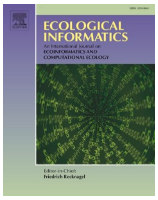
Ecological Informatics 83 (2024) 102826

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Transfer learning of species co-occurrence patterns between plant communities

Johannes Hirn[a], Verónica Sanz[b,c], José Enrique García[b], Marta Goberna[d], Alicia Montesinos-Navarro[a], José Antonio Navarro-Cano[d], Ricardo Sánchez-Martín[a], Alfonso Valiente-Banuet[e,f], Miguel Verdú[a,*]

[a] Centro de Investigaciones Sobre Desertificación (CIDE), CSIC-Universidad de Valencia-Generalitat Valenciana, Valencia, Spain
[b] Instituto de Física Corpuscular (IFIC), Universidad de Valencia-CSIC, Valencia, Spain
[c] Department of Physics and Astronomy, University of Sussex, Sussex, UK
[d] Departamento de Medio Ambiente y Agronomía, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), CSIC, Madrid, Spain
[e] Departamento de Ecología de la Biodiversidad, Instituto de Ecología, Universidad Nacional Autónoma de México, México D.F., Mexico
[f] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, México D.F., Mexico

ARTICLE INFO

Keywords:
Generative artificial intelligence
Patchy vegetation
Plant communities
Restoration ecology
Species co-occurrence
Variational autoencoders

ABSTRACT

Aim: The use of neural networks (NNs) is spreading to all areas of life, and Ecology is no exception. However, the data-hungry nature of NNs can leave out many small, valuable datasets. Here we show how to apply transfer learning to rescue small datasets that can be invaluable in understanding patterns of species co-occurrence.
Location: Semiarid plant communities in Spain and México.
Time period: 2016–2022.
Major taxa studied: Angiosperms.
Methods: Based on a large sample of plant species co-occurrence in vegetation patches in a semi-arid area of eastern Spain, we fit a generative artificial intelligence (AI) model that correctly reproduces which species live with which in these patches. Subsequently, we train the same type of model on two communities for which we only have smaller datasets (another semi-arid community in eastern Spain, and a tropical community in Mexico).
Results: When we transfer the knowledge learnt from the large dataset directly to the other two, the predictions improve for the community more similar to our reference one. As for the more dissimilar community, improving the accuracy of the transfer requires a further tuning of the model to the local data. In particular, the knowledge transferred relates primarily to species frequency and, to a lesser extent, to their phylogenetic relationships, which are known to be determinants of species interaction patterns.
Main conclusions: This AI-based approach can be performed for communities similar or not so similar to the reference community, opening the door to systematic transfer learning for accurate predictions on small datasets. Interestingly, this transfer operates by matching unrelated species between the origin and target datasets, implying that arbitrary datasets can then be transferred to, or even combined in order to augment each other, irrespective of the species involved, potentially allowing such models to be applied to a wide range of plant communities in different climates.

1. Introduction

Artificial intelligence (AI) is spreading to an increasing number of aspects of everyday life, including science Shen (2018). Ecology is no stranger to this trend, as new AI applications facilitate the understanding of entangled life Pichler and Hartig (2023), although not without the limits imposed by ecological complexity Lapeyrolerie and Boettiger (2023).

Comprehending the mechanisms behind species co-occurrence represents a fundamental challenge within the field of ecology, given its pivotal role in the preservation of biodiversity Chesson (2000). Ecologists have tried to explain how biodiversity is structured using information that might predict which species interact with each other Bascompte (2009). In particular, phylogenetic relatedness has been used

---

* Corresponding author at: Centro de Investigaciones sobre Desertificacion -CIDE- (CSIC/UV/GV), Carretera Moncada - Náquera, Km. 4,5, 46113 Moncada, Spain.
E-mail address: miguel.verdu@ext.uv.es (M. Verdú).

https://doi.org/10.1016/j.ecoinf.2024.102826
Received 7 February 2024; Received in revised form 11 September 2024; Accepted 11 September 2024

as a proxy for predicting who can live with whom under the assumption that closely-related species tend to behave similarly because they possess similar traits Morales-Castilla et al. (2015). However, a recent AI model has shown that many of the characteristics crucial for species survival (e.g., tolerance to abiotic conditions, resource requirements, competitive abilities, response to antagonists and mutualists or dispersal capacity) are not adequately captured by phylogeny while they are captured by the co-occurrences Siefert et al. (n.d.). This is because co-occurrence by itself can inform us on niche preferences as well as on the outcome of multiple ecological interactions.

Deep Neural Networks (NNs) have been used in related ecological applications to classify vegetation units based on species composition with ca. 80 % accuracy Černá and Chytrý (2005). Additionally, deep-learning models can predict vegetation temporal dynamics using climate data, as shown by Chen et al. (2021), who developed global gridded climate-vegetation models using monthly temperature and precipitation as inputs, with the normalized difference vegetation index (NDVI) representing vegetation greenness as the model output.

Species co-occurrences can be fitted with models that do not rely on NNs Harris (2016); Losapio et al. (2021). Although the mathematics behind such models may sound more familiar given the connection to the Lotka-Voltera equations and to Taylor series expansions, this very reliance on a specific functional shape limits their expressivity (i.e. their ability to fit arbitrarily complex non-linear functions) compared to Deep NNs. In the present article, we test whether a Deep NN (specifically a Generative AI model) trained on a larger plant co-occurrence dataset can still be applied to relatively small datasets, thanks to transfer learning.

In Hirn et al. (2022), a Generative AI model was proposed and applied to predict species co-occurrences in vegetation patches (i.e. clumps of plants from different specifies that grow in close proximity, but are well-separated from other such clumps/patches). Such patterns have been subsequently used to unravel the mechanisms behind the assembly of a plant community in South Eastern Spain. In that paper, two potential uses of Generative AI were suggested: 1) predicting the probabilities of already-observed patches under conditions different from those of the sampling, for instance to reconstruct successional trajectories, thereby guiding ecological restoration efforts in order to increase biodiversity, 2) predicting the probability of new, plausible vegetation patches that have not been seen in the training distribution because of limited sampling.

The present paper aims to overcome one of the limitations of the approach of Hirn et al. (2022), namely the fact that the accuracy of the method used in that paper depends on having datasets larger than those often available in the field of Ecology, limiting the applicability of these models in situations where sampling is costly and/or too time-consuming. Yet, we cannot afford to lose the information in small datasets, which contain valuable spatial and temporal information for the understanding of ecological processes. We therefore explore the use of AI to “rescue” information from these small datasets to understand the assembly processes of a particular community and help understand those of other communities Todman et al. (2023).

Transfer learning Baxter (1996); Caruana (1997); L. Pratt and Jennings (1996); L. Y. Pratt (1992) has been explicitly designed to reuse the knowledge acquired when training a model on a large dataset in order to apply it to a similar problem with a smaller dataset. More specifically, transfer learning uses parts of the training in one task for another one, saving up resources and requiring fewer observations. Transfer learning has been successfully performed to reconstruct the food web of Canadian mammals based on the trophic interactions observed in Europe Strydom et al. (2022). Although only 4% of the species were common between continents, the phylogenetic information served as a basis for successfully transferring knowledge between the two systems, indicating that a common evolutionary pattern is assembling both trophic networks.

The authors of Strydom et al. (2022) only considered direct transfer: directly applying a model trained on European mammals, and applying it to Canadian mammals without further training. Here we would like to explore various approaches to transferring the knowledge gained from one community to another one. We want to do this in at least two ways: a) try different options beyond phylogenetic distance in order to match species between different communities, and b) further train the models (or part of them) on the target datasets after matching.

In the case of species co-occurrence, transfer learning can be explored by studying the performance of the AI models on a new dataset, i.e. observations in a new site or new observations in the same site. For example, it is possible to re-train only the final layers of neurons in a neural network while freezing up the initial layers, as proposed in Fregier and Gouray (2021). As this re-training requires smaller amounts of data, transfer learning can shortcut part of the costly process of collecting field observations. In fact, the more similar the new site is to the reference one, the smaller the amount of new data presumably required to reach a given level of accuracy on the new site. Conversely, cases in which transfer learning is found to require only small amounts of new data would indicate that the two sites share similar assembly rules which are broadly captured by the AI. A desirable property of an AI model is the transferability of its knowledge to new situations. Thus, not only ecologists could save an enormous amount of fieldwork required to feed the models but also, we can add a new value to past, small datasets.

Our main goal here we test the transferability of the co-occurrence AI model developed in a Spanish plant community sampled in 2019, to two different communities for which we only have smaller, older datasets, one with the same Mediterranean climate (La Unión in Southeastern Spain, sampled in 2015) and one with a different climate (San Juan Raya in Central Mexico, sampled in 2007). These communities have in common that their vegetation grows in spatially delimited patches constituted by a variable number of individuals growing in close association, but differ in soil type, among other discriminating features. As for their species composition, there are no common species between the Mexican community and the two Spanish ones, while some infrequent species are common between the two Spanish communities (see species lists in Supporting Information A.2).

More specifically, we pose three questions:

- Q1: Compared to training models directly on relatively small datasets, is it possible to obtain better results by directly transferring a model trained on another (larger) dataset from a different plant community/location? This may obviously be dependent on the specific communities and dataset size, but we are looking for a proof of principle.
- Q2: Is it possible to find a procedure that improves on phylogenetic matching in order to build a dictionary for direct transfer between communities?
- Q3: For our example datasets, does further training on the target community (fine-tuning) improve the accuracy as compared to direct transfer?

2. Materials and methods

Fig. 1 presents a conceptual summary of the various steps we describe below.

2.1. Datasets

We selected three plant communities: the community of reference with the largest dataset (i.e., the largest number of vegetation patches) where the AI model of Hirn et al. (2022) was developed (Petrer, Spain), another community in the same climate (La Unión, Spain, ca. 100 km away) and a tropical community (San Juan Raya, Mexico, ca. 9300 km away). All three communities are governed by plant facilitation and consist of patchy vegetation, where each patch is usually composed of multiple species growing in close association, but surrounded by bare soil matrix.

The Petrer community (South Eastern Spain, $38^{\circ}29'$N; $0^{\circ}47'$W) is a

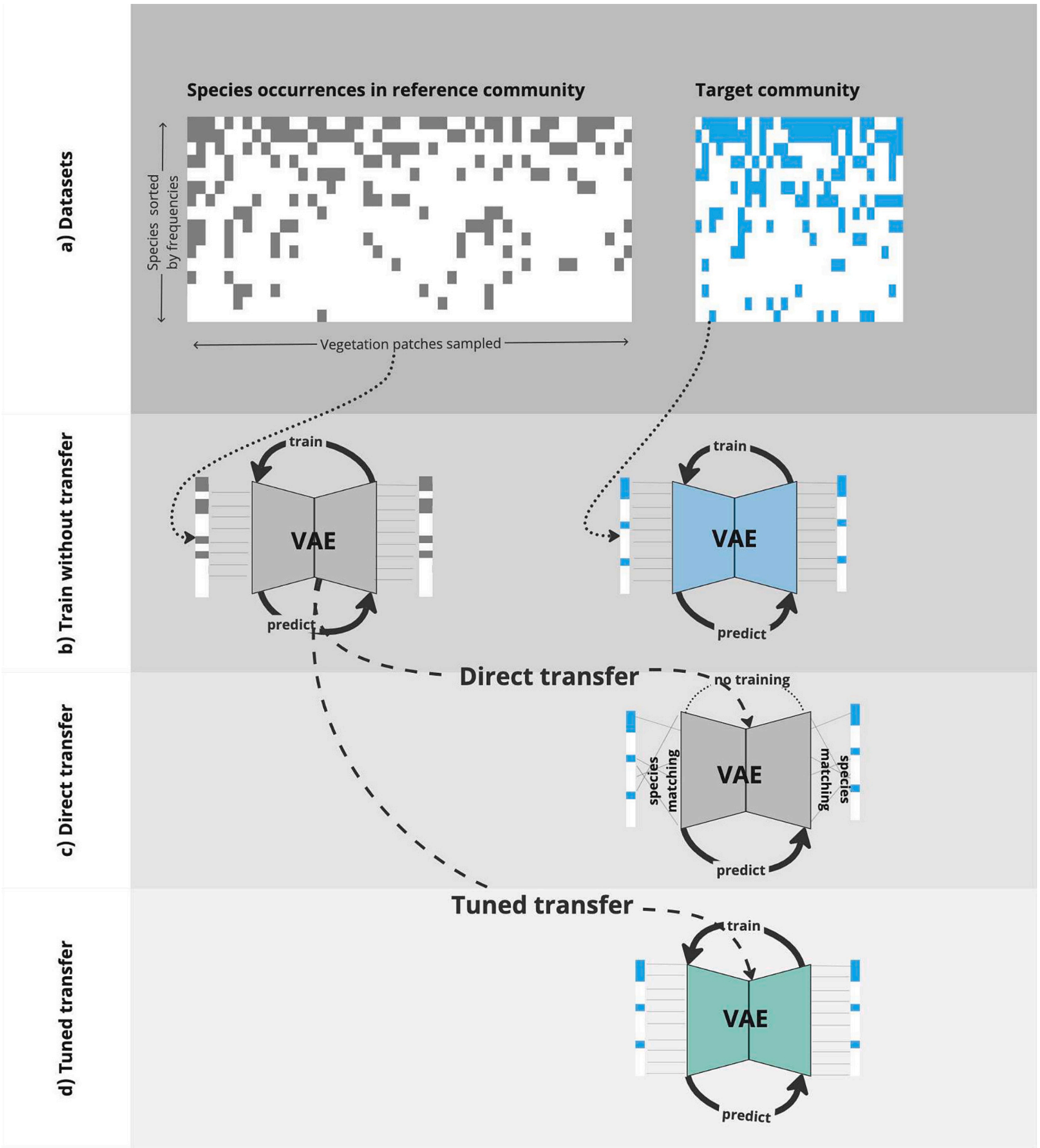


Fig. 1. Conceptual map depicting the various steps taken in the study, and the relations between them. a) Datasets: The reference community is the one with the largest dataset, while the target community has fewer observations. The datasets themselves consist of "barcodes", i.e. ones and zeroes representing the presence/absence of a given species in each patch. The species are ordered by frequency of occurrence, and only the 16 most frequent species are kept. b) Train without transfer: The AI model consists of a Variational Autoencoder (VAE) whose neural networks can be trained both on reference and target datasets, but the accuracy tends to increase with dataset size, and so the accuracy of a model trained on the (smaller) target dataset may be sub-par. c) Direct transfer: One can try to apply the model trained on the reference dataset directly to the target dataset, without further training. However, the success of this approach depends on designing a matching procedure between species in both communities, i.e. between rows in the barcodes of both datasets. In this study, we tested matching by phylogeny, by frequency, and interpolations in-between. d) Tuned transfer: One can also take the model that has already been trained on the reference dataset, and further train it on the target dataset. Since this tuning will modify the functions inside the model, we expect the exact species matching outside the network to be less relevant. In practice, we used the simplest species matching based on frequencies.

semi-arid Mediterranean ecosystem with a dry hot summer, and an average temperature of 16°C and a mean annual precipitation of 395 mm. Vegetation is distributed in patches, composed mainly of chamaephytes and tiny shrubs Sánchez-Martín et al. (2023) growing on limestone soils, see Fig. A2 in Supporting Information. 71 different species were recorded in that field sampling, see A.2.1 in Supporting Information.

La Unión community (South East Spain, 37°36′N; 0°50′W) has a very similar semi-arid Mediterranean climate with a dry hot summer, but is mostly composed of different species. Its mean annual temperature and precipitation are respectively 17.9°C and 316 mm. It is located in abandoned mine tailings along a low-altitude coastal mountain range. Vegetation patches colonize these disturbed systems triggered by nurse plants including trees, shrubs, perennial grasses and dwarf shrubs Navarro-Cano et al. (2018). Mine tailings have artificial soils that are the result of the refining process, with high metal concentrations and electrical conductivity and very low organic matter Colin et al. (2019), see Fig. A3 in Supporting Information. The data shows 156 different species, of which 29 also appear in Petrer, see A.2.2 in Supporting Information.

The San Juan Raya community (East-central Mexico, 18°19′N; 97°38′W) is composed of completely different species, and also has a different climate. While it is also a semi-arid ecosystem, rain there is concentrated during the hot summer. It is located in the Tehuacan-Cuicatlan Valley (Puebla state, Mexico), with an annual average rainfall of 380 mm and mean annual temperature of 21°C. The soils are deep and the lithology is made up of sandstones and the vegetation follows a patchy structure with xeric shrublands associated with columnar cacti Valiente-Banuet and Verdú (2008), see Fig. A4 in Supporting Information. The data shows 71 different species, none of which appear in the Petrer data, see A.2.3 in Supporting Information.

As was done in Hirn et al. (2022), we work with species presence/absence instead of abundance: this binary response variable represents the final outcome of complex interactions. To further simplify the data, we do not consider as patches the cases where only one species is present, since we are mostly interested in the interaction between species, not in the overall frequency of single species. Note that we define (relative) frequency as the proportion of patches in which a given species appears. Further, we also avoid very infrequent species: out of the dozens of species observed in each plant community (see Table 1), some only appear in a handful of patches, which is unlikely to provide enough replicates for our models to learn any patterns. In fact, for the three communities we consider, going beyond the 16 most frequent species would already entail species that appear in fewer than 10% of patches, see Fig. A5 in Supporting Information. For each community, we therefore arbitrarily limit ourselves to the 16 most frequent species and disregard the others. After this pruning, the final dataset used had fewer patches than the original data, see 1.

The plant species composition was recorded for 2635 vegetation patches in Petrer, 879 in La Unión and 892 in San Juan Raya (Table 1; see details on sampling procedures in Valiente-Banuet and Verdú (2008) for San Juan Raya, Navarro-Cano et al. (2018) for La Unión and Sánchez-Martín et al. (2023) for Petrer). Fig. A5 displays the relative frequency of occurrence of the 16 most common species in each community, as well as the distribution of the number of different species per patch. It should be noted that, of these 16 most common species in each community, none are shared between communities, i.e. even the two communities in Southeastern Spain do not share any of their 16 most common species, see Table A1. A more subjective way to appreciate the difference between the three communities is provided by the photographs Figs. A2, A3, A4. In our case, we transferred models from Petrer to the other two locations. The reason is that Petrer is the largest of our three datasets (2130 vegetation patches for the 16 species case).

2.2. *Training without transfer*

Following the steps in Hirn et al. (2022), we built a Generative AI model, specifically a Variational AutoEncoder (VAE) Kingma and Welling (2013), which takes inputs from the real data distribution and feeds them through an encoder neural network outputting a multivariate Gaussian distribution. A decoder then takes a sample point from the multivariate Gaussian (i.e. a point in the so-called latent space), and transforms it into the final output, i.e. some synthetic data that should resemble the input once the model has been trained.

The inputs are the patches themselves, with ones and zeroes representing the presence/absence of each species, with the default being that the species are ordered by frequency of occurrence for two main practical reasons: i) we do expect species frequency to be important in predicting interactions, and ii) with this choice, we can easily reduce or increase the number of species under consideration without affecting the order of species.

The training process for the VAE adjusts the model parameters in order to minimize a loss function. As is standard for a VAE, we took the loss function to be the sum of two contributions: i) The Kullback-Leibler (KL) divergence in latent space between the target distribution (a multidimensional unit Gaussian) and the data's distribution as encoded in the latent space–this KL contribution to the loss acts as a regularizer for the neural network–, and ii) A reconstruction error measuring the mismatch between the input and the final output. For the present binary classification task, we use the standard binary cross-entropy. As for the training procedure, we selected hyperparameters that enabled the NN to learn quickly, and kept those fixed for all the results of this paper: see A.3 for technical details about the model.

As a metric to compare models, we report the patch error rate, defined as the proportion of vegetation patches that are not perfectly reproduced by the VAE.

2.3. *Direct transfer (Q1 and Q2)*

After training an initial model on Petrer data, we expect that the model has learnt some of the basic rules governing species co-occurrence, and that a fraction of these rules may apply to new plant communities. The simplest way to test if any knowledge can be transferred from Peter to the other communities is to take models trained on Petrer and apply them directly to the datasets of La Unión and San Juan Raya, an approach often called zero-shot learning in computer vision Chang et al. (2008); Lampert et al. (2009); Larochelle et al. (2008); Palatucci et al. (2009).

Yet, before we could transfer from a model trained in Petrer to another plant community, we needed to decide how to match the species between the initial community and the target one, i.e. how to build the dictionary that translates from a given species in Petrer, to a species in another community.

One idea to build such a dictionary would be to use the logic of Strydom et al. (2022) and match species according to phylogenetic distance. Yet we also wanted to compare this approach with options (Q2), for instance with a dictionary that directly matches the species by order of frequency, since this is the way the species are ordered by default for each community separately.

In addition, to establish a baseline to compare specific matching dictionaries, we considered random matching of species. Yet, there are $16! \simeq 2{\cdot}10^{13}$ possible ways to sort 16 species. We chose a random sample of 60 possible such dictionaries to apply Petrer models to the two other communities, and computed the error rates in the target location. This gave us an estimate of the range of expected error rates from this direct transfer with random dictionaries. We next compared this results with three different approaches to build a dictionary which we first describe briefly here, then detail in the next paragraphs. Our three approaches to build dictionaries are:

- 1) Using phylogeny, i.e. additional biological information,
- 2) Simply matching species between the two communities in order of their relative frequencies,
- 3) A number of intermediate possibilities to interpolate between the purely phylogenic dictionary and the purely frequency-based dictionary.

For option 1), i.e. the dictionary by phylogenetic distance Strydom et al. (2022), we followed an iterative procedure: we selected the most

Table 1
Number of patches and species in the original data, and after restriction to the 16 most common species, for each of the three sampling locations.

	Original data	Original data	Original data	Data restricted to 16 main species	Data restricted to 16 main species
	Number of patches	Number of species	Species per patch (mean ± sd)	Number of patches	Species per patch (mean ± sd)
Petrer (Spain)	2635	82	3.53 ± 2.29	2130	3.23 ± 1.67
La Unión (Spain)	879	134	4.38 ± 4.51	498	3.31 ± 1.52
San Juan Raya (Mexico)	892	71	5.18 ± 2.37	840	4.25 ± 1.68

frequent species in the target dataset (among those that had not yet been matched), and looked for its closest relative in the reference dataset. If we needed to break a tie, we chose the most frequent species in the reference dataset. We then repeated the process for the next most frequent species in the target dataset, until we exhausted all 16 species. This algorithm yields a unique matching between any pair of communities, see Table A1. Phylogenetic distances were extracted from the phylogenetic tree constructed with the help of the 'V.Phylomaker' package in R (version 4.1.1), which contains a dated mega-tree of ca. 75,000 species of all vascular plant families Jin and Qian (2019).

For option 2), i.e. the dictionary by frequency, we first matched the most frequent species in both communities, then the second most frequent ones, and so on, until the least frequent species in both communities.

For option 3), i.e. the combined options, we considered 17 cases, labeled by an index n varying form 0 to 16. We selected the n most common species, and matched those by frequency. The remaining $16 - n$ species were then matched by phylognetic distance. The extreme cases ($n = 0$ and $n = 16$) reproduced options 1) and 2) respectively).

This served as a test (Q2) of whether phylogenetic information is relevant, whether it is more useful than relative frequencies, and allowed us to select the best matching procedure.

2.4. *Tuned transfer (Q3)*

We took models trained on the reference community, matched them to the target communities by frequency, and then trained them on the target community with the same hyperparameters described in A.3, including the learning rate, except that we kept the best model out of the first 50 epochs instead of 200, given that 50 epochs were more than enough for training to have converged when starting from a pre-trained model.

Note that this tuning of models is slightly different from fine-tuning as is commonly performed in computer vision, where a smaller learning rate is used, and some layers are either frozen altogether or are trained with a smaller training rate: here we did not freeze any layer, and instead directly updated the weights of the whole network with the same learning rate as during the initial training, as we did not detect any benefit from training slower. Data and code for all the above can be found at https://github.com/johanneshirn/eco_transfer

3. Results

3.1. *Training without transfer*

We obtained a median patch error rate of 30% in La Unión (trained on 249 patches), compared to 20% in San Juan Raya (420 training patches) and 7% in Petrer (1065 training patches). When comparing the accuracy of the models trained in different plant communities, it is important to remember that error rates may depend not only on dataset size, but also on the contents of the dataset, see Fig. 2: the three plant communities we studied differ in the species identity, richness and relative frequencies, see Tables 1, A1 and Fig. A5.

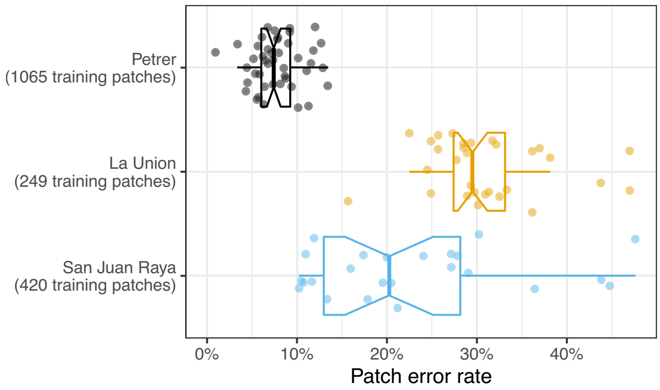


Fig. 2. Validation error rates for models trained without transfer on the datasets from the three locations.

3.2. *Direct transfer (Q1 and Q2)*

We now focus on transferring knowledge from Petrer to La Unión and San Juan Raya communities. As compared to the baseline (random matching dictionary between species of each community), a matching based on phylogenetic information reduces the error rate on average when transferring from Petrer to La Unión, see upper panel of Fig. 3. Not so in San Juan Raya, where the phylogenetic matching does not provide an improvement over the baseline on average (lower panel of Fig. 3). While this may sound surprising, it simply implies that there are better ways to build a matching dictionary than using pure phylogenetic distance. The height of the boxes in the left part of the plot of Fig. 3 also provides us with an indication of how much of an improvement other dictionaries might provide.

In our case, we already mentioned in 2.3 that we would consider other alternatives beyond a purely phylogenetic dictionary: a frequency based one, and a range of dictionaries interpolating between purely phylogeny-based and purely frequency-based. Indeed, we can see from Fig. 3 that in both locations, a matching that also includes some information about species frequencies in addition to phylogenetic information provides a better basis for transferring the information from the original community (Petrer) to the target ones (La Unión, San Juan Raya). Note however, that the relative importance of phylogeny and frequency in the matching, as well as the location of the optimum, differ between our two target datasets.

We therefore conclude that including frequency information when building the species dictionary between plant communities can improve the success rate of direct transfer as compare to using only phylogenetic distance (Q2). Also, we find that direct transfer worked better between climatically-similar ecosystems (i.e. from Petrer to La Unión led to a median patch error rate of 13%) than between more dissimilar ecosystems (i.e. transferring from Petrer to San Juan Raya led to a median patch error rate of 23%) (see Fig. 4). Fig. 4 shows that directly transferring a model may not always yield an improvement compared to directly training on the target dataset, so the answer to (Q1) depends on the specific case (i.e. which plant communities are considered on either side of the transfer, as well as the sizes of the respective datasets.

3.3. *Tuned transfer (Q3)*

Tuning the Petrer model to local data improves the accuracy compared to direct transfer (Q3), see Fig. 4. When tuning that Petrer model to local data in La Unión or San Juan Raya, we train the model on local data, as when training without transfer: the difference lies in the starting point (initialization of the model parameters). Instead of starting from a random initialization of the network, in the tuning case we start from a model that has already been trained on the reference community data (Petrer), which we match to the target community using frequencies (for simplicity). Fine-tuning in the closely-related community (La Unión) led to patch error rates around 7% while such a rate was 12% in the more dissimilar community (San Juan Raya), see Fig. 4.

4. Discussion

The explosion of AI methods is leading ecologists to search for complex patterns that were previously hard to detect Borowiec et al. (2022). To fully achieve the benefits of AI, big datasets are necessary Perry et al. (2022). This might not be a problem in areas where gigabytes of data are publicly available, such as those related to climate, satellite

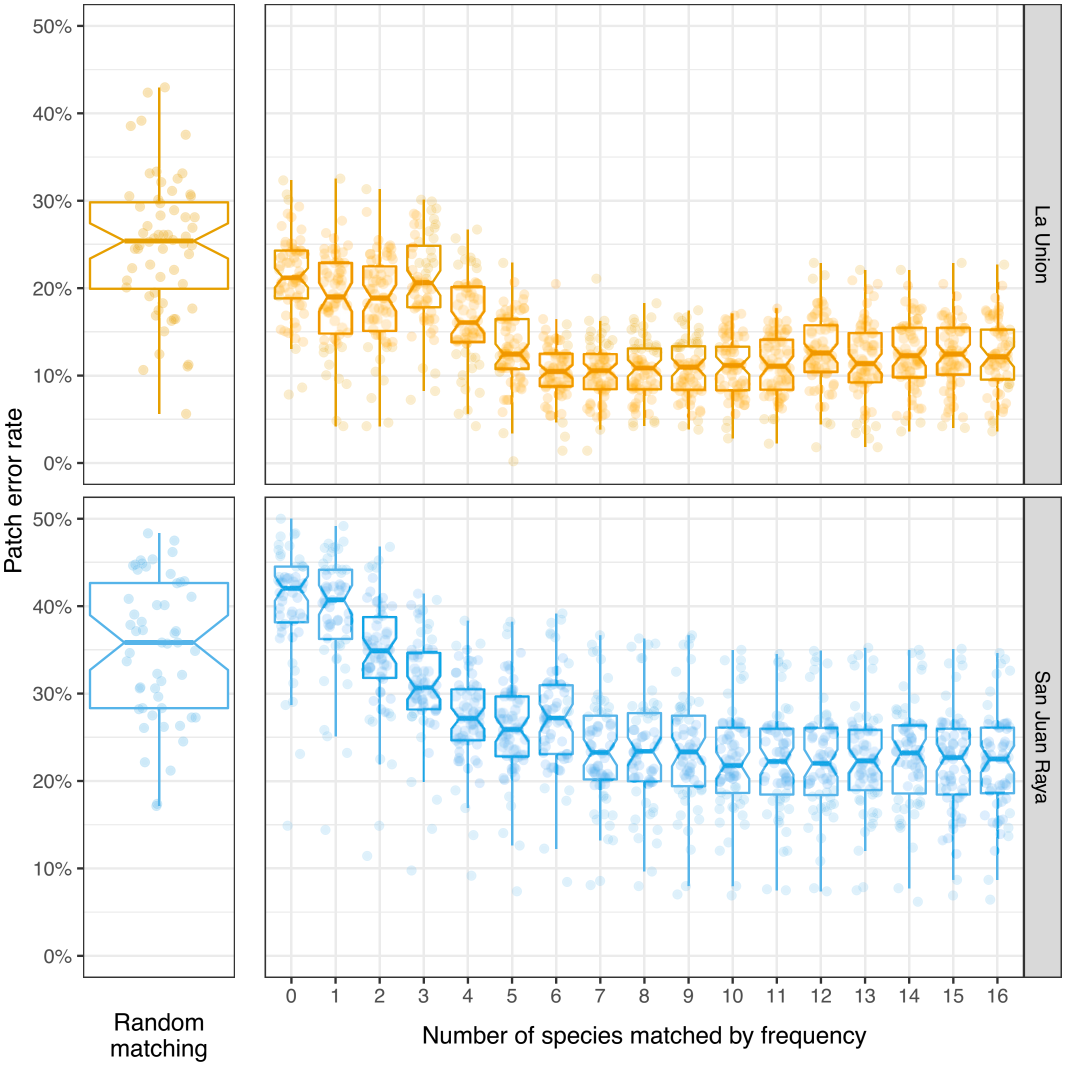


Fig. 3. Error rates for models trained on Petrer data, and directly applied to La Unión and San Juan Raya, depending on the procedure used to match species between locations. Left: A random matching of species used as a baseline. Right: Of the 16 selected species in each location, we match the n most frequent species between both locations by frequency, then match the 16-n remaining species by phylogeny. We thus get a purely phylogenetic match for $n = 0$, while $n = 16$ yields a match purely by frequencies. Top-right: When transferring from Petrer to La Unión, the match with $n = 6$, i.e. with a combination of frequency and phylogeny is significantly better than the two extremes (pure frequency and pure phylogeny) as confirmed by Wilcoxon signed-rank tests. Bottom-right: When transferring from Petrer to San Juan, pure phylogeny performed worse than the random matching, but combining phylogeny with frequency drastically reduced the error. In this system, matching by pure frequency results in an error rate that is not improved by any combination of frequency and phylogeny.

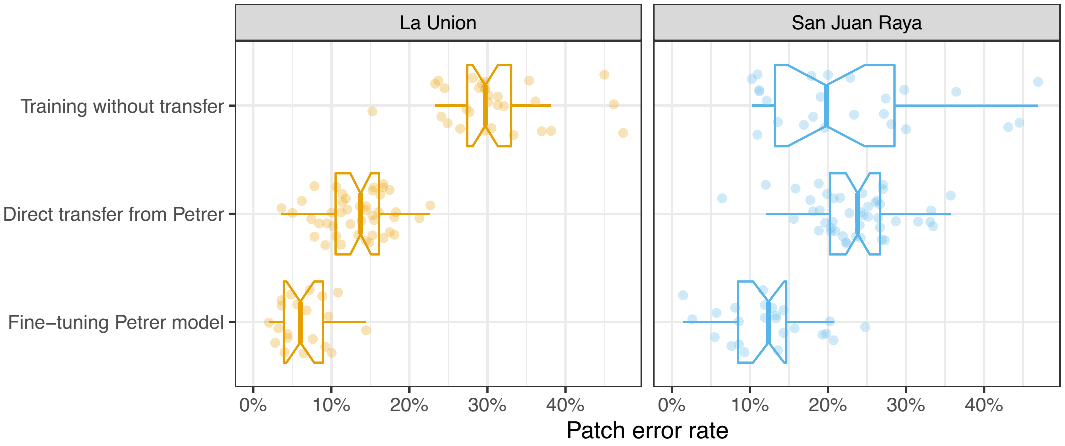


Fig. 4. Error rates on two target datasets for models trained: 1. without transfer, i.e. only on the data from the sampling location they are applied to, 2. on the Petrer data only, then transferred directly to the target data using the dictionary by frequencies, or 3. on the Petrer data first, then further trained on local data.

imagery or genetic sequences Peters et al. (2014). Yet field ecology, despite the advances in the publication of open-access datasets, is behind these disciplines in terms of sheer quantity of data, partly because of the difficulty of compiling large datasets under field conditions Liu et al. (2018). Here we have shown that smaller datasets (fewer than 1000 patches for the present communities) can benefit from the AI revolution if combined with a larger dataset (more that 2000 patches in the present case) and an adequate application of transfer learning.

In particular, we have shown that taking a model trained in a community of reference (Petrer) and tuning it to another community (La Unión or San Juan Raya) gives us the best of both worlds. Firstly, as compared to direct transfer, it allows for the possibility of including some knowledge from the target community, and how its pattern of species co-occurrence differ from those of the reference community. Secondly, as compared to training on the target community but starting from a random initialization of the neural network weights, it allows us to start from an ecologically sound set of parameters that already reproduces a plausible distribution of occurrence frequencies and species co-occurrence. Although knowledge transfer worked very well in both communities (i.e., correctly reproduced the species composition of most of the vegetation patches), it performed better in the community taxonomically and climatically more similar to the reference community. Altogether, these facts indicate that there are similarities in the rules of species co-occurrence, but, at the same time, shows the existence of idiosyncratic particularities in each community that our AI model has not yet learned.

The first question ecologists should ask is how large must be the field sampling for AI to achieve a reasonable accuracy. Our data on plant co-occurrence obtained in the community of reference show that the model reaches its minimum error rate (10 %) with 900 patches. This is to say that a model constructed with that amount of data will correctly reproduce the species composition of 90% of the patches. Below this sample size, the error rate of the model increases steeply. Note that this number should not be taken as a universal law, given that the accuracy of the model will depend on the biological complexity of the dataset.

Caution is needed when applying AI models to small datasets because overfitting could cause the impression that the model has perfectly learnt the patterns behind the data and can be used to predict new, unsampled, combinations of species. However, overfit models “memorize” the whole data, and therefore do not perform well on new, unobserved data Pichler and Hartig (2023). In our study, we guarded against this pitfall of overfitting by using a validation set and a standard early-stopping procedure. We are thus confident that our models could generalize their learning to other plant communities.

Using AI to “rescue” small datasets (sensu Christin et al. (2019)), we have shown that the knowledge about species co-occurrence obtained in our well-sampled plant community in Petrer, was successfully transferred to two other communities with smaller datasets. These datasets were not large enough for us to train a model without transfer with good accuracy, as they yielded error rates of 20% and 30% respectively on La Unión and San Juan Raya. However, after transferring the plant species co-occurrence patterns learnt in Petrer and tuning the model to local data, we could reduce the error rates to 7% and 12% respectively.

What were the co-occurrence patterns learnt in Petrer that were successfully transferred to the other two communities? To transfer optimally to a new community, we found that we needed to establish a matching of species that took into account species frequency in both communities, and possibly phylogenetic information in the closely-related community. This feature is consistent with other plant-plant interaction studies in which the concurrence of both variables improved the likelihood of models explaining which species interact with whom Alcántara et al. (2019); Verdú and Valiente-Banuet (2011). The rationale of this result is that phylogenetically-related species tend to behave similarly and therefore we could predict the outcome of interactions based on the phylogenetic position of each species. However, even more than phylogenetic relatedness, the relative frequency of species is the main determinant of most of the ecological interactions Vázquez et al. (2007) because frequency strongly determines the probability of co-occurrence for interacting species.

We have successfully transferred the ecological patterns learnt from plant species growing on limestone soils in a semi-arid dryland from South Eastern Spain to two communities that are also governed by plant facilitation but different climate and taxonomy. This outcome is very promising because it suggests that common rules are behind the assembly of communities. Therefore, transfer learning could be very useful in ecology to continue building on the shoulders of giants.

While there may be ways to get around them, two of the current limitations of our method are a) the inability to deal with infrequent species, and b) the requirement of defining in advance the number of species to be taken into account. As for the ability to transfer co-occurrence models to other plant communities, what we have presented here a first proof of principle: next steps would involve testing whether the idea can be applied to a wider range of ecosystems.

One possible avenue for future research would be to assemble a meta-dataset by joining together datasets from a large number of communities to train a general model. Such a model might then be applied via direct transfer to a wide variety of communities. Another promising application of the present technique of modeling interactions between species would be for studies of Dark Diversity. In that case, the unit of data would not be a single patch, but a community, and one would train the model to predict if a given species could be viable in a given community, even though it has not been observed there yet.

Funding statement

Financial support was provided by the projects TED2021-129682B-I00, PID2020-113157GB-I00 (funded by MCIN/AEI/10.13039/501100011033 and ‘ERDF A way of making Europe’) and CIPROM/2021/63 (Generalitat Valenciana). JANC and MG also received funding through PID2020-119634GB-I00.

Authors’ contributions

JH, VS and MV conceived the study. MG, AMN, JANC, RSM and AVB collected the data. JH analyzed the data. JH and MV wrote the first draft of the manuscript and all authors contributed to revisions.

CRediT authorship contribution statement

Johannes Hirn: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Verónica Sanz:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **José Enrique García:** Writing – review & editing. **Marta Goberna:** Writing – review & editing, Data curation. **Alicia Montesinos-Navarro:** Writing – review & editing, Data curation. **José Antonio Navarro-Cano:** Writing – review & editing, Data curation. **Ricardo Sánchez-Martín:** Writing – review & editing, Data curation. **Alfonso Valiente-Banuet:** Writing – review & editing, Data curation. **Miguel Verdú:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

The data and scripts that support the findings of this study are openly available at https://github.com/johanneshirn/eco_transfer

How to cite this article: J. Hirn, V. Sanz, J. E. García, M. Goberna, A. Montesinos-Navarro, J. A. Navarro-Cano, R. Sánchez-Martín, A.

Valiente-Banuet and M. Verdú (2023), Transfer learning for plant co-occurrence.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102826.

References

Alcántara, J.M., Garrido, J.L., Rey, P.J., 2019. Plant species abundance and phylogeny explain the structure of recruitment networks. New Phytol. 223, 366–376. https://doi.org/10.1111/NPH.15774.

Bascompte, J., 2009. Disentangling the web of life. Science 325, 416–419. https://doi.org/10.1126/SCIENCE.1170749.

Baxter, J., 1996. A Bayesian/Information Theoretic Model of Bias Learning, pp. 77–88. https://doi.org/10.1145/238061.238071.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. 13, 1640–1660. Retrieved from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13901.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28, 41–75. Retrieved from https://link.springer.com/article/10.1023/A:1007379606734.

Černá, L., Chytrý, M., 2005. Supervised classification of plant communities with artificial neural networks. J. Veg. Sci. 16 (4), 407–414. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1654-1103.2005.tb02380.x.

Chang, M.-W., Ratinov, L., Roth, D., Srikumar, V., 2008. Importance of Semantic Representation: Dataless classification, pp. 830–835. https://doi.org/10.5555/1620163.1620201.

Chen, Z., Liu, H., Xu, C., Wu, X., Liang, B., Cao, J., Chen, D., 2021. Modeling vegetation greenness and its climate sensitivity with deep-learning technology. Ecol. Evol. 11 (12), 7335–7345. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.7564.

Chesson, P., 2000. Mechanisms of maintenance of species diversity. Annu. Rev. Ecol. Syst. 31, 343–366. https://doi.org/10.1146/ANNUREV.ECOLSYS.31.1.343.

Christin, S., Hervet, E., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10, 1632–1644. https://doi.org/10.1111/2041-210X.13256.

Colin, Y., Goberna, M., Verdú, M., Navarro-Cano, J.A., 2019. Successional trajectories of soil bacterial communities in mine tailings: the role of plant functional traits. J. Environ. Manag. 241, 284–292. https://doi.org/10.1016/J.JENVMAN.2019.04.023.

Fregier, Y., Gouray, J.B., 2021. Mind2mind: transfer learning for gans. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 12829 LNCS, pp. 851–859. https://doi.org/10.1007/978-3-030-80209-7_91.

Harris, D.J., 2016. Inferring species interactions from co-occurrence data with markov networks. Ecology 97 (12), 3308–3314. Retrieved from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.1605.

Hirn, J., García, J.E., Montesinos-Navarro, A., Sánchez-Martín, R., Sanz, V., Verdú, M., 2022. A deep generative artificial intel-ligence system to predict species coexistence patterns. Methods Ecol. Evol. 13, 1052–1061. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13827https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13827https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13827.

Jin, Y., Qian, H., 2019. V.phylomaker: an r package that can generate very large phylogenies for vascular plants. Ecography 42, 1353–1359. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/ecog.04434https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.04434https://onlinelibrary.wiley.com/doi/10.1111/ecog.04434. https://doi.org/10.1111/ECOG.04434.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. https://doi.org/10.48550/arXiv.1312.6114. Retrieved from https://arxiv.org/abs/1312.6114v11.

Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to Detect Unseen Object classes by Between-Class Attribute Transfer, pp. 951–958. https://doi.org/10.1109/CVPR.2009.5206594.

Lapeyrolerie, M., Boettiger, C., 2023. Limits to ecological forecasting: estimating uncertainty for critical transitions with deep learning. Methods Ecol. Evol. 14, 785–798. https://doi.org/10.1111/2041-210X.14013.

Larochelle, H., Erhan, D., Bengio, Y., 2008. Zero-Data Learning of New Tasks, pp. 646–651. https://doi.org/10.5555/1620163.1620172.

Liu, Z., Peng, C., Work, T., Candau, J.N., Desrochers, A., Kneeshaw, D., 2018. Application of machine-learning methods in forest ecology: recent progress and future challenges. Environ. Rev. 26, 339–350. https://doi.org/10.1139/ER-2018-0034/ASSET/IMAGES/LARGE/ER-2018-0034F3.JPEG.

Losapio, G., Schöb, C., Staniczenko, P.P.A., Carrara, F., Palamara, G.M., De Moraes, C.M., Bascompte, J., 2021, February. Net-work motifs involving both competition and facilitation predict biodiversity in alpine plant communities. Proc. Natl. Acad. Sci. 118 (6), e2005759118. Retrieved 2024-05-31, from https://www.pnas.org/doi/10.1073/pnas.2005759118.

Morales-Castilla, I., Matias, M.G., Gravel, D., Araújo, M.B., 2015. Inferring biotic interactions from proxies. Trends Ecol. Evol. 30, 347–356. https://doi.org/10.1016/J.TREE.2015.03.014.

Navarro-Cano, J.A., Verdú, M., Goberna, M., 2018. Trait-based selection of nurse plants to restore ecosystem functions in mine tailings. J. Appl. Ecol. 55, 1195–1206. Retrieved from https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.13094.

Palatucci, M., Pomerleau, D., Hinton, G., Mitchell, T.M., 2009. Zero-Shot Learning with Semantic Output Codes, pp. 1410–1418. https://doi.org/10.5555/2984093.2984252.

Perry, G.L., Seidl, R., Bellvé, A.M., Rammer, W., 2022. An outlook for deep learning in ecosystem science. Ecosystems 25, 1700–1718. https://doi.org/10.1007/S10021-022-00789-Y.

Peters, D.P., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5. https://doi.org/10.1890/ES13-00359.1.

Pichler, M., Hartig, F., 2023. Machine learning and deep learning—a review for ecologists. Methods Ecol. Evol. 14, 994–1016. https://doi.org/10.1111/2041-210X.14061.

Pratt, L.Y., 1992. Discriminability-Based Transfer between Neural Networks, pp. 204–211. https://doi.org/10.5555/2987061.2987087.

Pratt, L., Jennings, B., 1996. A survey of transfer between connectionist networks. Connect. Sci. 8 (2), 163–184. Retrieved from. https://doi.org/10.1080/095400996116866.

Sánchez-Martín, R., Verdú, M., Montesinos-Navarro, A., 2023. Phylogenetic and functional constraints of plant facilitation rewiring. Ecology 104. https://doi.org/10.1002/ECY.3961.

Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. 54, 8558–8593. https://doi.org/10.1029/2018WR022643.

Siefert, A., Laughlin, D. C., & Sabatini, F. M. (n.d.). You shall know a species by the company it keeps: leveraging co-occurrence data to improvee cological prediction. doi: https://doi.org/10.1101/2023.02.15.528518.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., Poisot, T., 2022. Food web reconstruction through phylogenetic transfer of low-rank network representation. Methods Ecol. Evol. 13, 2838–2849. https://doi.org/10.1111/2041-210X.13835.

Todman, L.C., Bush, A., Hood, A.S.C., 2023. ‘Small data’ for big insights in ecology. Trends Ecol. Evol. 38, 615–622. https://doi.org/10.1016/j.tree.2023.01.015.

Valiente-Banuet, A., Verdú, M., 2008. Temporal shifts from facilitation to competition occur between closely related taxa. J. Ecol. 96, 489–494. https://doi.org/10.1111/J.1365-2745.2008.01357.X.

Vázquez, D.P., Melián, C.J., Williams, N.M., Blüthgen, N., Krasnov, B.R., Poulin, R., 2007. Species abundance and asymmetric inter-action strength in ecological networks. Oikos 116, 1120–1127. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.0030-1299.2007.15828.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.0030-1299.2007.15828.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.0030-1299.2007.15828.x.

Verdú, M., Valiente-Banuet, A., 2011. The relative contribution of abundance and phylogeny to the structure of plant facilitation networks. Oikos 120, 1351–1356. https://doi.org/10.1111/J.1600-0706.2011.19477.X.