BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography

Franziska Schrodt1,2,3,*, Jens Kattge1,2, Hanhuai Shan4,5, Farideh Fazayeli6, Julia Joswig7, Arindam Banerjee8, Markus Reichstein1, Gerhard Bönisch1, Sandra Diaz9, John Dickie7, Andy Gillison8, Anuj Karpatne9, Sandra Lavorel9, Paul Leadley10, Christian B. Wirth2,11, Ian J. Wright12, S. Joseph Wright13 and Peter B. Reich3,14

ABSTRACT

Aim Functional traits of organisms are key to understanding and predicting biodiversity and ecological change, which motivates continuous collection of traits and their integration into global databases. Such trait matrices are inherently sparse, severely limiting their usefulness for further analyses. On the other hand, traits are characterized by the phylogenetic trait signal, trait–trait correlations and environmental constraints, all of which provide information that could be used to statistically fill gaps. We propose the application of probabilistic models which, for the first time, utilize all three characteristics to fill gaps in trait databases and predict trait values at larger spatial scales.

Innovation For this purpose we introduce BHPMF, a hierarchical Bayesian extension of probabilistic matrix factorization (PMF). PMF is a machine learning technique which exploits the correlation structure of sparse matrices to impute missing entries. BHPMF additionally utilizes the taxonomic hierarchy for trait prediction and provides uncertainty estimates for each imputation. In combination with multiple regression against environmental information, BHPMF allows for extrapolation from point measurements to larger spatial scales. We demonstrate the applicability of BHPMF in ecological contexts, using different plant functional trait datasets, also comparing results to taking the species mean and PMF.

Main conclusions Sensitivity analyses validate the robustness and accuracy of BHPMF: our method captures the correlation structure of the trait matrix as well as the phylogenetic trait signal – also for extremely sparse trait matrices – and provides a robust measure of confidence in prediction accuracy for each missing entry. The combination of BHPMF with environmental constraints provides a promising concept to extrapolate traits beyond sampled regions, accounting for intraspecific trait variability. We conclude that BHPMF and its derivatives have a high potential to support future trait-based research in macroecology and functional biogeography.

Keywords Bayesian hierarchical model, gap-filling, imputation, machine learning, matrix factorization, PFT, plant functional trait, sparse matrix, spatial extrapolation, TRY.
INTRODUCTION

Functional trait measurements and analyses have been the focus of numerous studies in recent decades (e.g. Reich et al., 1997; Wright et al., 2004; Chave et al., 2009; Schrodt et al., 2015). However, due to the time and resources required and the sheer number of species on earth, only a small number of species and their traits could be captured to date, especially in tropical and remote ecosystems. In addition, trait data are highly dispersed among numerous datasets and are often not accessible to the wider scientific community. The integration of databases is thus becoming increasingly important for the consolidation of globally dispersed data, as a source of standardized data for further applications, such as model building and validation, and to coordinate future measurement efforts.

Combining trait observations from studies with different research foci produces matrices with substantial gaps. For example, the largest database for plant traits to date, TRY (Kattge et al., 2011), currently contains 215 datasets with 5.6 million trait entries for 1100 traits of 2 million individuals, representing 100,000 plant species. On average only 2 of the 1100 traits represented in TRY are measured for any individual, restricting the usefulness of combined datasets especially for multivariate analyses.

General characteristics of traits

Some characteristics inherent to functional traits may support statistical gap-filling of sparse trait matrices: a strong phylogenetic trait signal, functional and structural trade-offs between traits and trait–environment relationships.

The phylogenetic trait signal is an effective aid in predicting trait values (e.g. Lovette & Hochachka, 2006; Swenson, 2014): the closer two individuals are related, the more similar their traits will be – with exceptions due to convergent evolution and environmental diversification. This phylogenetic trait signal is reflected in the taxonomic hierarchy: on average individuals within a species are more similar than individuals within genera, families or phylogenetic groups (Kerkhoff et al., 2006; Swenson & Enquist, 2007). In general, most of the trait variance is observed between species (Kattge et al., 2011), differences between species are consistent across spatial scales (Kazakou et al., 2014) and mean trait values at native ranges are appropriate estimates at invaded areas (McMahon, 2002; Ordonez, 2014). Due to these frequently observed patterns, mean trait values of species and even genera are often used to fill gaps in trait matrices (Fried et al., 2012; Cordlandwehr et al., 2013).

Traits possessed by any individual are not independent – functional and structural trade-offs cause correlations between traits (Reich et al., 1997). This has been characterized at the global scale, for example for the leaf and wood economic spectra, respectively (Wright et al., 2004; Chave et al., 2009). Trait variability is also influenced by the environment which, on the one hand, causes large-scale patterns, for example latitudinal gradients of leaf N/P stoichiometry (Reich & Oleksyn, 2004) or changes in the competitive success of Drosophila species along a temperature gradient (Davis et al., 1998), and on the other hand small-scale intraspecific variation, for example in plant traits (Albert et al., 2010; Clark, 2010) and mammalian body size (Diniz-Filho et al., 2007).

Gap-filling in trait ecology

Many disciplines, for example psychology (e.g. Schafer & Graham, 2002), sociology (e.g. Johnson & Young, 2011) or biogeochemistry (e.g. Moffat et al., 2007; Lasslop et al., 2010) have developed their own set of accepted techniques for predicting missing values tailored to their data structure. In trait ecology, due to the relatively recent advent of large databases, gap-filling methods have been adopted from other disciplines, often without adjusting them to differences in data structure. In general three approaches are used within the community of trait ecologists: (1) deleting rows with missing cases or pair-wise analysis; (2) predicting trait values based on the taxonomic trait signal at species or genus level (mean traits); and (3) predicting values for individual missing entries based on structure, e.g. multiple imputation (MI) (Rubin, 1987; Su et al., 2011) and multivariate imputation using chained equations (MICE) (Rubin, 1987; van Buuren & Groothuis-Oudshoorn, 2011), or spatial correlations within the trait matrix (e.g. kriging; Lamsal et al., 2012; Räty & Kangas, 2012). These approaches provide useful imputations in some cases, but whilst deleting missing cases can introduce bias in model parameters (Nakagawa & Freckleton, 2008), taking the ‘mean’ adds new data points without adding new information, which results in incorrect confidence limits. The techniques mentioned in (3) on the other hand were developed to fill gaps in databases with 10–30% missing entries, but not for a sparsity as high as that observed in combined trait databases such as TRY.

A recent exercise comparing different approaches (amongst them ‘mean’ and MICE) to filling gaps in a plant trait matrix – although at the species rather than the individual level – showed that all methods were indeed only effective up to 30% gaps, that genus mean produced the least reliable results and that including ecological theory, for example by taking into account trait–trait correlations, will substantially improve the accuracy of gap-filling approaches (Taugourdeau et al., 2014). Ogle (2013) developed a hierarchical Bayesian model, which could be used for gap-filling of individual trait values based on taxonomic hierarchy and covariates. However, this method only allows for one trait to be predicted at a time, ignoring the correlation structure within a multi-trait matrix. In contrast, some phylogenetic imputation methods, as reviewed recently by Swenson (2014), do allow for multiple trait predictions but only give species or higher-order means and lack any accounting for intraspecific variability. While this may be resolved, for example by adding ‘twigs’ on the ends of phylogenies with terminal nodes as species (Swenson, pers. commun.) to the authors’ knowledge, this has not been implemented yet.

Here we present a new approach – Bayesian hierarchical probabilistic matrix factorization (BHPMF) – which imputes trait values based on the taxonomic hierarchy, structure within
the trait matrix and trait–environment relationships at the same
time as providing uncertainty estimates for each single trait
prediction. An extension of the method provides a concept for
out-of-sample predictions – the extrapolation of point measure-
ments to spatial scales beyond measured areas. We evaluate trait
predictions by BHPMF under various aspects and provide per-
spectives for future developments.

METHODS

Probabilistic matrix factorization

The method we present is a development of probabilistic matrix
factorization (PMF) (Salakhutdinov & Mnih, 2008). PMF is a
recommendation system developed on the example of predict-
ing users’ preferences in movies from other users’ movie ratings
(Netflix, 2009). Due to its good scalability and predictive accu-

racy, even for highly sparse datasets, PMF has become a standard
technique for imputing missing data (Koren et al., 2009).

PMF models a sparse matrix, such as the TRY database, as the
scalar (or inner, dot) product of two latent matrices with the aim
of finding a factorization that minimizes the error between pre-
dicted and observed data. This technique is closely related to
principal components analysis (PCA), which converts a set of
observations of correlated variables into a set of values of lin-
early uncorrelated variables called principal components. Like
PCA, PMF is efficient if the original matrix is of low rank, i.e. if
the axes of the original matrix provide strong correlations.

In the original case, each column represents a video and each
row a user, providing a video ranking (Salakhutdinov & Mnih,
2008). In the case of a trait matrix, videos are replaced by traits
and users by entities. In our case, entities are defined as individ-
ual plants. They could equally represent the level of an organ or
the average over different organisms, for example nitrogen
content of a single leaf, motility of a phytoplankton species or
the average beak length of several individuals of an avian species.
For simplicity, we refer to the individual plant as plant from now
on.

In a first step, independent latent vectors are generated over
each row (individual, \( u \)) and column (trait, \( v \)) of the plant \( \times \) trait
matrix \( X \in \mathbb{R}^{N \times M} \) with \( N \) rows and \( M \) columns (Fig. 1). Any
missing entry \((n, m)\) in the original matrix \( X \) can be predicted as
the inner product of these latent vectors

\[
x_{nm} = \langle u_n, v_m \rangle
\]

(Fig. 1).

PMF has been shown to be applicable to biological data, for
example in population genetics (Duforet-Frebourg and Blum,
2014). However, our first experiments indicated that for plant
traits the prediction accuracy of PMF was insufficient: the accu-
racy was worse than using species mean trait values to fill the
gaps (see Results). We therefore developed an extension of PMF,
which accounts for a plants’ taxonomic hierarchy to improve
prediction accuracy (Bayesian hierarchical PMF; BHPMF). The
concept was developed as HPMF in the context of machine
learning (Shan et al., 2012). We here introduce the additional
application of a Gibbs sampler in order to provide a measure of
uncertainty for each imputed trait value (BHPMF) (see ’Gibbs
sampler – uncertainty quantified trait prediction’), as well as an
extension to facilitate out-of-sample predictions (aHPMF).

Bayesian hierarchical probabilistic
matrix factorization

BHPMF exploits the taxonomic hierarchy of the plant kingdom
as a proxy for the phylogenetic trait signal, with the individual
plant being nested in species, species in genus, genus in family
and family in phylogenetic group (Fig. 2).

BHPMF sequentially performs PMF at the different hierar-
chical levels, using latent vectors of the neighbouring level \( \ell \)
as prior information at the current level. For example, trait data
averaged at species level are used to optimize latent vectors at
species level, which in turn act as priors for latent vectors at
the individual level, which finally are optimized against the
observed trait entries in the trait matrix (Fig. 2, equation 1).The
sequential approach across the taxonomic hierarchy turned out
to be most effective if applied iteratively top down and bottom
up.

After transformation of traits to approximate normal distri-
butions and z-score transformation, the cost function is devel-
oped as the sum of absolute deviations of predictions versus
observations for traits \( m \) of entities \( n \) (first summand in
above equation 1) and the sum of absolute deviations of posterior
and prior of the latent factors \( u \) and \( v \) (second and third summand of
above equation 1) across all hierarchical levels \( \ell \):

\[
E = \sum_{\ell=1}^{L} \sum_{nm} (x_{nm} - \langle u_n^{(\ell)}, v_m^{(\ell)} \rangle)^2
+ \lambda_u \sum_n \| u_n^{(\ell)} - u_n^{(\ell-1)} \|^2 + \lambda_v \sum_m \| v_m^{(\ell)} - v_m^{(\ell-1)} \|^2,
\]

(1)

Figure 1 Schematic of the probabilistic matrix factorization
(PMF) model. \( u \) denotes the latent vector on the individual plant
side, \( v \) the latent vector on the functional trait side, both of which
have a Gaussian normal distribution with a mean of 0 and a
variance of \( \sigma^2 \). Each missing entry \( X_{ij} \) can be approximated by
the product of the transposed latent vector \( U \) and the latent vector \( V \).

Global Ecology and Biogeography, © 2015 John Wiley & Sons Ltd
where $\lambda_n = \sigma^2 / \sigma^2_n$, $\lambda_v = \sigma^2 / \sigma^2_v$ with $\sigma$ being the standard deviation of the imputations. $\{ \}$ denotes the set of data at all L levels, and $\bar{\delta}^{X_{m,n}} = 1$ when the entry $(n, m)$ of $X^{(n)}$ is non-missing and 0 otherwise. Replace $\ell - 1$ with $\ell + 1$ and the parent node ($p(n)$) with the child node ($c(n)$) for the bottom-up approach. For details see BHPMF in Appendix S1 in Supporting Information.

**Gibbs sampler – uncertainty quantified trait prediction**

The parameters of our BHPMF model are optimized against the observations in the matrix using a Gibbs sampler (Fazayeli et al., 2014). The Gibbs sampler is a Markov chain Monte Carlo (MCMC) method, which samples the probability density distributions of model parameters (here the latent vectors) and model predictions (here entries in the plant x trait matrix) (see the grey inset in Fig. 2 and ‘Model evaluation’, as well as Gibbs sampler results in Appendix S10). The Gibbs sampler-inferred density distributions of trait values are then used to infer the most likely imputation value, as well as the associated uncertainty for each prediction.

**aHPMF – extrapolation from point measurements to regional scales**

If BHPMF is stopped at the species level, i.e. without accounting for trait variability specific to individual plants, the residual error represents the intraspecific variability and modelling/measurement errors. aHPMF focuses on explaining this residual trait variability based on environmental variables, such as soil and climate characteristics of their growth environment in order to enable out-of-sample prediction, i.e. trait predictions for individual plants where the only known factors are species identity and location but no traits have actually been measured on the given individual (Fig. 3).

To capture trait variability that can be attributed to environmental factors, we utilize a hierarchical regression framework, taking into account the taxonomic structure of plants to regularize the regression model. The regression framework takes as independent predictor variables the climatic and soil variables mentioned below at locations with georeferenced trait measurements. The residuals of BHPMF for the 13 plant traits of each observation are considered as the target dependent variables to be predicted. We treat each plant trait independently of every other while regressing them using climate and soil features.

In essence, combining BHPMF with least squares regression over the residuals against environmental factors, we can model the unknown value for species $n$ and trait $m$ in a probabilistic model as:

$$k_{nm} = \alpha u_n^T v_n + \beta w^T x + \epsilon_{nm}$$

where $(u_n, v_n)$ are the latent factors, with $u_n$ having a hierarchical prior from the taxonomy and $x$ being the environmental condition with $w$ as the regression coefficient. $\epsilon_{nm}$ is the zero mean Gaussian noise. Note that $\alpha, \beta$ are scalar parameters: for BHPMF set $(\alpha = 1, \beta = 0)$, for aHPMF set $(\alpha = 1, \beta = 1)$. For details see ‘aHPMF’ in Appendix S1.

**Data: traits, climate and soil**

We demonstrate the applicability of the methods introduced above on the example of a trait matrix derived from TRY. For details on data standardization see Kattge et al. (2011). The spatial distribution of measurement sites and detailed information on the original datasets are shown in Fig. S4.1 (Appendix S4) and Tables S3.1 & S3.2 in Appendix S3.

We extracted a matrix of 13 georeferenced traits consisting of 204,404 trait measurements on 78,300 individuals, spanning 14,320 species, 3793 genera, 358 families and 6 phylogenetic
groups. The sparsity ranged from 49.63% for leaf area to 92.33% for the leaf N to P ratio, with an average sparsity of 79.9% across the trait matrix (Table 1). All traits were log- and z-transformed to improve normality and equalize traits in the cost function during optimization.

For out-of-sample predictions by aHPMF, climate data for mean annual precipitation, mean annual temperature, isothermally and precipitation seasonality were extracted from the WorldClim dataset (Hijmans et al., 2005) and soil texture (sand, silt, clay) and soil organic carbon content in the top soil from the Harmonized World Soil Database v1.2 (FAO et al., 2012).

Model evaluation

We ran PMF, BHPMF and aHPMF on the test dataset extracted from TRY. Given the plant × trait matrix, we randomly selected 80% of entries for training (parameter setting), 10% for validation (parameter adjustment by optimizing performance) and 10% for test (independent performance testing after parameter adjustment and learning). This cross-validation improves model fidelity by ensuring that none of the observations are known by the model when performing new predictions. Test entries without training data in the same row would have highly inflated variance. Such cases were prevented by adjusting the splitting accordingly (see ‘BHPMF’ in Appendix S1).

We evaluated the predicted trait values, using the root mean squared error (RMSE; see equation S13 in Appendix S1) and the correlation coefficient ($R^2$) of z-transformed predicted versus observed traits as indicators of overall prediction accuracy. We compared the performance of PMF, BHPMF and aHPMF with a baseline of species mean trait values (MEAN), which uses the overall trait mean of all individual plants within a species for prediction. The effectiveness of capturing the phylogenetic trait signal was explored by performing BHPMF including increasingly detailed taxonomic information (Fig. 2).

In order to evaluate how well not only predicted versus measured but also trait-trait correlations are preserved in BHPMF, we performed standardized major axis (SMA) regression, the first principal component vector of a correlation matrix fitted through the data centroid (Taskinen & Warton, 2011), on the measured and imputed trait values for some key trait correlations. We also performed a Procrustes analysis with PROTEST (using the R package ‘vegan’) on a PCA of a subset of the original data versus a PCA based on the estimated values for artificially introduced gaps. Due to its good data cover, we performed this test on the RAINFOR extract from the TRY database (see below). Procrustes is a statistical shape analysis tool (least-squares orthogonal mapping) which compares two ‘superimposed’ matrices for overlap, with placement in space and object size being adjustable. We show how uncertainty in trait predictions is accounted for using the Gibbs sampler, comparing prediction confidence (SD) with prediction accuracy (RMSE).

The sensitivity of BHPMF to the fraction of gaps and the effect of using a global database to fill gaps in local or regional datasets were explored using two approaches. First by ‘cutting out’ a local dataset with high coverage, adding additional gaps (0, 10, 30, 60 and 80%; see Table S8.1 in Appendix S8) and second by using a regional gappy dataset, filling gaps in each of these ‘cut-outs’ using (1) the global data with information from the local/regional data and (2) just the local/regional data. For our local example, we extracted TRY trait data contributed by the RAINFOR group (Fyllas et al., 2009), which shows a good coverage (sparsity 11%) and covers most of the Amazon (Fig. S8.1 in Appendix S8). For our regional example, we extracted all of the European data (sparsity 72%) (Fig. S8.2 in Appendix S8). For details on methodology please refer to

---

**Table 1** Number of entries, sparsity and root mean square error (RMSE) of species mean (MEAN), probabilistic matrix factorization (PMF), Bayesian hierarchical PMF (BHPMF) and advanced hierarchical PMF (aHPMF) by trait, as well as $R^2$ values of the regression of imputed versus measured traits. The lowest RMSE and highest $R^2$ are shown in bold.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Entries</th>
<th>Sparsity</th>
<th>MEAN RMSE</th>
<th>$R^2$</th>
<th>PMF RMSE</th>
<th>$R^2$</th>
<th>BHPMF RMSE</th>
<th>$R^2$</th>
<th>aHPMF RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLA</td>
<td>33001</td>
<td>57.9</td>
<td>0.53</td>
<td>0.85</td>
<td>0.88</td>
<td>0.49</td>
<td>0.46</td>
<td>0.88</td>
<td>0.53</td>
<td>0.89</td>
</tr>
<tr>
<td>Plant height</td>
<td>16466</td>
<td>79.0</td>
<td>0.47</td>
<td>0.90</td>
<td>0.91</td>
<td>0.40</td>
<td>0.40</td>
<td>0.92</td>
<td>0.44</td>
<td>0.93</td>
</tr>
<tr>
<td>Seed mass</td>
<td>7311</td>
<td>90.7</td>
<td>0.37</td>
<td>0.91</td>
<td>0.77</td>
<td>0.40</td>
<td>0.36</td>
<td>0.92</td>
<td>0.36</td>
<td>0.92</td>
</tr>
<tr>
<td>SSD</td>
<td>17331</td>
<td>77.9</td>
<td>0.53</td>
<td>0.83</td>
<td>0.87</td>
<td>0.41</td>
<td>0.43</td>
<td>0.88</td>
<td>0.49</td>
<td>0.89</td>
</tr>
<tr>
<td>Leaf area</td>
<td>39438</td>
<td>49.6</td>
<td>0.50</td>
<td>0.87</td>
<td>0.91</td>
<td>0.41</td>
<td>0.37</td>
<td>0.93</td>
<td>0.41</td>
<td>0.93</td>
</tr>
<tr>
<td>Leaf N</td>
<td>26882</td>
<td>65.7</td>
<td>0.67</td>
<td>0.77</td>
<td>0.99</td>
<td>0.28</td>
<td>0.53</td>
<td>0.86</td>
<td>0.59</td>
<td>0.86</td>
</tr>
<tr>
<td>Leaf P</td>
<td>11975</td>
<td>84.7</td>
<td>0.72</td>
<td>0.69</td>
<td>0.78</td>
<td>0.62</td>
<td>0.52</td>
<td>0.83</td>
<td>0.62</td>
<td>0.83</td>
</tr>
<tr>
<td>Leaf N/area</td>
<td>8180</td>
<td>89.6</td>
<td>0.79</td>
<td>0.65</td>
<td>0.80</td>
<td>0.64</td>
<td>0.51</td>
<td>0.82</td>
<td>0.72</td>
<td>0.82</td>
</tr>
<tr>
<td>Leaf fresh mass</td>
<td>11484</td>
<td>85.3</td>
<td>0.47</td>
<td>0.89</td>
<td>0.71</td>
<td>0.71</td>
<td>0.27</td>
<td>0.96</td>
<td>0.39</td>
<td>0.96</td>
</tr>
<tr>
<td>Leaf N/P ratio</td>
<td>5999</td>
<td>92.3</td>
<td>0.76</td>
<td>0.69</td>
<td>0.90</td>
<td>0.44</td>
<td>0.49</td>
<td>0.85</td>
<td>0.67</td>
<td>0.84</td>
</tr>
<tr>
<td>Leaf C/dry mass</td>
<td>8123</td>
<td>89.6</td>
<td>0.70</td>
<td>0.74</td>
<td>0.884</td>
<td>0.35</td>
<td>0.61</td>
<td>0.61</td>
<td>0.62</td>
<td>0.78</td>
</tr>
<tr>
<td>Leaf δ 15N</td>
<td>9022</td>
<td>88.5</td>
<td>0.63</td>
<td>0.79</td>
<td>1.02</td>
<td>0.02</td>
<td>0.50</td>
<td>0.87</td>
<td>0.53</td>
<td>0.88</td>
</tr>
<tr>
<td>Average</td>
<td>15723</td>
<td>79.9</td>
<td>0.58</td>
<td>0.80</td>
<td>0.88</td>
<td>0.41</td>
<td>0.45</td>
<td>0.88</td>
<td>0.53</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Appendices S8 & S1. Finally, we provide an example for out-of-sample prediction, extrapolating leaf nitrogen concentration (leaf N) from point measurements to the whole species range of *Acer saccharum* using aHPMF.

Probabilistic matrix factorization and subsequent regression were developed and applied in MATLAB version 2012a (MATLAB, 2012). All other analyses were performed using the statistical platform R version 2.15 (R Core Team, 2014). The maps reported here were produced in ArcMap 10.1 (ArcGIS Desktop, 2011) and R, using the tree species distribution map of *A. saccharum* from the US Geological Survey (Little, 1971). R scripts to implement BHPMF are available from the authors by request.

**RESULTS**

**Predicted versus observed trait values**

To analyse prediction accuracy we compare RMSE and the coefficient of determination ($R^2$) for MEAN, PMF, BHPMF and aHPMF averaged across traits and for each trait separately (Table 1; for scatterplots of observed versus predicted for all traits see Fig. S9.1 in Appendix S9). On average, across all traits, BHPMF outperforms PMF, MEAN and aHPMF, with MEAN being significantly more accurate than PMF. This holds after statistical evaluation using a paired $t$-test with $P$-values smaller than $10^{-5}$ at all levels, and is supported by the evaluation of the correlation coefficient $R^2$ (Table 1). As the RMSE is calculated from $z$-transformed approximate normal distributions of traits, a RMSE of 0.45 for BHPMF indicates that the average error of predictions is about half a standard deviation, or about 10% of the 95% CI. BHPMF outperforms MEAN and PMF in all traits, while aHPMF shows the same or higher RMSE and higher $R^2$ than BHPMF for SLA, plant height, leaf dry matter content (LDMC), leaf carbon (C) per dry mass and leaf δ^{15}N (D15N) (Table 1). The advantage of BHPMF over MEAN is largest for ‘physiological traits’, such as leaf N and leaf phosphorus concentration (leaf P), and smaller for more ‘structural traits’ such as seed mass or plant height. The prediction accuracy of BHPMF varies across traits: from RMSE = 0.36 ($R^2 = 0.92$) for seed mass to RMSE = 0.61 ($R^2 = 0.61$) for leaf C content per dry mass. Interessingly, prediction accuracy is not related to the number of entries per trait (Table 1).

**Accounting for taxonomic hierarchy**

The RMSE of MEAN and BHPMF decreases with increasing taxonomic information, indicating that both methods can utilize the hierarchical structure to their advantage (Table S7.1 in Appendix S7). This is also supported by the scatter plot of measured versus predicted specific leaf area (SLA) and leaf N shown in Fig. 4. With increasing taxonomic information, the scatter plot approaches the 1:1 line, i.e. prediction accuracy improves.

**Figure 4** Scatter plots of predicted versus true values for two traits with increasing taxonomic information. Left column, leaf nitrogen concentration per dry weight; right column, specific leaf area. Row 1, no phylogenetic information is used; row 2, only the phylogenetic group is used; row 3, phylogenetic group and family are used; row 4, phylogenetic group, family and genus are used; row 5, phylogenetic group, family, genus and species are used. Predictions are based on Bayesian hierarchical probabilistic matrix factorization. The data are presented in $z$-transformed space. Dotted lines indicate the 1:1 correlation.

**Trait–trait correlations**

Although the presence of strong trait–trait correlations is a prerequisite for the accuracy of BHPMF, such correlations are not provided a priori and are thus not part of the objective function used (equation 1). This turns them into a suitable evaluation measure. An important quality criterion is to what extent the imputed values reflect the observed bivariate correlations, as this is a first indication of the extent to which the overall correlation structure of the n-dimensional trait matrix is maintained by imputation. Our dataset shows on average strong trait–trait correlations, with some exceptions (Fig. S9.5 in Appendix S9). BHPMF and MEAN capture these general trait–trait correlations, but BHPMF reproduces extreme values more accurately than MEAN and is therefore generally better at capturing the shape of the scatter of observed trait data, which is confirmed by more similar SMA $R^2$ values (Fig. S9.2 in Appendix S9). Looking at the multivariate preservation of trait–trait correlations using Procrustes analysis, our results indicate again that BHPMF does...
not significantly alter the correlation structure of the gap-filled matrix (Fig. 5). The first four principal component axes explain 83.4% and 83.4% of the variability in the dataset for the original and gap-filled data, respectively. None of the principal component axes are significantly different between the gappy and gap-filled data for any of the traits. The traits stem specific density and leaf carbon differ – but not significantly – along the third and fourth axes (see Fig. S9.3 in Appendix S9).

Uncertainty quantified predictions

The Gibbs sampler provides a probability distribution for every single prediction, as shown in the example of Gibbs sampler-generated density plots of BHPMF-estimated LDMC, leaf N and SLA for A. saccharum and Pinus sylvestris trees (Fig. S10.1 in Appendix S10). This distribution can be exploited to calculate indices for the best estimate (e.g. mean) and variability (e.g. SD). This provides an additional means to evaluate our imputation model by comparing prediction confidence (SD) with prediction accuracy (RMSE): when we are confident about our predictions (small SD), these predictions should also be accurate (small RMSE) and vice versa. Figure S10.2 in Appendix S10 shows that this is indeed the case for the whole 13-trait dataset, implying that our model is appropriate. This remains true when we evaluate the Gibbs sampler on each trait separately (Fig. S10.3 in Appendix S10).

Gap-filling of regional/local data using BHPMF

As expected, increasing the number of gaps in the RAINFOR dataset generally resulted in a decrease of prediction accuracy (Fig. 6), although less so for structural traits, such as stem-specific density (SSD) and plant height. Reproducibility was high in all cases (Fig. 6). Prediction accuracy of BHPMF was generally approximately equal, no matter whether the regional (RforR) or global (WforR) datasets were used to fill the gaps (Figs 6 & S8.3 in Appendix S8). This was particularly the case if gap sizes were large (above 10%), whilst RforR outperformed WforR for the imputations of plant height, leaf N, SLA and leaf carbon only where additional gap sizes were small (0 and 10%).

Out-of-sample prediction (aHPMF)

We illustrate the extension of BHPMF towards out-of-sample prediction with the example of leaf N across the species range of A. saccharum (Fig. 7).
BHPMF was stopped at species level, followed by a multivariate regression of residuals of predicted versus observed traits at the level of the individual plants against environmental conditions. Trait predictions follow the trends in both measured leaf N and environmental conditions. However, the variability within aHPMF-predicted leaf N is low (18.69–18.80 mg g$^{-1}$) compared with the range of actually measured values (8.65–28.23 mg g$^{-1}$). While the amplitude of predicted traits for the grid elements is much smaller than the observed ranges, it is notable that *A. saccharum* occurs over a wide range of environmental conditions and grid averages may not reflect variation at local scales. The flat response surface Reich & Oleksyn (2004) found for the correlation between leaf N in the genus *Acer* and mean annual temperature may further support the validity of our results.

**DISCUSSION**

We demonstrate that BHPMF provides accurate and robust uncertainty-quantified trait predictions, even for sparse matrices with up to 80% missing entries. BHPMF outperforms the species MEAN baseline in all aspects: RMSE and $R^2$ of predicted versus observed entries are smaller and larger, respectively, for each individual trait (Table 1) and trait–trait correlations are better retrieved (Fig. S9.2 in Appendix S9). Prediction accuracy is high (small RMSE) when prediction uncertainty is small (small SD; Fig. S10.2 in Appendix S10), providing a measure of confidence for prediction accuracy. In addition, aHPMF provides a concept to extrapolate from point measurements to species ranges accounting for intraspecific variability (Fig. 7). These results give rise to three major questions: (1) Why does BHPMF provide accurate and robust trait predictions even for sparse trait matrices? (2) Is the prediction accuracy of BHPMF sufficient for applications in ecological contexts? (3) What are the prospects for BHPMF?

*Why are BHPMF predictions robust and accurate?*

BHPMF is a Bayesian hierarchical approach that simultaneously takes the taxonomic trait signal, the correlation structure within the trait matrix and environmental constraints into account to fill gaps in trait matrices. A comparison of BHPMF-predicted trait values with results based on PMF, MEAN and aHPMF indicates the relevance of all three aspects.

PMF has been shown to be accurate even for the imputation of sparse datasets (Koren et al., 2009). However, in the case of our test dataset, PMF performs worse than the baseline MEAN approach. This indicates that our test dataset is not of sufficiently low rank to allow for robust trait predictions based on the correlation structure in the matrix alone.

BHPMF converts PMF into a hierarchical Bayesian model. This has two effects: (1) the higher levels of taxonomy (e.g. phylogenetic group, family) provide almost complete information (very low sparsity), which enables efficient PMF; (2) approximations of the matrices at higher taxonomic level provide excellent prior information for the approximations at lower levels (e.g. genus, species), thus constraining imputations due to the taxonomic signal in trait variation at all levels.

This is achieved without a priori assuming a phylogenetic signal in the trait variability, but rather by opening the door for our model to extract a signal, if it should be there. Thus, in some cases, BHPMF might not put any constraint on the imputed...
trait values from the taxonomy side, whereas in others, this signal might be stronger, hence the constraint from the taxonomy. The hierarchical taxonomic structure in combination with a consistent phylogenetic trait signal is therefore the key to facilitate robust gap-filling by BHPMF – despite high sparsity – at the level of the individual observations. On the other hand, BHPMF outperforms the MEAN approach in all aspects. This indicates that the capture by PMF of the correlation structure of the matrix on top of the phylogenetic trait signal is the key to providing accurate predictions.

A comparison of BHPMF with aHPMF indicates that explicitly taking environmental constraints into account surprisingly adds little or no improvement to BHPMF: the average $R^2$ was only higher in 5 out of 13 traits with the RMSE being consistently smaller for BHPMF compared with aHPMF (Table 1). At the individual level, PMF was replaced by multiple regressions against environmental constraints, based on the assumption that the phylogenetic trait signal is mainly observed at the species level whilst environmental constraints add to the trait variability at the individual level.

The fact that BHPMF largely outperforms aHPMF is an indication that, by taking into account trait–trait correlations, PMF seems to be more appropriate than the multiple regression at the individual level. Also, BHPMF implicitly takes environmental information into account via the correlation of measured to predicted traits. This environmental constraint is related to the immediate environment experienced by the observed plant and propagated via trait–trait correlations to the predicted traits, while environmental information incorporated in the multiple regression model by necessity represents only the average over larger scales. Taking environmental information explicitly into account may therefore not substantially improve BHPMF-based gap-filling. Rather, it may be a tool to extrapolate traits from point measurements to the regional scale, i.e. perform out-of-sample predictions.

Is the prediction accuracy sufficient for applications in ecological contexts?

Our results indicate that BHPMF outperforms the species MEAN baseline in all aspects, but is the prediction accuracy indeed appropriate for use in ecological contexts, and to what extent are the results a special case of our test dataset?

The test dataset is an extract of 78,300 individual observations and the 13 best covered traits from the TRY database, still with 80% of trait entries missing. Our dataset is not typical for common datasets of traits, which generally originate from specific measurement campaigns with only 5–30% missing entries. In these cases, gap-filling may not be such a challenge and common approaches may be sufficient. However, given the experience that BHPMF substantially outperforms both PMF and MEAN, we also expected improved trait predictions in such cases. Indeed, our sensitivity test using ‘cut outs’ from Europe and data across the Amazon showed that BHPMF was able to impute gaps even in smaller and extremely sparse databases.

When taking advantage of using the global database to fill gaps in a smaller ‘cut out’, possible confounding factors introduced by global trait variability influencing the sometimes more constrained trait space of a local dataset should be considered. For example in the case of the RAINFOR cut-out, plant height was better predicted using just the local dataset. One likely explanation is that a large amount of new information was included from the global database, amounting to more than 60% more data with a high standard deviation within species. For SSD, on the other hand, only 30% more data were contributed by the global dataset, which were also less plastic within species compared with plant height. Thus, a global dataset may easily introduce ‘false’ information in the case of plastic traits that are highly influenced by local environmental conditions but provide a lot of valuable additional information in the case of traits that are mainly determined by phylogeny. We recommend, depending on the number and sparsity of the data, using both approaches wherever possible – local and global gap-filling – and select the best-fitting approach depending on the target trait and aim.

BHPMF trait prediction uncertainties are well correlated with their error, turning uncertainty into a good surrogate for prediction error (Fig. S10.2 in Appendix S10). This offers the opportunity to select trait predictions with high probability of low errors or weight results in further analyses according to their uncertainty.

Perspectives

Statistical approaches to gap-filling and improvement of ecological understanding of species occurrence and dynamics have been criticized for being too complex and including parameters and assumptions without appropriate prior validation (Lavine, 2010). In contrast, BHPMF is a simple and generic model, which has the advantage of incorporating factors known to influence trait expression, such as phylogenetic trait signals, trait–trait correlations and trade-offs without implicitly requiring prior assumptions. This simplicity opens the opportunity to add complexity to improve trait predictions, for example by replacing the taxonomic hierarchy by a hierarchy of phylogenetic distances. BHPMF has been explicitly developed for gap-filling (matrix completion), not for the prediction of trait values in individuals where no trait has ever been measured before, i.e. out-of-sample predictions. Advancing BHPMF to account for phylogenetic distances instead of taxonomy will improve opportunities for efficient out-of-sample predictions, at least in well-resolved clades.

Uncertainty quantified predictions support the validity of the BHPMF approach, where predictions with low uncertainty (small SD) also show high accuracy (small RMSE), and vice versa. Using the Gibbs sampler to get an indication as to where uncertainties are largest merits special attention. Data coverage for traits such as SLA, which are relatively fast and easy to collect, is much better than for other traits. In addition, some traits are highly variable across ecosystems and plant populations. Using the Gibbs sampler, one can statistically define where more sampling effort is probably going to significantly improve our
understanding of variation in any specific trait and where the currently available data are sufficient.

The combination of BHPMF with environmental covariates may not necessarily improve gap-filling of trait matrices, but seems to be a promising concept for the extrapolation of traits from point measurements to regional scales. The concept of out-of-sample prediction presented here illustrates a fundamental problem in trait ecology, separating the phylogenetic signal and environmental impact on intraspecific trait variation. The regression of trait values against the environment after gap-filling by BHPMF provides an advantage over using non-gap-filled data of an improved number of data points, which might be essential as trait predictions are often limited by data availability (Verheijen et al., 2013). BHPMF-filled trait matrices could thus become an invaluable tool for the parameterization and validation of global vegetation models.

The TRY database is a typical example of a combination of datasets which have been collected for different purposes. Other disciplines besides plant ecology which have also started to combine their (trait) data are faced with the same problem of sparsity and restrictions in the context of multivariate analyses. BHPMF provides an opportunity to extract the entire information content from such databases combining data from various aspects measured at locations all over the world.

Conclusion

BHPMF is a hierarchical Bayesian implementation of probabilistic matrix factorization, which for the first time simultaneously utilizes the taxonomic trait signal, the correlation structure within trait matrices and – implicitly through trait–environment relationships – environmental constraints for gap-filling of trait matrices. We demonstrate using the example of different plant trait datasets that BHPMF provides robust and accurate predictions even for sparse matrices. In addition, Gibbs sampler-calculated uncertainties indicate how accurate each imputed trait value is, and thus how to treat it in further analyses. The combination of BHPMF with environmental information provides the opportunity to extrapolate from point measurements to continuous trait surfaces across large spatial scales whilst accounting for intraspecific trait variability. We therefore conclude that BHPMF-based gap-filling and trait prediction has a high potential to support future trait-based research in macroecology and functional biogeography.

ACKNOWLEDGEMENTS

F.S. was supported by the University of Minnesota, Institute on the Environment via the grant to P.R.: ‘Transformational steps in synthesis science’, the Max Planck Institute for Biogeochemistry and iDiv, the German Centre for Integrative Biodiversity Research. H.S. was supported by NSF grants IIS-0812183, IIS-0916750, IIS-1029711, IIS-1017647, and NSF CAREER award IIS-0953274. This study has been performed with support by the TRY initiative on plant traits (https://www.try-db.org). TRY is hosted and developed at the Max Planck Institute for Biogeochemistry, with support from DIVERSITAS and iDiv. We thank Ulrich Weber for help with the preparation of soil data and Maarten Braakhekke, Nate Swenson and two anonymous referees for helpful comments and suggestions.

REFERENCES


Verheijen, L.M., Brovkin, V., Aerts, R., Bönisch, G., Cornelissen, J.H.C., Kattge, J., Reich, P.B., Wright, I.J. & van Bodegom, P.M. (2013) Impacts of trait variation through observed

Additional references are in the supplementary file at: (weblink)

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article at the publisher’s web-site.

Appendix S1 Supplementary methods.
Appendix S2 Definition of traits used in this study.
Appendix S3 References for contributing databases and number of traits contributed.
Appendix S4 Map of TRY measurement sites.
Appendix S5 Location of *Acer saccharum* range map and soil and climate across the range of *Acer saccharum*
Appendix S6 Correlation between traits and environmental variables used in aHPMF.
Appendix S7 Root mean squared error comparison between MEAN, BHPMF and aHPMF across the taxonomic hierarchy.
Appendix S8 Sensitivity analysis.
Appendix S9 Bi- and multivariate relationships between traits, measured and imputed trait values.
Appendix S10 Gibbs sampler results.
Appendix S11 Additional references of data contributors.
Appendix S12 Author contributions.

**BIOSKETCH**

Franziska Schrodt is a post-doctoral researcher at the Max Planck Institute for Biogeochemistry in Jena and the German Centre for Integrative Biodiversity Research iDIV Leipzig, Jena, Halle. Her work focuses on the application of machine learning and nonlinear statistical tools to the study of biogeochemical patterns. She is especially interested in plant functional trait/biodiversity–environment correlations and the associated implications for ecosystem structure and functioning.

Editor: José Alexandre Diniz-Filho