OIKOS

Research

Unifying community detection across scales from genomes to landscapes

Stephanie F. Hudon, Andrii Zaiats, Anna Roser, Anand Roopsind, Cristina Barber, Brecken C. Robb, Britt A. Pendleton, Meghan J. Camp, Patrick E. Clark, Merry M. Davidson, Jonas Frankel-Bricker, Marcella Fremgen-Tarantino, Jennifer Sorensen Forbey, Eric J. Hayden, Lora A. Richards, Olivia K. Rodriguez and T. Trevor Caughlin

S. F. Hudon (https://orcid.org/0000-0001-5255-9381) ✉ (stephaniehudon@boisestate.edu), A. Zaiats (https://orcid.org/0000-0001-8978-4152), A. Roser, A. Roopsind, C. Barber, B. C. Robb, B. A. Pendleton, M. M. Davidson, J. Frankel-Bricker, M. Fremgen-Tarantino, J. S. Forbey, E. J. Hayden, O. K. Rodriguez and T. T. Caughlin, Boise State Univ., Boise, ID, USA. AR also at: Center for Natural Climate Solutions, Conservation International, Arlington, VA, USA. – M. J. Camp, Washington State Univ., Pullman, WA, USA. – L. A. Richards, Univ. of Nevada, Reno, Reno, NV, USA. – P. E. Clark, Northwest Watershed Research Center, USDA Agricultural Research Service, Boise, ID, USA.

Oikos
00: 1–13, 2021
doi: 10.1111/oik.08393

Subject Editor and
Editor-in-Chief: Dries Bonte
Accepted 5 March 2021

Biodiversity science encompasses multiple disciplines and biological scales from molecules to landscapes. Nevertheless, biodiversity data are often analyzed separately with discipline-specific methodologies, constraining resulting inferences to a single scale. To overcome this, we present a topic modeling framework to analyze community composition in cross-disciplinary datasets, including those generated from metagenomics, metabolomics, field ecology and remote sensing. Using topic models, we demonstrate how community detection in different datasets can inform the conservation of interacting plants and herbivores. We show how topic models can identify members of molecular, organismal and landscape-level communities that relate to wildlife health, from gut microbes to forage quality. We conclude with a future vision for how topic modeling can be used to design cross-scale studies that promote a holistic approach to detect, monitor and manage biodiversity.

Keywords: biodiversity, Latent Dirichlet Allocation, metabolomics, metagenomics, sagebrush, wildlife conservation

Introduction

Understanding biodiversity will require crossing disciplinary boundaries to link biological organization across scales. Early efforts to quantify biodiversity focused on the organismal scale of plants and animals (Simpson 1949). However, modern biodiversity research encompasses molecular scales, as well as scales beyond individual organisms, including biotic and abiotic features within landscapes, regions and continents. Studying biodiversity at microscopic and macrosystem scales has led to insights with relevance for human health (Mohajeri et al. 2018), global sustainability (Bennett et al. 2015) and wildlife conservation (Trevelline et al. 2019). As recognition of the importance of biodiversity has increased, so have methods for analyzing biodiversity, from

NORDIC SOCIETY OIKOS

www.oikosjournal.org

molecular approaches such as next generation sequencing for genomic data to airborne sensors that can measure large-scale landscape features. These discipline-specific methods limit analysis of biodiversity patterns that may be nested within or interact among scales. The lack of interdisciplinary cohesion in biodiversity studies with different terminology and varying scales of interest is a barrier to understanding biological processes vital to biodiversity conservation. One step toward overcoming this lack of cohesion is to identify patterns in data across disciplines that can then be discussed with a common language (Mosher et al. 2020).

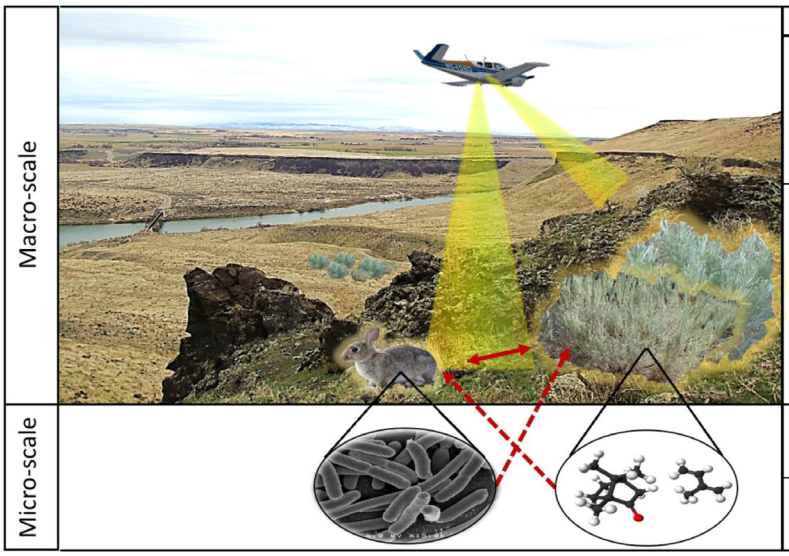
Community organization is a unifying pattern in biodiversity data across scales. Ecological communities of species that occupy the same space at the same time are a major focus of empirical and theoretical work in community ecology (Vellend 2010). More recently, ideas from community ecology have been extended to other biological disciplines, including detecting co-occurring microbes (Nemergut et al. 2013), functional genes (Burke et al. 2011) and landscape features (Räsänen et al. 2016). Despite differences in community assembly and study techniques, co-occurring features can reveal ecologically meaningful patterns in metabolites, microbial taxa, plant and animal species, and spectral bands from land surface reflectance. Community detection across multiple scales opens the possibility to study cross-scale interrelationships. For example, metabolite features within plants influence the microbial organisms of individual herbivores (Kohl et al. 2014) and reflectance features of plants can predict herbivore population dynamics across landscapes (Fauchald et al. 2017). The overarching importance of communities in ecology and evolution has led to a multitude of methods to detect communities in ecological data (Legendre and Legendre 2012).

A common challenge of detecting communities is mixed membership, when single features and single samples can potentially be assigned to more than one community. The degree of mixed membership in communities depends on whether features arrange themselves as discrete members of different communities (Clements 1936), or as fluid entities with membership in multiple communities (Gleason 1926). Within cellular units, biomolecular processes such as mutation and differential gene expression can promote mixing of metabolic and genetic features. Within a landscape, processes such as dispersal and anthropogenic disturbances lead to mixing of species and obscure boundaries between communities (Lortie et al. 2004). Another challenge for community detection is the tradeoff between sampling extent and resolution, a methodological choice that can affect community membership results. For example, in metagenomics, the benefit of deep sequencing must be weighed against the cost of generating more reads. Similarly, in remote sensing, larger pixels capture more surface area than smaller pixels, but the higher resolution of smaller pixels improves detection of land cover features (Kennedy et al. 2009).

Altogether, mixed membership of features within sampling units and communities is common. Nevertheless, many analytical methods, such as clustering and ordination techniques (McCune et al. 2002), lack a probabilistic interpretation of community membership, which limits the potential for model transferability and prediction in novel environments. One solution is topic modeling of community membership, which has revolutionized multivariate analysis by enabling a single feature or sampling unit to belong to multiple communities. The term 'topic model' arises from text mining, where models are used to assign co-occurring words in documents to underlying subjects ('topics'; Barde and Bainwad 2017). Topics are referred to as 'latent' because they are not known before hand and must be inferred from the data. Latent Dirichlet Allocation (LDA) is a topic modeling approach that can identify communities of features, while allowing for mixed membership of features across communities as well as mixtures of communities within individual sampling units (Valle et al. 2014). LDA was first developed in population genetics, motivated by the need to use genotypes as features that could group individuals into populations, while allowing for admixture (i.e. the presence of several distinct genotypes/genomes in a single population) (Pritchard et al. 2000). Several years later, LDA was independently developed as a tool to uncover latent structure in text data and broadly adopted by the machine learning community (Blei et al. 2003). Since then, LDA has resulted in transformative biological insights across disciplines including annotating unknown chemicals in fermented beverages (van der Hooft et al. 2016), characterizing functional roles of gene regions (Chen et al. 2010) and identifying communities of bird species in citizen science data (Valle et al. 2018). Beyond single discipline applications, we contend that topic modeling has unrealized potential to unify biodiversity science across scales.

Here we demonstrate how to apply LDA across multiple scales to inform conservation of herbivores. We focus on the sagebrush steppe ecosystem that once covered ~ 1 million km^2 of land in the western United States but is increasingly threatened by wildfires and invasive species (Requena-Mullor et al. 2019). Sagebrush (*Artemisia* spp.) are the dominant plant species in these ecosystems and are critical for two sagebrush obligate species: the pygmy rabbit *Brachylagus idahoensis* and the greater sage-grouse (*Centrocercus urophasianus*, hereafter sage-grouse). Both herbivores are considered species of conservation concern across the Intermountain West. However, efforts to conserve and reintroduce populations of pygmy rabbits and sage-grouse have had mixed success due to problems that range from lack of consideration of local diet adaptations (Oh et al. 2019) to ecosystem fragmentation (Cross et al. 2018).

Management of threatened species, including pygmy rabbits and sage-grouse, will benefit from a deeper and more functional understanding of the biological communities that impact individual health and population dynamics. We use four case studies from the sagebrush steppe ecosystem to show how LDA can assess community mixtures of 1) metabolites from leaf material of individual sagebrush plants, 2) microbial species from fecal pellets of pygmy rabbits, 3) plant species from field plots within sagebrush patches and 4) spectra from pixels across a sagebrush landscape (Fig. 1). At



Feature	Sampling unit	Spatial extent
Reflectance in spectral bands	Pixels	Landscape
Plant taxa	Field plots	Habitat
Chemicals	Leaf material	Individual plants
Microbial taxa	Fecal pellet	Individual animals

Figure 1. Illustration of how communities are measured in sampling units that span micro- and macro-scales. In the sagebrush steppe ecosystem, these communities are linked across scales. Microbial taxa in fecal pellets from individual herbivores interact with chemical features in leaf material when herbivores consume individual plants. Metabolite features in leaf material consumed by herbivores are dependent on the abundance of individual plant taxa detected within field plots. The distribution of plant taxa can be detected with spectral bands in pixels of aerial imagery obtained remotely within landscapes.

the micro-scale, microbial features in herbivores (Kohl et al. 2016) interact with metabolite concentrations in the gut after herbivores consume sagebrush (White et al. 1982). At the macro-scale, features of herbivores and sagebrush are dependent on metabolite concentrations of plant taxa within habitat patches (Frye et al. 2013, Ulappa et al. 2014) and those plant taxa can be detected with aerial remote sensing platforms (Olsoy et al. 2020). Ultimately, the community patterns that emerge from analyzing features across scales could deepen our understanding of plant–herbivore interactions and identify molecular, organismal, and landscape targets for management in changing landscapes.

Overview of Latent Dirichlet Allocation

The overall objective of LDA is to identify latent communities of co-occurring features in data. Communities are latent because they are not directly observed in relative abundance data; instead, communities represent a hidden structure that can be uncovered with statistical modeling. For example, consider a book as a sampling unit, filled with words as features. The co-occurrence of particular words (e.g. 'spaceship,' 'alien,' 'planet') indicate that book is likely to represent a particular topic, or community (e.g. science fiction). LDA assigns both individual words and individual books to latent communities. LDA factorizes relative abundance data into two matrices, one representing membership of communities in sampling units and the other representing feature membership in communities (Box 1). Input relative abundance data can either represent binary (zero or one) or multinomial (count) data (Valle et al. 2018). The number of communities in LDA can either be set in advance or estimated from the data (Albuquerque et al. 2019) LDA output includes probabilities of community membership for each sampling unit and feature. As a generative model (Box 2), LDA can account for missing data, predict relative abundance at new sites and represent uncertainty in community membership. For further technical details on LDA, we refer readers to recent reviews by Sankaran and Holmes (2017) and Valle et al. (2014).

Applying Latent Dirichlet Allocation across data sets

We analyzed each of our datasets with LDA models in RStudio (ver. 3.4.4) to detect communities within sampling units. Our models applied a Bayesian framework from the 'Rlda' package (Albuquerque et al. 2019). We used a binomial version of the LDA (Valle et al. 2018) to detect communities from occurrence data on metabolites and spectral reflectance, and a multinomial parametrization (Blei et al. 2003, Valle et al. 2014) for the analyses of count data on microbial taxa and leaf area index. We provide detailed methods for each case study in the Supporting information.

Case study 1. Reflectance of spectral bands within pixels at the landscape scale

Our first case study uses LDA to assess patterns in spectral data obtained from remotely sensed images. In this case study, communities represent co-occurring wavelengths of spectral reflectance. Understanding impacts of global change on sagebrush ecosystems will require measurements over areas larger than that provided by field plots alone. We used a binomial version of LDA to detect communities or patterns

Box 1. Modeling framework

Latent Dirichlet Allocation (LDA) is a statistical model for identifying latent (unobserved) commeunities. Input data structure for LDA consists of an abundance matrix organized with sampling units (m) as rows and features (n) as columns. Sampling units contain measurements of features. Features measured in sampling units could include species abundance, chemical concentration or sets of reads from DNA sequencing. Count data on abundance of features in each sample unit is modeled using a multinomial version of LDA:

Abundance of feature n in sample unit m	$Y_{m,n} \sim \text{Multinomial}\left(\phi_{z[m,n]}, z_{m,n}\right)$
Abundance of community z in sample unit m	$z_{m,n} \sim \text{Multinomial}\left(\theta_m, S_{\max}\right)$
Membership of features in communities	$\phi_k \sim \text{Dirichlet}\left(\beta\right)$
Membership of communities in sample units	$\theta_m \sim \text{Dirichlet}\left(\gamma\right)$

where ($Y_{m,n}$) represents the observed abundance of n-th feature in m-th sample unit. Each entry in the data matrix is assigned to a community type, which is estimated as a latent variable $z_{m,n}$ and depends on the distribution of communities across sample unit m (θ_m), and the maximum possible abundance of features in a site ($S_{\max}$). The Dirichlet distribution is a probability distribution for proportional data (Douma and Weedon 2019) that enables mixed membership. The θ and ϕ parameter matrices reveal the probability that sample units and features, respectively, belong to k communities. The hyperparameters, β and γ, represent the degree of mixed membership in the Dirichlet distribution and are often specified to initiate the model (Supporting information). An alternate parameterization of LDA for binary data assumes that observations are drawn from a binomial distribution but is otherwise similar to the multinomial model (Valle et al. 2018). LDA can be fit with frequentist maximum likelihood estimation or with Bayesian approaches, such as Gibbs sampling (Hornik and Grün 2011). See the Supporting Information for more details, including R scripts with examples of LDA fit to multiple types of data.

in reflectance from aerial imagery of a sagebrush steppe landscape (data available from National Ecological Observatory Network 2019, Fig. 2a).

Using LDA, we were able to detect ecological patterns related to changing composition of plants. We identified two communities of spectral features characteristic of vegetation (Fig. 2a). Based on visual interpretation of concurrently collected red–green–blue (RGB) imagery, the first community (Community 1) represents juniper trees *Juniperus* spp. while the second community is associated with low-growing shrubs (Fig. 2b). Juniper range expansion threatens wildlife species (Severson et al. 2017). The patchiness of Community 1 suggests fine-scale variation in juniper cover during the early stages of woody encroachment (Fig. 2b). The more uniform representation of Community 2 (Fig. 2b, right panel) is attributable to a dominant but sparse canopy of shrubs documented in ground observations. Our results demonstrate how high-resolution hyperspectral data can detect and map juniper encroachment in sagebrush steppe. Ultimately, patterns of remotely-sensed, spectral features could be used to monitor ecological change in landscapes where herbivores forage (Frye et al. 2013, Ulappa et al. 2014).

Case study 2. Plants within field plots at the habitat scale

Our second case study uses LDA to detect communities of plant species in field plots using measurements of leaf–area–index (LAI). LAI is the relative size of one leaf over a unit of ground surface (Ewert 2004). Measurements of LAI in drylands relate to food availability for herbivores (Olsoy et al. 2015). We quantified LAI in field plots within a Wyoming big sagebrush habitat (*Artemisia tridentata* ssp. *wyomingensis*) (Fig. 3).

Results from applying LDA suggest that this habitat type contained six plant species communities (Fig. 3). We report on the composition of three of these communities due to their ecological significance. Community 1 and 3 were dominated by the presence of Wyoming big sagebrush and a native bunchgrass Sandberg bluegrass *Poa secunda*, respectively. Wyoming big sagebrush and Sandberg bluegrass are of particular importance because their presence indicates habitats favorable for herbivores (Beck et al. 2009). Community 6 was dominated by cheatgrass *Bromus tectorum*, an invasive annual (Fig. 3a) that indicates degraded ecosystems less suitable for herbivores (Steenvoorden et al. 2019).

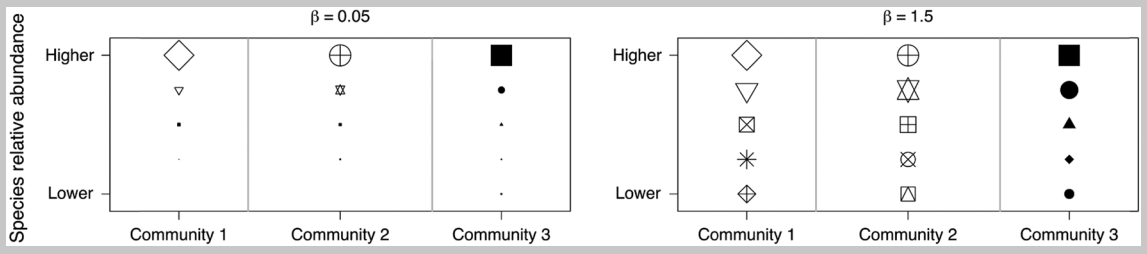
Our results show that communities in this habitat are dominated by Wyoming big sagebrush and Sandberg bluegrass, with low probability of the invasive cheatgrass community (Fig. 3b). These results are visible at the level of a single sampling unit (1 m^2, Fig. 3c). Our leaf-level analysis could be used to quantify fine-scale suitability for particular wildlife species. For example, plots monitored after fires with high probability of Wyoming big sagebrush and low probability of cheatgrass might indicate successful post-fire restoration (Baker 2006), including the regeneration of suitable forage for herbivores (Beck et al. 2012).

Box 2. Latent Dirichlet Allocation as a generative model for biodiversity

LDA belongs to a broad class of models known as generative models that define joint probabilities for latent variables and observed features (Bernardo et al. 2007), including a data generating mechanism for observed data based on probability distributions. In contrast, commonly used models for multivariate data in ecology (e.g. ordination; Legendre and Legendre 2012) tend to be non-generative, describing patterns in community data without a probabilistic explanation. A key advantage of generative models is the ability to simulate data that is consistent with observed data. Data simulation from generative models is increasingly considered best practice for statistical analyses (Gabry et al. 2019). Comparing observed data to simulations improves overall understanding of data, can identify potential pathologies in statistical models, and assists the design of more efficient sampling schemes.

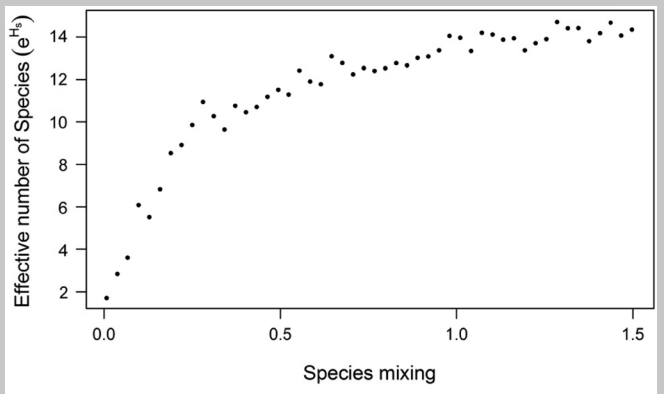
Generative models can also provide a link to ecological theory (Harris et al. 2017). In community ecology, long-standing debates on biodiversity metrics for alpha, beta and gamma diversity have relied upon simulations for understanding these metrics (Legendre et al. 2005, Baselga 2010, Veech and Crist 2010). However, simulation models used to explore general properties of biodiversity metrics are often not the same models used to analyze observed data; LDA presents an opportunity to better integrate simulation experiments and statistical models.

To demonstrate the potential for LDA to simulate biodiversity data, we generated 50 fake datasets from the multinomial model described in Box 1. For each dataset, we varied the hyperparameter for membership of species (i.e. features) in latent communities (β in Box 1) from β = 0.01–1.5 A heuristic explanation for this hyperparameter is the degree of species mixing within communities. A lower β value corresponds to a minimal mixing of species in communities and results in a few high-abundance species. In contrast, with increasing β, there is a high degree of species mixing resulting in higher probability of a more uniform distribution of species abundances. Note, the datasets were generated with the site mixing hyperparameter (γ in Box 1) set to zero, resulting in no mixing across communities (high species turnover). In practice, however, the observed patterns of species abundances within and across sites jointly depend on both β and γ hyperparameters.



The figure above shows two simulated datasets on species relative abundance, generated by LDA. The left panel shows three communities from a simulated dataset with low species mixing in and the right panel shows three communities from a simulated dataset with high mixing. Each icon represents one species, with the relative size of each icon indicative of relative abundance of that species within a community.

After simulating datasets using LDA, we then calculated alpha diversity, the variation in species composition within sites, for each of the simulated communities using the expected number of species ($e^{H_{\text{Shannon}}}$). Results from this simulation experiment demonstrate a strong relationship between a commonly-used metric for alpha diversity and LDA. These results demonstrate how using LDA as a generative model could provide a way to better understand fundamental concepts in community ecology by linking statistical models for observed data with simulation experiments.



The hyperparameter for species mixing in LDA (β in Box 1) provides a generative model for biodiversity metrics, including alpha diversity. Each dot represents one simulation, with a different hyperparameter for species mixing.

Case study 3. Metabolites within leaves at the plant scale

Our third case study uses LDA to detect patterns in metabolite features, specifically volatile monoterpenes, between two sagebrush taxa. While several herbivores rely on sagebrush as forage year-round, the volatile monoterpene features of this plant influence selection by herbivores at the species, patch and plant scale (Frye et al. 2013). Although there are known concentration-dependent consequences of individual monoterpenes, the unique mixtures of metabolites in plants may better explain intake by herbivores (Nobler et al. 2019). Moreover,

foraging herbivores consume mixtures of metabolites, not individual metabolites. Approaches that focus on the presence or concentration of a specific metabolite likely miss differences in the relative ratios of compounds that better determine diet selection by herbivores and predict interactions with the microbial features (e.g. case study 4 below) in herbivore guts. In this case study, communities represent 'chemical bouquets,' or groups of secondary metabolites.

We found that LDA detected communities of monoterpenes that were relevant to herbivore diet selection in two different sagebrush taxa (Fig. 4). Specifically, there were three communities that contained unique individual monoterpenes known to predict foraging by herbivores. While the identity of several monoterpenes quantified are unknown (Unk indicate unknown compounds), the suite of monoterpenes in Community 1, 3 and 4 were present across many of the plant samples (Fig. 4b). At the individual sampling unit (plant) level, three-tip sagebrush *Artemisia tripartita* had a high probability of Community 4, whereas Wyoming big sagebrush was dominated by Communities 1 and 3 (Fig. 4c). Concentrations of Unk 21.0 (dominant in Community 4) and Unk 21.5 (dominant in Community 3) have previously been found to predict diet selection by free-ranging sage-grouse (Fremgen-Tarantino et al. 2020) and β-pinene (dominant in Community 1) was avoided by captive mountain cottontails *Sylvilagus nuttallii* (Nobler et al. 2019). Our results demonstrate how LDA can reveal communities of metabolite features that predict foraging decisions by herbivores. A potential application of LDA could be to improve post-fire restoration by reseeding with plants that have similar chemical community profiles to plants consumed and preferred by threatened herbivores.

Case study 4. Microbial taxa within fecal pellets at the herbivore scale

Chemical communities in herbivore forage, including plants in the wild and artificial pellets in captivity, can modify microbial species composition within animal guts (Kohl et al. 2014, Sandifer et al. 2015, Mohajeri et al. 2018). Our fourth case study uses LDA to identify patterns in microbial taxonomic features detected in fecal pellets of pygmy rabbits over time. For this case study, communities represent groups of co-occurring microbial taxa. Specifically,

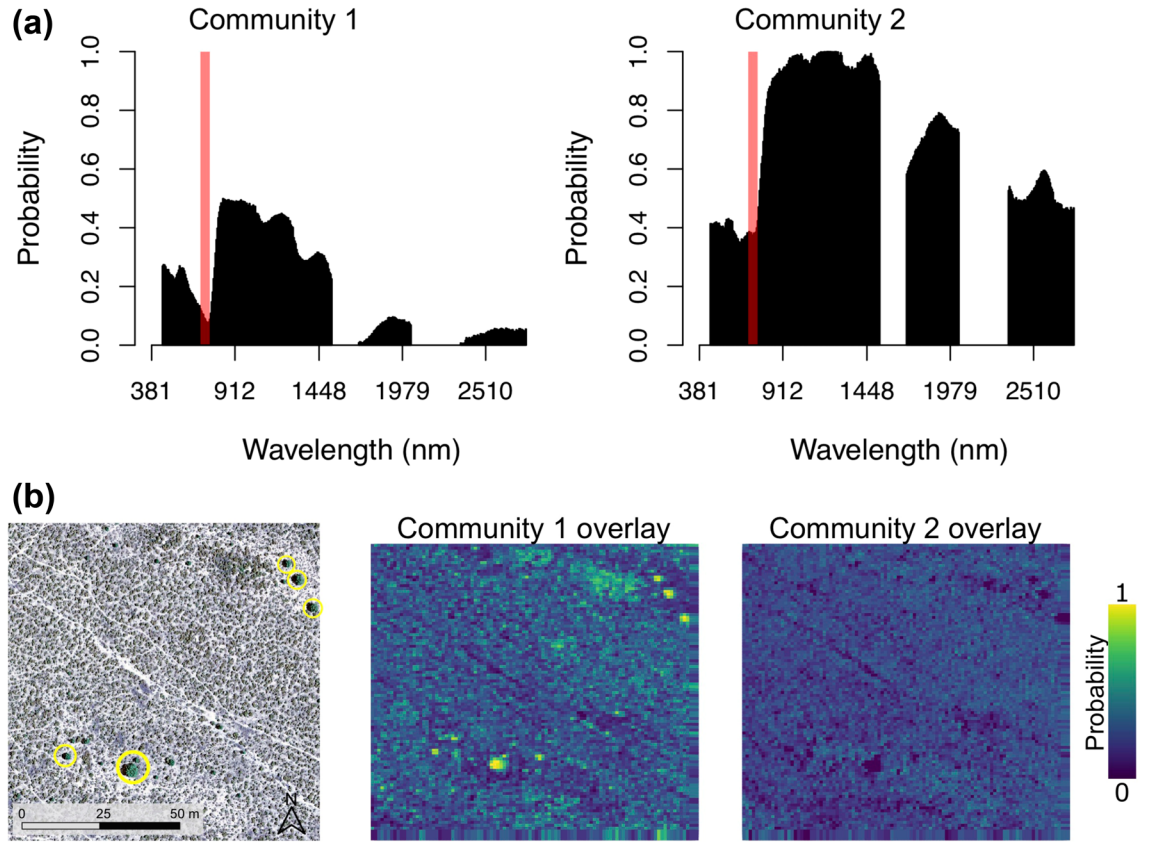


Figure 2. LDA applied to a subset of hyperspectral (1 m^2 resolution) orthomosaic from a sagebrush steppe ecosystem (Onaqui, UT, USA). (a) The probability of each wavelength of reflected light (nm) belonging to two communities. The rapid change in reflectance between 690 nm and 750 nm (the 'red edge') is representative of changes in plant photosynthetic activity. (b) Red, green, blue (RGB) image of the area with encroaching juniper trees *Juniperus* spp. circled in yellow (left image) with community 1 overlay (middle) and community 2 overlay (right) in the same area outlining a high probability that junipers belong to Community 1. Colors approaching yellow in the community overlays indicate higher probability of pixel membership from a particular spectral feature.

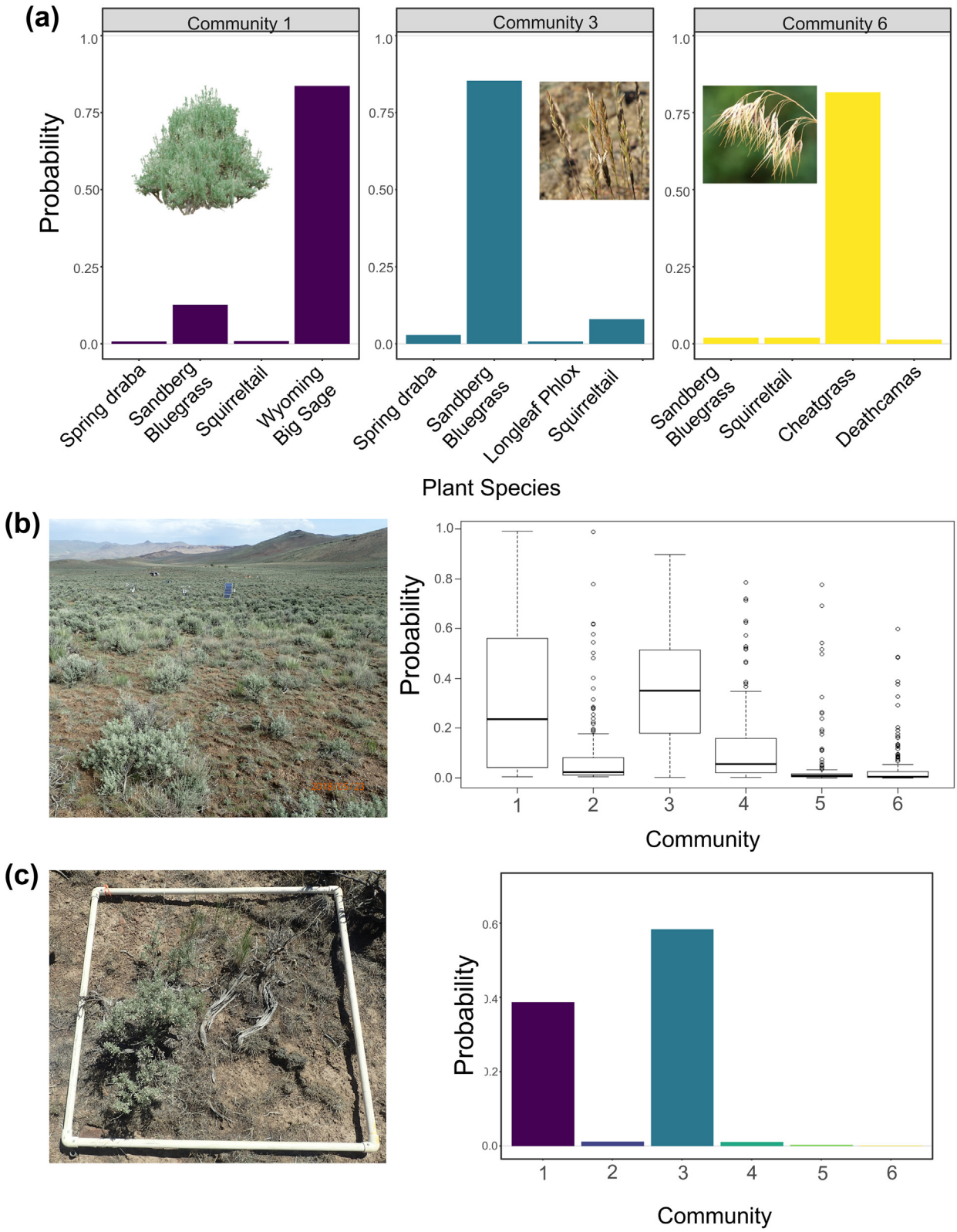


Figure 3. The results of LDA analysis on leaf area index (LAI) in a Wyoming big sagebrush habitat. (a) The probability of plant species occurring within three communities with an image of the dominant species, *Artemisia tridentata* ssp. *wyomingensis*, in inset. (b) A landscape level photo (left) and the probability of presence of the six most common communities within the habitat sampling units (right). (c) A representative photo of a single 1m^2 field plot sampling unit (left) and the probability of presence of each community within a single plot (right).

we analyzed how the taxa of the fecal microbiome from this obligate sagebrush herbivore would change as they were transitioned from a natural diet containing Wyoming big sagebrush to a captive diet, containing commercial rabbit food, over a seven-day period. Fecal samples from the rabbits on day 1 (sagebrush diet) and day 10 (captive diet) were collected and analyzed using shotgun metagenomics. We used LDA to identify communities of bacteria at the genus level (Fig. 5). The anaerobes, *Clostridium* and *Bacteroides*, were common features of these bacterial communities (Fig. 5a). Communities 3 and 8 show the highest probability of being found within all fecal samples (Fig. 5b). Community 3 was dominated by *Bacteroides* and had a higher probability of being present when the rabbits were on a natural diet, whereas Community 8, which was dominated by *Clostridium* species, was more prevalent after a week of transitioning to a captive diet (Fig. 5c). Some *Clostridium* species are associated with enteritis (inflammation of the small intestine) and increased mortality in wild and captive animals (Paul and Friend 2019) whereas other *Clostridium* species may improve animal health (Liu et al. 2019). These preliminary results suggest that LDA can be used to monitor changes in bacterial communities associated with dietary shifts, and potential health, in sagebrush-dependent herbivores. Because microbial function is largely driven by communities, rather than individual species, community-level analyses (e.g. LDA) are crucial for identifying physiologically-relevant changes in herbivore metagenomes.

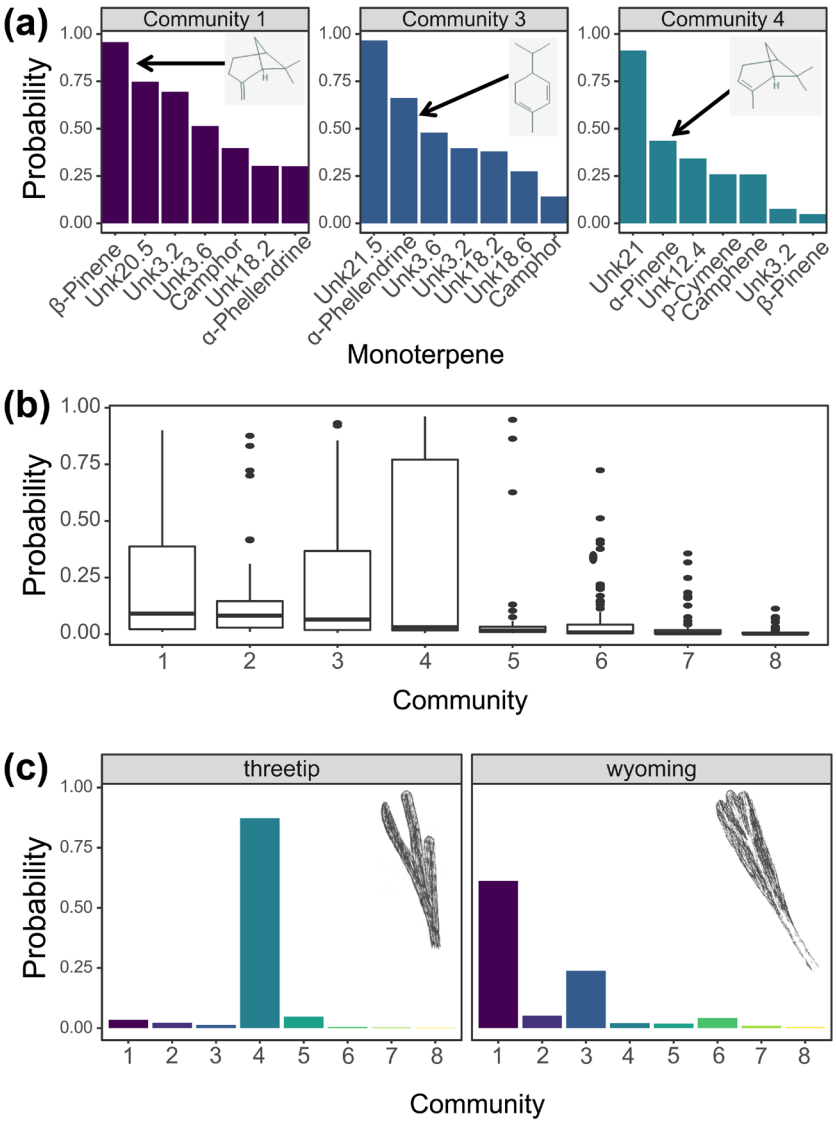


Figure 4. Results of LDA analysis of monoterpenes from leaves of sagebrush plants consumed by herbivores. (a) Probability of monoterpenes occurring within three metabolite communities with an image of the molecular structure of the dominate known monoterpene in inset. (b) Probability of the eight most common metabolite communities across all sagebrush samples. (c) Probability of the eight metabolite communities occurring within an individual three-tip *Artemisia tripartita* and a Wyoming big sagebrush *Artemisia tridentata* ssp. *wyomingensis* sampling unit with an image of the leaf morphology of each species in inset.

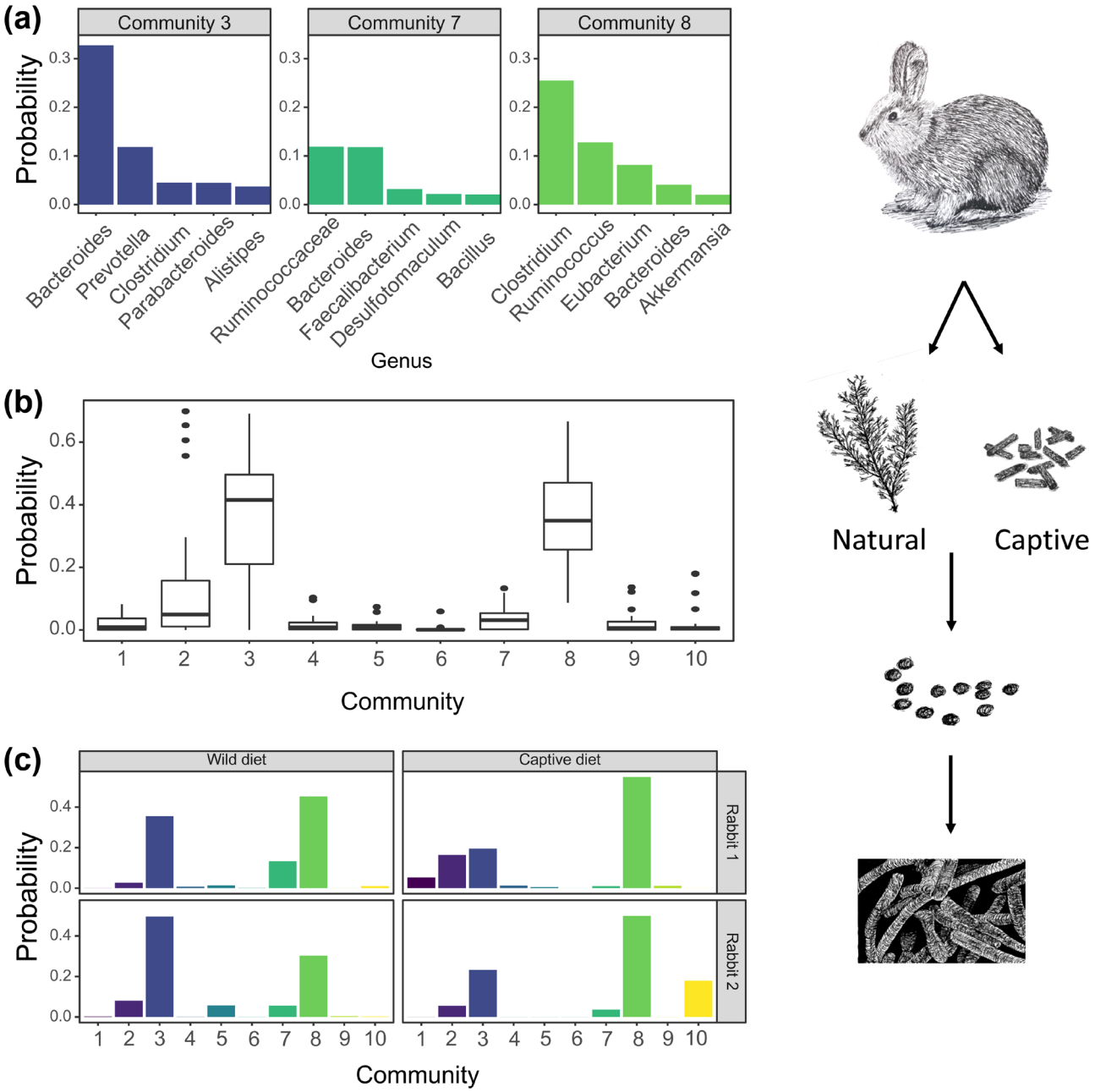


Figure 5. LDA analysis using genus level taxonomy counts from metagenomics of fecal samples collected from pygmy rabbits *Brachylagus idahoensis* (shown top left). (a) Probability of microbial features within the three most prevalent communities detected in fecal samples, each dominated by different microbial taxa. (b) Probability of the ten identified microbial communities within fecal samples from the pygmy rabbit (n=22). (c) Probability of the ten microbial communities in fecal samples from the pygmy rabbit sampling units consuming a natural diet (primarily Wyoming big sagebrush *Artemisia tridentata* ssp. *wyomingensis*) and after seven days on an artificial pellet diet in captivity.

Discussion

As biodiversity science grows to encompass scales from molecular to continental, the need for integrative approaches has increased as well. We have demonstrated the potential for community detection to unite patterns of biodiversity across disciplines. Latent Dirichlet Allocation, a topic model that can represent mixed membership of features, enabled us to quantify communities across molecular, organismal, and landscape scales. Our results have potential relevance for conservation of threatened herbivores in the sagebrush steppe ecosystem, including the pygmy rabbit. At the landscape scale, LDA detected juniper encroachment, a driver of habitat degradation in sagebrush steppe, from aerial remote sensing data. At the plant scale, LDA enabled discrimination between plant species assemblages, with relevance for habitat structure, including the availability of quality forage for herbivores and the presence of invasive species. At the molecular scale,

LDA identified mixtures of secondary metabolites that can differentiate plant species and predict diet selection by herbivores. At the microbial scale, LDA quantified shifts in bacterial communities that are predictive of disease and survival, and respond to diet transitions of herbivores. Co-analyzing datasets with LDA improved comprehension of biodiversity across scales for members of our multidisciplinary research team, leading us to develop a more holistic view of plant-herbivore ecology. Common models for disparate datasets, including LDA, will enable collaborative studies that can better inform cross-scale strategies for conservation.

One realization that emerged from co-analyzing our data is the importance of herbivore gut microbiomes for uniting scales. We argue that studying gut microbiomes has great potential to develop a more complete understanding of herbivore ecology, particularly if multiple scales are incorporated into analyses. Herbivores, such as the pygmy rabbits in our study, make foraging decisions at individual metabolite, leaf, plant and landscape scales (Ulappa et al. 2014, Nobler et al. 2019). In turn, foraging herbivores can influence patterns of habitat structure and plant species composition (Eldridge et al. 2016). Over long periods of time, we would expect that gut microbes mediate feedback loops between plants and herbivores, with ecological and evolutionary implications (Ley et al. 2008, Kohl and Dearing 2016). In a practical sense, the gut microbiome links these disparate scales and represents the net sum of forage availability and quality across landscapes (Fig. 2), habitat patches (Fig. 3) and within plants (Fig. 4). Herbivore foraging has wide-ranging consequences for above-ground (Frye et al. 2013, Ulappa et al. 2014, Fremgen-Tarantino et al. 2020) and below-ground (Chomel et al. 2016) ecological processes, therefore a more holistic understanding of co-occurring plant, metabolite and microbial communities in the guts of herbivores is needed. Topic models, such as LDA, present an opportunity to describe microbial community structure (Chen et al. 2012). The development of analytical tools that integrate hierarchies of scale and complex network structure will further enable researchers to uncover how microbial communities might interact with communities at other scales, from molecular to landscape.

Given the importance of the gut microbiome, we envision designing future studies to collect data from multiple biological units at the same time and place with a focus around fecal collections. Data collection focused around herbivore fecal pellets could involve sampling feces from herbivores for metagenomic and metabolite analyses while simultaneously collecting leaf tissue from plants browsed by herbivores for metabolite content (parent and detoxification products), and mapping the GPS location where pellets and plant samples are collected. Subsequently, research teams could assess how communities of microbes, plant-derived metabolites and plant species detected in feces are influenced by variation in plant species availability at the scale of foraging plots. Remote sensing data, such as hyperspectral aerial images, could then be applied to detect temporal and spatial variation in the composition of plant species and foliar chemistry across the landscape (Fine et al. 2021). This type of data collection will require extensive interdisciplinary coordination, but will lead to a more connected understanding of coupled biodiversity among scales. Long-term ecological research sites, such as the NEON network, provide a valuable starting point for this type of study where collection and analyses of herbivore metagenomics and metabolites from plants could add substantial value to existing data on plant diversity and soil microbial communities. In the context of planning field studies, LDA could be applied as a generative model to simulate data and estimate appropriate sample sizes for statistical inference (Box 2).

LDA has distinct advantages over other clustering techniques. A key strength of LDA is its probabilistic nature, enabling detection of novel communities. For example, when analyzing metabolomic data sets from tandem mass spectrometry analysis, LDA identified relevant substructures from co-occurrence of mass fragments and neutral losses in 70% of spectra analyzed, in contrast to other clustering techniques that only found hits for 25 and 6% of spectra respectively (van der Hooft 2016). In some cases, LDA may also improve classification accuracy. For example, LDA-based methods outperformed simple harmonization methods based on semantic affinity scores for identifying latent land cover communities from different source maps (Li et al. 2020). More broadly, continuous representations of community membership from LDA enable explanatory analyses that would not be possible with clustering methods that assume discrete membership (Knott et al. 2020). In addition, LDA is a generative model that can simulate data to improve links between ecological theory and statistical analyses (Box 2). Altogether, probabilistic topic models for community detection are poised to generate novel insights from biodiversity data.

Continued advances in community detection across scales

While our case studies of community membership represent separate analyses, topic modeling is well-poised to address long-standing questions of whether different taxonomic units co-occur in space and time (Heino 2010). Correlated topic models represent covariance between communities (Blei and Lafferty 2007), using mathematical relationships that are similar to existing models for co-occurrence between species in joint species distribution models (Pollock et al. 2014). Correlated topic models could quantify whether species from different trophic levels co-occur in space (often referred to as spatial concordance or cross-taxon congruence; Pearson and Carroll 1999, Su et al. 2004). Analyses of cross-taxon congruence typically involves a two-step process, first to identify latent communities and second to analyze correlations between them (Heino 2010). Correlated topic models present an opportunity to combine both of these steps into a single statistical model. The flexibility of probabilistic models, such as LDA, could prove invaluable for extending questions of cross-taxon congruence beyond species to secondary

metabolites, genes, landscape features and other levels of biological organization.

Biodiversity data commonly includes features and communities that change over time and in response to environmental covariates. Newly developing topic modeling approaches could improve our capacity for inference on drivers of community membership. Dynamic topic models are currently used in text mining to account for changing community membership (Blei and Lafferty 2006), while dynamic mixture models enable realized proportions of communities to change over time (Wei et al. 2007). In ecology, LDA has recently been applied to interpret how gradual changes in rodent communities over a 40-year period were related to environmental drivers (Christensen et al. 2018). As a statistical approach conceptually related to regression models for proportional data (Douma and Weedon 2019), LDA can provide insights on community dynamics across time and space.

Conclusions

Coordinated studies of community structure across scales will enable researchers to address fundamental questions in ecology and evolution. One such question relates to the long-standing debate over whether biological features, from genes to species assemblages, are organized by neutral processes or deterministic ecological and evolutionary forces (Kreitman 1996, Lynch 2007, Lowe and McPeek 2014). For example, convergent communities of microbes in the soil and guts of herbivores exposed to similar plant metabolite communities across broad biogeographical scales would provide powerful evidence for the role of non-neutral processes. Alternately, random associations between overlain communities could suggest neutral theory as an explanation for observed assemblages. Common models for community structure will provide detailed and cohesive insight into the complex interactions among plants, animals and microbes co-occurring across landscapes. Altogether, we anticipate that interdisciplinary collaboration, facilitated by the common modeling language of LDA, will have payoffs for biodiversity studies that must address complex problems across scales.

Data availability statement

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.8w9ghx3mf> (Zaiats et al. 2021).

Acknowledgements – We thank J. Connelly, D. D. Musil, L. Cross, L. A. Shipley and the GUTT seminar working group in addition to financial support from a NASA grant 80NSCCC17K0738, a Idaho State Board of Education grant IGEM19-002, a Semiconductor Research Corporation grant SRC 2018-SB-2842, Pittman-Robertson 683 funds from the Idaho Department of Fish and Game, Sigma Xi Grants-In-Aid, Bureau of Land Management grant no. L09AC16253, the USDA Agricultural Research Service and National Science Foundation grants IOS-1258217, DEB-1146194, DEB-1146368, OIA-1826801, OIA-1757324, OIA-1738865 and ECCS-1807809. Thanks to James Hudon for drawings in Fig. 4 and 5.

Author contributions

Stephanie F. Hudon and **Andrii Zaiats** contributed equally to this publication. **Stephanie F. Hudon**: Conceptualization (equal); Formal analysis (equal); Project administration (equal); Writing – original draft (equal); Writing – review and editing (equal). **Andrii Zaiats**: Conceptualization (equal); Formal analysis (equal); Project administration (equal); Writing – original draft (equal); Writing – review and editing (equal). **Anna Roser**: Conceptualization (equal); Formal analysis (equal); Writing – original draft (equal); Writing – review and editing (equal). **Anand Roopsind**: Conceptualization (equal); Formal analysis (equal); Writing – original draft (equal). **Cristina Barber**: Conceptualization (equal); Formal analysis (equal); Writing – original draft (equal). **Brecken C. Robb**: Conceptualization (equal); Writing – original draft (equal); Writing – review and editing (equal). **Britt A. Pendleton**: Conceptualization (equal); Writing – original draft (equal); Writing – review and editing (equal). **Meghan J. Camp**: Conceptualization (equal); Writing – original draft (equal). **Patrick E. Clark**: Conceptualization (equal); Writing – original draft (equal). **Merry M. Davidson**: Conceptualization (equal); Writing – original draft (equal). **Jonas Frankel-Bricker**: Conceptualization (equal); Writing – original draft (equal). **Marcella Fremgen-Tarantino**: Conceptualization (equal); Writing – original draft (equal). **Jennifer Forbey**: Conceptualization (equal); Funding acquisition (equal); Writing – original draft (equal). **Eric J. Hayden**: Conceptualization (equal); Funding acquisition (equal); Writing – original draft (equal). **Lora A. Richards**: Conceptualization (equal); Writing – original draft (equal). **Olivia K. Rodriguez**: Conceptualization (equal); Writing – original draft (equal). T. **Trevor Caughlin**: Conceptualization (equal); Formal analysis (equal); Supervision (equal); Writing – original draft (equal); Writing – review and editing (equal).

References

Albuquerque, P. H. M. et al. 2019. Bayesian LDA for mixed-membership clustering analysis: the Rlda package. – Knowl.-Based Syst. 163: 988–995.

Baker, W. L. 2006. Fire and restoration of sagebrush ecosystems. – Wildl. Soc. Bull. 34: 177–185.

Barde, B. V. and Bainwad, A. M. 2017. An overview of topic modeling methods and tools. – In: 2017 Int. Conf. on intelligent computing and control systems (ICICCS). IEEE.

Baselga, A. 2010. Multiplicative partition of true diversity yields independent alpha and beta components; additive partition does not. – Ecology 91: 1974–1981.

Beck, J. L. et al. 2009. Recovery of greater sage-grouse habitat features in wyoming big sagebrush following prescribed fire. – Restor. Ecol. 17: 393–403.

Beck, J. L. et al. 2012. Consequences of treating wyoming big sagebrush to enhance wildlife habitats. – Rangel Ecol. Manage. 65: 444–455.

Bennett, E. M. et al. 2015. Linking biodiversity, ecosystem services and human well-being: three challenges for designing research for sustainability. – Curr. Opin. Environ. Sustain. 14: 76–85.
Bernardo, J. M. et al. 2007. Generative or discriminative? Getting the best of both worlds. – Bayesian Stat. 8: 3–24.
Blei, D. M. and Lafferty, J. D. 2006. Dynamic topic models. – In: Proc. 23rd Int. Conf. on machine learning. Assoc. for Computing Machinery, Pittsburgh, PA, USA.
Blei, D. M. and Lafferty, J. D. 2007. A correlated topic model of science. – Ann. Appl. Stat. 1: 17–35.
Blei, D. M. et al. 2003. Latent Dirichlet Allocation. – J. Mach. Learn. Res. 3: 993–1022.
Burke, C. et al. 2011. Bacterial community assembly based on functional genes rather than species. – Proc. Natl Acad. Sci. USA 108: 14288–14293.
Chen, X. et al. 2010. Probabilistic topic modeling for genomic data interpretation. – In: 2010 IEEE Int. Conf. on bioinformatics and biomedicine (BIBM). IEEE.
Chen, X. et al. 2012. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. – IEEE/ACM Trans. Comput. Biol. Bioinform. 9: 980–991.
Chomel, M. et al. 2016. Plant secondary metabolites: a key driver of litter decomposition and soil nutrient cycling. – J. Ecol. 104: 1527–1541.
Christensen, E. M. et al. 2018. Long-term community change through multiple rapid transitions in a desert rodent community. – Ecology 99: 1523–1529.
Clements, F. E. 1936. Nature and structure of the climax. – J. Ecol. 24: 252.
Cross, T. B. et al. 2018. The genetic network of greater sage-grouse: range-wide identification of keystone hubs of connectivity. – Ecol. Evol. 8: 5394–5412.
Douma, J. C. and Weedon, J. T. 2019. Analysing continuous proportions in ecology and evolution: a practical introduction to beta and Dirichlet regression. – Methods Ecol. Evol. 10: 1412–1430.
Eldridge, D. J. et al. 2016. Ecosystem structure, function and composition in rangelands are negatively affected by livestock grazing. – Ecol. Appl. 26: 1273–1283.
Ewert, F. 2004. Modelling plant responses to elevated CO_2: how important is leaf area index? – Ann. Bot. 93: 619–627.
Fauchald, P. et al. 2017. Arctic greening from warming promotes declines in caribou populations. – Sci. Adv. 3: e1601365.
Fine, P. V. A. et al. 2021. Exploring the links between secondary metabolites and leaf spectral reflectance in a diverse genus of Amazonian trees. – Ecosphere 12: e03362.
Fremgen-Tarantino, M. R. et al. 2020. Winter foraging ecology of Greater sage-grouse in a post-fire landscape. – J. Arid Environ. 178: 104154.
Frye, G. G. et al. 2013. Phytochemistry predicts habitat selection by an avian herbivore at multiple spatial scales. – Ecology 94: 308–314.
Gabry, J. et al. 2019. Visualization in Bayesian workflow. – J. R. Stat. Soc. Ser. A Stat. Soc. 182: 389–402.
Gleason, H. A. 1926. The individualistic concept of the plant association. – Bull. Torrey Bot. Club 53: 7–26.
Harris, K. et al. 2017. Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process. – Proc. IEEE 105: 516–529.
Heino, J. 2010. Are indicator groups and cross-taxon congruence useful for predicting biodiversity in aquatic ecosystems? – Ecol. Indic. 10: 112–117.
Hornik, K. and Grün, B. 2011. topicmodels: an R package for fitting topic models. – J. Stat. Softw. 40: 1–30.
Kennedy, R. E. et al. 2009. Remote sensing change detection tools for natural resource managers: understanding concepts and tradeoffs in the design of landscape monitoring projects. – Remote Sens. Environ. 113: 1382–1396.
Knott, J. A. et al. 2020. Community-level responses to climate change in forests of the eastern United States. – Global Ecol. Biogeogr. 29: 1299–1314.
Kohl, K. D. and Dearing, M. D. 2016. The woodrat gut microbiota as an experimental system for understanding microbial metabolism of dietary toxins. – Front. Microbiol. 7: 1165.
Kohl, K. D. et al. 2014. Gut microbes of mammalian herbivores facilitate intake of plant toxins. – Ecol. Lett. 17: 1238–1246.
Kohl, K. D. et al. 2016. Microbial detoxification in the gut of a specialist avian herbivore, the greater sage-grouse. – FEMS Microbiol. Lett. 363: fnw144.
Kreitman, M. 1996. The neutral theory is dead. Long live the neutral theory. – Bioessays 18: 678–683.
Legendre, P. and Legendre, L. 2012. Numerical ecology. – Elsevier.
Legendre, P. et al. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. – Ecol. Monogr. 75: 435–450.
Ley, R. E. et al. 2008. Evolution of mammals and their gut microbes. – Science 320: 1647–1651.
Li, Z. et al. 2020. Land cover harmonization using Latent Dirichlet Allocation. – Int. J. Geogr. Inf. Sci. 0: 1–27.
Liu, L. et al. 2019. Probiotic clostridium butyricum improves the growth performance, immune function and gut microbiota of weaning rex rabbits. – Probiotics Antimicrob. Proteins 11: 1278–1292.
Lortie, C. J. et al. 2004. Rethinking plant community theory. – Oikos 107: 433–438.
Lowe, W. H. and McPeek, M. A. 2014. Is dispersal neutral? – Trends Ecol. Evol. 29: 444–450.
Lynch, M. 2007. The evolution of genetic networks by non-adaptive processes. – Nat. Rev. Genet. 8: 803–813.
McCune, B. et al. 2002. Analysis of ecological communities. – MjM Software Design Gleneden Beach, OR.
Mohajeri, M. H. et al. 2018. The role of the microbiome for human health: from basic science to clinical applications. – Eur. J. Nutr. 57: 1–14.
Mosher, B. A. et al. 2020. Successful molecular detection studies require clear communication among diverse research partners. – Front. Ecol. Environ. 18: 43–51.
National Ecological Observatory Network. 2019. Data Product DP3.30006.001, Spectrometer orthorectified surface directional reflectance – mosaic. Provisional data downloaded from <http://data.neonscience.org>, February 2019.
Nemergut, D. R. et al. 2013. Patterns and processes of microbial community assembly. – Microbiol. Mol. Biol. Rev. 77: 342–356.
Nobler, J. D. et al. 2019. Preferences of specialist and generalist mammalian herbivores for mixtures versus individual plant secondary metabolites. – J. Chem. Ecol. 45: 74–85.
Oh, K. P. et al. 2019. Conservation genomics in the sagebrush sea: population divergence, demographic history and local adaptation in sage-grouse (*Centrocercus* spp.). – Genome Biol. Evol. 11: 2023–2034.
Olsoy, P. J. et al. 2015. Fearscapes: mapping functional properties of cover for prey with terrestrial LiDAR. – BioScience 65: 74–80.

Olsoy, P. J. et al. 2020. Mapping foodscapes and sagebrush morphotypes with unmanned aerial systems for multiple herbivores. – Landscape Ecol. 35: 921–936.
Paul, G. C. and Friend, D. G. 2019. Clostridial enterotoxemia and coccidiosis in weanling cottontail rabbits (*Sylvilagus audubonii*, *Sylvilagus floridanus*, *Sylvilagus nuttallii*) from Colorado, USA. – J. Wildl. Dis. 55: 189–195.
Pearson, D. L. and Carroll, S. S. 1999. The influence of spatial scale on cross-taxon congruence patterns and prediction accuracy of species richness. – J. Biogeogr. 26: 1079–1090.
Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). – Methods Ecol. Evol. 5: 397–406.
Pritchard, J. K. et al. 2000. Inference of population structure using multilocus genotype data. – Genetics 155: 945–959.
Räsänen, A. et al. 2016. The role of landscape, topography and geodiversity in explaining vascular plant species richness in a fragmented landscape. – Boreal Environ. Res. 21: 53–70.
Requena-Mullor, J. M. et al. 2019. Integrating anthropogenic factors into regional-scale species distribution models-A novel application in the imperiled sagebrush biome. – Global Change Biol. 25: 3844–3858.
Sandifer, P. A. et al. 2015. Exploring connections among nature, biodiversity, ecosystem services and human health and well-being: opportunities to enhance health and biodiversity conservation. – Ecosyst. Serv. 12: 1–15.
Sankaran, K. and Holmes, S. P. 2017. Latent variable modeling for the microbiome. – Biostatistics 20: 599–614.
Severson, J. P. et al. 2017. Effects of conifer expansion on greater sage-grouse nesting habitat selection. – J. Wildl. Manage. 81: 86–95.
Simpson, E. H. 1949. Measurement of diversity. – Nature 163: 688–688.
Steenvoorden, J. et al. 2019. The potential importance of unburned islands as refugia for the persistence of wildlife species in fire-prone ecosystems. – Ecol. Evol. 9: 8800–8812.
Su, J. C. et al. 2004. Beyond species richness: community similarity as a measure of cross-taxon congruence for coarse-filter conservation. – Conserv. Biol. 18: 167–173.
Trevelline, B. K. et al. 2019. Conservation biology needs a microbial renaissance: a call for the consideration of host-associated microbiota in wildlife management practices. – Proc. R. Soc. B 286: 20182448.
Ulappa, A. C. et al. 2014. Plant protein and secondary metabolites influence diet selection in a mammalian specialist herbivore. – J. Mammal. 95: 834–842.
Valle, D. et al. 2014. Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. – Ecol. Lett. 17: 1591–1601.
Valle, D. et al. 2018. Extending the Latent Dirichlet Allocation model to presence/absence data: a case study on North American breeding birds and biogeographical shifts expected from climate change. – Global Change Biol. 24: 5560–5572.
van der Hooft, J. J. J. et al. 2016. Topic modeling for untargeted substructure exploration in metabolomics. – Proc. Natl Acad. Sci. USA 113: 13738–13743.
Veech, J. A. and Crist, T. O. 2010. Diversity partitioning without statistical independence of alpha and beta. – Ecology 91: 1964–1969.
Vellend, M. 2010. Conceptual synthesis in community ecology. – Q. Rev. Biol. 85: 183–206.
Wei, X. et al. 2007. Dynamic mixture models for multiple time-series. – In: Proc. 20th Int. joint conference on artifical intelligence. Morgan Kaufmann Publishers Inc., pp 2909–2914
White, S. M. et al. 1982. Preference of pygmy rabbits *Brachylagus idahoensis* for various populations of big sagebrush *Artemisia tridentata*. – J. Range Manage. 35: 724–726.
Zaiats,. A et al. 2021. Data from: Unifying community detection across scales from genomes to landscapes. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.8w9ghx3mf>.