A generative deep learning approach for global species distribution prediction

Yujing Yan[1*], Bin Shao[2*], Charles C. Davis[1]

[1] Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, Cambridge MA 02138, USA

[2] Department of Molecular and Cellular Biology, Harvard University, Cambridge MA 02138, USA

***Corresponding authors:**

Yujing Yan: yjyan7@gmail.com

Bin Shao: shaobinlx@gmail.com

Abstract

Anthropogenic pressures on biodiversity necessitate efficient and highly scalable methods to predict global species distributions. Current species distribution models (SDMs) face limitations with large-scale datasets, complex interspecies interactions, and data quality. Here, we introduce EcoVAE, a framework of autoencoder-based generative models trained separately on nearly 124 million georeferenced occurrences from taxa including plants, butterflies and mammals, to predict their global distributions at both genus and species levels. EcoVAE achieves high precision and speed, captures underlying distribution patterns through unsupervised learning, and reveals interspecies interactions via *in silico* perturbation analyses. Additionally, it evaluates global sampling efforts and interpolates distributions without relying on environmental variables, offering new applications for biodiversity exploration and monitoring.

Main

Anthropogenic pressures have intensified the need for efficient and scalable methods to predict species distributions for attaining a comprehensive picture of biodiversity. Over the past two decades, species distribution modeling (SDM) has become an essential tool for this purpose [1–3],typically using species occurrence data and environmental variables to predict distributions through statistical and machine learning (ML) algorithms [4–7]. While the mobilization of vast amounts of specimen records and the rapid accumulation of observational data have greatly promoted the development of SDMs [8–10], several challenges still persist.

First, current SDMs struggle to handle large-scale datasets in our big data era [8–12], especially for modeling species assemblages. Traditional methods can only address these tasks through computationally intensive “stacking methods” that have limited scalability. Second, most SDMs overlook complex interspecies interactions, limiting their ecological relevance and utility in modeling community dynamics [13–16]. Third, the accuracy of current techniques depends heavily on input data quality and resolution [17,18]. Although platforms like GBIF and eBird provide valuable datasets for exploration, these are often biased by varying observation efforts across taxonomic groups and regions [19–21]. Finally, reliance on environmental variables introduces additional issues, including collinearity and limited availability in certain regions, further constraining model accuracy and applicability[22].

Generative models are a type of deep learning model that captures complex and nonlinear relationships between input variables. They have been widely adopted in various fields, including natural language processing [23], image generation [24], data capturing [25], and biomedicine [26]. Among these models, autoencoders are designed to compress and reconstruct data in an unsupervised way, making it effective for data denoising, interpolation, and handling randomly missing data [27].

Here, we present for the first time an autoencoder-based framework, named Ecological Variational Autoencoder (EcoVAE), to predict species distributions for the first time using large-scale, unstructured, and sparse occurrence data. To demonstrate the effectiveness of our framework, we trained a customized EcoVAE on a massive global dataset including nearly 34 million georeferenced vouchered occurrences from the GBIF platform spanning 13,125 plant genera and 127,281 species. EcoVAE learns the patterns of global plant distributions without any dependency on environmental variables. It efficiently processes this large-scale dataset with fast computational times, accommodates varying rates of missing data and biases, and enables the study of complex interactions within communities. Remarkably, EcoVAE can accurately reconstruct full plant distributions across all genera using as little as 20% of randomly selected occurrence records. We further demonstrate the broad applicability of EcoVAE by applying it to 68 million occurrence records of butterflies and 22 million records of mammals at both the genus and species levels. Additionally, our model predictions offer an unsupervised approach to assess collection completeness across different taxa on a global scale. These results demonstrate the unprecedented capacity of deep learning methods to decode and predict biodiversity patterns at global scales.

EcoVAE applies a unique masked approach to model the global species distributions using well-curated occurrences data based on vouchered specimen records at the rank of genus or species (Fig. 1a). We treat genera as the unit of inference in the modeling process due to their computational efficiency and biological relevance. Genera represent coherent, morphologically similar, and often monophyletic groups of species, providing a practical compromise between taxonomic detail and manageable computational demands. The input data was grouped into grids of 0.1', where plant observations within each grid were summarized into vectors. The richness of genera per grid varied widely, ranging from 10^1 to 10^3 (Supplementary Fig. 1a). Our model consists of an encoder that learns a low-dimensional representation of the input data and a decoder that reconstructs the presence of genera per grid. For model training, we randomly masked 50% of the genera presence data and the model was trained to predict these masked genera based on the remaining observed data (Fig. 1b). This process allows the model to interpolate sparse observations and estimate the true, unobserved plant distributions.

We evaluated the performance of our model in three randomly selected regions in North America, Europe, and Asia, and applied the remaining global data for training (Fig. 1c, Supplementary Table 1). Our model demonstrated a 10-fold increase in computational speed compared to popular traditional SDMs, including logistic regression and random forest, when predicting the distribution of a single genus using all other genera as input (Fig. 1d). This difference is even greater when predicting distributions of multiple genera (Fig. 1e).

We further assess the model's accuracy in prediction. Here, we calculated the predicted genera counts per grid for the test regions and compared them to the actual observations. The results demonstrated very high Pearson correlation coefficients of 0.98, 0.99 and 0.99 for test regions in North America, Europe, and Asia, respectively. At the species level, our model achieves correlation coefficients of 0.95, 0.98, and 0.98 across the three regions (Supplementary Fig. 2). We also calculated the total number of genera present in each grid and observed high correlations between the model's predictions and the actual data (Fig. 1f).

Next, we used the Area Under the Receiver Operating Characteristic (AUROC) curve to evaluate EcoVAE's accuracy in modeling the distribution of each genus. For the masked genera, the mean AUROC was 0.82 for North America, 0.83 for Europe, 0.85 for Asia, which demonstrates the robust performance of our model to infer missing information from incomplete datasets (Fig. 1g). For example, *Lonicera* has a localized distribution in North America and our model correctly predicted this pattern despite this genus being masked in the input, with an overlap rate of 0.90 (Methods). Similar performance was observed for *Lamium* in Europe (0.89) and for *Rhus* in Asia (0.80), which exhibit more scattered distribution patterns (Fig. 1i). It is important to note that the three regions we selected randomly differ substantially in plant distributions, area size, and genera counts per grid (Supplementary Fig. 1b-d, Supplementary Table 1). Nevertheless, our model performed equally well across them, which highlights the wide applicability of EcoVAE to diverse geographic contexts. Herbarium specimen records represent a sparse sampling of actual plant distributions, and data completeness varies significantly across regions [28,29]. To address this inherent limitation, we analyzed the impact of data sparsity on our model's performance. We tested our model using only 1%, 5%, 10%, 20%, and 30% of the input genera, and evaluated its performance based on the AUROC for the remaining genera. With only 1% of the input data, the model's performance was relatively low and the mean

AUROC was 0.56. However, increasing the input to only 5% improved the mean AUROC to 0.68. The mean AUROC further rose to 0.77 when we used 20% of the input data, close to the performance seen with 50% (AUROC of 0.81). These results demonstrate that our trained model can effectively use as little as 20% of the available data to reconstruct the full generic distribution with high precision (Fig. 1h).

We extended our modeling framework to other major clades with high conservation values, i.e., butterflies and mammals, and evaluated its performance using a similar approach with three test regions (Supplementary Table 2). For butterflies, we found that our model achieves high accuracy in predicting the number of genera per grid, with the Pearson's correlation coefficients of 0.96, 0.99, 0.80 for genera counts per grid for the test regions (Supplementary Fig. 3). The AUROC scores for genus-level predictions were 0.79, 0.84, and 0.75 for these regions. At the species level, the model achieved comparable results for North America and Europe, but the AUROC decreased to 0.68 for Asia, which may reflect its incompleteness of vouchered occurrences at the species level. For mammals, the model performed best in the test region of North America at both genus and species levels (Supplementary Fig. 4). In contrast, the sparser data in Asia posed challenges for reconstructing full species distributions. Overall, our results demonstrate that EcoVAE generalizes effectively across diverse taxa and geographies.

One important application of species modeling is interpolating occurrences where data were lacking. We assume that the prediction error of EcoVAE reflects the completeness of the occurrence records: if the records are incomplete, the model will struggle to reconstruct the input data effectively. We estimated the prediction error globally (Methods) and found that regions with high prediction error overlap with known "darkspots" of biodiversity collection [30,31]. For example, the highest prediction errors for plants were observed in South Asia, Southeast Asia, the Middle East, and Central Africa. South America showed higher prediction errors compared to North America (Fig. 2b). Notably, despite generally sparse records from high-latitude regions, the prediction error remained low, suggesting that the occurrence records in these areas are nearly complete, which allows the model to reflect true species distributions (Fig. 2b) more accurately. For butterflies, the highest prediction errors were observed in South America and parts of

Southeast Asia (Supplementary Fig. 5). Interestingly, Middle Africa exhibited low prediction errors, in contrast to the patterns observed for plants. For mammals, the prediction error was generally smaller, likely due to the low genus diversity. However, regions in South America and Central Asia displayed comparatively high prediction errors, highlighting the need for further investigation efforts in these regions.

We then assessed the interpolation power of EcoVAE on i.) a region in southeastern North America with relatively incomplete herbarium records but rich observation data from iNaturalist, and ii.) a region in South Asia with sparse online occurrence records of both kinds (Fig. 2a). We applied the same model structure to train a full global model based on all available plant voucher records and generated the new predictions in this test region (Methods). In North America, we calculated an overlap index for each genus with iNaturalist observations, defined as the ratio of predictions that are absent from input data but present in the iNaturalist data. We found that our model performed best for genera with a moderate number of observations, while abundant data results in diminishing returns from interpolation (Supplementary Fig. 6). Using the genus *Sassafras* as an example, we found that the new predictions largely overlap with the iNaturalist data (Fig. 2b).

For regions like South Asia, both georeferenced vouchered and observational datasets are sparser. We selected genera that showed significant expansion in our model's new predictions. For example, the *Desmodium* genus in the Fabaceae family only have vouchered specimens in the eastern Himalayan/Nepal region in GBIF, but our model predicts its much wider distribution across western and southern India (Fig. 2d), which aligns better with field surveys and floristic investigations [32,33]. Similarly, for *Melicope* in Rutaceae, the original observations were localized in southern India, but our new predictions included occurrences in Myanmar and lowlands of Nepal, which is also confirmed by third-party observations (Fig. 2d) [34,35]. For *Adonis*, our new predictions suggested a broad distribution across the Himalayan region, consistent with various local floras and checklists describing the widespread nature of the genus from Pakistan to temperate regions in China (Fig. 2d) [36]. These results highlight the power of our model to uncover plant distribution patterns in regions where observational data are limited.

Another important aspect of distribution modeling is understanding the community response to changing distributions of organisms within it. Here, we interrogated our full global model to study genus-to-genus interactions. We selected a test region in Australia with abundant occurrences data and a good representation of major biomes (Supplementary Fig. 7, Supplementary Table 1). We conducted *in silico* perturbation analysis, where each grid cell in the region was artificially altered by introducing a genus to areas where it was previously absent. By comparing the perturbed predictions to unperturbed models (Fig. 2e), we assessed the invasive potential of one genus on others' distributions. We focused on statistically significant interactions for downstream analysis (Methods, Supplementary Fig. 8).

Examination of our genus network revealed that genera with high out-degree (those influencing others significantly) tend to have low in-degree (being influenced by others), suggesting asymmetric interactions [37] (Fig. 2f). Genera with broader ranges tend to interact with a larger number of genera (Supplementary Fig. 9). We revealed that certain families, including Poaceae, Cyperaceae, and Amaranthaceae, are more sensitive to disturbance in the study region (Fig. 2g). These families were significantly influenced by members from Poaceae, Asteraceae, and Fabaceae, which are globally well-represented in naturalized and invasive floras (Supplementary Fig. 10) [38]. Thus, our model reveals patterns of genus interactions, providing insights into community dynamics that may not be directly observable in the original co-occurrence records.

In this work, we present EcoVAE, a generative deep learning framework for modeling global plant distributions with high precision and speed. EcoVAE evaluates global sampling efforts, interpolates distributions, and reveals interspecies interactions via *in silico* perturbation analyses, offering novel applications for biodiversity exploration and monitoring. It demonstrates that species distributions can be reconstructed using co-occurrence information alone, even with incomplete data, capturing ecological patterns often missed by traditional approaches. EcoVAE complements current SDMs by providing a scalable framework for global analyses that can guide more targeted ecological studies. For instance, it can identify under-sampled regions or unexpected patterns, directing SDM efforts and field surveys to areas most in need of investigation. Furthermore, while we have demonstrated the high performance of EcoVAE on taxa such as plants, butterflies, and mammals, it can be easily extended to other taxa including birds and invertebrates. We envision that EcoVAE will advance biodiversity investigations, especially in under-sampled regions with limited environmental data, and ultimately support global biodiversity monitoring efforts aligned with the Convention on Biological Diversity [39]. Future integration of additional data, such as geographic or climate variables, could potentially improve performance and reveal insights into organism distributions and environmental change.

Methods

Data preprocessing

We downloaded the world plant, butterfly and mammal distribution data from the GBIF platform (Supplementary Note 1). We used the R package "CoordinateCleaner" [40] to remove records located in the sea, on country or major area centroids, capitals, or in major biodiversity facilities. We modeled the distribution both at the genus and species level. The cleaned observation dataset includes 33.8 million observations of plant, 67.6 million observations of butterfly and 21.9 million observations for mammal (Supplementary Table 2). The grid size was set to 0.1' x 0.1' and we summed all the observation data in each grid. For any genus or species with more than one observation, we set the value to 1 (existence) in contrast to 0 (non-existence). For plant observations, we only kept genera or species that occurs in over 20 grid cells and grid cells with more than 5 different genera. The finalized data contains 277,133 grids, while the three test regions include 11,292, 3,878 and 1,210 grids for North America, Europe, and Asia respectively. For butterfly and mammal observations, we only kept genera or species that occur in more than 5 grid cells, and grid cells with more than 1 genus or species.

Model structure

We developed an Ecological Variational Autoencoder (EcoVAE) model which aims to reconstruct the full plant distribution based on partial observations. The core VAE architecture comprises an encoder and a decoder.

1. Encoder function: the encoder is implemented as a sequential network with two hidden layers of Gelu activated linear transformations (dimension: 128). The encoder maps the input data (e.g., dimension: 13,125 for plant genera) into a latent space characterized by mean (μ) and log variance ($log\ (\sigma^2)$) parameters (dimension: 32).

$$(\mu, log\ log\ (\sigma^2)\ ) = f_{encoder}(x) = GELU\left(W_2\left(GELU(W_1x + b_1)\right) + b_2\right)$$

Where W1, b1, W2, b2 are the weights and biases of the two hidden layers, and GELU denotes the Gaussian Error Linear Unit activation function.

2. Reparameterization: the latent representation is obtained through a reparameterization step that ensures differentiability by sampling from a Gaussian distribution parameterized by these μ and $log\ (\sigma^2)$.

$$z = \mu + \sigma \odot \epsilon$$

where ϵ follows normal distribution, $\epsilon \sim N(0,\ I)$.

3. Decoder function: the decoder, mirroring the encoder's structure, reconstructs the input data from the latent space, aiming to minimize reconstruction error. The output dimension is equivalent to the input dimension (13,125), ensuring that the reconstructed output mirrors the input feature set.

$$\hat{x} = f_{decoder}(z) = GELU\big(W_4\big(GELU(W_3 z + b_3)\big) + b_4\big)$$

Where W3, b3, W4, b4 are the corresponding weights and biases for the decoder layers.

This mathematical framework enables the EcoVAE to compress high-dimensional data into a lower-dimensional latent space and subsequently reconstruct the original data with minimized reconstruction error.

Model training and evaluation

During the model training process, we used a unique masking strategy where 50% of input data are randomly set to zero to simulate missing data scenarios. The model uses a weighted reconstruction loss function that is based on mean squared error (MSE). A weighting factor of 0.5 modifies the contribution of masked and unmasked genera to the reconstruction loss, providing a balanced approach to learning from both visible and obscured portions of the data. We used the Adam optimizer with a fixed learning rate of 0.001, and the models were trained for 15 epochs with a batch size of 512.

To evaluate the model's performance, we excluded data from three randomly selected regions in Europe, North America, and China, during the training phase (Supplementary Table 1). After model training, we conducted a masking procedure where we randomly selected 50% of all genera or species from these regions and set their corresponding observational data to zero. These masked data were then used as inputs to evaluate the model's ability to reconstruct the observational data for the masked genera or species from the unmasked genera. We quantitatively measured the model's performance using the Area Under the Receiver Operating Characteristic (AUROC) and MSE metrics (Supplementary Fig. 2). For each genus or species in the test regions, we selected an equal number of top predicted grids as in the original data. We calculated the overlap rate as the fraction of predictions that had a true occurrence record either within that grid or in neighboring grids.

Model benchmarking

To evaluate the time efficiency of our EcoVAE model in processing large-scale input data, we conducted a series of benchmarking experiments in comparison with two popular SDM methods, i.e., random forest and logistic regression [18]. We tested the model's performance under varying input dimensions by randomly selecting genera from the input genus presence matrix. For these experiments, we used a masking strategy where 80% of the input columns were masked to simulate missing data. A binary mask was generated using a probability threshold proportional to the desired masking percentage, ensuring that columns in the input data were set to zero with the defined probability. We used the training and test data split as previously described. A random genus with presence in at least five grids was randomly chosen as the prediction target: the model should predict the presence of this genus across all grids in the training and test data based on the masked input matrix. We trained the random forest classifier using the *sklearn.ensemble.RandomForestClassifier* function with default parameters. We trained the logistic regression model with the *sklearn.linear_model. LogisticRegression* function with the following parameters: max_iter=1000, solver='saga', penalty='l1'. Time measurements were recorded from the start of data preparation to the completion of the prediction phase. The time consumption was calculated over 10 iterations to benchmark the models' time efficiency.

Model application

-Data interpolation

We trained the EcoVAE model as previously described on all available occurrence data for interpolating unobserved plant distribution (full global model). To evaluate the model's prediction error, we used the unmasked global data as input and applied a threshold to binarize the output genus presence matrix, ensuring that the total number of occurrences was doubled. For each grid, we then calculated the ratio of observed genera not represented in the output matrix, which we defined as the prediction error. Based on the global distribution of prediction error, we selected two regions to evaluate the performance of data interpolation, i.e., North America and South Asia. For these regions, we used a similar strategy to generate the binarized output for downstream analysis.

For the North America region, we used observation data collected from iNaturalist to verify the accuracy of model prediction instead of traditional data splitting method. The ratio between iNaturalist observational data and vouchered specimen data is 7:1 and we would expect that iNaturalist data have a better geographic coverage than herbarium specimens for many species. We calculated overlapping rate between predicted species occurrences and actual observations to quantify the performance of our prediction. For each species, we first extracted the observed and predicted occurrences based on the genus index. For both datasets, we retained only the presence points. The observed points were buffered by 0.1 degrees to account for geographic uncertainties. We then converted the presence data into spatial objects using the "sf" package in R [41]. Overlap between predicted occurrences and observed occurrences, as well as overlap between predicted and original input occurrences, was calculated using spatial intersections (`st_intersects`). The key metric, the overlapping rate, was calculated by dividing the number of predicted points that overlapped with observed points but not with original input points by the total number of predicted points. This rate reflects the proportion of new predicted occurrences that align with observed data but were not part of the input data, providing a measure of prediction accuracy and novelty. We also compared the predicted overlapping rate with the overlapping rate calculated between the same number of randomly generated points in the study area and observations.

For the South Asia region, we assessed the occurrence of each genus both before and after data interpolation. We focused on genera that initially occurred in more than 5 grids and whose distribution region has expanded most for downstream analysis. Due to the lack of georeferenced observational data in this region, we compared our prediction with the distribution described in plant atlas and related literature.

-Simulation of genera interaction

To simulate the impact of a specific genus i on all other genera within a targeted region, we initially identified all grid cells lacking genus i. The observational data from these grids were utilized as input for the model, and the corresponding reconstructed data served as the background dataset (x_background). Then we introduced observations of genus i into these grids and generated perturbed model outputs (x_perturb). By comparing the plant distributions between x_background and x_perturb, we were able to identify genera that exhibited significant changes, thereby quantifying the ecological influence of genus i on the plant community dynamics within the region.

To assess species interactions after species additions, we first fit linear regression models to compare grid numbers of all genera before and after addition of a specific genus i. Specifically, we used the ‘lm’ function in R to model the relationship between grid numbers before and after addition of genus i. For each model, we calculated 99.99% confidence intervals using the ‘predict’ function with the interval parameter of "prediction" and level parameter of 0.9999. We defined the significant interaction between genus i and j if the predicted grid number for genus j falls outside the bounds of the confidence intervals after addition of genus i. In such circumstances, we identified j as an “outlier” and defined it as a “sensitive genus”. Z-scores were calculated for each genus by normalizing the residuals, computed as the difference between actual and predicted values. We then performed frequency analysis of the impactful and vulnerable genera based on all significant interactions. To further explore genus interactions at the family level, we utilized the "plantlist" package [42] to classify genera into families and analyzed the proportion of sensitive genera within each family. We selected the most sensitive family based on the following criteria: it includes more than 5 genera and at least 35% of the genera are classified as "sensitive" (significantly impacted by at least one other genus).

Author Contributions

YY and SB conceptualized the study, conceived EcoVAE, collected and analyzed the data. YY and SB wrote the manuscript with key contributions from CCD. All authors approved the manuscript.

Acknowledgements

YY and SB were supported by Harvard University. CCD was supported by Harvard University, LVMH Research, Dior Science, and National Science Foundation grants DEB-1355064 and DEB-0544039.

References

1. Guisan, A. *et al.* Predicting species distributions for conservation decisions. *Ecology Letters* **16**, 1424–1435 (2013).
2. Franklin, J. Species distribution models in conservation biogeography: developments and challenges. *Diversity and Distributions* **19**, 1217–1223 (2013).
3. Franklin, J. Species distribution modelling supports the study of past, present and future biogeographies. *Journal of Biogeography* **50**, 1533–1545 (2023).
4. Elith*, J. *et al.* Novel methods improve prediction of species' distributions from

occurrence data. *Ecography* **29**, 129–151 (2006).

5. Chollet Ramampiandra, E., Scheidegger, A., Wydler, J. & Schuwirth, N. A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation. *Ecological Modelling* **481**, 110353 (2023).
6. Franklin, J. Species distribution modeling. in *Mapping Species Distributions: spatial inference and prediction* 3–20 (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2009).
7. Zurell, D. *et al.* Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology* **22**, 2651–2664 (2016).
8. Franklin, J., Serra-Diaz, J. M., Syphard, A. D. & Regan, H. M. Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography* (2016) doi:10.1111/geb.12501.
9. Runting, R. K., Phinn, S., Xie, Z., Venter, O. & Watson, J. E. M. Opportunities for big data in conservation and sustainability. *Nat Commun* **11**, 2003 (2020).
10. Di Cecco, G. J. *et al.* Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. *BioScience* **71**, 1179–1188 (2021).
11. Davis, C. C. The herbarium of the future. *Trends in Ecology & Evolution* **38**, 412–423 (2023).
12. Hedrick, B. P. *et al.* Digitization and the Future of Natural History Collections. *BioScience* **70**, 243–251 (2020).
13. Araújo, M. B. & Luoto, M. The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* **16**, 743–753 (2007).
14. Van der Putten, W. H., Macel, M. & Visser, M. E. 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 of London. Series B, Biological sciences* **365**, 2025–2034 (2010).
15. Thuiller, W. *et al.* Navigating the integration of biotic interactions in biogeography. *Journal of Biogeography* **51**, 550–559 (2024).
16. Wisz, M. S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological*

Reviews **88**, 15–30 (2013).

17. Guisan, A. *et al.* Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* **13**, 332–340 (2007).

18. Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J. & Elith, J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs* **92**, e01486 (2022).

19. Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* **19**, 10–15 (2014).

20. Garcia-Rosello, E., Gonzalez-Dacosta, J., Guisande, C. & Lobo, J. M. GBIF falls short of providing a representative picture of the global distribution of insects. *Systematic Entomology* **48**, 489–497 (2023).

21. Pender, J. E. *et al.* How sensitive are climatic niche inferences to distribution data sampling? A comparison of Biota of North America Program (BONAP) and Global Biodiversity Information Facility (GBIF) datasets. *Ecological Informatics* **54**, 100991 (2019).

22. Dormann, C. F. *et al.* Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 027–046 (2013).

23. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. Preprint at https://doi.org/10.48550/arXiv.2203.02155 (2022).

24. Ramesh, A. *et al.* Zero-Shot Text-to-Image Generation. Preprint at https://doi.org/10.48550/arXiv.2102.12092 (2021).

25. Weaver, W. N., Ruhfel, B. R., Lough, K. J. & Smith, S. A. Herbarium specimen label transcription reimagined with large language models: Capabilities, productivity, and risks. *American J of Botany* **110**, e16256 (2023).

26. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053–1058 (2018).

27. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2022).

28. Daru, B. H. *et al.* Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol* **217**, 939–955 (2018).

29. Daru, B. H. & Rodriguez, J. Mass production of unvouchered records fails to represent global biodiversity patterns. *Nat Ecol Evol* **7**, 816–831 (2023).

30. Daru, B. H. Predicting undetected native vascular plant diversity at a global scale. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2319989121 (2024).

31. Ondo, I. *et al.* Plant diversity darkspots for global collection priorities. *New Phytologist* **244**, 719–733 (2024).

32. Joshi, B. R., Hakim, M. M. & Patel, I. C. The biological active compounds and biological activities of Desmodium species from Indian region: a review. *Beni-Suef Univ J Basic Appl Sci* **12**, 1 (2023).

33. Desmodium gangeticum (L.)DC. Species. *India Biodiversity Portal* https://indiabiodiversity.org/species/show/229507.

34. Wood, K. R., Appelhans, M. S. & Wagner, W. L. Melicope oppenheimeri, section Pelea (Rutaceae), a new species from West Maui, Hawaiian Islands: with notes on its ecology, conservation, and phylogenetic placement. *PK* **69**, 51–64 (2016).

35. Hartley TG. On the Taxonomy and Biogeography of Euodia and *Melicope* (Rutaceae). in *Allertonia* vol. 8(1) (National Tropical Botanical Garden, Lawaʻi, Kauaʻi, Hawaiʻi., 2001).

36. Fu, D. Adonis, in *Flora of China* **6**: 389-391. (2001).

37. Bascompte, J., Jordano, P. & Olesen, J. M. Asymmetric Coevolutionary Networks Facilitate Biodiversity Maintenance. *Science* **312**, 431–433 (2006).

38. Pyšek, P. *et al.* Naturalized alien flora of the world: species diversity, taxonomic and phylogenetic patterns, geographic distribution and global hotspots of plant invasion. *Preslia* **89**, 203–274 (2017).

39. Díaz, S. *et al.* Set ambitious goals for biodiversity and sustainability. *Science* **370**, 411–413 (2020).

40. Zizka, A. *et al.* CoordinateCleaner : Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* **10**, 744–751 (2019).

41. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**, 439 (2018).

42. Zhang, J. Plantlist: Looking Up the Status of Plant Scientific Names Based on The Plant List Database. (2019). R package version

(4.3.3) https://github.com/helixcn/plantlist/

Code Availability

Our trained model and codes are available from GitHub: https://github.com/lingxusb/EcoVAE

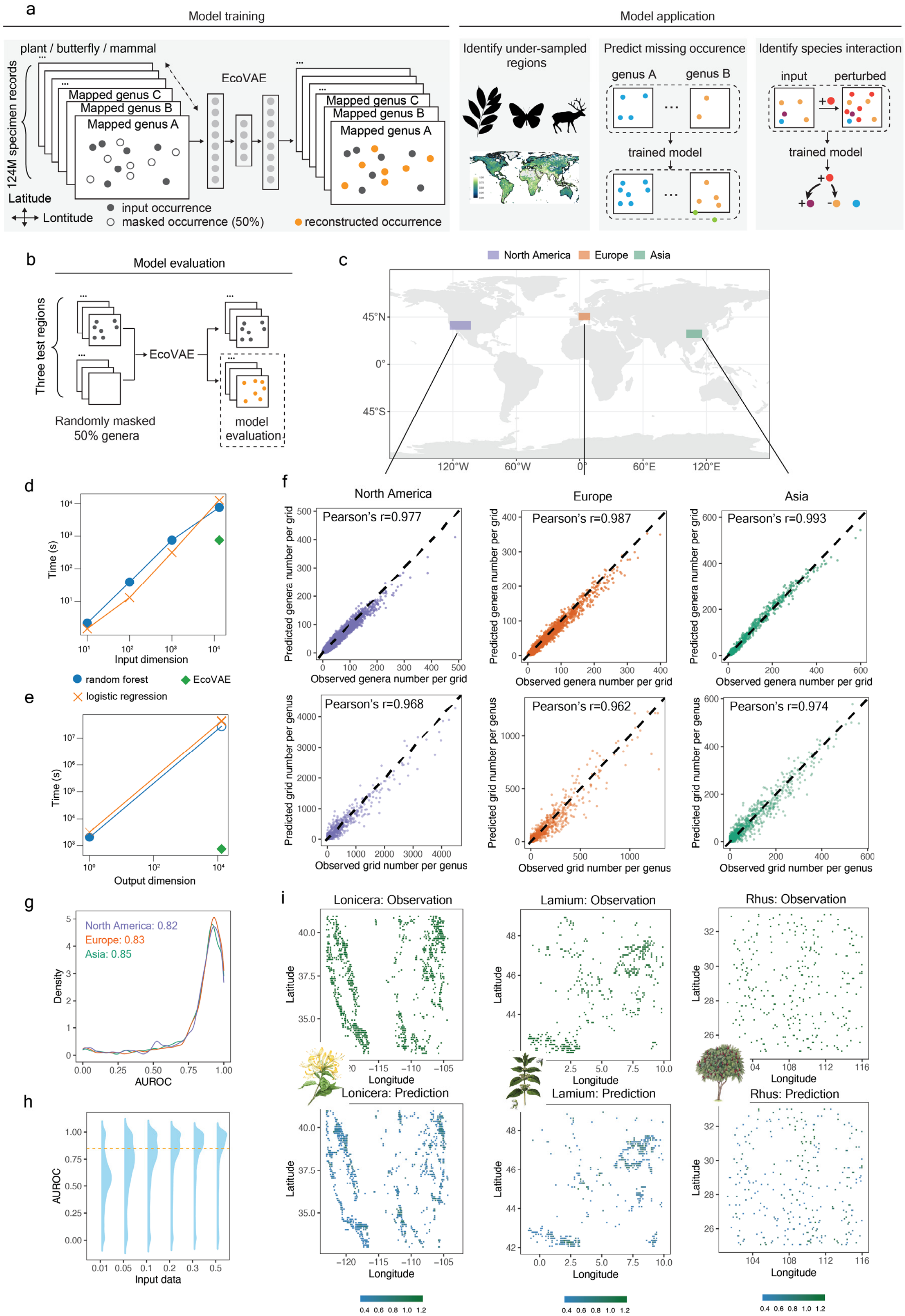
Additional Information

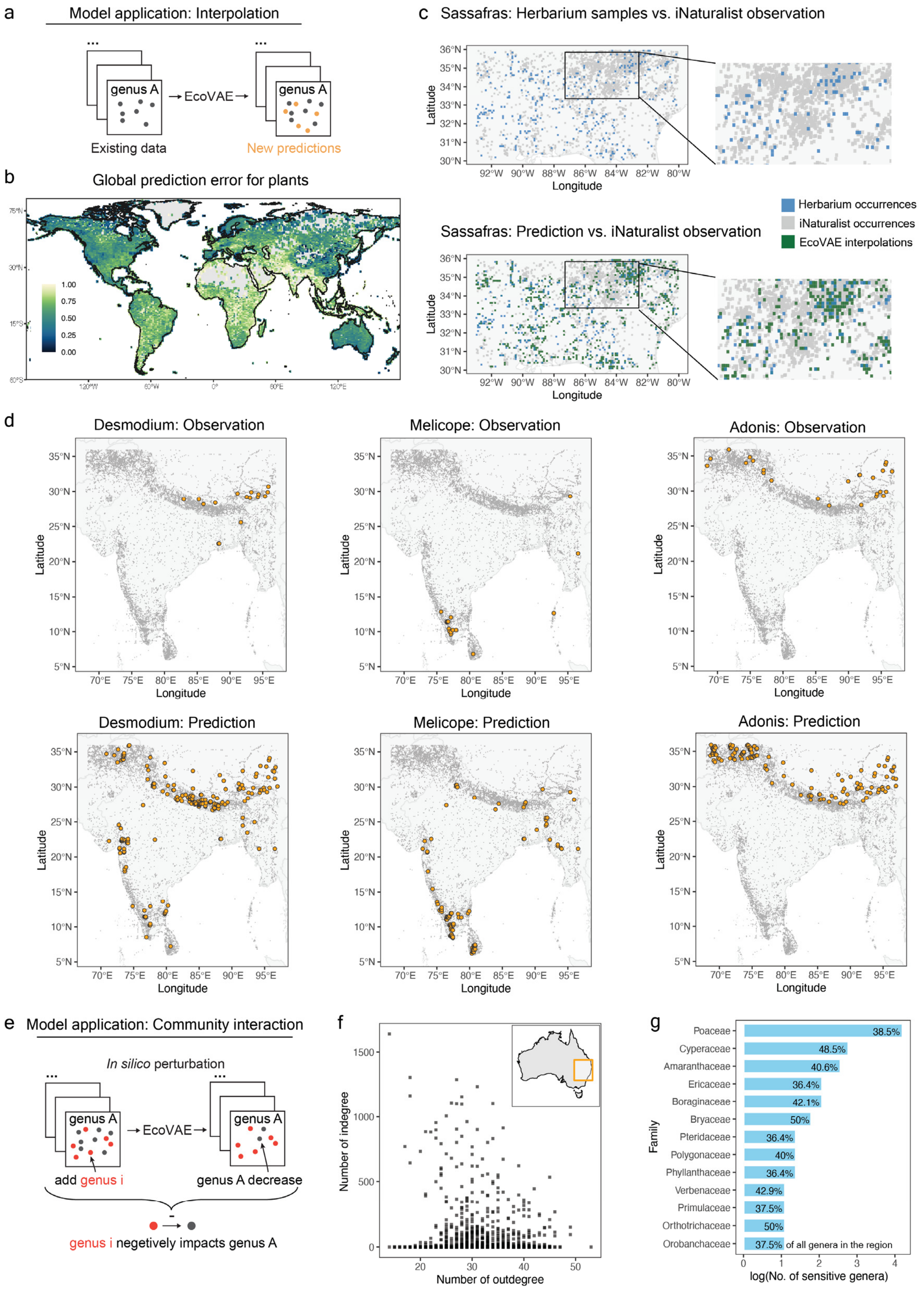
Supplementary information

Figure Legends

Figure 1. Framework and evaluation of EcoVAE model performance. a, Schematic representation of the model training and application pipeline. **b,** Overview of the model evaluation process. **c,** Map showing the locations of the three testing regions. **d,** Comparison of time consumption between two machine learning methods and EcoVAE with the increase of input dimension. **e,** Comparison of time consumption between two machine learning methods and EcoVAE with the increase of output dimension. **f,** Correlation between observed genera counts per grid (or observed grid counts per genus, upper panels) and predicted genera counts per grid (or predicted grid counts per genus, lower panels) across the three testing regions, with high Pearson correlation values. The black dashed lines indicate identity lines. **g,** Density plot showing the Area Under the Receiver Operating Characteristic (AUROC) for three testing regions. **h,** Relationship between the ratio of masked data and AUROC values. **i,** Comparison of the distribution of observations (upper panels) and predictions (lower panels) for randomly selected genus within each test region.

Figure 2. Applications of EcoVAE model. a, Schematic illustration of the interpolation process using EcoVAE. **b,** Global distribution of relative collection completeness, represented by the value of prediction error of EcoVAE. Darker color represents lower prediction error and higher completeness, while lighter color represents higher prediction error and lower completeness. **c,** Comparison between original herbarium specimen records and EcoVAE interpolation results of genus *Sassafras* in North America. **d,** Interpolation results of three example genera in South Asia. Gray dots show the distribution of all georeferenced vouchered occurrences in the study area. **e,** Schematic illustration of studying community interactions using EcoVAE model. **f,** Relationship between the number of outdegree and indegree for the genus-to-genus interactions. Each dot represents a single genus. **g,** Log number of sensitive genera across the most sensitive plant families identified by EcoVAE.


a
Model training
Model application
plant / butterfly / mammal
124M specimen records
Mapped genus C
Mapped genus B
Mapped genus A
EcoVAE
Latitude
Lontitude
input occurrence
masked occurrence (50%)
reconstructed occurrence
Identify under-sampled regions
Predict missing occurence
genus A
genus B
trained model
Identify species interaction
input
perturbed
trained model
b
Model evaluation
Three test regions
EcoVAE
Randomly masked 50% genera
model evaluation
c
North America
Europe
Asia
45°N
0°
45°S
120°W
60°W
0°
60°E
120°E
d
Time (s)
Input dimension
random forest
EcoVAE
logistic regression
e
Time (s)
Output dimension
f
North America
Europe
Asia
Pearson's r=0.977
Pearson's r=0.987
Pearson's r=0.993
Predicted genera number per grid
Observed genera number per grid
Pearson's r=0.968
Pearson's r=0.962
Pearson's r=0.974
Predicted grid number per genus
Observed grid number per genus
g
North America: 0.82
Europe: 0.83
Asia: 0.85
Density
AUROC
h
AUROC
Input data
0.01
0.05
0.1
0.2
0.3
0.5
i
Lonicera: Observation
Lamium: Observation
Rhus: Observation
Lonicera: Prediction
Lamium: Prediction
Rhus: Prediction
Latitude
Longitude
0.4 0.6 0.8 1.0 1.2


a
Model application: Interpolation
genus A
EcoVAE
genus A
Existing data
New predictions
b
Global prediction error for plants
1.00
0.75
0.50
0.25
0.00
c
Sassafras: Herbarium samples vs. iNaturalist observation
Latitude
Longitude
Herbarium occurrences
iNaturalist occurrences
EcoVAE interpolations
Sassafras: Prediction vs. iNaturalist observation
36°N
35°N
34°N
33°N
32°N
31°N
30°N
92°W 90°W 88°W 86°W 84°W 82°W 80°W
d
Desmodium: Observation
Melicope: Observation
Adonis: Observation
Desmodium: Prediction
Melicope: Prediction
Adonis: Prediction
35°N
30°N
25°N
20°N
15°N
10°N
5°N
70°E 75°E 80°E 85°E 90°E 95°E
e
Model application: Community interaction
In silico perturbation
genus A
EcoVAE
genus A
add genus i
genus A decrease
-
genus i negetively impacts genus A
f
Number of indegree
Number of outdegree
1500
1000
500
0
20
30
40
50
g
Family
Poaceae 38.5%
Cyperaceae 48.5%
Amaranthaceae 40.6%
Ericaceae 36.4%
Boraginaceae 42.1%
Bryaceae 50%
Pteridaceae 36.4%
Polygonaceae 40%
Phyllanthaceae 36.4%
Verbenaceae 42.9%
Primulaceae 37.5%
Orthotrichaceae 50%
Orobanchaceae 37.5% of all genera in the region
0 1 2 3 4
log(No. of sensitive genera)