Received: 24 May 2021 | Accepted: 19 January 2022

DOI: 10.1111/2041-210X.13827

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

A deep Generative Artificial Intelligence system to predict species coexistence patterns

Johannes Hirn[1] | José Enrique García[1] | Alicia Montesinos-Navarro[2] | Ricardo Sánchez-Martín[2] | Veronica Sanz[1,3] | Miguel Verdú[2]

[1]Instituto de Física Corpuscular (IFIC, Universidad de Valencia-CSIC), Valencia, Spain

[2]Centro de Investigaciones Sobre Desertificación (CIDE, CSIC-Universidad de Valencia-Generalitat Valenciana), Valencia, Spain

[3]Department of Physics and Astronomy, University of Sussex, Brighton, UK

Correspondence
Miguel Verdú
Email: miguel.verdu@ext.uv.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: RTI2018-099672-J-I00 and FPU17/00629

Handling Editor: Timothée Poisot

Abstract

1. Predicting coexistence patterns is a current challenge to understand diversity maintenance, especially in rich communities where these patterns' complexity is magnified through indirect interactions that prevent their approximation with classical experimental approaches.
2. We explore cutting-edge Machine Learning techniques called Generative Artificial Intelligence (GenAI) to predict species coexistence patterns in vegetation patches, training generative adversarial networks (GAN) and variational AutoEncoders (VAE) that are then used to unravel some of the mechanisms behind community assemblage.
3. The GAN accurately reproduces real patches' species composition and plant species' affinity to different soil types, and the VAE also reaches a high level of accuracy, above 99%. Using the artificially generated patches, we found that high-order interactions tend to suppress the positive effects of low-order interactions. Finally, by reconstructing successional trajectories, we could identify the pioneer species with larger potential to generate a high diversity of distinct patches in terms of species composition.
4. Understanding the complexity of species coexistence patterns in diverse ecological communities requires new approaches beyond heuristic rules. Generative Artificial Intelligence can be a powerful tool to this end as it allows to overcome the inherent dimensionality of this challenge.

KEYWORDS
artificial intelligence, direct interactions, generative adversarial networks, indirect interactions, species coexistence, variational AutoEncoders

1 | INTRODUCTION

Understanding how species coexist has always been a central problem in ecology as it is at the core of diversity maintenance (Chesson, 2000). The complexity of the coexistence patterns is magnified in diverse communities where coexistence is not only a signal of paired interactions, but also of indirect interactions (Strauss, 1991). Thus, the probability that two species coexist depends on the presence of a third, fourth, fifth or nth species. Experimental approaches have been commonly used to explore coexistence patterns, although their focus on pairs or on a few sets of species cannot reproduce the frequent situation occurring in natural communities where species interact with many other species (van Kleunen et al., 2014). This is unavoidable, as the number of indirect interactions increases exponentially with the number of species considered, precluding the experimental quantification of all of them. The importance of indirect interactions structuring communities is well known (Schöb et al., 2013; Simmons et al., 2019) and therefore, other tools are currently needed to assess coexistence patterns in species-rich ecological communities where tens, hundreds or even thousands of species coexist.

Machine learning is able to detect complex patterns beyond heuristic rules and traditional statistics (Bzdok et al., 2018). The development of Neural Networks, a form of Machine Learning, can detect intricate patterns produced by high-order interactions such as those produced among genes into regulatory networks (Libbrecht & Noble, 2015) or genetic, clinical and histological variables used to diagnose cancer (Kourou et al., 2015). In ecology, deep learning has often been used to assist researchers in processing large datasets produced by automatic monitoring of populations and ecosystems by applying deep neural networks (Christin et al., 2019; Joseph, 2020). However, the possibilities of this methodology are much wider and the road is paved to study high-dimensional problems related to ecological interactions (Desjardins-Proulx et al., 2017; Poisot et al., 2021; Strydom et al., 2021).

Species coexistence in nature follows complex, unknown patterns that Machine Learning should be able to capture thanks to its high degree of expressivity (i.e. the capacity of a model to express complex relations) (Balamurugan et al., 2019; Chen et al., 2017; Harris, 2015; Raghu et al., 2017; Tang et al., 2018). Here, we explore the use of a set of cutting-edge Machine Learning techniques called Generative Artificial Intelligence (GenAI) (Ruthotto & Haber, 2021) to predict species coexistence patterns that could be later used to unravel the mechanisms behind community assemblage. The word Generative indicates the ability of these techniques to create new, unseen situations from a limited dataset of examples. Among the Generative AI methods, we choose the two most powerful ones because they are based on Deep Learning techniques and hence are fast to train in big datasets yet able to capture subtle relations. These types of algorithms are called Generative Adversarial Networks (GAN) and Variational AutoEncoders (VAE), both with their own strengths and weaknesses (Ruthotto & Haber, 2021). On the one hand, GANs consist of two models that are simultaneously trained so that a generative model G captures the distribution of the data, and a discriminative model D estimates the probability that a sample comes from the training data rather than from G (Figure 1 left). The training for G maximizes the likelihood that the discriminative model makes a mistake (Goodfellow et al., 2014). On the other hand, VAEs are generative machine learning models that combine a pair of neural networks that aim to first compress and then mirror the input data given a set of latent coordinates (Kingma & Welling, 2013) (Figure 1 right). VAEs incorporate nonlinear relationships and allow users to define the dimensionality of the latent space.

The loss function for a VAE is the sum of the reconstruction error (difference between the generated and input data), and the Kullback–Leibler term. This last term is the divergence between a sample's posterior distribution in latent space and a reference distribution which acts as a prior on the latent space (Battey et al., 2021). As explained in Ruthotto and Haber (2021), GANs have shown great ability to generate realistic avatars after moderate training, whereas the more complex structure of VAEs allows us to ask deeper questions, at the level of the latent space, which enables exploring what the Generative model has learnt and how to connect it to human variables (Iten et al., 2020).

In this study, we introduce the application of generative artificial intelligence to predict plant coexistence patterns and illustrate its potential in a facilitation-driven community where plants tend to grow together, forming vegetation patches (Montesinos-Navarro et al., 2019). In facilitation-driven patches, indirect interactions occur, and the coexistence of species within a patch strongly depends on the composition of the neighbourhood (Castillo et al., 2010; Schöb et al., 2013). We characterize plant species composition of vegetation patches, which could be used to estimate the probability of species co-occurrence across them. With an unlimited number of patches sampled, this probability would become a theoretical distribution of all species co-occurrence, and thus a manifestation of the underlying rules dictating the patch composition. Hence, we develop a Machine Learning method able to model a continuous probability distribution from a finite set of observations. Based on the observed composition of patches, the model is able to generate new patches whose composition cannot be derived in simple ways. We trained two Generative Artificial Intelligence systems (GAN and VAE) to generate fake but likely compositions of patches (hereafter fake patches) and validate them by comparing the patterns observed in the field. First, we assess whether the fake patches mirrored (a) the relative abundance of patches with a given species composition and (b) the affinity of gypsum specialist plants to different soil types. Then, we used the GAN to (c) assess the relative contribution of direct and indirect interactions in determining the probability of species co-occurrence in vegetation patches. Once we validated the models, we used GAN and VAE to produce patterns that are hard to validate in the field; specifically, we forecast the amount of potential species compositions in fake patches following the succession triggered by a pioneer species. Finally,

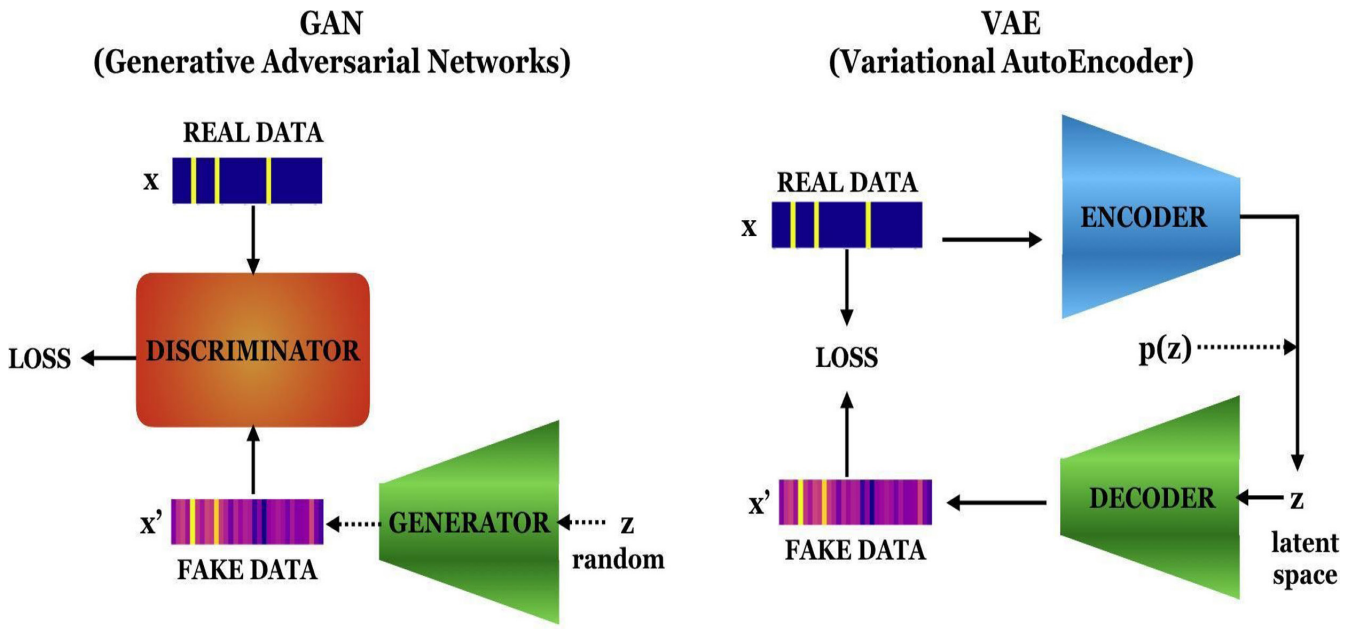

FIGURE 1 Schematic description of the generative adversarial networks (GAN) and variational AutoEncoders (VAE) architectures. Real data are represented with the label *x*, and the generated data by *x*', whereas *z* denotes an external variable randomly generated. Generator, encoder and decoder are made of layers of artificial neurons

we provide guidelines to construct personalized GenAI models and the code to run them.

2 | MATERIALS AND METHODS

2.1 | Input data

The species composition of 5,153 vegetation patches was characterized in four dryland plant communities (hereafter sites) situated within a radius of 20 km in Alicante (southeast Spain). Within each site, the vegetation patches were distributed in two adjacent soil types (hereafter gypsum and limestone subsites) located <10 m apart, minimizing the potential effect of dispersal limitation of species between subsites. The sampling design comprised 80 plots (150 × 150 cm) randomly distributed in each subsite, except one subsite with 79 plots. Inside each plot, we identified and registered all the species present in each vegetation patch. Permission for fieldwork was not necessary. A patch is composed of at least two individuals of different species surrounded by bare ground, with a mean surface area of 512 ± 982 cm^2.

Vegetation patches are expressed as arrays of presence/absence of plant species in an ordered list of *n* species where 1 would denote presence and 0 absence (e.g. *x* = [1, 0, 0, 0, 1, ..., 0]). There are as many arrays as vegetation patches sampled (*N* = 5153) so that our database is then a list of vectors $\{x_1, x_2, x_3, ..., x_N\}$ in a $\Re^n$ space, which includes a number of species (we will choose the most abundant species for illustrative purposes), and also the soil type (1 = gypsum or 0 = limestone) in which that vegetation patch was observed. These vectors will be the input given to the GenAI networks to learn co-occurrence patterns among species.

2.2 | Generative Artificial Intelligence systems

In the following, we describe the two techniques employed in this study, GAN and VAE.

2.2.1 | Generative adversarial networks (GAN)

We trained different GANs, denoted by GANn, with different dimensionalities *n* indicating the number of plant species considered in each of them (most abundant), plus the soil type: GAN8, GAN16 and GAN32. During the training, the GAN takes each real patch and creates a fake patch. At the beginning, the fake patches are very different from the real ones, but the GAN trains adversarially until the fake and the real patches are indistinguishable from each other. At that point, the GAN has reached the ability to recapitulate the existing patterns, that is, not only can it produce the initial real patches, but also any new fake patches which represent suitable possibilities. This is the generative feature of GANs, the ability to generate an infinite number of fake patches that were not found in the original dataset but reflect likely species' composition.

We use the *Python* library *fastai 2.1.5* to train a basic Wasserstein GAN with 10 dimensions in the input space, one extra layer in the generator and one in the critic, using ReLU activation functions with a negative slope of 0.2. The GAN is trained with RMSProp optimizer and 2×10^{-4} learning rate on 2D square images each representing the composition of a single patch. We construct these 2D images by taking as one direction the pattern of zeroes and ones describing the absence and presence of a given species or soil type in that patch, and repeating that pattern along a second dimension to form

a square. The reason to form these 2D inputs from one-dimensional vectors is to speed up the training of the GAN and VAE model, which are optimized for 2D inputs.

After training for 2,000 epochs, when the loss of the model reached a plateau, we produced 300,000 fake patches by feeding the GAN a 10-dimensional standard normal noise. We make sure the output could be translated into the original binary prediction for the absence/presence of species by removing the edges along the repeated dimension of the image, averaging along the remainder of that dimension, and finally using a threshold of 0.5 (the average between presence and absence of a species) to discretize the output.

To estimate the value of the systematic error of our procedure, we have performed the training procedure 14 times using the same number of epochs, but with independent, random initialization each time. The error bars in our figures depict two standard deviations around the mean, all computed over these 14 GAN runs.

Since the GANs are trained on real patches with at least two species of plants (to avoid training on the numerous patches that contain only a single species), we also reject fake patches containing fewer than two species (about 7.5%).

2.2.2 | Variational AutoEncoder (VAE)

We explore the ability of VAE to learn subtle species interactions in a space of a large dimensionality. Instead of training a GAN with a small number of features (species) we train a VAE with information of the most abundant 32 species. The VAE learns by looping around an encoder and decoder which transforms the real data into fake data. Our VAE takes as an input a rectangular greyscale 2D image. We extend the 1D line of zeros and ones representing the absence/presence of a species or soil type into a second dimension by repeating it eight times.

For this case, we build a convolution VAE using the *Python* library *Keras 2.3.1* and three 2×2 convolution layers with stride 2, with successive numbers of filters 128, 256 and 512, then reduce this to fit a 128-dimensional latent space. This architecture is customary to this size of images. The GAN is trained with Adam optimizer and 1×10^{-4} learning rate, partly optimized to balance accuracy goals with a reasonable amount of epochs and computational resources. We obtain 99.90% accuracy on a pixel-by-pixel level for our best model, which translates into a 99.19% accuracy at patch level. Indeed, to turn our rectangular monochrome images back into information about plants, we remove the edges along the repeated dimension of the image, average along the remainder of that dimension, and finally use a threshold to discretize the output. This value was set to 0.5, the average between presence and absence of a species.

In supervised Machine Learning methods, the accuracy and other measures of performance are obtained through a train/test separation in the dataset. But the validation of Generative methods (GAN and VAE in this paper) is adapted to unsupervised methods of learning, and is done as follows: the Generative algorithm trains by examining the real patches and adjusting the network parameters to produce avatars as closely as possible to the original patch. When we quote a 99% accuracy, this corresponds to the statement that we trained the Generative algorithm so it is able to transform a real patch into an avatar which is equal to the real patch 99% of the time. But once trained, the Generative algorithm will be able to generate new avatars by providing as input random numbers (instead of a real patch) and as output an avatar which, if trained correctly, should represent a realistic possibility. Note, though, that there are multiple options for measures of performance in GenAI models (Shmelkov et al., 2018), and ours just follows the intuitive and simple notion of similitude of images.

2.3 | Ecological validation: Patch species composition and plant soil affinity

To validate whether the GAN can produce species' co-occurrence patterns similar to those observed in the system, we focus on two features: The relative abundance of the different species compositions of patches and the affinity of certain species to a given soil type. We use GAN to generate 300 K fake patches, and then compare their features with those of the real patches characterized in the field. First, we quantified the relative abundance of patches with a given species composition, and compared the relative abundances between real and fake patches. Second, we tested whether the model trained without any information about soil type correctly identifies the affinity of gypsum specialists to gypsum soils.

2.4 | Contribution of direct and indirect interactions to species coexistence

Direct and indirect interactions among species can be quantified using conditional probabilities. To study direct pairwise interactions, one can compute the probability P(A|B) = probability that species A is present when B is present. In Figure 4a, we represent this conditional probability P(A|B) as a function of the relative abundance of species A in the GAN8 analysis, P(A). Different colours represent different choices of species A, and the circle sizes are related to the relative amount of a particular combination AB. The values of P(A) are found in the range of 25%–45%, as GAN8 is trained with the most abundant species.

If there were no interactions between A and B, P(A) = P(A|B), a situation which would follow the dashed trend line. Instead we observe that for a given species A, circles of the same colour along the vertical axis, P(A|B) lies outside that line, which corresponds to sizable interactions between A and B. Points above the dashed line indicate enhanced coexistence, and below depressed coexistence. Note that the quantity P(A|B) is not symmetric, that is, P(A|B) − P(B|A) is not necessarily zero, as situations when A or B are pioneers may be different. For example for A = *Fumana thymifolia* and B=*Brachypodim retusum*, P(A|B) = 0.5 and P(B|A) = 0.4.

To study indirect interactions, we can compute conditional probabilities involving three or more plants. In the two lower panels of Figure 4 we represent third- and fourth-order (indirect) interactions via conditional probabilities of presence of species A in patches where B, C and D are already present. In Figure 4b, points above the diagonal P(A|B) = P(A|BC) indicate that, in general terms, the presence of a third species enhances the co-occurrence of pairs, and points below imply that the third species suppresses the co-occurrence of pairs. The analysis can be carried over to higher-order interactions, as shown in Figure 4c, but paying the price of a lower probability represented by small circles.

2.5 | Forecasting the final composition of patches triggered by a pioneer species

Once we trained the VAES as described in Section 2.2 above, we started by considering a patch with a single pioneer, and then added to this patch a level of random noise. By processing this input through the VAE, we would obtain a number of possible configurations with other species, and their abundance would give us information on how likely that configuration would be.

3 | RESULTS

3.1 | Ecological validation: Patch species composition and plant soil affinity

The patches generated by the GAN8 and VAE models do indeed reproduce real patches' species composition, but also extend to produce new unseen but likely possibilities (Figure 2). In particular, the fake patches generated by GAN reproduce a similar abundance of patches with a given composition compared to the patches found in the field (Figure 2a). Furthermore, GAN is also able to produce new types of configurations beyond those used to train it (Figure 2b). While the real data (blue line in Figure 2b) shows a plateau in the number of patches with given species composition due to the limited amount of field observations, the GAN results (red line in Figure 2b) can exceed that amount, showing its ability to generate new possible species' composition in fake patches. Note that we find similar levels of capacity for learning for GAN8, GAN16 and GAN32, despite increasing dimensionality, when we compare abundances and interaction distributions between the real and fake generated patches. Yet for visualization we show results with the GAN8 model. A detailed comparison between the real data and the GAN results can be found at datapane (https://datapane.com/u/johannes/reports/gan), and the database and codes can be found at github (https://github.com/jegarcian/AI4Ecology) as well as zenodo (jegarcian, 2022; https://doi.org/10.5281/zenodo.5903355). These results indicate that the GAN training is not leading to mode collapse (Lala et al., 2018), or GAN overfitting, as the GAN is able to produce as much variety of the real distribution and a tail of less likely possibilities.

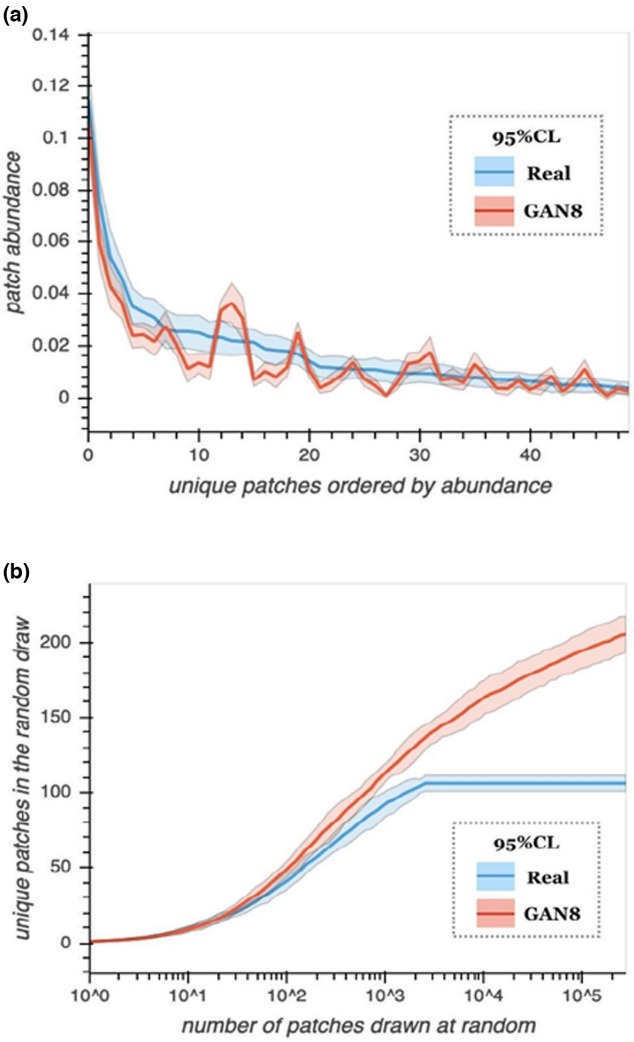

FIGURE 2 (a) Each patch ID represents a given species composition and they are ordered by abundance. Blue line represents the real abundance of patches, while the red line shows the abundance of patches generated by the GAN8 mode. (b) Number of patches with a given species composition in random samples of increasing size for both real and GAN8 patches. Abundances are shown with their 95% CL ranges

From an ecological perspective, the GAN accurately reproduces even without soil information, the affinity of gypsum specialists to gypsum soils (Figure 3). The prediction was good enough for the whole range of plant species' soil affinities, including species with high, medium or low affinity to the two different types of soils.

3.2 | Contribution of direct and indirect interactions to species coexistence

Direct interactions precluding species coexistence (dots below the diagonal in Figure 4a) were much more frequent than those promoting coexistence (dots below the diagonal in Figure 4a; $t = -17.01$; $p < 0.001$; $n = 839$), suggesting that most of the pairs of species seldom co-occur, either because species tend to live in different

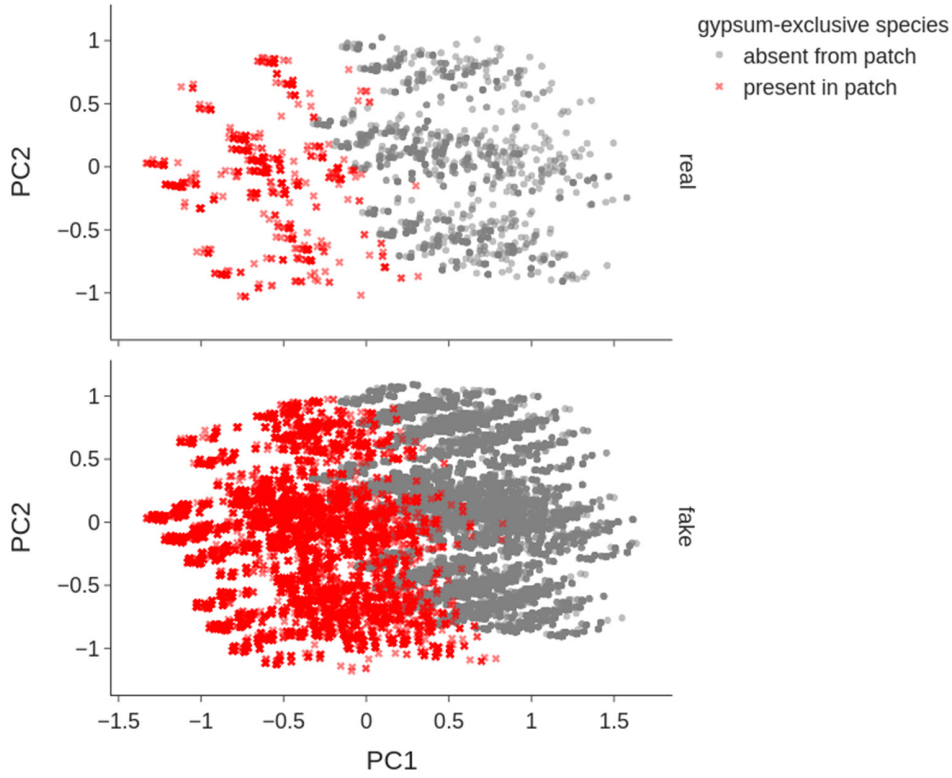

FIGURE 3 Projection onto two-dimensional of the 16-dimensional distribution of real patches and fake patches for GAN16 with soil type info withheld, highlighting the location of patches containing gypsum-exclusive species (*Teucrium libanitis* and *Helianthemum squamatum*). The upper plot depicts the first two principal components of a PCA calculated on the real patches summarizing the presence/absence of the 16 most common plant species (991 unique combinations), with information about soil type withheld. The bottom plot is obtained after applying the same projection to the fake patches (5,679 unique patch compositions) generated by GAN16 trained without any information about soil type. One can see that, although the GAN16 comes up with many fake patches with original combinations of species, all of those that contain gypsum-exclusive species are located in the same area as the real ones

habitats (e.g. soil types) and/or to exclude competitively each other. Pairs of species with low probability to coexist were not affected by the presence of a third species but coexistence of those with high probability to coexist tended to be suppressed in the presence of a third species. Similarly, the effect of a fourth species reduced the positive effects of the third species, as we exemplify below.

For example, let us focus on the purple vertical set of points at $x = 0.5$ in Figure 4b, which correspond to species A = *Fumana thymifolia*. From Figure 4a, we know that this species is present in about 40% of the patches. All these points in Figure 4b around $x = 0.5$ correspond to the coexistence with B = *Stipa tenacissima* or B = *Brachypodium retusum*, which boost the presence of *F. thymifolia* from 40% to 50%. But when another, third species C appears, the presence of *F. thymifolia* swings again in a wide range from a highly suppressed 10% (lower points at $x = 0.5$) due to the presence of *Teucrium libanitis* or *Helianthemum squamatum*, to enhanced to 60% (higher points at $x = 0.5$) due to the presence of *Stipa tenacissima*.

With the help of the GAN, we can go further than interactions among three and four species. Figure 4c shows the distribution of indirect interactions among three and four species. We observe a larger density of points close to the diagonal, indicating the slow weaning of the indirect interactions. However, there are still many outliers indicating strong fourth-order interactions. For example, let us focus on the light green dots in Figure 4c, corresponding to A = *Helianthemum syriacum*. From Figure 4a, we know that direct interactions do suppress the presence of this species, which on its own appears 42% of the time, but in lower frequencies when another species is present. From Figure 4b, we see that triple interactions do not overcome this suppression, with P(A|BC) always below 40%. But then in Figure 4c we see how the presence of yet another species, a fourth-order interaction, can change this trend, with a set of the light green combinations found above 40%. In particular, the fourth-order interactions resulting in an enhancement to 55% of the *Helianthemum syriacum* abundance are due to the co-occurrence with *Stipa tenacissima*, *Helianthemum squamatum* and *Fumana thymifolia*.

3.3 | Forecasting the final composition of patches triggered by a pioneer species

We also train VAEs with 8, 16 and 32 species. After training, the VAE reaches a high level of accuracy, above 99%. This accuracy means

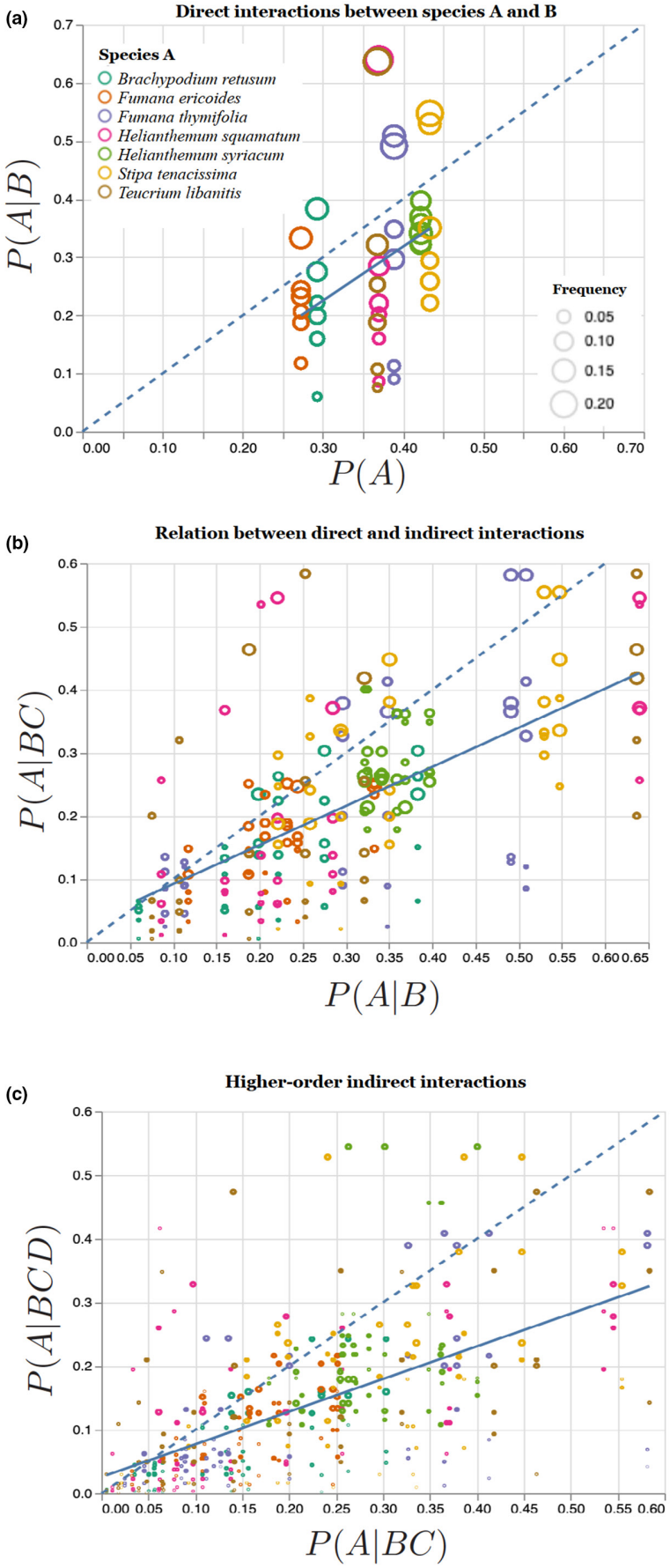

FIGURE 4 The strength of direct and indirect species interactions. (a) Relative abundance of species A P(A) vs the abundance a when another species B is already present in the patch, P(A|B). (b) Relation between the abundance of species A when B is present P(A|B) with the abundance of A when both B and C are present. The colour coding in both plots corresponds to species A, and B, C and D are varied. (c) Relation between triple (P(A|BC)) and higher-order interactions, represented by the conditional probability that species A is present when species B, C and D are already in the patch. In all the plots, the size of the circles indicates the relative abundance of a particular combination in the overall population. The dashed line is the diagonal $x = y$, which would correspond to the case of independent probabilities (no interactions). The solid blue line corresponds to a linear fit to the data

that the VAE is able to produce fake patches which strongly resemble the original patches. In particular, if we input a real patch composition, the VAE transformed patch would be identical to the input configuration 99% of the time.

We can exploit the VAE ability to represent the probability distribution in its latent space by, for example, inputting a pioneer species into this space and observing how the VAE generated probability distributions of generated patches with this pioneer, and thus evaluate whether there are better pioneer species. The results of VAE8 are shown in Figure 5, where we represent the distribution of unique patches generated by a single pioneer species introduced into the VAE's latent space. On the top of the plot, we observe that both *Helianthemum squamatum* and *Teucrium libanitis* as pioneer species produce a few independent types of patches with high probability (20%–40% each) and seldom any other, quickly saturating close to 100% after about 10 unique patches. On the other hand, using *Fumana ericoides* and *Helianthemum syriacum* as pioneer species produces a wide range of distinct patches, each with a low probability (5% or less). These last two species therefore seem to encourage a wider biodiversity.

4 | DISCUSSION

Species do interact in complex ways, with non-negligible indirect interactions leading to high boosting effects (Bairey et al., 2016). Therefore, a simple set of rules involving two species would not capture the whole set of patterns emerging in a community. This complexity and the inherent dimensionality of this problem motivate us to find a new approach to describe coexistence patterns, beyond heuristic rules (Bzdok et al., 2018). Here we show that unsupervised machine learning methods based on generative artificial intelligence correctly predict a range of characteristics related to species coexistence. Just feeding the models with the species composition of 5,153 vegetation patches in gypsum and limestone soils, we obtained correct predictions on (a) the relative abundance of patches with different species composition; (b) the plant species' soil affinity; and (c) the role of indirect interactions of third and fourth order into the coexistence of pairs of species. Furthermore, based on its ability to recapitulate existing patterns, the model should be able to predict the species composition of patches not registered in the field. This ability allows the model to generate realistic predictions on complex patterns that would be hard to detect in the field, such as the ecological succession trajectory given the colonization of a particular pioneer species.

In the context of species coexistence, the ability of Generative Artificial Intelligence to identify interactions of high order is especially relevant. Here we have identified third- and fourth-order interactions as an application with our dataset, but extensions to fifth order and sixth order would be possible with a larger dataset.

The salient picture of these analyses is that a high-order interaction tends to buffer the positive effects of the immediately lower-order interaction. For example, third-order interactions tend to promote exclusion between pairs of species that tend to coexist. Similarly, fourth-order interactions may depress the positive effects of coexistence produced by third-order interactions. Although this is the general trend, there are also indirect interactions that positively promote coexistence (points above the diagonal in Figure 4) or that have no effect on it (points close to the diagonal in Figure 4). The final outcome of the high-order interaction is absolutely dependent on the identity of the third (or fourth) species involved.

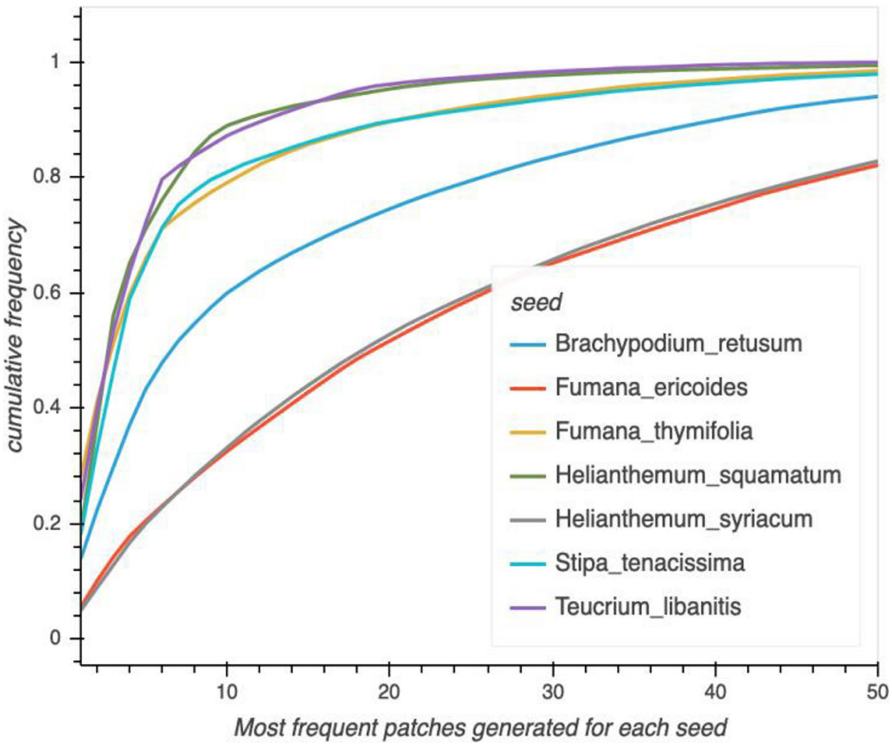

FIGURE 5 Cumulative distribution of unique patches generated by a single pioneer species introduced into the VAE's latent space

To get a mechanism behind these patterns, future research could supervise the learning process by including phenotypic, phylogenetic or other relevant information that can act as a proxy of other processes. For example, we could check whether the probability of two species with a low affinity to a particular stressful type of soil increases with the presence of a soil-specialist plant species, or whether the probability of exclusion between two closely related species decreases with the presence of a third closely related species. More generally, the predictions of the VAE could then help guide the restoration of ecosystems by planting those species that trigger succession, enhance the presence of many other species or favour the occurrence of a particular, endangered species.

To get correct answers, the researcher should follow several steps. First, the use of big data should be adequate for the particular question under scrutiny as data accumulated in large public datasets are theory laden (Devictor & Bensaude-Vincent, 2016). Second, sampling size should be enough to capture replicates of the interactions of order *n*th. As a general rule, for *n* species, the number of interactions of order *k* will be $n!/(n-k)!$. For example, if we are interested in third-order interactions in a community with 20 species, we will have $20!/(20-3)! = 6{,}840$ possible combinations that should be replicated. Third, proper training and validation are needed to decide whether the model is good enough to be used (see Christin et al., 2021). To further facilitate the application of this method to other datasets, the code used in this study has been made accessible in a repository (https://github.com/jegarcian/AI4Ecology) and detailed steps added to the documentation. Steps include as follows: pre-processing of the collected data, GAN/VAE training and data analysis, each of them has their own *notebook* which runs almost independently.

ACKNOWLEDGEMENTS

The authors thank the Yesaires team for making the fieldwork of quantification of species gypsum affinity possible. R.S.-M. was supported by the Ministry of Science and Innovations (FPU grant FPU17/00629). Financial support was provided by the projects RTI2018-099672-J-I00 and PID2020-113157GB-I00 (funded by MCIN/AEI/10.13039/501100011033 and 'ERDF A way of making Europe').

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

M.V. and A.M.-N. conceived the idea, V.S., J.H. and J.E.G. designed the methodology; R.S.-M. and A.M.-N. collected the data; V.S., J.H. and J.E.G. analysed the data; M.V. and V.S. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13827.

DATA AVAILABILITY STATEMENT

Data and code to run the analyses can be found at zenodo (jegarcian, 2022; https://doi.org/10.5281/zenodo.5903355), datapane (https://datapane.com/u/johannes/reports/gan/) and github (https://github.com/jegarcian/AI4Ecology).

ORCID

Johannes Hirn https://orcid.org/0000-0003-0267-2479
José Enrique García https://orcid.org/0000-0002-0279-0523
Alicia Montesinos-Navarro https://orcid.org/0000-0003-4656-0321
Ricardo Sánchez-Martín https://orcid.org/0000-0001-5272-3276
Veronica Sanz https://orcid.org/0000-0001-8864-2507
Miguel Verdú https://orcid.org/0000-0002-9778-7692

REFERENCES

Bairey, E., Kelsic, E. D., & Kishony, R. (2016). High-order species interactions shape ecosystem diversity. *Nature Communications*, *7*, 1–7. https://doi.org/10.1038/ncomms12285

Balamurugan, S. A. A., Chitra, P. K. A., & Geetha, S. (2019). Multi label learning approaches for multi species avifaunal occurrence modelling: A case study of south eastern Tamil Nadu. *International Journal of Business Intelligence and Data Mining*, *15*, 449–477. https://doi.org/10.1504/IJBIDM.2019.102804

Battey, C. J., Coffing, G. C., & Kern, A. D. (2021). Visualizing population structure with variational autoencoders. *G3*, *11*(1), jkaa036. https://doi.org/10.1093/g3journal/jkaa036

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*, 233–234. https://doi.org/10.1038/nmeth.4642

Castillo, J. P., Verdú, M., & Valiente-Banuet, A. (2010). Neighborhood phylodiversity affects plant performance. *Ecology*, *91*, 3656–3663. https://doi.org/10.1890/10-0720.1

Chen, D., Xue, Y., Fink, D., Chen, S., & Gomes, C. P. (2017). Deep multi-species embedding. arXiv preprint arXiv:1609.09353. Retrieved from https://arxiv.org/abs/1609.09353

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

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*, 1632–1644. https://doi.org/10.1111/2041-210X.13256

Christin, S., Hervet, É., & Lecomte, N. (2021). Going further with model verification and deep learning. *Methods in Ecology and Evolution*, *12*, 130–134. https://doi.org/10.1111/2041-210X.13494

Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the Netflix problem. *PeerJ*, *5*, e3644. https://doi.org/10.7717/peerj.3644

Devictor, V., & Bensaude-Vincent, B. (2016). From ecological records to big data: The invention of global biodiversity. *History and Philosophy of the Life Sciences*, *38*(4), 1–23. https://doi.org/10.1007/s40656-016-0113-2

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *Processing Systems*, *27*, 2672–2680.

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

Iten, R., Metger, T., Wilming, H., Del Rio, L., & Renner, R. (2020). Discovering physical concepts with neural networks. *Physical Review Letters*, *124*(1), 010508.

jegarcian. (2022). jegarcian/AI4Ecology: Methods in Ecology and Evolution. *Zenodo*, https://doi.org/10.5281/zenodo.5903355

Joseph, M. B. (2020). Neural hierarchical models of ecological populations. *Ecology Letters*, *23*, 734–747. https://doi.org/10.1111/ele.13462

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Lala, S., Shady, M., Belyaeva, A., & Liu, M. (2018). Evaluation of mode collapse in generative adversarial networks. *Proceedings of the IEEE High Performance Extreme Computing*, *10*, 25–27.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920

Montesinos-Navarro, A., Storer, I., & Perez-Barrales, R. (2019). Benefits for nurse and facilitated plants emerge when interactions are considered along the entire life-span. *Perspectives in Plant Ecology, Evolution and Systematics*, *41*, 125483. https://doi.org/10.1016/j.ppees.2019.125483

Poisot, T., Ouellet, M. A., Mollentze, N., Farrell, M. J., Becker, D. J., Albery, G. F., Gibb, R. J., Seifert, S. N., & Carlson, C. J. (2021). Imputing the mammalian virome with linear filtering and singular value decomposition. arXiv preprint arXiv:2105.14973. Retrieved from https://arxiv.org/abs/2105.14973

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:2847–2854. Retrieved from https://arxiv.org/pdf/1606.05336.pdf

Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. arXiv e-prints, arXiv:2103.05180.

Schöb, C., Armas, C., & Pugnaire, F. I. (2013). Direct and indirect interactions co-determine species composition in nurse plant systems. *Oikos*, *122*, 1371–1379. https://doi.org/10.1111/j.1600-0706.2013.00390.x

Shmelkov, K., Schmid, C., & Alahari, K. (2018). How good is my GAN? In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds), *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 218–234). Springer.

Simmons, B. I., Cirtwill, A. R., Baker, N. J., Wauchope, H. S., Dicks, L. V., Stouffer, D. B., & Sutherland, W. J. (2019). Motifs in bipartite ecological networks: Uncovering indirect interactions. *Oikos*, *128*, 154–170. https://doi.org/10.1111/oik.05670

Strauss, S. Y. (1991). Indirect effects in community ecology: Their definition, study and importance. *Trends in Ecology & Evolution*, *6*, 206–210. https://doi.org/10.1016/0169-5347(91)90023-Q

Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz, N. R., Higino, G., Mercier, B., González, A., Gravel, D., Pollock, L. A., & Poisot, T. (2021). A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B*, *376*, 20210063. https://doi.org/10.1098/rstb.2021.0063

Tang, L., Xue, Y., Chen, D., & Gomes, C. (2018). Multi-entity dependence learning with rich context via conditional variational auto-encoder. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, no. 1). Retrieved from https://arxiv.org/abs/1709.05612

van Kleunen, M., Dawson, W., Bossdorf, O., & Fischer, M. (2014). The more the merrier: Multi-species experiments in ecology. *Basic and Applied Ecology*, *15*, 1–9. https://doi.org/10.1016/j.baae.2013.10.006

How to cite this article: Hirn, J., García, J. E., Montesinos-Navarro, A., Sánchez-Martín, R., Sanz, V. & Verdú, M. (2022). A deep Generative Artificial Intelligence system to predict species coexistence patterns. *Methods in Ecology and Evolution*, 13, 1052–1061. https://doi.org/10.1111/2041-210X.13827