Received: 31 March 2021 | Accepted: 17 July 2021

DOI: 10.1111/2041-210X.13687

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

# A new joint species distribution model for faster and more accurate inference of species associations from big community data

Maximilian Pichler | Florian Hartig

Theoretical Ecology, University of Regensburg, Regensburg, Germany

**Correspondence**
Maximilian Pichler
Email: maximilian.pichler@biologie.uni-regensburg.de

**Handling Editor:** Gavin Simpson

## Abstract

1. Joint species distribution models (JSDMs) explain spatial variation in community composition by contributions of the environment, biotic associations and possibly spatially structured residual covariance. They show great promise as a general analytical framework for community ecology and macroecology, but current JSDMs, even when approximated by latent variables, scale poorly on large datasets, limiting their usefulness for currently emerging big (e.g. metabarcoding and metagenomics) community datasets.
2. Here, we present a novel, more scalable JSDM (sjSDM) that circumvents the need to use latent variables by using a Monte Carlo integration of the joint JSDM likelihood together with flexible elastic net regularization on all model components. We implemented sjSDM in PyTorch, a modern machine learning framework, which allows making use of both CPU and GPU calculations. Using simulated communities with known species–species associations and different number of species and sites, we compare sjSDM with state-of-the-art JSDM implementations to determine computational runtimes and accuracy of the inferred species–species and species–environment associations.
3. We find that sjSDM is orders of magnitude faster than existing JSDM algorithms (even when run on the CPU) and can be scaled to very large datasets. Despite the dramatically improved speed, sjSDM produces more accurate estimates of species association structures than alternative JSDM implementations. We demonstrate the applicability of sjSDM to big community data using eDNA case study with thousands of fungi operational taxonomic units (OTU).
4. Our sjSDM approach makes the analysis of JSDMs to large community datasets with hundreds or thousands of species possible, substantially extending the applicability of JSDMs in ecology. We provide our method in an R package to facilitate its applicability for practical data analysis.

KEYWORDS

big data, co-occurrence, machine learning, metacommunity, regularization, statistics

## 1 | INTRODUCTION

Understanding the structure and assembly of ecological communities is a central concern for ecology, biogeography and macroecology (Vellend, 2010). The question is tightly connected to important research programs of the field, including coexistence theory (see Chesson, 2000; e.g. Levine et al., 2017), the emergence of diversity patterns (e.g. Pontarp et al., 2019) or understanding ecosystem responses to global change (Urban et al., 2016).

The statistical analysis of spatial community data is currently dominated by two major ecological frameworks: metacommunity theory (see Leibold et al., 2004) and species distribution models (SDMs, Elith & Leathwick, 2009). Metacommunity theory formed in the last two decades as the study of the spatial processes that give rise to regional community assembly (e.g. Leibold & Chase, 2017; Leibold et al., 2004). Most current metacommunity analyses rely on ordination (e.g. Leibold & Mikkelson, 2002) or variation partitioning (e.g. Cottenie, 2005) techniques, which disentangle abiotic and spatial contributions to community assembly (see Leibold & Chase, 2017). SDMs are statistical models that link abiotic covariates to species occurrences. They are widely used in spatial ecology, for example to study invading species (Gallien et al., 2012; Mainali et al., 2015) or species responses to climate change (Thuiller et al., 2006).

A key limitation of both variation partitioning and SDMs, noted in countless studies, is that they do not account for species interactions. Both approaches essentially assume that species depend only on space and the environment (Cottenie, 2005; Dormann et al., 2012; Peres-Neto & Legendre, 2010; Wisz et al., 2013), whereas we know that in reality, species can also influence each other through competition, predation, facilitation and other processes (Gilbert & Bennett, 2010; Van der Putten et al., 2010; see Mittelbach & Schemske, 2015; see Leibold & Chase, 2017).

Joint species distribution models (JSDM) recently emerged as a novel analytical framework promising to integrate species interactions into metacommunity and macroecology (Leibold et al., 2021). JSDMs are similar to SDMs in that they describe species occurrence as a function of the environment, but additionally consider the possibility of species–species associations. By an association, we mean that two species tend to appear together more or less often than expected from their environmental responses alone. Whether those association originate from biotic interactions (e.g. competition, predation, parasitism, mutualism) or other reasons (e.g. unmeasured environmental predictors) needs to be carefully considered (see Blanchet et al., 2020; Dormann et al., 2018; König et al., 2021; Poggiato et al., 2021). Still, when appropriately interpreted, JSDMs combine the essential processes believed to be responsible for the assembly of ecological communities—environment, space and biotic interactions—and they could be applied to large scale as well as for regional metacommunity analyses (e.g. Gilbert & Bennett, 2010; Leibold & Chase, 2017; Mittelbach & Schemske, 2015).

Recent interest in JSDMs was further fuelled by the emergence of high-throughput technologies that are currently revolutionizing our capacities for observing community data (e.g. Pimm et al., 2015). We can now detect hundreds or even thousands of species from environmental DNA (eDNA) or bulk-sampled DNA (Yu et al., 2012; Bohmann et al., 2014; Cristescu, 2014; Deiner et al., 2017; see Bálint et al., 2018; Barsoum et al., 2019; Humphreys et al., 2019; Tikhonov, Duan, et al., 2020) in a given sample, and next generation sequencing (NGS) has become cheap enough that this process could be replicated at scale. Other emerging technologies will likely also produce large amounts of community data, such as automatic species recognition (Guirado et al., 2018; e.g. Tabak et al., 2019) from acoustic recordings. Recent studies have used these methods to generate community inventories of fish (see Desjonquères et al., 2019; e.g. Picciulin et al., 2019), forest wildlife (e.g. see Wrege et al., 2017), bird communities (Fritzler et al., 2017; Lasseck, 2018; Wood et al., 2019) or bats (e.g. Mac Aodha et al., 2018). Jointly, these developments mean that large spatial community datasets will become available in the near future, and ecologists have to consider how to best analyse them.

Joint species distribution models would seem the natural analytical approach for these emerging new data, given their ability to separate the essential processes for spatial community assembly. Current JSDM software, however, has severe limitations for processing such large (and/or wide) datasets. Early JSDMs were based on the multivariate probit (MVP) model (Chib & Greenberg, 1998), which describes species–species associations via a covariance matrix (e.g. Clark et al., 2014; Golding et al., 2015; Hui, 2016; Ovaskainen et al., 2010; Pollock et al., 2014). The limitation of the MVP approach is that it scales poorly for species-rich data, as the number of parameters in the species–species covariance matrix increases quadratically with the number of species (see Warton et al., 2015).

The current solution to this problem is latent variable models (LVMs), which replace the covariance matrix with a small number of latent variables (see Warton et al., 2015). The LVM reparameterization makes the estimation of MVP models computationally more efficient (see Niku et al., 2019; Norberg et al., 2019; Ovaskainen, Tikhonov, Norberg, et al., 2017; Tikhonov et al., 2017; Tikhonov, Duan, et al., 2020; Warton et al., 2015). That, however, does not mean that simultaneously estimating species' abiotic preferences and species–species associations with LVMs is fast. Integrating out the latent variables requires MCMC sampling or numerical approximations (e.g. Laplace, variational inference, see Niku et al., 2019), which is computationally costly and can fail to converge. For communities with hundreds of species, computational runtimes of current LVMs can still exceed hours or days (e.g. Tikhonov, Duan, et al., 2020; Wilkinson et al., 2019). This poses severe limitations for analysing eDNA data, which can include thousands of species or operational taxonomic units (OTUs, e.g. Frøslev et al., 2019). Moreover, LVMs also scale disadvantageously with the number of sites, because each site introduces additional parameters in the latent variables (Bartholomew et al., 2011; Skrondal & Rabe-Hesketh, 2004). Thus, the advantage of the LVM over the full-MVP model decreases with increasing numbers of sites (on the order of thousands). An important challenge for

the field is therefore to make JSDMs fast enough for big datasets (Krapu & Borsuk, 2020).

A second question for JSDM development is the accuracy of inferred species associations. Surprisingly little is known about this question. Most existing JSDM assessments (Norberg et al., 2019; e.g. Tobler et al., 2019; Wilkinson et al., 2019) concentrate on runtime, predictive performance or on aggregated measures of accuracy that do not necessarily capture the error of the estimated species-species association structure (but see Zurell et al., 2018). From a statistical perspective, however, it is clear that estimating a large species covariance matrix with limited data must have considerable error, and it would be desirable to better understand the dependence on this error on the structure of the data and the chosen modelling approach.

The LVM approach, specifically, not only makes the models faster, but also reduces the number of free parameters (see Warton et al., 2015), which should theoretically reduce the variance (and thus the error) of the species–species covariance estimates, but possibly at the cost of a certain bias. The trade-off between bias and variance is controlled by the number of latent variables—when the number of latent variables is similar to the number of species, the LVM will be as flexible (and unbiased) as the full-MVP model. The fewer latent variables are used, the stronger the reduction in variance and the potential increase in bias. In practice, the number of latent variables is usually chosen much smaller than the number of species (the highest value we saw was 32 with hundreds of species in Tikhonov, Duan, et al., 2020), which means that JSDMs fitted currently by LVMs could show biases due to the regularization induced by the LVM structure (Stein, 2014).

While trading off some bias against a reduction in variance is fundamental to all regularization approaches, and no concern as such, it seems important to understand the nature of the bias that is created by the LVM structure and examine if alternative regularization structures are more appropriate. Similar to LVMs, spatial models for large data often use a low-rank approximation of the covariance matrix (e.g. Stein, 2007, 2014; e.g. Sang et al., 2011). For Gaussian process models, it has been shown that this approximation captures the overall structure well (in the sense that the magnitude of covariances is captured well), but at the costs of larger errors in local structures (see Stein, 2014). We conjecture that LVMs with a small number of latent variables behave analogous—with a few latent variables, it will be difficult to model a specific covariance structure without unintentionally introducing other covariances elsewhere, but it could be possible to generate a good approximation of the overall correlation level between species.

Here, we propose a new method for estimating JSDMs, called scalable JSDM, that addresses many of the above-mentioned problems. By using a Monte Carlo approach [originally proposed by Chen et al. (2018)] that can be outsourced to graphical processing units (GPUs), sjSDM is able to fit JSDMs with a full covariance structure extremely fast, without having to resort to latent variables. To address the issue of overfitting due to the increased number of parameters compared to state-of-the-art latent variable JSDMs, we introduce a new regularization approach, which directly targets the covariance matrix of the full-MVP model. Additionally, we propose a method for optimizing the regularization strength based on tuning the parameter under a k-fold cross-validation.

To demonstrate the beneficial properties of the new model structure, we assess: (a) its computational runtime on GPUs and CPUs, (b) the accuracy of inferred species–species associations and species' environmental responses and (c) its predictive performance. We compare the performance of sjSDM to several state-of-the-art JSDM software packages (Hmsc, gllvm and BayesComm; see also Harris, 2015; Clark et al., 2017; Vieilledent & Clément, 2019), as well as results from a recent JSDM comparison (Wilkinson et al., 2019). Finally, to illustrate the applicability of our approach to wide community data, we additionally applied our model to a community eDNA dataset containing 3,649 fungi OTUs over 125 sites.

## 2 | MATERIALS AND METHODS

### 2.1 | The structure of the JSDM problem

Species–environment associations are classically addressed by SDMs, which estimate the expected probability of the presence of a species as a function of the environmental predictors. The functional response to the environment can be expressed by GLMs, or by more flexible (i.e. nonlinear and/or nonparametric) approaches such as generalized additive models, boosted regression trees or Random Forest (Elith & Leathwick, 2009; e.g. Ingram et al., 2020).

A JSDM generalizes this approach by including the possibility of residual species–species correlations (in the literature usually called species–species associations). The most common JSDM structure is the MVP model, which describes the site by species matrix $Y_{ij}$ ($Y_{ij} = 1$ if species $j = 1, \ldots, J$ is present at site $i = 1, \ldots, I$ or $Y_{ij} = 0$ if species $j$ is absent) as a function of the environmental covariates $X_{in}$ ($n = 1, \ldots, N$ covariates), and the covariance matrix (species associations) $\Sigma$ accounts for correlations in $e_{ij}$:

$$Z_{ij} = \beta_{j0} + \sum_{n=1}^{N} X_{in} * \beta_{nj} + e_{ij},$$

$$Y_{ij} = 1(Z_{ij} > 0),$$

$$\boldsymbol{e}_i \sim \mathrm{MVN}(0, \Sigma). \quad (1)$$

For the results, we normalized the fitted species–species covariance matrix $\Sigma$ to a correlation matrix.

### 2.2 | Current approaches to fit the JSDM structure

The model structure described in Equation 1 can be fitted directly using the probit link, and the first JSDMs used this approach (Chib & Greenberg, 1998; Pollock et al., 2014; see Wilkinson et al., 2019). Fitting the full-MVP model directly, however, has two drawbacks:

first, calculating likelihoods for large covariance matrices is computationally costly. Second, the number of parameters in the covariance matrix for $j$ species increases quadratically as $j * (j-1)/2$. For 50 species, for example, we would have to estimate 2,250 covariance parameters.

Because of these problems, a series of papers (Ovaskainen et al., 2016; Warton et al., 2015) introduced the LVM (see Skrondal & Rabe-Hesketh, 2004) to the JSDM problem. The latent variable JSDM approximates the species–species covariance by introducing a number of latent covariates (=latent variables), which act exactly like real environmental covariates, except that their values are estimated as well. Species that react (via their factor loadings) similarly or differently to the latent variables thus show positive or negative associations respectively (see Ovaskainen, Tikhonov, Norberg, et al., 2017; Warton et al., 2015; Wilkinson et al., 2019 for details). The factor loadings can be translated into a species–species covariance matrix: $\Sigma = \lambda * \lambda^T$ ($\lambda$ = matrix of factor loadings). The latent variables can be interpreted as unobserved environmental predictors, but they can also be viewed as a purely technical approach to regularized low-rank reparameterization of the covariance matrix. One advantage of the LVM is that the latent variables can be used for constrained (LVM with environmental predictors) and unconstrained ordination (LVM without environmental predictors; Warton et al., 2015). The complexity of the association structure can be set via the number of latent variables (usually to a low number, see Warton et al., 2015).

## 2.3 | An alternative approach to fit the JSDM structure

Because LVMs still have computational limitations, and because of the need for flexible regularization discussed in the Introduction, we propose a different approach to fit the model structure in Equation 1. The full-MVP assumes that the observed binary occurrence vector $\boldsymbol{Y}_i \in \{0,1\}^J$ arises as the sign of the latent Gaussian variable $\boldsymbol{Y}_i^* \sim N(\boldsymbol{X}_i\beta, \Sigma)$:

$$Y_{ij} = \mathbb{1}\left(Y_{ij}^* > 0\right), \tag{2}$$

where $\beta$ is the environmental coefficient matrix and $\sum$ the covariance matrix. Then the probability to observe $\boldsymbol{Y}_i$ is:

$$Pr\left(\boldsymbol{Y}_i|\boldsymbol{X}_i\beta,\Sigma\right) = \int_{A_{iJ}} \dots \int_{A_{i1}} \phi_J\left(\boldsymbol{Y}_i^*;\boldsymbol{X}_i\beta,\Sigma\right) d\boldsymbol{Y}_{i1}^* \dots d\boldsymbol{Y}_{iJ}^*, \tag{3}$$

with the interval $A_{ij}$ defined as:

$$A_{ij} = \begin{cases} (-inf, 0] & Y_{ij} = 0 \\ [0, +inf] & Y_{ij} = 1 \end{cases}, \tag{4}$$

and $\phi$ being the density function of the multivariate normal distribution. The main computational issue of the full-MVP (Equation 3) is that calculating the probability of $\boldsymbol{Y}_i$ requires to integrate over $\boldsymbol{Y}_i^*$, which has no closed analytical expression for more than two species ($J > 2$). This makes the evaluation of the likelihood computationally costly when $J \gg 1$ and motivates the search for an efficient numerical approximation of Equation 3.

To see how this approximation can be achieved, note that Equation 3 can be expressed more generally as:

$$\mathscr{L}\left(\beta,\Sigma;Y_i,X_i\right) = \int_{\Omega} \prod_{j=1}^{J} Pr\left(Y_{ij}|X_i\beta+\xi\right) Pr\left(\xi|\Sigma\right) d\xi. \tag{5}$$

In sjSDM, we approximate this integral by $M$ Monte Carlo samples from the multivariate normal species–species covariance. With the covariance term being integrated out, we can calculate the remaining part of the likelihood as in a univariate case, and use the average of the $M$ samples to get an approximation of Equation 5

$$\mathscr{L}\left(\beta,\Sigma;\boldsymbol{Y}_i,\boldsymbol{X}_i\right) \approx \frac{1}{M}\sum_{m=1}^{M}\prod_{j=1}^{J} P\left(Y_{ij}|\boldsymbol{X}_i\beta+\xi_m\right),$$

$$\xi_m \sim \text{MVN}(0,\Sigma). \tag{6}$$

This approximation of the MVP was first proposed by Chen et al. (2018) in the context of fitting deep neural networks with an MVP response structure. Its most notable computational advantage over other existing approximations to the MVP problem, such as the Geweke-Hajivassiliou-Keane (GHK) algorithm (Hajivassiliou & Ruud, 1994), is that the Monte Carlo sampling in Equation 6 can be parallelized. This is especially efficient when performing calculations on GPUs rather than CPUs, due to their much higher number of cores (see also Golding, 2019, who similarly uses GPUs to improve an expensive computational tasks in ecology). The GHK algorithm, on the other hand, is based on a recursive and thus non-independent importance sampling procedure, which means that the sampling cannot be parallelized.

For sjSDM, we implemented this approximation, which was previously only used in the deep learning literature, to the generalized linear MVP, which means that we conform to the model structure typically used in this field and can profit from all benefits associated with parametric models. The only difference to a standard MVP is that we use an approximation of the probit link, which we found to be numerically more stable than the analytical probit link (see Supporting Information S1 for details).

We implemented the method in an R package (https://github.com/TheoreticalEcology/s-jSDM), using the Python package PyTorch, which is designed for deep learning (Paszke et al., 2019), and the R package reticulate, which allows us to run PyTorch from within R (Ushey et al., 2019). This setup allows us to leverage various sophisticated numerical algorithms from PyTorch, including the possibility to switch between efficiently parallelized CPU and GPU calculations, and the ability to obtain analytical gradients (via automatic derivatives) of the MVP likelihood with the latent covariance structure marginalized out via the Monte Carlo ensemble. The combination of efficient parallelization and analytical derivatives of the

Monte Carlo approximated likelihood makes finding the maximum likelihood estimate (MLE) for the full-MVP model extremely fast, despite the large number of parameters to optimize.

Outsourcing the Monte Carlo approach to a GPU solves the issue of computational speed (as we show below), but it does not yet solve the problem that the covariance matrix has a very large number of parameters, which raises the problem of overfitting when the method is used on small datasets. To address this, we penalized the actual covariances in the species–species covariance matrix, as well as the environmental predictors, with a combination of ridge and lasso penalty (elastic net, see Zou & Hastie, 2005, more details below). Our R package includes a function to tune the strength of the penalty for each model component separately via cross-validation.

The here-tested implementation of sjSDM only considers binary (presence–absence) data, but there are several ways to extend the approach to count and proportional data as well. To a large extent, these are already implemented in our R package. Currently supported are count (Poisson distribution with log-link), presence–absence (binomial distribution with logit and probit links) and normal data (multivariate normal distribution). Moreover, the model-based ordination that is popular for latent variable JSDMs is currently not implemented in sjSDM and probably challenging to achieve, since the model is fit without latent environmental variables. However, new ordination techniques with a focus on co-occurrence patterns (e.g. Popovic et al., 2019) could complement sjSDM in practical analyses.

## 2.4 | Benchmarking our method against state-of-the-art JSDM implementations

To benchmark our approach, we fit sjSDM to six datasets from a recent JSDM benchmark study by Wilkinson et al. (2019) (Table S1). Covariates were centred and standardized. To be able to compare our results to theirs across different hardware, we also reran BayesComm, the fastest JSDM in their study, with the same parameters as in the study by Wilkinson et al. (2019) on our hardware.

Additionally, we simulated new data from an MVP (Equation 2), varying the number of sites from 50 to 500 (50, 70, 100, 140, 180, 260, 320, 400, 500) and the number of species as a percentage (10%, 30% and 50%) of the sites (e.g. the scenario with 100 sites and 10% results in 10 species). In all simulations, the species' environmental preference was described for five environmental covariates (beta), which was randomly selected. Each scenario was sampled five times. Here, all species had species–species associations, that is the species–species covariance matrices were not sparse (for details, see Supporting Information S1).

To compare our model to existing JSDM software packages, we selected BayesComm (version 0.1-2, Golding & Harris, 2015), the fastest MVP-based JSDM according to the study by Wilkinson et al. (2019), and two state-of-the-art latent-variable JSDM implementations: Hmsc (version 3.0-4, Tikhonov et al., 2019b), which uses MCMC sampling, and gllvm (version 1.2.1, Niku et al., 2020), which uses variational Bayes and Laplace approximation to fit the model. We used the default parameter settings for all three methods which were in line with other recent JSDM benchmarks (details see Supporting Information S1).

Since GPUs might be not commonly available, we calculated sjSDM results both on the CPU and on the GPU. To estimate the influence of the number of Monte Carlo samples on the error of the MVP approximation, we used 100 Monte Carlo samples for each species when run on the CPU and 1,000 Monte Carlo samples for each species when run on the GPU for sjSDM. In the following, we will refer to sjSDM when run on the GPU as GPU-sjSDM, and when run on the CPU as CPU-sjSDM

To assess the predictive performance of the models, we calculated the average area under the curve (AUC) over all species and five independent replicates for each scenario of a hold-out dataset (same size as the dataset used for fitting the model). The AUC measures the capability of the model to distinguish between absence and presence of species. To calculate the accuracy of the estimated species associations and environmental coefficients, we used root mean squared error and the accuracy of the coefficients' signs, again averaged over all species and replicates.

To additionally explore the ability of sjSDM to infer community assembly processes from more realistic ecological data, we simulated communities from the process-based ecological model used by Leibold et al. (2021) and compared the inferred species–species association structures with the true structures for sjSDM, BayesComm, Hmsc and gllvm. For details, see Supporting Information S1.

## 2.5 | Regularization to infer sparse species–species associations

For the benchmark described above, we simulated data under the assumption that all species interact. While this assumption may or may not be realistic, it is generally desirable for a method to work well also when there is only a small number of associations, that is when the species–species covariance matrix is sparse. We were particularly interested in this question, because we conjectured that the LVM approach imposes correlations on the species–species associations that makes it difficult for LVMs to fit arbitrarily sparse covariance structures.

We therefore simulated data under the same scenarios as before, but with 95% sparsity in the species–species associations. To adjust our model to such a sparse structure, we applied an elastic net shrinkage (Zou & Hastie, 2005) to all off-diagonals of the covariance matrix. Following Zou & Hastie, 2005, the parameters lambda (the regularization strength) and alpha (the weighting between LASSO and ridge) of the elastic net were tuned via fivefolded cross-validation in 40 random steps. As species are correlated within sites, we blocked the CV in sites. For real data, one could additionally consider a spatial blocking (Roberts et al., 2017) to account for correlations between sites (e.g. by using the blockCV package, Valavi et al., 2019).

For the cross-validation, we used 2,000 samples for the MVP approximation in sjSDM, because we found that the approximation error can introduce stochasticity in the tuning process. For BayesComm, Hmsc, and gllvm, we used the default settings (see details and additional comments in Supporting Information S1). For Hmsc, following Tikhonov, Opedal, et al., 2020, associations with >95% posterior probability being positive or negative were set to zero.

To measure the accuracy of inferred species–species associations for this benchmark, we normalized the covariance matrices to correlation matrices and calculate the true skill statistic (TSS = Sensitivity + Specificity − 1, Allouche et al., 2006) by transforming the true and predicted associations into two classes: all absolute associations smaller than 0.01 were assigned to class '0' and all absolute associations >0.01 were assigned to class '1'. That way, a two-class classification problem was obtained and the TSS was calculated.

### 2.6 | Case study – Inference of species–species associations from eDNA

To demonstrate the practical applicability of our approach, we fitted our model to an eDNA community dataset from a published study that sampled 130 sites across Denmark (for details on the study design, see Brunbjerg et al., 2017; for data and bioinformatics, see Frøslev et al., 2019). On each site, eight environmental variables were recorded: precipitation, soil pH, soil organic matter, soil carbon content, soil phosphorous content and mean Ellenberg's indicator values (light condition, nutrient status and soil moisture) based on the plant community. Frøslev et al., 2019 identified 10,490 OTUs by eDNA sequencing (81 samples per site). We followed Frøslev et al., 2019 and removed five sites with <4 OTU presences (low species richnesss). We used only OTUs occurring at least three times over the remaining 125 sites, which reduced the overall number of OTUs from 10,490 to 3,649 OTU. All eight environmental variables were used in our analysis as main effects on the linear scale. The final dataset consisted of 3,649 OTU co-occurrences over 125 sites with eight environmental variables.

For this analysis, we set the regularization for the z-transformed environmental predictors to lambda = 0.1 and alpha = 0.5 (equal weighting of ridge and LASSO regularization). The regularization for the covariances of the species–species associations was tuned over 40 random steps (independent samples from the hyper-parameter space) and with leave-one-out cross-validation. For each of the resulting 40 × 125 = 5,000 evaluations, we fitted a GPU-sjSDM in 150 iterations (with a batch size of 12 and 125 site, one iteration consists of 100 optimization steps, see Bottou, 2010), 3,649 × 3,649 weights for the covariance matrix (see Supporting Information S1 for details about the parametrization of the covariance matrix in sjSDM), with batch size of 8 and learning rate of 0.001 (the size of the update of the parameters in one optimization step).

## 3 | RESULTS

### 3.1 | Method validation and benchmark against state-of-the-art JSDMs

#### 3.1.1 | Computational speed

On a GPU, our approach (GPU-sjSDM) required under 3-s runtime for any of our simulated data with 50–500 sites and 5–250 species. When run on CPUs only (CPU-sjSDM), runtimes increased to a maximum of around 2 min (Figure 1a; Figure S1). In comparison, Hmsc had a runtime of around 7 min for our smallest scenario and increase in runtime exponentially when the number of species exceeded 40 (Figure 1a). BayesComm was slightly faster than Hmsc, but scaled worse than Hmsc to large data sizes (Figure S1). gllvm achieved low runtimes, equivalent and sometimes better

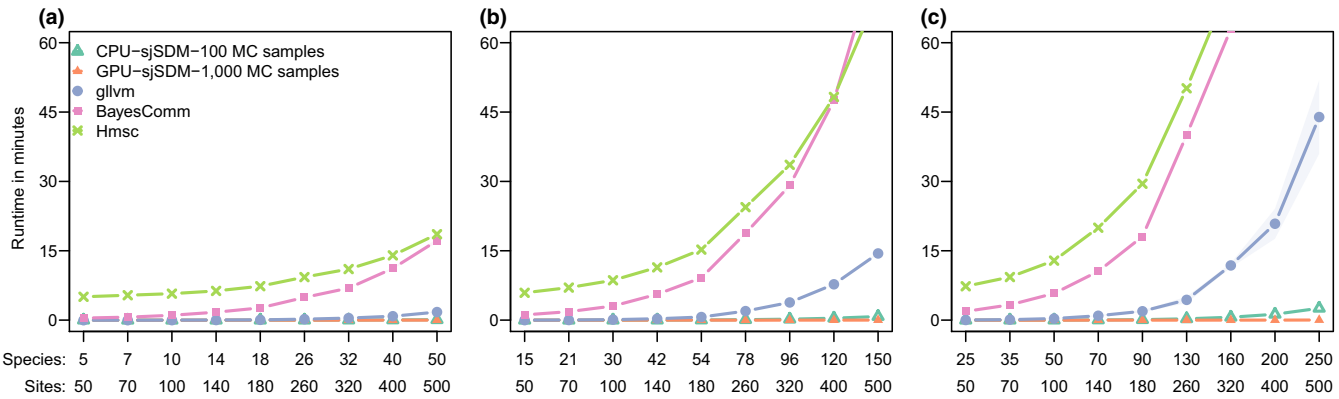


**FIGURE 1** Runtime benchmarks for GPU-sjSDM, CPU-sjSDM, gllvm, BayesComm and Hmsc fitted to simulated data with 50–500 sites (dense species–species association matrices) and the number of species set to (a) 0.1, (b) 0.3 and (c) 0.5 times the number of sites. All values are averages from five simulated datasets. To estimate the inference error of the Monte Carlo approximation, GPU-sjSDM was fitted with 1,000 and CPU-sjSDM with 100 MC samples for each species. sjSDM, Scalable joint species distribution model

than our method for small data (<50 species), but for larger data, runtime started to increase exponentially as well, leading to runtimes of approximately 45 min for our most demanding scenario (Figure 1a).

Because of the runtime limitations of the other approaches, we calculated big data benchmarks only for GPU-sjSDM. The overall runtimes for GPU-sjSDM increased from under 1 min for 5,000 sites to a maximum of around 4.5 min for 30,000 sites (Figure 2). GPU-sjSDM showed greater runtime increases when increasing numbers of sites, while the numbers of species (300, 500 and 1,000 species in each scenario) had only small effects on runtimes (Figure 2).

For the empirical benchmarking datasets from the study by Wilkinson et al. (2019), CPU-sjSDM achieved a 3.8 times lower runtime for the bird dataset and 23 times lower runtime for the butterfly dataset, and GPU-sjSDM achieved a 500 times lower runtime for the bird dataset and a 150 times lower runtime for the butterfly dataset compared to BayesComm, the fastest JSDM in the study by Wilkinson et al. (2019) (Table 1).

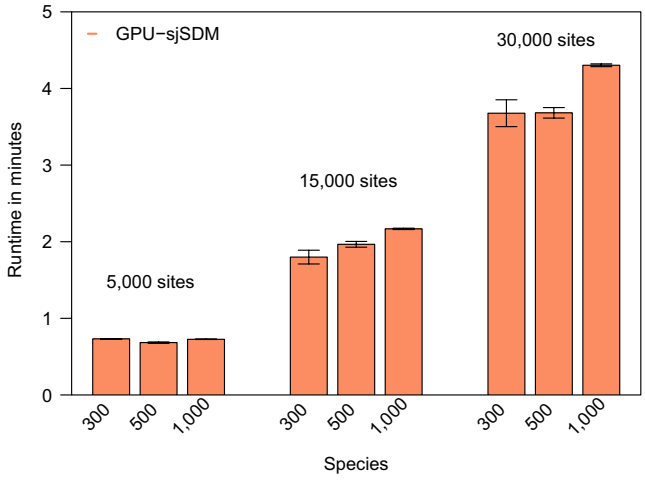


**FIGURE 2** Benchmark results for sjSDM on big community data. We simulated communities with 5,000, 15,000 and 30,000 sites and for each set of 300, 500 and 1,000 species. sjSDM, scalable joint species Distribution model

### 3.1.2 | Accuracy of the inference about species–environment and species–species associations

For simulated data with dense species–species association structures, BayesComm and sjSDM consistently achieved higher accuracy in the inferred species–species associations than the LVMs Hmsc and gllvm (Figure 3a–c). The accuracy of all methods decreased with an increasing proportion of species, to around 70% for the full-MVP models (sjSDM and BayesComm) and 60% for the LVMs (Figure 3a–c). Even for communities with 300 to 1,000 species, sjSDM achieved accuracies of 69% and higher (Table S4).

For environmental preferences (measured by RMSE), Hmsc showed slightly higher inferential performance when the number of sites was low (Figure S4a,b) while all models performed approximately equal for a high number of sites (Figure S4a,b).

For simulated data with sparse species–species association structures (95% sparsity), sjSDM achieved the highest TSS (up to 0.35–0.38 with 30% and 50% species, see Figure 3d–f). Hmsc showed for 10% species the second highest TSS (Figure 3d–f), but for 30% and 50% species together with gllvm the lowest TSS (a maximum of 0.1 TSS for 30% and 50% species). BayesComm showed in average the lowest TSS for 10% species, but for 30% and 50% species the second highest TSS (Figure 3d–f). The inferential performance regarding the environmental predictors showed the same pattern as for dense species–species associations. All models improved their environmental accuracy (Figure S4c) and reduced RMSE as the number of sites increased (Figure S4d).

Fitting sjSDM to data simulated with the process-based simulation model used in the study by Leibold et al. (2021), we find, similar to Leibold et al. (2021), that important signals of the underlying processes, including biotic interactions, can be recovered by sjSDM (Figure S10). Our results also hint towards certain advantages of MVP JSDMs over LVMs for this task, although we caution that this question will require further exploration. For details, see Supporting Information S1.

### 3.1.3 | Predicting species occurrences

All models performed similarly in predicting species occurrences in the simulation scenarios, with predictive accuracies of around 0.75 AUC (Figure 4).

**TABLE 1** Model runtimes in hours. Results for BayesComm against our new approach scalable joint species distribution model (sjSDM) (CPU and GPU version)

| | Wilkinson et al. (2019) | | Our approach | |
|---|---|---|---|---|
| **Dataset** | **Size (site × species)** | **BayesComm** | **CPU-sjSDM** | **GPU-sjSDM** |
| Birds (Harris, 2015) | 2,752 × 370 | 3.5 | 0.97 | 0.007 |
| Butterflies (Ovaskainen et al., 2016) | 2,609 × 55 | 0.15 | 0.01 | 0.001 |
| Eucalypts (Pollock et al., 2014) | 458 × 12 | <0.01 | <0.001 | <0.001 |
| Frogs (Pollock et al., 2014) | 104 × 9 | <0.002 | <0.001 | <0.001 |
| Fungi (Ovaskainen et al., 2010) | 800 × 11 | <0.02 | <0.001 | <0.001 |
| Mosquitos (Golding, 2015) | 167 × 16 | <0.01 | <0.001 | <0.001 |

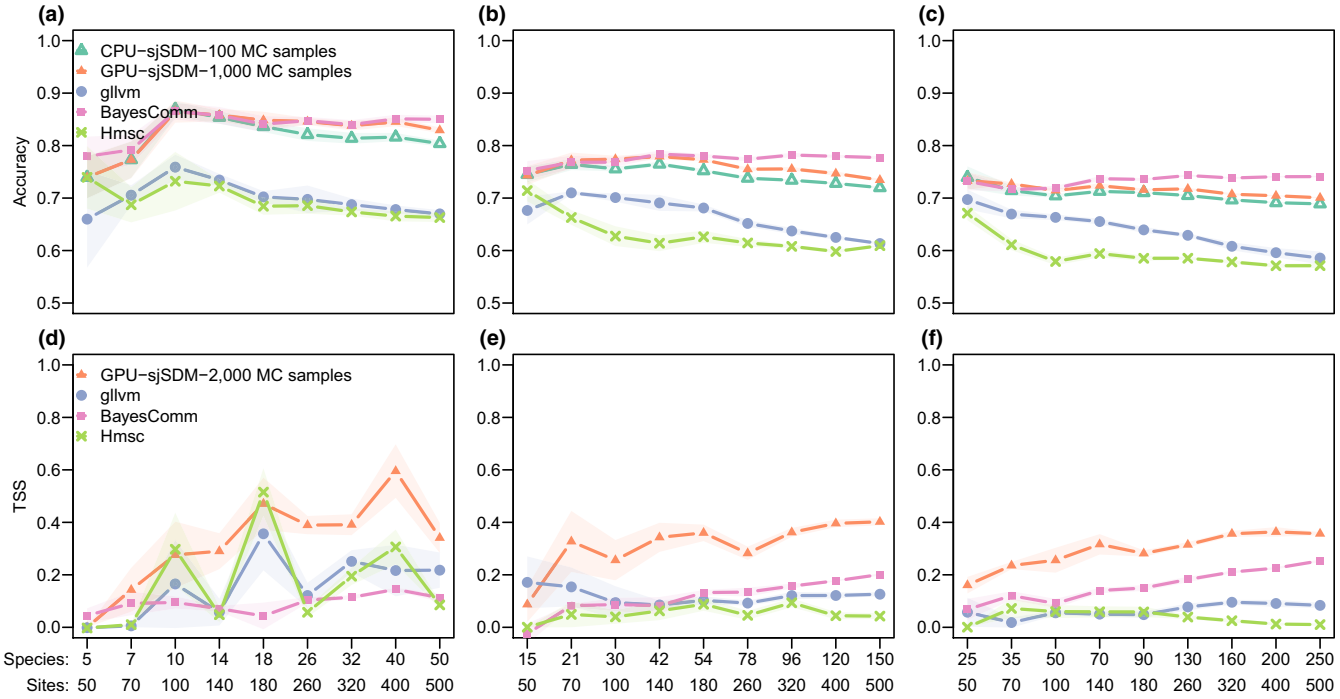


**FIGURE 3** Inference performance of the inferred sparse and non-sparse species–species associations. Models were fitted to simulated data with 50 to 500 sites. All values are averages from five simulated datasets. (a–c) The upper row shows the accuracies of matching signs (positive or negative covariance) for the estimated and true dense species–species association matrix. (d–f) The lower row shows the accuracy of inferring non-zero species associations for sparse association matrices (95% sparsity), measured by the true skill statistic (absolute associations smaller than 0.01 were assigned the class '0' and absolute associations >0.01 were assigned the class '1'). The number of species for were set to 0.1 (a, d), 0.3 (b, e) and 0.5 (c, f) times the number of sites. To estimate the inference error of the Monte Carlo (MC) approximation, GPU-sjSDM was fitted with 1,000 and CPU-sjSDM with 100 MC samples for each species

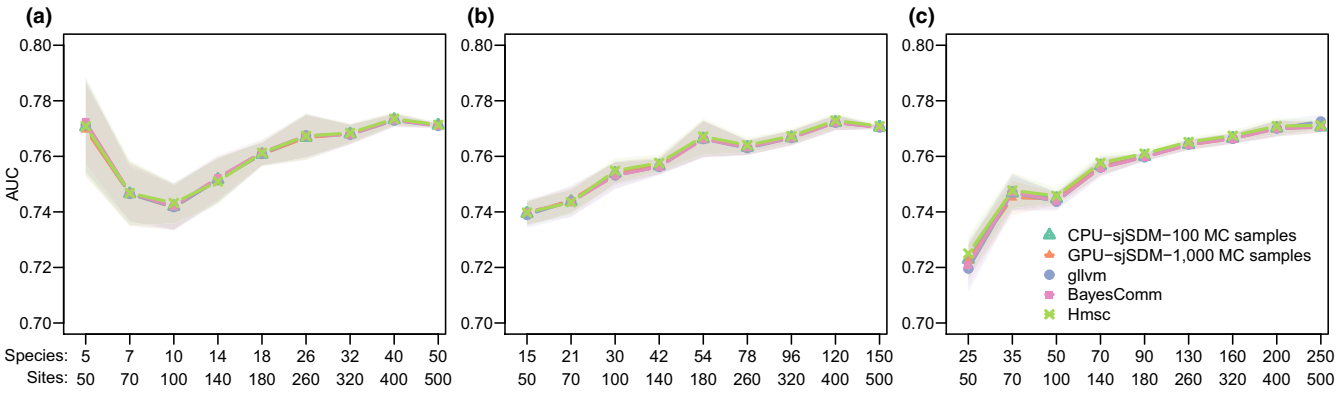


**FIGURE 4** Predictive performance in simulated species distributions for GPU-sjSDM and CPU-sjSDM with gllvm, BayesComm and Hmsc as references. Species distribution scenarios with (a) 50–500 sites and 10%, (b) 30% and (c) 50% species were simulated, on which the models were fitted (training). Models predicted species distributions for additional 50–500 sites (testing). Area under the curve (AUC) was used to evaluate predictive performance on hold-out. sjSDM, scalable soint species distribution model

## 3.2 | Case Study – Inference of species–species associations from eDNA

In our eDNA case study with 3,649 OTUs over 125 sites, we found that without regularization, sjSDM inferred the strongest negative OTU–OTU covariances among the most abundant species and the strongest positive OTU–OTU associations among the rarest OTUs (Figure 5a,b). When optimizing the regularization strength for the OTU–OTU associations via a leave-one-out cross-validation, positive and negative OTU–OTU associations changed somewhat, but the overall pattern stayed qualitatively constant (Figure 5a,b). For the environmental covariates (a weak non-optimized regularization was used), we found that most OTUs showed the highest dependency on Ellenberg F (moisture), Ellenberg L (light availability) and Ellenberg N (nitrogen).

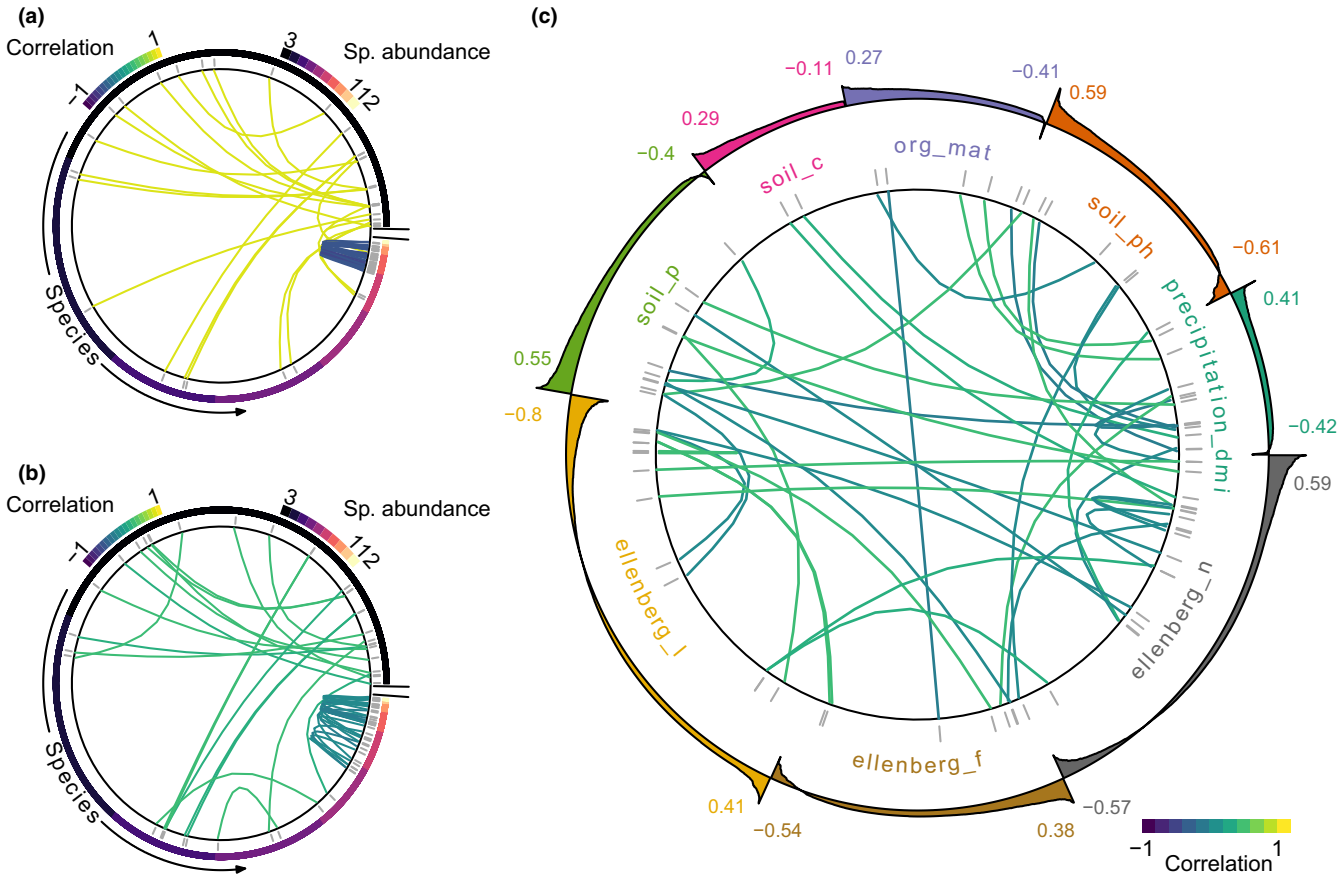


**FIGURE 5** Inferred operational taxonomic unit (OTU) associations and environmental preferences for the eDNA community data. The left column (panels a–c) shows OTU–OTU associations for (a) no regularization and (b) tuned regularization, with the 3,649 OTUs sorted according to their summed abundance over 125 sites. The large panel (c) shows the covariance structure of (b), but with OTUs sorted after their most important environmental coefficients (largest absolute environmental effect size; the outer ring shows the environmental effect distribution for the OTUs within the group)

## 4 | DISCUSSION

Joint species distribution models extend standard species distribution models by also accounting for species–species associations. Current JSDM software, however, exhibits computational limitations for large community matrices, which limits their use for big community data that are created by novel methods such as eDNA studies and metabarcoding. Here, we presented sjSDM, a new numerical approach for fitting JSDMs that uses Monte Carlo integration of the model likelihood, which allows moving calculations to GPUs. We show that this approach is orders of magnitude faster than existing methods (even when run on the CPU) and predicts as well as any of the other JSDM packages that we used as a benchmark. To avoid overfitting, especially when fitting sjSDM to hitherto computationally unrealistic eDNA datasets with thousands of species, we introduced a flexible elastic net regularization on species associations and environmental preferences. sjSDM inferred the signs of full association matrices and identified zero/non-zero entries in sparse species–species associations across a wide range of scenarios better than all tested alternatives. Advantages of BayesComm and sjSDM over LVM-based JSDMs (Hmsc and gllvm) occurred for all species–species associations structures tested, while improvement of sjSDM over BayesComm was in particular visible for sparse species–species associations.

### 4.1 | Computational performance

Whereas runtimes for Hmsc, BayesComm and gllvm started to increase exponentially when the number of species exceeded around 100, sjSDM scaled close to linearly with the number of species regardless of whether we used GPU or CPU computations (Figure 1a). A further advantage of sjSDM is that, unlike in particular the MCMC algorithms used in BayesComm and Hmsc, it is highly parallelizable, which allows using efficiently the advantages of modern computer hardware such as GPUs. These two properties, scalability and parallelizability, make sjSDM the first and currently only JSDM software package that seems capable of analysing big eDNA datasets (Humphreys et al., 2019; Tikhonov, Duan, et al., 2020; Wilkinson et al., 2019) on standard computers with acceptable runtimes.

We concede that runtimes of the different JSDM implementations may depend on hyper-parameters such as the number of MCMC

iterations in BayesComm or Hmsc, or the number of MC samples in sjSDM. Changing these parameters could affect results; however, increasing or decreasing MCMC iterations would only linearly shift the runtime curves (Figure 1; Figure S1). When we compare such a linear shift with the strong nonlinearity scaling of BayesComm and Hmsc, it seems unlikely that changes to the hyper-parameters could qualitatively change the results. Moreover, sjSDM uses a Monte Carlo approximation of the likelihood and runtime, thus naturally depends on number of Monte Carlo samples. Yet, all other tested methods use approximations as well to obtain the inference. Neither our inferential results nor other indicators give us reasons to think that the approximation made by sjSDM is worse than that of competing algorithms. Specifically, increasing the number of Monte Carlo samples for each species in sjSDM from 100 to 1,000 increased the inferential performance moderately (Figure 3a–c). Also, the excellent inferential accuracy of sjSDM across various tests does not suggest a large approximation error. We are therefore confident that our Monte Carlo approximation is acceptable in general, and not worse than the approximations made in other packages.

State-of-the-art JSDM implementations offer a variety of extensions such as the inclusion of phylogeny, space and traits (e.g. Hmsc, Tikhonov, Opedal, et al., 2020). Here, we used sjSDM only for estimating a simple MVP structure, which is arguably the most generic version of a JSDM that is implemented by all packages. In principle, however, the algorithm used in sjSDM could be extended to include other structures that have been proposed in the literature. The sjSDM package already supports alternative responses and link functions (e.g. normal, Poisson or binomial), and has an option to add spatial model components (e.g. via spatial eigenvectors). Also the option to include traits by using the fourth-corner approach as in gllvm (Brown et al., 2014; Niku et al., 2019) could be added. A crucial question for all these extensions is if they interact beneficially with our MLE approximation, that is if we can optimize the MLE without having to resort to other integration methods (such as MCMC or Laplace approximations) for the added structures, which would negate the speed advantage of sjSDM. For example, we found that the approximation used by sjSDM does not interact well with the addition of conditional autoregressive (CAR) terms in the model structure.

## 4.2 | Inferential performance

All JSDM implementations showed similar performance in correctly inferring environmental responses, but the MVP approaches, sjSDM and BayesComm, achieved significantly higher accuracy in inferring the correct signs of species–species associations (Figure 3a–c) and identifying sparse structures (Figure 3d–f). It should be noted here that we tuned the regularization of sjSDM to improve the performance for sparse associations and the other JSDM might also benefit from tuning the regularization. BayesComm and Hmsc allow more restrictive priors to be specified on the covariance matrix (BayesComm) or on the factor loadings (Hmsc). However, the long runtimes of these JSDM implementations place time constraints on testing different prior specification. Moreover, BayesComm already achieved high TSS for sparse associations with default specifications, indicating superiority of highly parametrized JSDM over LVM for sparse structures (Figure 3d–f).

We speculate that the LVMs' lower performance for the inferred species associations originates from the constraints imposed by the LVM structure, which creates some bias that showed in particular for dense species association structures (compare Figure 2a; Figure S2). This is not particularly surprising, as similar phenomena have been found also for other approaches to covariance regularizations, for example in spatial models (Stein, 2014). It is difficult to estimate how important these biases are in practical applications, because we still know too little about the typical structure of species associations in real ecological data (Ovaskainen, Tikhonov, Dunson, et al., 2017). One might expect that associations in data generated by high-throughput technologies, which detect species already at very low densities, would be relatively sparse, or consist of a mix between sparse and non-sparse blocks for rare and common species (cf. Calatayud et al., 2019). Moreover, one would expect that LVMs would be particularly efficient if species associations follow the structure implemented in the LVMs. To test this, we also simulated data from an LVM structure, and fitted these data with sjSDM and the two LVMs (gllvm and Hmsc). Our results show that the LVMs indeed perform better than for such data than for our previously used general covariance matrices, but not better than sjSDM (Figures S7, S8, and S9).

A slight disadvantage of sjSDM is that it is more complicated to obtain parameter uncertainties, compared to JSDM implementations based on MCMC sampling such as BayesComm and Hmsc. The R implementation of sjSDM calculates Wald confidence intervals for all environmental predictors using PyTorch's automatic differentiation feature. However, we have currently no analytical option to calculate confidence intervals for the species–species associations. If these are needed, we propose using bootstrap samples.

## 4.3 | Implications and outlook for ecological data analysis

The JSDM structure has the potential to become the new default statistical approach for species and community observations that originate from eDNA and similar big community data. However, to fulfil this promise, we need statistical algorithms that scale to big datasets and deliver accurate inference, in particular for a large number of species or operational taxonomic units. Our results show that a combination of a scalable and parallelizable Monte Carlo approximation of the likelihood, together with a shrinkage regularization of the species–species covariance, can achieve both goals.

Our results also suggest that regularization of the species–species covariance is particularly crucial to obtain reasonable inference for such data. In principle, all software packages that we compared could include additional regularization methods, such as the elastic net employed in our approach. Better understanding the

use of such statistical approaches is one promising route for further research. Another option would be to impose ecologically motivated structures on the species–species covariance matrix (e.g. Bystrova et al., 2021; Clark et al., 2017; Taylor-Rodríguez et al., 2017).

Another interesting question is how ecologist should use and interpret JSDMs, once they scale to big data. Many recent studies have stressed that JSDMs may improve predictions (e.g. Norberg et al., 2019), and indeed, from ecological theory, one would expect that species associations are important for accurate species occurrence predictions (Dormann et al., 2012; Norberg et al., 2019; Wisz et al., 2013). Despite different accuracy in inferring true species associations (Figure 3), we found similar predictive performances (Figure 4) for all tested JSDMs. It should be noted, however, that the AUC metric we used captures only marginal predictive performance, and a closer relationship between inferential and predictive performance might have arisen when using joint predictive performance measures (Wilkinson et al., 2021).

Another open question in the context of predictions is the relative importance of including the association structure, compared to a more detailed description of the environmental model components. Without systematic benchmarks, where model structures on both biotic and abiotic predictions are flexibly adopted (e.g. via machine learning approaches such as in Chen et al., 2018), and where indicators of joint predictive performance (Wilkinson et al., 2021) are used that are sensitive to covariances, it is difficult to examine whether increases in predictive performance of JSDMs are really due to their exploitation of a stable association structure, or simply arise from the higher model complexity of JSDMs, which allows fitting the data more flexibly.

When turning to inference, the new information that JSDMs deliver to ecologists are species–species covariance estimates (Leibold et al., 2021). These could be used, for example, to test if the strength or structure of species associations varies with space or environmental predictors; or if spatial species associations correlate with local trophic or competitive interactions or traits (see generally Poisot et al., 2015). For regional studies, there is the prospect of extending the traditional variation partitioning (environment and space; Cottenie, 2005) to include biotic associations by using JSDMs (Leibold et al., 2021). Our results regarding the moderate, but significantly better than random accuracy of inferred covariance structures, even on datasets with hundreds of species, are encouraging for such a research program.

Recently, however, concerns about the usefulness of JSDM for examining species interactions have emerged. For instance, it has been criticized that the species–species associations inferred by JSDM cannot always be linked to ecological interactions because of their symmetric nature (Blanchet et al., 2020; Poggiato et al., 2021; Zurell et al., 2018), that the associations may absorb missing environmental covariates (Poggiato et al., 2021) or that JSDM associations can be scale dependent (see König et al., 2021; although this also applies to ecological interactions, see Poisot et al., 2015). We acknowledge these observations but do not share all concerns. JSDM estimate associations between species after accounting for the environment. Such associations are not necessarily causal or mechanistic, and they are naturally also influenced by unmeasured predictors, scale and other factors, but they can also be caused by real species interactions, as shown in the study by Leibold et al. (2021) and confirmed by us for sjSDM (Figure S10). Thus, when interpreted with due care, JSDMs provide useful ecological information beyond pure niche models. If more high-resolution dynamic data were available, we could use more precise (causal) methods to infer the direction of interactions (Barraquand et al., 2021; Momal et al., 2020), which likely match much closer to actual species interactions. Yet, for the static community data that make up the bulk of the data available to ecologists today, these methods are not applicable, but JSDMs are and can provide additional information compared to existing alternatives.

## 5 | CONCLUSIONS

We presented sjSDM, a new method to fit JSDMs, and benchmarked it against state-of-the-art JSDM software. sjSDM is orders of magnitudes faster than current alternatives, and it can be flexibly regularized, which leads to overall superior performance in inferring the correct species association structure. We emphasize that the superior scaling holds also when using CPU computations, and that the possibility to move calculations on a GPU is only a further advantage of the algorithm. We provide our tool in an R package (https://github.com/TheoreticalEcology/s-jSDM, available for Linux, MacOS and Windows), with a simple and intuitive interface and the ability to switch easily between linear and nonlinear modelling, as well as between CPU and GPU computing. The R package also includes extensions for considering abundance data as well as spatial coordinates, and to partition the importance of space, environment and species associations for predicting the observed community composition.

### ACKNOWLEDGEMENTS

The authors thank Douglas Yu and Yuanheng Li, as well as Gavin Simpson and four anonymous reviewers for their valuable comments and suggestions.

### CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

### AUTHORS' CONTRIBUTIONS

F.H. and M.P. jointly conceived and designed the study; M.P. implemented the sjSDM software, ran the experiments and analysed the data. Both authors contributed equally to discussing and interpreting the results, and to the preparation of the manuscript.

### DATA AVAILABILITY STATEMENT

The processed datasets for runtime benchmarking (case study 1) are available as Supporting Information for Wilkinson et al., 2019. The eDNA dataset is available at https://github.com/tobiasgf/man_vs_machine. The analysis and the version of the R package sjSDM

used in this analysis are available in an online repository (Pichler & Hartig, 2021). The latest version of the sjSDM R package can be found at https://github.com/TheoreticalEcology/s-jSDM.

## PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13687.

## ORCID

*Maximilian Pichler* https://orcid.org/0000-0003-2252-8327
*Florian Hartig* https://orcid.org/0000-0002-6255-9059

## REFERENCES

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Bálint, M., Pfenninger, M., Grossart, H.-P., Taberlet, P., Vellend, M., Leibold, M. A., Englund, G., & Bowler, D. (2018). Environmental DNA time series in ecology. *Trends in Ecology & Evolution*, *33*, 945–957. https://doi.org/10.1016/j.tree.2018.09.003

Barraquand, F., Picoche, C., Detto, M., & Hartig, F. (2021). Inferring species interactions using Granger causality and convergent cross mapping. *Theoretical Ecology*, *14*, 87–105. https://doi.org/10.1007/s12080-020-00482-7

Barsoum, N., Bruce, C., Forster, J., Ji, Y.-Q., & Yu, D. W. (2019). The devil is in the detail: Metabarcoding of arthropods provides a sensitive measure of biodiversity response to forest stand composition compared with surrogate measures of biodiversity. *Ecological Indicators*, *101*, 313–323. https://doi.org/10.1016/j.ecolind.2019.01.023

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, *23*, 1050–1063. https://doi.org/10.1111/ele.13525

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*, 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Physica-Verlag HD.

Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., & Gibb, H. (2014). The fourth-corner solution–using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, *5*, 344–352. https://doi.org/10.1111/2041-210X.12163

Brunbjerg, A. K., Bruun, H. H., Moeslund, J. E., Sadler, J. P., Svenning, J.-C., & Ejrnæs, R. (2017). Ecospace: A unified framework for understanding variation in terrestrial biodiversity. *Basic and Applied Ecology*, *18*, 86–94. https://doi.org/10.1016/j.baae.2016.09.002

Bystrova, D., Poggiato, G., Bektaş, B., Arbel, J., Clark, J. S., Guglielmi, A., & Thuiller, W. (2021). Clustering species with residual covariance matrix in joint species distribution models. *Frontiers in Ecology and Evolution*, *9*, 601384.

Calatayud, J., Andivia, E., Escudero, A., Melián, C. J., Bernardo-Madrid, R., Stoffel, M., Aponte, C., Medina, N. G., Molina-Venegas, R., Arnan, X., Rosvall, M., Neuman, M., Noriega, J. A., Alves-Martins, F., Draper, I., Luzuriaga, A., Ballesteros-Cánovas, J. A., Morales-Molino, C., Ferrandis, P., … Madrigal-González, J. (2019). Positive associations among rare species and their persistence in ecological assemblages. *Nature Ecology & Evolution*, *1*, 40–45.

Chen, D., Xue, Y., & Gomes, C. P. (2018). End-to-end learning for the deep multivariate probit model. *arXiv:1803.08591 [cs, Stat]*.

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, *31*, 343–366. https://doi.org/10.1146/annurev.ecolsys.31.1.343

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*, 347–361. https://doi.org/10.1093/biomet/85.2.347

Clark, J. S., Gelfand, A. E., Woodall, C. W., & Zhu, K. (2014). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, *24*, 990–999. https://doi.org/10.1890/13-1015.1

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., & Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, *87*, 34–56. https://doi.org/10.1002/ecm.1241

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics: Meta-analysis of metacommunities. *Ecology Letters*, *8*, 1175–1182. https://doi.org/10.1111/j.1461-0248.2005.00820.x

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*, 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*, 5872–5895.

Desjonquères, C., Gifford, T., & Linke, S. (2019). Passive acoustic monitoring as a potential tool to survey animal and ecosystem processes in freshwater environments. *Freshwater Biology*. https://doi.org/10.1111/fwb.13356

Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., Moretti, M. D., Pagel, J., Pinkert, S., Schleuning, M., Schmidt, S. I., Sheppard, C. S., Steinbauer, M. J., Zeuss, D., & Kraan, C. (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography*, *27*, 1004–1016. https://doi.org/10.1111/geb.12759

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., & Singer, A. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography*, *39*, 2119–2131. https://doi.org/10.1111/j.1365-2699.2011.02659.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fritzler, A., Koitka, S., & Friedrich, C. M. (2017). Recognizing bird species in audio files using transfer learning. *LEF (Working Notes)*, *14*.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Hansen, A. J., Læssøe, T., & Heilmann-Clausen, J. (2019). Man against machine: Do fungal fruitbodies and eDNA give similar biodiversity assessments across broad environmental gradients? *Biological Conservation*, *233*, 201–212. https://doi.org/10.1016/j.biocon.2019.02.038

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N. E., & Thuiller, W. (2012). Invasive species distribution models – How violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography*, *21*, 1126–1136. https://doi.org/10.1111/j.1466-8238.2012.00768.x

Gilbert, B., & Bennett, J. R. (2010). Partitioning variation in ecological communities: Do the numbers add up? *Journal of Applied Ecology*, *47*, 1071–1082. https://doi.org/10.1111/j.1365-2664.2010.01861.x

Golding, N. (2015). *Mosquito community data for Golding et al. 2015 (parasites & vectors)*. figshare.

Golding, N. (2019). greta: Simple and scalable statistical modelling in R. *Journal of Open Source Software*, *4*, 1601. https://doi.org/10.21105/joss.01601

Golding, N., & Harris, D. J. (2015). *BayesComm: Bayesian community ecology analysis*.

Golding, N., Nunn, M. A., & Purse, B. V. (2015). Identifying biotic interactions which drive the spatial distribution of a mosquito community. *Parasites & Vectors*, *8*, 367. https://doi.org/10.1186/s13071-015-0915-1

Guirado, E., Tabik, S., Rivas, M. L., Alcaraz-Segura, D., & Herrera, F. (2018). Automatic whale counting in satellite images with deep learning. *bioRxiv*.

Hajivassiliou, V. A., & Ruud, P. A. (1994). Classical estimation methods for LDV models using simulation. *Handbook of Econometrics*, *4*, 2383–2441.

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, *6*, 465–473. https://doi.org/10.1111/2041-210X.12332

Hui, F. K. C. (2016). boral – Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, *7*, 744–750.

Humphreys, J. M., Murrow, J. L., Sullivan, J. D., & Prosser, D. J. (2019). Seasonal occurrence and abundance of dabbling ducks across the continental United States: Joint spatio-temporal modelling for the Genus Anas. *Diversity and Distributions*, *25*, 1497–1508.

Ingram, M., Vukcevic, D., & Golding, N. (2020). Multi-output Gaussian processes for species distribution modelling. *Methods in Ecology and Evolution*, *11*, 1587–1598. https://doi.org/10.1111/2041-210X.13496

König, C., Wüest, R. O., Graham, C. H., Karger, D. N., Sattler, T., Zimmermann, N. E., & Zurell, D. (2021). Scale dependency of joint species distribution models challenges interpretation of biotic interactions. *Journal of Biogeography*, *48*, 1541–1551. https://doi.org/10.1111/jbi.14106

Krapu, C., & Borsuk, M. (2020). A spatial community regression approach to exploratory analysis of ecological data. *Methods in Ecology and Evolution*, *11*, 608–620. https://doi.org/10.1111/2041-210X.13371

Lasseck, M. (2018). Audio-based bird species identification with deep convolutional neural networks. *Working Notes of CLEF*, 2018.

Leibold, M. A., & Chase, J. M. (2017). *Metacommunity ecology*. Princeton University Press.

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., Holt, R. D., Shurin, J. B., Law, R., Tilman, D., Loreau, M., & Gonzalez, A. (2004). The metacommunity concept: A framework for multi-scale community ecology. *Ecology Letters*, *7*, 601–613. https://doi.org/10.1111/j.1461-0248.2004.00608.x

Leibold, M. A., & Mikkelson, G. M. (2002). Coherence, species turnover, and boundary clumping: Elements of meta-community structure. *Oikos*, *97*, 237–250. https://doi.org/10.1034/j.1600-0706.2002.970210.x

Leibold, M. A., Rudolph, J., Blanchet, F. G., Meester, L. D., Gravel, D., Hartig, F., Peres-Neto, P., Shoemaker, L., & Chase, J. M. (2021). The internal structure of metacommunities. *bioRxiv*, 2020.07.04.187955.

Levine, J. M., Bascompte, J., Adler, P. B., & Allesina, S. (2017). Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*, *546*, 56–64. https://doi.org/10.1038/nature22898

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., & Jones, K. E. (2018). Bat detective–Deep learning tools for bat acoustic signal detection. *PLOS Computational Biology*, *14*, e1005995.

Mainali, K. P., Warren, D. L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B. B., & Parmesan, C. (2015). Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. *Global Change Biology*, *21*, 4464–4480. https://doi.org/10.1111/gcb.13038

Mittelbach, G. G., & Schemske, D. W. (2015). Ecological and evolutionary perspectives on community assembly. *Trends in Ecology & Evolution*, *30*, 241–247. https://doi.org/10.1016/j.tree.2015.02.008

Momal, R., Robin, S., & Ambroise, C. (2020). Tree-based inference of species interaction networks from abundance data. *Methods in Ecology and Evolution*, *11*, 621–632. https://doi.org/10.1111/2041-210X.13380

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2020). *gllvm: Generalized linear latent variable models*. CRAN.

Niku, J., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, *10*, 2173–2182.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*, e01370. https://doi.org/10.1002/ecm.1370

Ovaskainen, O., Abrego, N., Halme, P., & Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, *7*, 549–555. https://doi.org/10.1111/2041-210X.12501

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*, 2514–2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Sæther, B.-E., & Abrego, N. (2017). How are species interactions structured in species-rich communities? A new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences*, *284*, 20170768. https://doi.org/10.1098/rspb.2017.0768

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*, 561–576. https://doi.org/10.1111/ele.12757

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Garnett, R. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (Vol. *32*, pp. 8024–8035). Curran Associates, Inc.

Peres-Neto, P. R., & Legendre, P. (2010). Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography*, *19*, 174–184. https://doi.org/10.1111/j.1466-8238.2009.00506.x

Picciulin, M., Kéver, L., Parmentier, E., & Bolgan, M. (2019). Listening to the unseen: Passive acoustic monitoring reveals the presence of a cryptic fish species. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *29*, 202–210. https://doi.org/10.1002/aqc.2973

Pichler, M., & Hartig, F. (2021). Pichler & Hartig 2021 – A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Zenodo*. https://zenodo.org/record/5131594

Pimm, S. L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R., & Loarie, S. (2015). Emerging technologies to conserve biodiversity. *Trends in Ecology & Evolution*, *30*, 685–696. https://doi.org/10.1016/j.tree.2015.08.008

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. *Trends in Ecology & Evolution*. https://doi.org/10.1016/j.tree.2021.01.002

Poisot, T., Stouffer, D. B., & Gravel, D. (2015). Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, *124*, 243–251. https://doi.org/10.1111/oik.01719

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*, 397–406.

Pontarp, M., Bunnefeld, L., Cabral, J. S., Etienne, R. S., Fritz, S. A., Gillespie, R., Graham, C. H., Hagen, O., Hartig, F., Huang, S., Jansson, R., Maliet, O., Münkemüller, T., Pellissier, L., Rangel, T. F., Storch, D., Wiegand, T., & Hurlbert, A. H. (2019). The latitudinal diversity gradient: Novel understanding through mechanistic eco-evolutionary models. *Trends in Ecology & Evolution*, *34*, 211–223. https://doi.org/10.1016/j.tree.2018.11.009

Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K., & Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, *10*, 1571–1583. https://doi.org/10.1111/2041-210X.13247

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Sang, H., Jun, M., & Huang, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *The Annals of Applied Statistics*, *5*, 2519–2548. https://doi.org/10.1214/11-AOAS478

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Stein, M. L. (2007). Spatial variation of total column ozone on a global scale. *The Annals of Applied Statistics*, *1*, 191–210. https://doi.org/10.1214/07-AOAS106

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, *8*, 1–19. https://doi.org/10.1016/j.spasta.2013.06.003

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., … Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, *10*, 585–590. https://doi.org/10.1111/2041-210X.13120

Taylor-Rodríguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S., & Gelfand, A. E. (2017). Joint species distribution modeling: Dimension reduction using Dirichlet processes. *Bayesian Analysis*, *12*, 939–967. https://doi.org/10.1214/16-BA1031

Thuiller, W., Lavorel, S., Sykes, M. T., & Araújo, M. B. (2006). Using niche-based modelling to assess the impact of climate change on tree functional diversity in Europe. *Diversity and Distributions*, 49–60. https://doi.org/10.1111/j.1366-9516.2006.00216.x

Tikhonov, G., Abrego, N., Dunson, D., & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, *8*, 443–452. https://doi.org/10.1111/2041-210X.12723

Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D., & Ovaskainen, O. (2020). Computationally efficient joint species distribution modeling of big spatial data. *Ecology*, *101*(2). https://doi.org/10.1002/ecy.2929

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, *11*, 442–447.

Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedal, O., & Dallas, T. (2019b). *Hmsc: Hierarchical model of species communities*. CRAN. https://CRAN.R-project.org/package=Hmsc

Tobler, M. W., Kéry, M., Hui, F. K. C., Guillera-Arroita, G., Knaus, P., & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, *100*, e02754. https://doi.org/10.1002/ecy.2754

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Peer, G., Singer, A., Bridle, J. R., Crozier, L. G., De Meester, L., Godsoe, W., Gonzalez, A., Hellmann, J. J., Holt, R. D., Huth, A., Johst, K., Krug, C. B., Leadley, P. W., Palmer, S. C. F., Pantel, J. H., … Travis, J. M. J. (2016). Improving the forecast for biodiversity under climate change. *Science*, *353*. https://doi.org/10.1126/science.aad8466

Ushey, K., Allaire, J., & Tang, Y. (2019). *Reticulate: Interface to 'Python'*. CRAN. https://CRAN.R-project.org/package=reticulate

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, *10*, 225–232.

Van der Putten, W. H., Macel, M., & Visser, M. E. (2010). Predicting species distribution and abundance responses to climate change: Why it is essential to include biotic interactions across trophic levels. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 2025–2034.

Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, *85*, 183–206. https://doi.org/10.1086/652373

Vieilledent, G., & Clément, J. (2019). *jSDM: Joint species distribution models*. CRAN. https://CRAN.R-project.org/package=jSDM

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, *10*, 198–211. https://doi.org/10.1111/2041-210X.13106

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2021). Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution*, *12*, 394–404. https://doi.org/10.1111/2041-210X.13518

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N. M., … Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, *88*, 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Wood, C. M., Gutiérrez, R. J., & Peery, M. Z. (2019). Acoustic monitoring reveals a diverse forest owl community, illustrating its potential for basic and applied ecology. *Ecology*, *100*, e02764. https://doi.org/10.1002/ecy.2764

Wrege, P. H., Rowland, E. D., Keen, S., & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: Examples from forest elephants. *Methods in Ecology and Evolution*, *8*, 1292–1301. https://doi.org/10.1111/2041-210X.12730

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*, 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, *41*, 1812–1819. https://doi.org/10.1111/ecog.03315

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Pichler, M., & Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, 00, 1–15. https://doi.org/10.1111/2041-210X.13687