks. *Funct. Ecol.* **30**, 1062–1072 (2016).
- King, A. E. & Blesh, J. Crop rotations for increased soil carbon: Perenniality as a guiding principle. *Ecol. Appl.* **28**, 249–261 (2018).
- Lensi, R., Claysjoserand, A. & Monrozier, L. J. Denitrifiers and denitrifying activity in size fractions of a mollisol under permanent pasture and continuous cultivation. *Soil Biol. Biochem.* **27**, 61–69 (1995).
- Chotte, J. L., Schwartzmann, A., Bally, R. & Monrozier, L. J. Changes in bacterial communities and *Azospirillum* diversity in soil fractions of a tropical soil under 3 or 19 years of natural fallow. *Soil Biol. Biochem.* **34**, 1083–1092 (2002).
- Or, D., Smets, B. F., Wraith, J. M., Dechesne, A. & Friedman, S. P. Physical constraints affecting bacterial habitats and activity in unsaturated porous media: A review. *Adv. Water Res.* **30**, 1505–1527 (2007).
- Sexstone, A. J., Revsbech, N. P., Parkin, T. B. & Tiedje, J. M. Direct measurement of oxygen profiles and denitrification rates in soil aggregates. *Soil Sci. Soc. Am. J.* **49**, 643–651 (1985).
- Keiluweit, M., Wanzek, T., Kleber, M., Nico, P. & Fendorf, S. Anaerobic microsites have an unaccounted role in soil carbon stabilization. *Nat. Commun.* **8**, 1771. <https://doi.org/10.1038/s41467-017-01406-6> (2017).
- Gregory, A. S. *et al.* Long-term management changes topsoil and subsoil organic carbon and nitrogen dynamics in a temperate agricultural system. *Eur. J. Soil Sci.* **67**, 421–430 (2016).
- Hirsch, P. R. *et al.* Starving the soil of plant inputs for 50 years reduces abundance but not diversity of soil bacterial communities. *Soil Biol. Biochem.* **41**, 2021–2024 (2009).
- Barré, P. *et al.* Long-term bare fallow experiments offer new opportunities for the quantification and the study of stable carbon in soil. *Biogeoosci. Discuss.* **7**, 4887–4917 (2010).
- Bacq-Labreuil, A. *et al.* Effects of cropping systems upon the three-dimensional architecture of soil systems are highly contingent upon texture. *Geoderma* **332**, 73–83 (2018).

26. Köhne, J. M., Schlüter, S. & Vogel, H.-J. Predicting solute transport in structured soil using pore network models. *Vadose Zone J.* **10**, 1082–1096 (2011).
27. Vogel, H.-J., Weller, U. & Schlüter, S. Quantification of soil structure based upon Minkowski functions. *Comput. Geosci.* **36**, 1236–1245 (2010).
28. Zhang, X. & Ren, L. Lattice Boltzmann model for agrochemical transport in soils. *J. Contam. Hydrol.* **67**, 27–42 (2003).
29. Zhang, X., Deeks, L. K., Bengough, A. G., Crawford, J. W. & Young, I. M. Determination of soil hydraulic conductivity with lattice Boltzmann method and soil thin-section technique. *J. Hydrol.* **306**, 59–70 (2005).
30. Brito, I., Goss, M. J., de Carvalho, M., Chatagnier, O. & van Tuinen, D. Impact of tillage system on arbuscular mycorrhiza fungal communities in the soil under Mediterranean conditions. *Soil Tillage Res.* **121**, 63–67 (2012).
31. Lu, X., Lu, X. & Liao, Y. Effect of tillage treatment on the diversity of soil arbuscular mycorrhizal fungal and soil aggregate-associated carbon content. *Front. Microbiol.* **9**, 2986. <https://doi.org/10.3389/fmicb.2018.02986> (2018).
32. Pishgar, R., Dominic, J. A., Sheng, Z. & Tay, J. H. Denitrification performance and microbial versatility in response to different selection pressures. *Bioresour. Technol.* **281**, 72–83 (2019).
33. de Bruyn, J., Nixon, L., Fawaz, M., Johnson, M. & Radosevich, M. Global biogeography and quantitative season dynamics of Gemmatimonadetes in soil. *Appl. Environ. Microbiol.* **77**, 6295–6300 (2011).
34. Jones, F. *et al.* Novel European free-living, non-diazotrophic *Bradyrhizobium* isolates from contrasting soils that lack nodulation and nitrogen fixation genes: A genome comparison. *Sci. Rep.* **6**, 25858. <https://doi.org/10.1038/srep25858> (2016).
35. Allison, S. D. Cheaters, diffusion and nutrients constrain decomposition by microbial enzymes in spatially structured environments. *Ecol. Lett.* **8**, 626–635 (2005).
36. Harrington, K. I. & Sanchez, A. Eco-evolutionary dynamics of complex social strategies in microbial communities. *Commun. Integr. Biol.* **7**, e28230. <https://doi.org/10.4161/cib.28230> (2014).
37. Allison, S. D., Lu, L., Kent, A. G. & Martiny, A. C. Extracellular enzyme production and cheating in *Pseudomonas fluorescens* depend upon diffusion rates. *Front. Microbiol.* **5**, 169. <https://doi.org/10.3389/fmicb.2014.00169> (2014).
38. Allison, S. D., Weintraub, M. N., Gartner, T. B. & Waldrop, M. P. Evolutionary-economic principles as regulators of soil enzyme production and ecosystem production. *Soil Enzymol.* **22**, 229–243 (2011).
39. Markovic, D. & Gros, C. Power-laws and self-organised criticality in theory and nature. *Phys. Rep.* **536**, 41–47 (2014).
40. Obour, P. B., Jensen, J. L., Lamandé, M., Watts, C. W. & Munkholm, L. J. Soil organic matter widens the range of water contents for tillage. *Soil Tillage Res.* **182**, 57–65 (2018).
41. Poulton, P., Johnston, J., Macdonald, A., White, R. & Powlson, D. Major limitations to achieving “4 per 1000” increases in soil organic carbon stock in temperate regions: Evidence from long-term experiments at Rothamsted Research, United Kingdom. *Glob. Change Biol.* **24**, 2563–2584 (2018).
42. Barré, P. *et al.* Quantifying and isolating stable soil organic carbon using long-term bare fallow experiments. *Biogeosciences* **7**, 3839–3850 (2010).
43. Vogel, H.-J. A numerical experiment on pore size, pore connectivity, water retention, permeability, and solute transport using network models. *Eur. J. Soil Sci.* **51**, 99–105 (2000).
44. Vogel, H.-J. Topological characterization of porous media. In *Morphology of Condensed Matter: Physics and Geometry of Spatially Complex Systems. Chapter 3* (eds Mecke, K. & Stoyan, D.) 75–92 (Springer, Berlin, 2002).
45. van der Vorst, H. A. BI-CGSTAB—A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems. *Siam J. Sci. Stat. Comp.* **13**, 631–644 (1992).
46. Harrison, D. E. F. & Pirt, S. J. The influence of dissolved oxygen concentration on the respiration and glucose metabolism of *Klebsiella aerogenes* during growth. *J. Gen. Microbiol.* **46**, 193–211 (1967).
47. Coleman, K. *et al.* Simulating trends in soil organic carbon in long-term experiments using RothC-26.3. *Geoderma* **81**, 29–44 (1997).
48. Johnston, A. E., Poulton, P. R. & Coleman, K. Soil organic matter: Its importance in sustainable agriculture and carbon dioxide fluxes. *Adv. Agron.* **101**, 1–57 (2009).
49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27. <https://doi.org/10.1186/s40168-017-0237-y> (2017).
51. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* **12**, 59–60 (2015).
52. Huson, D. H. *et al.* MEGAN Community Edition—interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* **12**(6), e1004957. <https://doi.org/10.1371/journal.pcbi.1004957> (2016).
53. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2015).
54. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: An effective distance metric for microbial community comparison. *ISME J.* **5**, 169–172 (2011).
55. Cutler, D. R. *et al.* Random forests for classification in ecology. *Ecology* **88**, 2783–2792 (2007).
56. Dhariwal, A. *et al.* MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* **45**, W180–W188 (2017).
57. Bowman, J. S. & Ducklow, H. W. Microbial communities can be described by metabolic structure: A general framework and application to a seasonally variable, depth-stratified microbial community from the Coastal West Antarctic Peninsula. *PLoS ONE* **10**(8), e0135868. <https://doi.org/10.1371/journal.pone.0135868> (2015).
58. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
59. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform.* **11**, 538. <https://doi.org/10.1186/1471-2105-11-538> (2010).
60. Evans, S. N. & Matsen, F. A. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 569–592 (2012).
61. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. <https://doi.org/10.1186/s13059-014-0550-8> (2014).
62. Hammer, Ø., Harper, D. A. T. & Ryan, P. D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontol. Electron.* **4**, 9 (2001).
63. Mitra, S., Gilbert, J. A., Field, D. & Huson, D. H. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J.* **4**, 1236–1242 (2010).
64. Anderson, M. J. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**, 245–253 (2006).
65. McArdle, B. H. & Anderson, M. J. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **82**, 290–297 (2001).
66. Legendre, P. & Anderson, M. J. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.* **69**, 1–24 (1999).
67. Anderson, M. J. & Robinson, J. Generalized discriminant analysis based on distances. *Aust. N. Zeal. J. Stat.* **45**, 301–318 (2003).
68. Neal, A. L. *et al.* Land-use influences phosphatase gene microdiversity in soils. *Environ. Microbiol.* **19**, 2740–2753 (2017).

## Acknowledgements

The authors are grateful to the DeepSoil consortium (including Illumina, Great Abington, U.K., Penny Hirsch at Rothamsted Research and Janet Jansson at the Pacific Northwest National Laboratory, WA, USA) for assistance in generating metagenomic data from the soils described in this manuscript, and to David Hughes and Ajit Singh, Department of Computational and Analytical Sciences at Rothamsted for data and software code curation associated with metagenomic analyses; we are also grateful for the comments of an anonymous reviewer who encouraged us to present a more coherent text structure. This research was supported by the United Kingdom's Biotechnology and Biological Science Research Council (BBSRC)-funded *Soil to Nutrition* strategic program (BBS/E/C/00010310) and jointly by the Natural Environment Research Council and BBSRC as part of the *Achieving Sustainable Agricultural Systems* research program (NE/N018125/1 LTS-M). Access to the Highfield Ley-Arable experiment is supported by the UK's Long-Term Experiment National Capability funded by BBSRC (BBS/E/C/00010300).

## Author contributions

All authors played a part in conceptualizing the study and methodology development; A.B.-L. generated and analyzed the X-ray computed tomography of soil pore space under supervision of K.R. and S.J.M.; I.C. manages the Highfield Conversion experiment and also performed nucleic acid extractions, library preparation for sequencing and preliminary metagenomic analysis; A.L.N. performed analyses of metagenomic data and prepared the original draft manuscript and was responsible for data presentation; X.Z. was responsible for programming and software development associated with modeling of hydraulic conductivity and soil anoxic volume, implementation of computer code and supporting algorithms; K.C. was responsible for implementation of the RothC-26.3 computer code and supporting algorithms associated with modeling organic carbon in the soils; all authors participated in the development of the final version of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-67631-0>.

**Correspondence** and requests for materials should be addressed to A.L.N.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020