Learning the syntax of plant assemblages

César Leblanc

cesar.leblanc@inria.fr

Inria, LIRMM, Université de Montpellier, CNRS, Montpellier, France https://orcid.org/0000-0002-5682-8179

Pierre Bonnet

CIRAD https://orcid.org/0000-0002-2828-4389

Maximilien Servajean

LIRMM, AMIS, University of Montpellier Paul Valéry, CNRS, Montpellier, France

Wilfried Thuiller

Univ. Grenoble Alpes - CNRS

Milan Chytrý

Masaryk University https://orcid.org/0000-0002-8122-3075

Svetlana Aćić

University of Belgrade, Faculty of Agriculture, Department of Botany, Nemanjina 6, Belgrade-Zemun, Serbia

Olivier Argagnon

Conservatoire botanique national méditerranéen, Hyères, France

Idoia Biurrun

Department of Plant Biology and Ecology, University of the Basque Country UPV/EHU https://orcid.org/0000-0002-1454-0433

Gianmaria Bonari

Department of Life Sciences, University of Siena, Siena, Italy

Helge Bruelheide

Martin Luther University Halle-Wittenberg https://orcid.org/0000-0003-3135-0356

Juan Campos

Department of Plant Biology and Ecology, University of the Basque Country UPV/EHU, Bilbao, Spain

Andraž Čarni

Research Centre of the Slovenian Academy of Sciences and Arts, Jovan Hadži Institute of Biology, Novi trg 2, 1000 Ljubljana, Slovenia

Renata Ćušterevska

Institute of Biology, Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Arhimedova Str. 3, 1000 Skopje, Republic of North Macedonia

Michele De Sanctis

Department of Environmental Biology, University Sapienza of Rome https://orcid.org/0000-0002-7280-6199

Jürgen Dengler

Vegetation Ecology Research Group, Institute of Natural Resource Management (IUNR), Zurich University of Applied Sciences (ZHAW) https://orcid.org/0000-0003-3221-660X

Tetiana Dziuba

Department of Geobotany and Ecology, M.G. Kholodny Institute of Botany, National Academy of Sciences of Ukraine, Kyiv, Ukraine

Jesper Moeslund

Department of Ecoscience, Aarhus University, C. F. Møllers Allé 6-8, DK-8000 Aarhus C, Denmark

Emmanuel Garbolino

Mines Paris PSL-ISIGE, 35 rue Saint-Honoré, 77300 Fontainebleau, France

Ute Jandt

Martin Luther University Halle-Wittenberg https://orcid.org/0000-0002-3177-3669

Florian Jansen

University of Rostock https://orcid.org/0000-0002-0331-5185

Jonathan Lenoir

CNRS https://orcid.org/0000-0003-0638-9582

Aaron Haase

Department of Evolutionary Biology, Ecology, and Environmental Sciences, University of Barcelona, Barcelona, Spain

Remigiusz Pielech

Institute of Botany, Faculty of Biology, Jagiellonian University in Kraków

Jozef Sibik

Plant Science and Biodiversity Center of Slovak Academy of Sciences, Dúbravská cesta 9, SK-845 23 Bratislava, Slovak Republic

Zvjezdana Stančić

Faculty of Geotechnical Engineering University of Zagreb, Hallerova aleja 7, HR-42000 Varaždin, Croatia

Domas Uogintas

State Scientific Research Institute Nature Research Centre, Akademija 2, Vilnius, Lithuania https://orcid.org/0000-0002-3937-1218

Thomas Wohlgemuth

Swiss Federal Institute for Forest, Snow and Landscape Research

Alexis Joly

INRIA Sophia-Antipolis

Article

Keywords:

Posted Date: April 7th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6304381/v1

Additional Declarations: There is **NO** Competing Interest.

Learning the syntax of plant assemblages

César Leblanc[1,2,*], Pierre Bonnet[2,+], Maximilien Servajean[3,+], Wilfried Thuiller[4], Milan Chytrý[5], Svetlana Aćić[6], Olivier Argagnon[7], Idoia Biurrun[8], Gianmaria Bonari[9], Helge Bruelheide[10,11], Juan Antonio Campos[12], Andraž Čarni[13,14], Renata Ćušterevska[15], Michele De Sanctis[16], Jürgen Dengler[17,18], Tetiana Dziuba[19], Jesper Erenskjold Moeslund[20], Emmanuel Garbolino[21], Ute Jandt[10,11], Florian Jansen[22], Jonathan Lenoir[23], Aaron Pérez Haase[24,25], Remigiusz Pielech[26], Jozef Sibik[27], Zvjezdana Stančić[28], Domas Uogintas[29], Thomas Wohlgemuth[30], and Alexis Joly[1,+]

[1]Inria, LIRMM, Université de Montpellier, CNRS, Montpellier, France
[2]AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France
[3]LIRMM, AMIS, University of Montpellier Paul Valéry, CNRS, Montpellier, France
[4]Université Grenoble Alpes, Univiversité Savoie Mont Blanc, CNRS, LECA, F-38000 Grenoble, France
[5]Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic
[6]University of Belgrade, Faculty of Agriculture, Department of Botany, Nemanjina 6, Belgrade-Zemun, Serbia
[7]Conservatoire botanique national méditerranéen, Hyères, France
[8]Dept. Plant Biology and Ecology, University of the Basque Country UPV/EHU, Apdo. 644, 48080 Bilbao, Spain
[9]Department of Life Sciences, University of Siena, Siena, Italy
[10]Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, Germany
[11]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[12]Department of Plant Biology and Ecology, University of the Basque Country UPV/EHU, Bilbao, Spain
[13]Research Centre of the Slovenian Academy of Sciences and Arts, Jovan Hadži Institute of Biology, Novi trg 2, 1000 Ljubljana, Slovenia
[14]University of Nova Gorica, School for Viticulture and Enology, Vipavska 13, 5000 Nova Gorica, Slovenia
[15]Institute of Biology, Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Arhimedova Str. 3, 1000 Skopje, Republic of North Macedonia
[16]Department of Environmental biology, Sapienza University of Rome, Rome, Italy
[17]Vegetation Ecology Research Group, Institute of Natural Resource Sciences (IUNR), Zurich University of Applied Sciences (ZHAW), Grüentalstr. 14, 8820 Wädenswil, Switzerland
[18]Bayreuth Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Universitätsstr. 30, 95447 Bayreuth, Germany
[19]Department of Geobotany and Ecology, M.G. Kholodny Institute of Botany, National Academy of Sciences of Ukraine, Kyiv, Ukraine
[20]Department of Ecoscience, Aarhus University, C. F. Møllers Allé 6-8, DK-8000 Aarhus C, Denmark
[21]Mines Paris PSL-ISIGE, 35 rue Saint-Honoré, 77300 Fontainebleau, France
[22]Faculty of Agricultural and Environmental sciences, Justus-von-Liebig-Weg 6, 18053 Rostock, Germany
[23]UMR CNRS 7058 "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN), Université de Picardie Jules Verne, Amiens, France
[24]Department of Evolutionary Biology, Ecology, and Environmental Sciences, University of Barcelona, Barcelona, Spain
[25]Biodiversity Research Institute (IRBio), University of Barcelona, Barcelona, Spain
[26]Institute of Botany, Faculty of Biology, Jagiellonian University in Kraków, Poland
[27]Plant Science and Biodiversity Center of Slovak Academy of Sciences, Dúbravská cesta 9, SK-845 23 Bratislava, Slovak Republic
[28]Faculty of Geotechnical Engineering University of Zagreb, Hallerova aleja 7, HR-42000 Varaždin, Croatia
[29]State Scientific Research Institute Nature Research Centre, Akademija 2, Vilnius, Lithuania
[30]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zürcherstr. 111, 8903 Birmensdorf, Switzerland
[*]cesar.leblanc@inria.fr
[+]these authors contributed equally to this work

ABSTRACT

To address the urgent biodiversity crisis, it is crucial to understand the nature of plant assemblages. The distribution of plant species is not only shaped by their broad environmental requirements, but also by micro-environmental conditions, dispersal limitations, and direct and indirect species interactions. While predicting species composition and habitat identity is essential for conservation and restoration purposes, it thus remains challenging. In this study, we propose a novel approach inspired by advances in large language models to learn the "syntax" of abundance-ordered plant species sequences in communities. Our method, which captures latent associations between species across diverse ecosystems, can be fine-tuned for diverse tasks. In particular, we show that our methodology is able to outperform other approaches to (i) predict species that might occur in an assemblage given the other listed species, despite being originally missing in the species list (+16.53% compared to co-occurrence matrices and +6.56% compared to neural networks) and (ii) classify habitat types from species assemblages (+5.54% compared to expert systems and +1.14% compared to deep learning). The proposed application has a vocabulary that covers over ten thousand plant species from Europe and adjacent countries and provides a powerful methodology for improving biodiversity mapping, restoration, and conservation biology.

Introduction

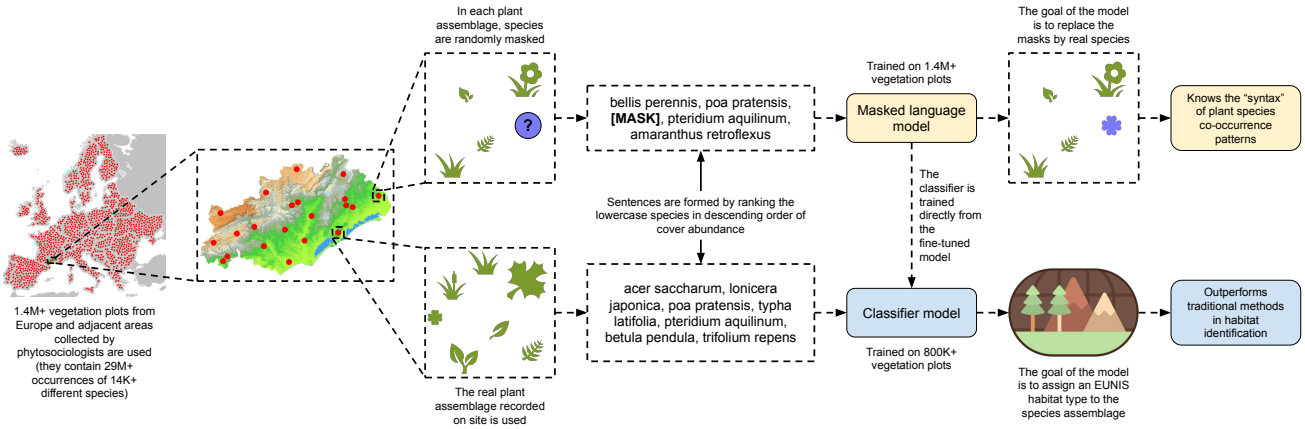


Figure 1. The proposed approach leverages large language models (LLMs) to capture the latent dependencies between plant species in diverse ecosystems. By training on over 1.4M vegetation plots, 29M species occurrences and 14K species from Europe and adjacent regions, the model learns the "syntax" of sentences formed by abundance-ordered plant species sequences, allowing it to predict missing (i.e., [MASK]) taxa in sequences of species. The resulting foundation model can be further fine-tuned to assign EUNIS habitat types to vegetation plots, outperforming traditional methods

Understanding vegetation patterns and plant assemblages is central to ecology, as co-occurring species ultimately determine the structure and function of ecosystems[1]. Plant species rarely exist in isolation[2]; instead, they form complex assemblages influenced by biotic and abiotic conditions[3–5]. These assemblages represent the emergent properties of ecosystems, where each species contributes to and is influenced by the broader assemblage[6]. Identifying and analyzing these intricate patterns is crucial for understanding the underlying mechanisms governing biodiversity and ecosystem stability and dynamic[7,8]. Despite progress, unraveling these patterns remains challenging, given the high dimensionality and complexity of community assembly[9]. In this study, we attempt to decode the "syntax" of plant community structure, aiming to provide new insights on the composition of vegetation across diverse ecosystems. In this context, "syntax" refers to the implicit rules and patterns that govern how plant species co-occur and interact to form structured assemblages, similar to how syntax in language defines the arrangement of words to create meaningful sentences. Just as language syntax reveals relationships between words based on their positions and roles, the "syntax" of plant assemblages represents the hidden shared environmental preferences, direct and indirect interactions, and organization underlying species assemblages (i.e., just as the ordering of words in a sentence matters, the ranking of species names in a community matters as well). We focus particularly on how this approach can be used to improve habitat type identification, offering insights that could enhance ecological classification and conservation efforts.

The analysis of species communities is often done by leveraging presence-absence matrices of species co-occurrences[10], which record how many times two different species were observed together in the same vegetation plot. This traditional approach allows for global analyses of co-occurrence patterns in vegetation plots found in a dataset, making it suitable for

detecting broad patterns, such as clusters of species with a high tendency of co-occurrence[11–13]. However, this method is often biased towards common species[14], as they have higher occurrence frequencies across vegetation plots, leading to inflated co-occurrence estimates. This can obscure the detection of rare or specialized species interactions[15], which may play critical ecological roles but are underrepresented in presence-absence matrices.

To address this limitation, alternative approaches such as fidelity indices[16] quantify species' specificity to particular habitat types rather than relying solely on their co-occurrence frequencies, making them particularly useful for distinguishing diagnostic species from widely distributed ones. While such methods might offer an improvement over raw co-occurrence counts, they remain constrained by predefined habitat classifications and do not fully capture the hierarchical and context-dependent nature of species associations. In addition, most co-occurrence matrices only account for species presence or absence in the assemblage, but the relative abundance of species within plant assemblages, which is often important for habitat and vegetation classification[17], is not taken into account. Notably, statistical interdependencies, which reflect biotic interactions, often exhibit asymmetric, transitive, and hierarchical patterns[18,19] that are beyond the scope of classical co-occurrence approaches but can be captured by novel and more sophisticated AI-based abundance-order language models. These new models use a transformer-type deep learning architecture based on self-attention mechanisms[20] (which allow the model to weight the importance of each species in relation to all others in a given assemblage, much like how one might focus on key words in a sentence to understand its meaning). This allows such a model to account for bi-directional dependencies (asymmetry, i.e., if species A influences species B but species B does not necessarily influence species A) and aggregate indirect relationships across assemblages (transitivity, i.e., if species A influences species B and species B influences species C then species A influences species C). It can also learn hierarchical patterns in the assemblage, such as which species are often abundant and how they can influence other species that are often less abundant.

A concrete application of the model evaluated in our study is the classification of European habitat types based on ordered species assemblages. Europe hosts a rich diversity of vascular plant species, contributing to a great number of unique habitats[21] shaped by both biotic and abiotic factors and protected by the European Habitats Directive. However, this biodiversity faces many problems, including, but not limited to, the effects of various kinds of agricultural activities (e.g., intensification for more productive farming and abandonment of traditional land use) and modifications of natural systems (e.g., dredging and sea defense works)[22]. All habitats protected by the Habitat Directive are listed in Annex I of this directive[23] and with the new EU restoration law, a large proportion of these habitats have to be in favorable state in the near future[24]. A major challenge is that in many EU countries, only a fraction of these habitats have been mapped, making it difficult to monitor their development and condition. Moreover, even when mapped, their ecological quality often remains unknown, further complicating conservation and management efforts. Here, we try to patch this major knowledge gap.

For the purpose of this study, habitats were defined as terrestrial, freshwater or marine areas characterized by geographic, abiotic and biotic features[25]. We leveraged the European Nature Information System (EUNIS)[26] maintained by the European Environment Agency (EEA). This hierarchical classification system covers all types of habitats and contains at least five levels of complexity[27]. We retained the first three levels: broad habitat groups (level one), habitat groups (level two), and habitat types (level three). Our work especially focused on the level three of eight broad habitat groups.

Habitat distribution modeling typically involves linking information on plant species composition (such as a full list of vascular plant species with estimates of cover abundance) and environmental covariates (such as whether a community is located on a coastal dune[28] or within a specific terrestrial ecoregion[29]) to habitat type occurrences. This approach helps identify the habitat type of vegetation plots. There are two basic types of methodologies used for vegetation classification based on species composition[30]: expert systems[31] and machine learning[32]. The former leverage explicitly defined logical rules and emulate the process of expert classification done by humans[33], whereas the latter are tools for induction of the independent knowledge base.

Expert systems, even though they are still the most used tools to assign plots to vegetation types[34], do not consistently align with the basic requirements for vegetation classification[35]:

- they tend to overfit by learning the detail in the training data too well. Thus, minor changes in a vegetation plot (e.g., a small difference in the cover of an individual species) can considerably alter the result of the classification procedure, making those expert systems not robust.

- some of them involve sets of external criteria (e.g., environmental or geographical attributes of vegetation plots in addition to species composition) to classify some vegetation types, making those expert systems not simple.

- they are often based on one specific nomenclatural and taxonomic dataset, but using vegetation plots from different origins might result in different names for the same entity or identical names for different entities (depending on the taxonomic concepts and determination literature used in a particular region or period), making those expert systems not consistent.

Modern deep learning techniques have great potential for modeling habitat distributions[36]. In particular, experiments with feedforward neural networks have shown that they have the ability to capture complex information about the plant species composition of vegetation plots to classify plant communities[37]. One limitation of such models, however, is that their architecture induces an intrinsic inductive bias in the sense that they process each plant species as if it is equally different from all the others[38]. Thus, they cannot accurately model complex relationships between plant species. Therefore, they are not really suitable for modeling ecological systems and identifying habitat types where the interdependencies between plant species are complex[39]. While classical approaches offer interpretable and mathematically grounded methods for ecological modeling[40], they may lack the capacity to learn latent patterns (i.e., underlying structures, correlations, or dependencies within the data that are not explicitly observable such as subtle co-occurrence relationships between plant species, hierarchical community structures, or environmental gradients that shape species assemblages) from high-dimensional data.

In contrast, transformers[41], a different kind of deep learning model, go beyond local processing and exploit global attention mechanisms for increased performance. Although transformers were leveraged in various fields of biology (e.g., the extraction of morphological traits[42] or the prediction of protein structures[43]), their use in vegetation classification is still largely unexplored. Such models should allow the segmenting of habitats in a much more efficient manner than current methods. In particular, large language models (LLMs) have not yet been embraced by the global community of ecologists despite their ability to find patterns and correlations in noisy biological data[44].

The goal of this work is to enhance the understanding of species assemblages and facilitate habitat identification within Europe through the use of the potential of LLMs. To achieve this goal, we introduce a novel computational pipeline centered around Pl@ntBERT[45], a model based on BERT[46] (i.e., Bidirectional Encoder Representations from Transformers, a deep learning model originally designed for natural language understanding). Consequently, it means that without any further adaptation (i.e., fine-tuning), Pl@ntBERT would be only pre-trained in a self-supervised manner on very large volumes of common text data unrelated to vegetation (i.e., BookCorpus and English Wikipedia) and would be some kind of Swiss army knife solution (i.e., this model would work for the most common language tasks, such as sentiment analysis or named entity recognition, as long as they don't require a deep knowledge of the domain). However, to make it ecologically meaningful, we pre-train it (i.e., we make the model learn the general structure in the data) on an in-domain dataset named the European Vegetation Archive (EVA)[47], an integrated database of European vegetation plots. This adaptation allows Pl@ntBERT to develop a statistical representation of the vegetation assemblages, capturing implicit relationships between species that commonly co-occur, and boost the performance of the downstream task (i.e., habitat type identification).

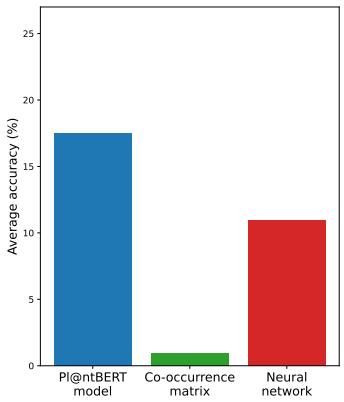
The next step is to train the model for a supervised classification task: assigning habitat types to species assemblages. We use the EUNIS classification system, a widely used European framework that organizes vegetation into hierarchical habitat types based primarily on dominant species composition, ecological structure, and environmental conditions. The EUNIS typology provides a standardized way to classify and compare habitats across Europe, making it a key reference for conservation and land management. Unlike traditional expert systems, which rely on manually defined classification rules, or classical machine learning approaches, which process species independently without considering their ecological interdependencies, Pl@ntBERT learns to infer habitat types by recognizing patterns in species composition and their statistical relationships. This approach enhances classification accuracy, mitigates inconsistencies in taxonomic nomenclature, and provides a scalable solution for habitat identification, including for habitats under threat of collapse.

Results

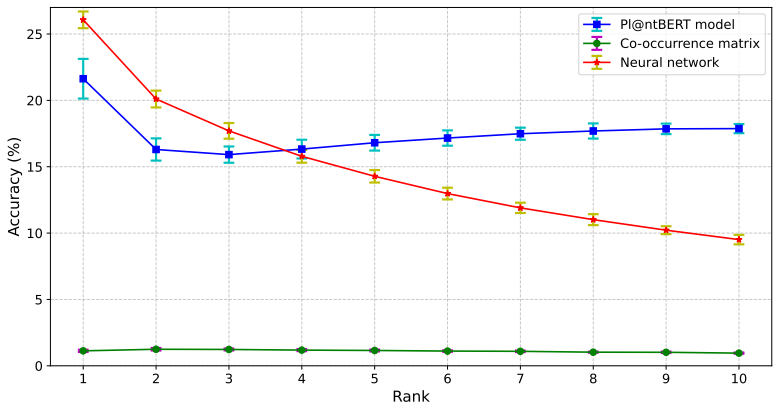
The syntax of species assemblages

Understanding the structure of species assemblages requires capturing both direct and indirect relationships between co-occurring species. To measure Pl@ntBERT's ability to capture these complex relationships from abundance-ordered species communities, we evaluated it on a so called masking or fill-mask task (i.e., a species is removed from the assemblage, and the accuracy of the model in recovering the right species is measured). This approach is conceptually related to the notion of dark diversity[48], as it aims to identify missing species that, based on the ecological context, are expected to be present but are absent in a given assemblage. For this evaluation, we tested different versions of Pl@ntBERT, which vary in how they tokenize species names. Refer to the Methods section for more details about these different versions. Naturally, the "term" versions (i.e., both small and large models), that split species names into two tokens (i.e., one for the genus name and one for the species epithet), perform better when it comes to replacing masked tokens in a sentence, because each mask only hides a half of a species name (i.e., either the genus name or the species epithet). As a result, it is easier for these models to figure out what the other half of the binomial name is (e.g., "*thinopyrum junceum*, **[MASK]** *marina*, *pancratium maritimum*"). On the contrary, each mask of the "species" versions of Pl@ntBERT hides completely a species name, meaning that the model has to choose between over 14K different species to replace the mask (e.g., "*thinopyrum junceum*, **[MASK]**, *pancratium maritimum*").

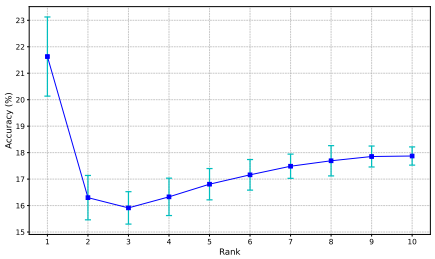
To assess how well Pl@ntBERT captures species relationships beyond simple co-occurrences, we conducted a comparative evaluation against two alternative approaches: (1) a naive Bayes model[49] using only the species co-occurrence matrix and (2) a



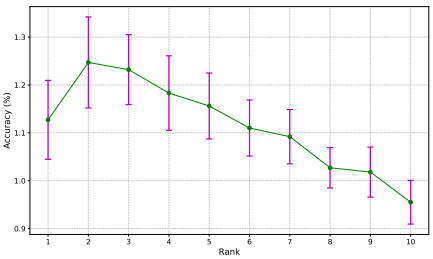
(a) Overall accuracy



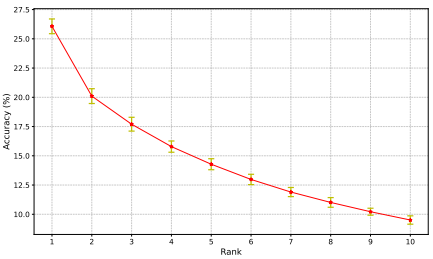
(b) Rank accuracy of the Pl@ntBERT model, the co-occurrence matrix, and the neural network



(c) Focus on the Pl@ntBERT model



(d) Focus on the co-occurrence matrix



(e) Focus on the neural network

Figure 2. Overall masking accuracy (micro-averaged over the ten cross-validation folds) of the three methods (2a), breakdown of the rank accuracy (2b), and specific focus of the results obtained by the large-species model (2c), the co-occurrence matrix (2d), and the neural network (2e). Only the labeled vegetation plots for which over ten species were recorded were kept in the test set. For each remaining vegetation plot (n = 705 479), the ten most abundant species were masked one by one and the accuracy corresponding to each rank was computed. Note the difference in *y* axis in the three graphs of Figures 2c, 2d, and 2e. Figure 2b shows the three error bars displayed on the same *y* axis.

classical deep learning model[50] based on a feedforward neural network (see Figure 2). This comparison allowed us to determine whether Pl@ntBERT's ability to encode species assemblages translates into improved predictive power when identifying missing species in vegetation plots. The graphs (see Figure 2b) show that the Pl@ntBERT model clearly outperforms the co-occurrence matrix at every rank (i.e., at every position that species can occupy in the vegetation plot when they are sorted by cover-abundance). Moreover, the co-occurrence matrix tends to perform worse when the species is less abundant (see Figure 2d). The neural network is very good for the most dominant species, even outperforming the Pl@ntBERT model on the first ranks. However, when the species become less abundant, it quickly loses its predictive power (see Figure 2e). In contrast, the Pl@ntBERT model tends to perform better for rare species than for common species (see Figure 2c). Indeed, the accuracy of its predictions drops sharply when the first ranked species (most abundant) are masked (from around 22% to around 16% for species ranked second to third) but then slowly increases for species ranked after (and stabilizes around 18% for species ranked tenth). This indicates that, as the first species is the one contributing the most to the assemblage structure and identity, it is easy for our model to find it if it has complete knowledge of the assemblages (i.e., all other species), especially the second and third species. Moreover, it shows that the presence of abundant species is essential but not sufficient to determine the habitat. However, the assemblage of the first three species (and also the assemblage of only the second and third species) is often sufficient to determine the habitat. This emphasizes the critical role that species abundance plays in accurately predicting missing species in an assemblage. As it is often the rarer and less abundant species that are missing from vegetation-plot records, this experiment highlights the importance of using models like Pl@ntBERT to capture nuanced relationships between species.

The task of finding missing species from highly diverse, incomplete plant assemblages benefits significantly from the ability to capture complex relationships, leverage extensive textual data for contextual understanding, and learn rich, abstract data

representations. A comparison between the results obtained by the Pl@ntBERT model, the co-occurrence matrix, and the neural network (see Supplementary Figure S12 online) shows that large language model clearly outperforms the other two approaches in this regard. LLMs provide a holistic view that aids in recognizing patterns and improving species identification. The co-occurrence matrix relies on simple frequency counts of species pairs appearing together in the training dataset[51] and the neural network relies on one-hot encoded assemblages of co-occurring species[52], which lack the contextual understanding necessary to accurately predict the masked tokens in a complex and domain-specific dataset such as plant species names. Whatever the broad habitat groups (e.g., Vegetated man-made habitats, Wetlands, Forests and other wooded land), Pl@ntBERT consistently outperforms the co-occurrence matrix by a factor of more than ten and, except for Littoral biogenic habitats and Coastal habitats, the neural network by a factor of almost two (overall accuracy of 17.49% for the Pl@ntBERT model, of 0.96% for the co-occurrence matrix, and of 10.93% for the neural network, see Figure 2a).

Furthermore, we show that Pl@ntBERT is able to perform better than both the co-occurrence matrix and the neural network when detecting species patterns (see Supplementary Figure S29 online). In scenarios where three species A, B, and C occur together more than 100 times in a vegetation plot but where species A and species C never occur together without species B, Pl@ntBERT is often able to predict that the species B is required for the presence of the other two species, unlike the other methods. In contrast, the co-occurrence matrix and the neural network repeatedly predict common species (e.g., *Dactylis glomerata*, which is the most frequent species in the dataset, or *Phragmites australis*), even in cases where they are not plausible candidates, showing a tendency to favoring species that appear many times in the dataset rather than recognizing specific ecological patterns. Pl@ntBERT's success demonstrates its capacity to learn the complex "syntax" of plant assemblages and correctly identify species occurrence relationships, even in complicated ecological contexts. This further emphasizes the model's potential to improve vegetation surveys and habitat assessments by providing more accurate and context-sensitive species predictions. Indeed, observer errors (e.g., overlooking errors[53] and misidentification errors[54]) may result in species richness being artificially underestimated[55].

The task of finding a missing species in an assemblage is a complex problem, as the hypothesis space is large. Indeed, when asked to replace a [MASK] token in a sentence describing a vegetation plot, the model Pl@ntBERT must select from over 14,000 different vascular plant species. However, the perplexity[56] of the base model indicates that it mostly hesitates between around 12 species when it has to replace the mask. More importantly, an experiment shown in Supplementary Figure S15 indicates that:

- when the Pl@ntBERT model (the large-species version) does not replace the [MASK] token by the correct species, it actually outputs a species coming from the same vegetation class[57] (i.e., the same broad unit in a hierarchical classification system that group plant communities based on shared floristic composition, ecological characteristics, and biogeography) over 39% of the time. For comparison, a random approach (i.e., predicting a random species to replace the [MASK] token) would result in a species coming from the same vegetation class around 3.5% of the time.

- when the Pl@ntBERT model (the large-species version) does not replace the [MASK] token by the correct species, it actually outputs a species that is characteristic of the habitat type (level 3) of the vegetation plot 49% of the time, of the habitat group (level 2) 66% of the time, and of the broad habitat group (level 1) 76% of the time. For comparison, a random approach would result in a species being characteristic of the habitat type of the vegetation plot 0.3% of the time, of the habitat group 2.3% of the time and of the broad habitat group 7.0% of the time.

In addition, a comparison of the vocabularies of different models can be found in Supplementary Table S18. For example, *verticillatoinundata*, a species epithet, is divided into eight pieces ([ve, ##rti, ##ci, ##lla, ##to, ##in, ##unda, ##ta]) by BERT and into seven pieces ([ver, ##tic, ##illa, ##to, ##in, ##und, ##ata]) by SciBERT[59] (i.e., a BERT model trained on scientific text). In contrast, this term appears in the in-domain vocabulary of Pl@ntBERT, as well as around 10,000 other genus names and species epithets. Species names are specific, meaningful biological entities. Splitting them into multiple smaller components (referred to as "subwords" in machine learning terminology) blocks the model's ability to recognize these tokens as representing a unified biological entity. Instead of treating the entire species name as a single, coherent unit, the model sees it as a collection of unrelated fragments, which reduces its ability to capture biological relationships. An example of the benefits of domain adaptation is shown in Figure 3. It shows that Pl@ntBERT (i.e., a fine-tuned BERT), compared to a vanilla BERT (i.e., the standard, pre-trained BERT model not specialized for plant-related data), really "understands" plant species compositions. A visualization of the attention in Pl@ntBERT can be found in Supplementary Figure S8. This makes the model more accessible and shows at multiple scales which species in a vegetation plot most influence the predictions.

Identifying habitat types

To optimize the hyperparameters (i.e., learning rate and batch size) and identify the set of parameters yielding the most accurate model, we first fine-tuned all versions of Pl@ntBERT using the first fold as a test set and the remaining nine folds as a training set. All results obtained during this fine-tuning process can be found in Supplementary Table S4. Table 1 gives an overview

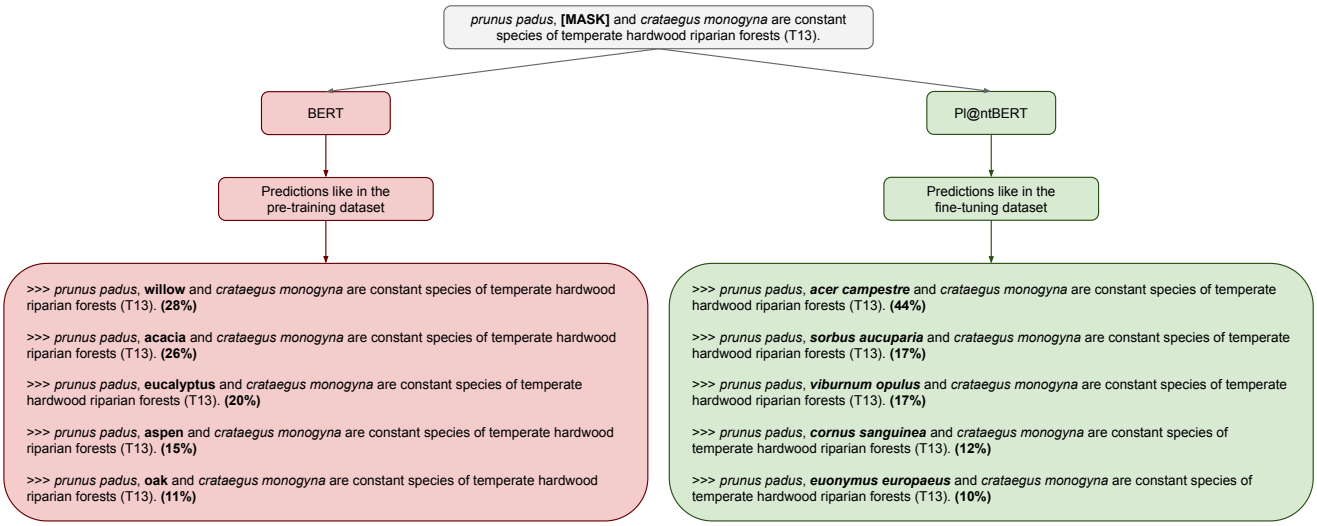
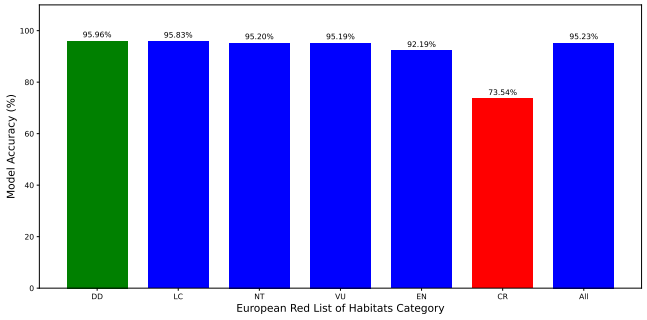
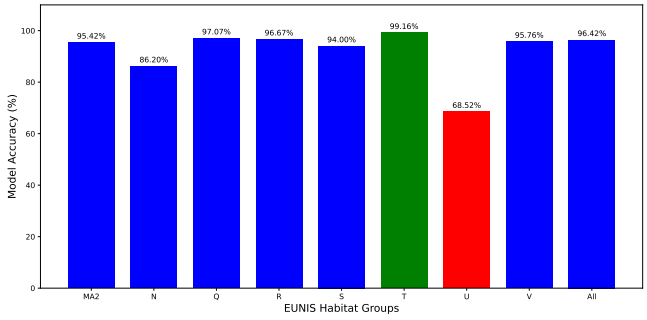


Figure 3. Comparison of the top five predictions for the BERT (large-uncased version) and Pl@ntBERT (large-species version trained on folds 1-9) models for our sample text of "*Prunus padus*, **[MASK]** and *Crataegus monogyna* are constant species of temperate hardwood riparian forests (T13).". On the one hand, the candidates from BERT are all trees, which shows that the model "understood" we are in a forest. However, all of them are common plant names (and not scientific names of taxa) and, except for the oak which is the last candidate, are not found within the *T13* habitat type. On the other hand, the candidates from Pl@ntBERT are all scientific names of constant species[58] from the required habitat type.

of the results obtained in the text classification task, and Supplementary Figure S5 provides more details. Among all tested models, Pl@ntBERT-large-species appears as the clear winner when it comes to identifying habitat types, outperforming all other models, whether it is on top-1 accuracy (i.e., the first candidate output by the model is the real habitat type, or level 3 habitat), top-3 accuracy (i.e., among the three first candidates output by the model is the real habitat type, or level 3 habitat), group accuracy (i.e., the first candidate output by the model belongs to the real habitat group, or level 2 habitat), or broad accuracy (i.e., the first candidate output by the model belongs to the real broad habitat group, or level 1 habitat). It also outperforms models that, in addition to species composition, use the abiotic environment and geographic location as classification criteria. The different versions of the expert system EUNIS-ESy and the different models of hdm-framework, as statistical and general-purpose machine learning approaches, are not capable of matching domain-adapted models such as Pl@ntBERT for specialized tasks in vegetation classification.



(a) Results across the European Red List of Habitats categories (DD: Data Deficient, LC: Least Concern, NT: Near Threatened, VU: Vulnerable, EN: Endangered, CR: Critical Endangered). The best accuracy is in green and the worst accuracy is in red.



(b) Results across the EUNIS broad habitat groups (MA2: Marine, N: Coastal, Q: Wetlands, R: Grasslands, S: Heathlands, T: Forests, U: Inland, V: Man-made). The best accuracy is in green and the worst accuracy is in red.

Figure 4. Accuracy obtained by the Pl@ntBERT-large-species model on different typologies (results averaged over the ten cross-validation folds)

Pl@ntBERT (the large-species version) achieves an accuracy of 92% when asked to classify a vegetation plot into one of

Table 1. Comparison between Pl@ntBERT and several habitat identification alternatives: the expert system EUNIS-ESy[60] and the tabular deep learning models from hdm-framework[61]. The models from hdm-framework were used with the settings recommended by the authors. The algorithms from EUNIS-ESy were implemented in the statistical computing environment R[62]. All results were averaged over the same ten cross-validation folds. A ∼ indicates that the cell is not applicable or relevant for the corresponding model. "Group accuracy" denotes the accuracy of the models on level 2 of the EUNIS hierarchy (i.e., habitat groups) and "Broad accuracy" denotes the accuracy of the models on level 1 of the EUNIS hierarchy (i.e., broad habitat groups). The predictions were always made at level 3 of the EUNIS hierarchy (i.e., habitat types) and the higher hierarchical levels were then inferred by removing one or two characters from the EUNIS habitat code. EUNIS-ESy uses the exact cover abundance of each species instead of its rank in a vegetation plot. As this expert system also requires plot-location criteria (country name, terrestrial ecoregion, coastline, coastal dune, degrees of latitude and longitude, elevation, and dataset name) to perform classification, and hdm-framework performs better with information about plot location as well (the same predictors except the dataset name), we added those covariates. hdm-framework was also evaluated purely based on species composition for a fair comparison with Pl@ntBERT, which does not use any additional variables than the species composition. The bold entries are the best-performing model for each metric. More information about the metrics can be found in Supplementary Text S6.

Framework	Model	Fine-tuning			
		Accuracy (%)	Top-3 accuracy (%)	Group accuracy (%)	Broad accuracy (%)
Predictors: species composition, abiotic environment, and geographic location					
EUNIS-ESy	v2020-06-08	82.68	∼	84.34	90.72
	v2021-06-01	86.44	∼	88.26	94.64
hdm-framework	MLP[63]	90.84	98.90	93.94	95.79
	RFC[64]	80.37	95.73	87.85	92.13
	XGB[65]	88.81	98.95	93.00	95.69
	TNC[66]	81.50	92.13	87.11	90.70
	FTT[67]	88.84	97.28	92.65	94.92
Predictors: species composition					
hdm-framework	MLP	90.00	98.73	93.36	95.27
	RFC	80.34	95.66	87.82	92.00
	XGB	88.11	98.75	92.60	95.29
	TNC	80.64	91.73	86.40	89.98
	FTT	87.92	97.06	92.08	94.40
Pl@ntBERT (ours)	large-species	**91.98**	**99.10**	**94.79**	**96.42**

the 227 habitat types present in the dataset. More details on how some habitat groups are sometimes confused with other habitat groups can be found in Supplementary Figure S13. As shown in Figure 4, when assessing the risk of habitat collapse (after converting the predictions from EUNIS habitat types to European Red List of Habitats categories), Pl@ntBERT achieves an overall micro-accuracy of 96.5%. Furthermore, our transformer-based method outperforms all other approaches in the accuracy of identifying conservation status (see Figure 4a) and broad habitat groups (see Figure 4b). As a result, Pl@ntBERT can be seen as a powerful tool to inform and catalyze action for biodiversity conservation and policy change. More details about the distribution of the European Red List of Habitats categories across the dataset can be found in Supplementary Figure S27. We use this model to map all the unlabeled vegetation plots from the dataset, and we compare the output with the map of all labeled vegetation plots from the dataset in Supplementary Figure S33 (with a further breakdown on each individual broad habitat group from the fill-mask dataset in Supplementary Figure S34).

Some other experiments found in Supplementary Figure S17 show that the most important species for identifying the habitat type of a vegetation plot are the first ones in the cover-abundance rank. Indeed, over all the vegetation plots of the dataset containing ten species or more, Pl@ntBERT-large-species achieves an accuracy of 92.2%. When removing the first species (i.e., the most abundant) of each vegetation plot, the accuracy drops by 35 percentage points to 57.2%). When removing the last species (i.e., the least abundant) of each vegetation plot, the accuracy almost stays the same and only drops by 0.43 percentage points (91.7%). When removing a random species from each vegetation plot, the accuracy decreases by 3.0 percentage points to 89.2%. This discrepancy likely arises because dominant species shape the ecological structure of habitats. These results

highlight the strong influence of dominant species in habitat type identification, while rare species contribute minimally to the model's predictive performance.

Open science

To facilitate the reproducibility of our study and the reuse of codes and models, we develop, share and maintain a generic, free, and open-source deep learning framework facilitating the training and evaluation of predictive models of habitats from *in-situ* observation data and the inference on new and unseen vegetation-plot records. The framework, coded in the programming language Python[68] and powered by the parallel computing platform CUDA[69] for accelerated training and inference, is accessible to various user profiles (including non-deep learning experts who want to easily identify European habitat types) at the following link: https://github.com/cesar-leblanc/plantbert. A user guide on how to install the framework and run the basic tasks (i.e., data curation, fill-mask training, text classification training, and inference) can be found in Supplementary Text S20 and some examples of how the model works can be found Supplementary Text S23. If the user only have a few vegetation plots from which they want to find potentially missing species or identify the habitat type, a quicker way to test the framework is to visit the tool available here: https://huggingface.co/spaces/CesarLeblanc/plantbert_space. A demo can be found in Supplementary Figure S19.

Discussion

The Pl@ntBERT model has been created to offer novel insights into how vegetation patterns can be encoded and classified, contributing to advancements in plant ecology and conservation biology[70]. It introduces an innovative approach by leveraging NLP techniques on top of abundance-ordered species lists from specific sites aimed at capturing complex species relationships such as transitive or sequential dependencies. As a result, it can model the species composition of hundreds of terrestrial, freshwater, and marine habitat types that contain plants, including most of the threatened, vulnerable, and endangered ecosystems found across Europe and adjacent areas[71]. It addition, this approach can be expanded worldwide, e.g., by applying it to the global vegetation plot database sPlot[72].

The model has been primarily designed to predict missing species in an assemblage (which can also be used for predicting species pools of plant assemblages[73]), e.g., in incomplete monitoring projects[74], leveraging masked language modeling to infer statistically probable species compositions, hence enhancing species completeness and improving vegetation surveys. This capability is especially relevant in cases where survey data may be incomplete or where one or more species could be overlooked due to sampling limitations or observer bias. By simulating the expected species pool, Pl@ntBERT offers a means to improve the ecological relevance of data used for habitat assessments, management, and reporting. This predictive function can support the identification of indicator species and enhance the detection of key ecological patterns that may be otherwise underrepresented. However, although Pl@ntBERT can predict missing species in incomplete assemblages, caution is needed when interpreting these predictions. In some cases, a species' absence from a vegetation plot might be due to observer bias or sampling limitations, in which case its predicted presence could be justified. But some absent species belong to dark diversity (i.e., species expected to occur based on ecological conditions but that are genuinely missing due to dispersal limitations), competition, or other constraints. In such cases, attempting to "correct" field surveys by adding model-predicted species risks misrepresenting reality and creating fictional plots, which could introduce more error than it solves. From an ethical standpoint, modifying field data in this way might also be controversial, as it could lead to unintended biases in conservation and management decisions. Incomplete data are an inherent part of ecological research, and rather than filling gaps artificially, it might sometimes be preferable to acknowledge and work with these uncertainties.

The second key application of Pl@ntBERT is its capacity to classify plant species records into EUNIS habitat types. This ability addresses an essential need in habitat identification and conservation planning, where the ability to classify survey data is foundational for monitoring biodiversity and guiding restoration efforts. Traditional methods have largely relied on manual expertise or rigid algorithms that cannot capture the complex patterns and overlook associations that occur in large ecological datasets. By leveraging transformer-based architectures and fine-tuning them with domain-specific botanical datasets, Pl@ntBERT offers a more refined and accurate approach. It is also worth noting that some vegetation plots in the EVA database may represent transitional or ecotonal habitats that do not fit neatly into a single EUNIS type. Such cases introduce ambiguity in classification and may contribute to an underestimation of Pl@ntBERT's true accuracy, as the model is trained to assign exactly one habitat type, that might be ecologically reasonable but could differ from the labeled category. It is also important to consider potential regional biases due to uneven plot densities in EVA. Some habitat types may be disproportionately represented in well-surveyed regions, leading the model to learn patterns that reflect data availability rather than true ecological distributions. This could result in higher accuracy for frequently sampled habitats and reduced performance for underrepresented ones.

By learning the context to translate plant species into a modelled ecological process within an ecosystem, Pl@ntBERT is able to improve vegetation models for identifying habitat types. This domain adaptation helps the model automatically understand that some species occur only in very specific assemblages, whilst others can tolerate and thrive in a wide range of

ecosystems. Therefore, predictions are influenced not only by the actual occurrence of a given species but also by the relative probability of the presence of this species. However, some habitat types, such as those listed in Annex I, are not solely defined by vegetation but rather by geomorphological or geolocational parameters (e.g., springs, cliffs, and dune slacks). These features are unlikely to be predictable by Pl@ntBERT, as they do not necessarily correlate with species composition alone. Similarly, certain species-poor habitats present challenges for classification since their low species richness limits the available signal for distinguishing between communities. Moreover, in few cases, it is impossible to distinguish some habitats by plant species composition and relative abundance alone, because their species composition can be the same even if they occur in different regions. This is one of the main reasons why attribute data were incorporated in expert-based systems like EUNIS-ESy, rather than relying purely on species presence.

The relative position of the species within a vegetation plot (i.e., their abundance compared to the other species) is key to habitat type identification and fragmentary records completion (even more than the exact cover-abundance information of each individual species). When surveying plant species, it might be hard, whatever the level of expertise, to accurately collect the exact abundance of plants in a vegetation plot[75]. However, recording the relative abundance of the most abundant species is much easier and often sufficient. However, the spatial scale was not explicitly considered when selecting data for domain adaptation (fill-mask task) and training (text classification task). Since plant species typically co-occur at small spatial scales (a few meters), including plots from larger spatial scales may introduce noise by grouping species that do not actually form a coherent community. For example, a few meters' difference in elevation or soil moisture can lead to entirely different plant communities, yet a model trained on large-scale data may incorrectly associate species that do not truly co-occur. The larger the spatial scale used, the messier the ecological signal becomes. We did not account for this explicitly because EVA contains a limited number of plots, and we aimed to retain as many as possible, assuming that vegetation scientists conducted relevés with spatial scale in mind. However, future work should investigate how different spatial resolutions impact model performance.

The use of large language models for understanding vegetation patterns is particularly interesting because these models can learn and interpret the "syntax" of plant species assemblages. Like natural languages are composed of words following grammatical rules, plant assemblages can be thought of as following certain ecological "rules" that dictate how species co-occur and interact[76]. By leveraging the bi-directional architecture of BERT, Pl@ntBERT can effectively learn these intricate patterns, by capturing relationships between species in both forward and backward directions, which provides a more comprehensive view of assemblage composition[77]. This allows the model to understand not only direct associations but also higher-order dependencies within complex assemblages[78]. Such a syntactic approach enables Pl@ntBERT to represent ecological interdependencies with a level of detail that is challenging for traditional statistical methods, offering a novel way of encoding the relationships that define biodiversity[79]. Through this perspective, Pl@ntBERT provides a more nuanced understanding of the "grammar" underlying ecosystem composition and dynamics, ultimately contributing to better conservation and habitat management strategies, and possibly to a better fundamental understanding of nature. However, as it is a large language model, Pl@ntBERT can only learn from existing datasets and cannot anticipate novel species assemblages that may emerge in response to climate change, species invasions, or land-use changes. This is particularly relevant for neoecosystems, where new combinations of native and non-native species form as environmental conditions shift. Pl@ntBERT cannot infer future biodiversity patterns beyond what is already recorded in datasets, meaning that ongoing field surveys and expert input remain essential. Ecologists will need to continuously document new assemblages and update training data to keep the model relevant in a rapidly changing world. This underscores that Pl@ntBERT is not a replacement for field expertise but rather a tool to assist researchers in making sense of complex ecological patterns.

When it comes to vegetation classification, having a good understanding of how and why Pl@ntBERT assigns a EUNIS habitat type to a given vegetation plot is essential if we want researchers and practitioners to trust the results[80]. Integrated gradients[81], a method to calculate how important each input feature (i.e., species) is to the prediction, were used to explain how positively or negatively a species contributes to the classification of a vegetation plot. A more detailed overview of species attributions on a vegetation plot can be found in Supplementary Figure S28. It is interesting to see how a change in diagnostic, constant, or dominant taxa can change the model behavior. This study shows that the most abundant species in a vegetation plot (i.e., the first species of the sentence) is often the one that contributes the most to the classification, which reflects the experience with probabilistic keys for identifying vegetation types[82]. One of the advantages of this model is that it brings vegetation science closer to a wider circle of people.

Other experiences, whose details can be found in Supplementary Figure S22, corroborate these findings. When removing the information on abundance (i.e., by forming sentences with species in random order), the performance of Pl@ntBERT significantly drops. For example, the accuracy of the text classification task decreased by 14% compared to the classical approach. This drop was more important than when we kept the information on abundance but removed 30% of the species by random selection, meaning that capturing the relative abundance is more important than recording all plant species. Similarly, when it comes to finding which species is hiding behind a mask in a vegetation plot, Pl@ntBERT went from correctly assigning the correct species in over 17% of the cases when the species were sorted to less than 7% of the cases when the species were

not sorted. This means that plant assemblages are defined not only by the species present but also by their order of abundance because abundance influences community structure, ecological interactions, and ecosystem functioning. Abundance influences functional diversity, which is critical for ecosystem processes. Species with higher abundance often have significant roles in ecosystem functioning due to their traits and interactions with other species[83].

While Pl@ntBERT demonstrates promising results in identifying vegetation patterns and assigning habitat types based on species co-occurrence, one key limitation of the current model is that it does not explicitly account for the vertical structure of plant communities. Some habitats are characterized not only by their species composition but also by their layering structure, which plays a crucial role in defining their ecological identity. Thus, a possible improvement would be to introduce explicit hierarchical encoding of vegetation strata within Pl@ntBERT's input data. This could be achieved by adopting a standardized syntax, such as: "Tree layer: *Fagus sylvatica*, *Quercus robur*; Shrub layer: *Carpinus betulus*, *Fagus sylvatica*, *Corylus avellana*; Herb layer: *Anemone nemorosa*, *Hyacinthoides non-scripta*, *Mercurialis perennis*". By integrating layering information into Pl@ntBERT's training, the model could better capture functional differences between habitats (especially those that are defined as much by their structural complexity as by species composition alone), improve classification accuracy, and potentially enhance its ability to predict missing species within specific strata. Additionally, this hierarchical representation could facilitate better interpretability, as users could analyze species associations within distinct vertical layers rather than treating all species as equally co-occurring in a single homogeneous space. Future work should explore how to best format and standardize stratification data, as well as whether habitat-specific differences in layering (e.g., grasslands vs. forests) require different encoding strategies. Incorporating structural information into Pl@ntBERT could significantly refine its ecological modeling capabilities, making it a more powerful tool for vegetation science and conservation applications.

Finally, as a perspective, an interesting approach could be to directly train a habitat type classifier on the output of a species distribution model (SDM) instead of relying solely on real vegetation plots (e.g., by ranking the species in descending order of the probability of occurrence). SDM, which have been widely used for predicting species occurrences based on environmental variables[84,85], provide a solid foundation for such tasks. Building on this, modern deep-learning techniques, often referred to as Deep-SDMs, have already shown great potential for modeling species distributions[86,87], and in particular for vascular plant species[88,89]. Hence, a next step could involve leveraging the vast number of geolocated plant species occurrences available on citizen science platforms[90,91]. These platforms provide far more plant occurrence data than traditional vegetation-plot datasets[92], and their communities can be very engaged[93,94]. Those communities are not experts in botany and thus they may capture the most common and iconic species but miss the rare and difficult to recognize ones, so using Pl@ntBERT to complete and fill citizen science data could be useful. By utilizing this wealth of data, it may be possible to develop very high-resolution, multi-modal species distribution models. These predicted assemblages could then be used to infer habitat types. A pipeline based on computer vision (convolutional neural networks[95]) and natural language processing (transformers[96]) and focusing on (i) image classification (plant assemblages created with satellite images and rasterized environmental data), (ii) fill-mask (predicted species translated into a modeled ecological process) and (iii) text classification (habitats assigned to sentences describing species compositions) could become a powerful workflow for understanding and monitoring biodiversity dynamics, and going from habitat identification models to Habitat Distribution Models (HDMs).

Methods

A visualization of the methodology used in this paper can be seen in Figure 1, a more complete overview in Supplementary Figure S26 and a detailed description of each step in Supplementary Figures S9, S10, and S11. An explanation of all acronyms and terms can be found in Supplementary Texts S30 and S31.

Leveraging vegetation plots

The data used for training the Pl@ntBERT model were extracted from the European Vegetation Archive (EVA)[47]. EVA is a database of vegetation plots, i.e., records of plant taxon co-occurrence which have been collected by vegetation scientists at particular sites and times. The EVA data was extracted on May 22nd, 2023. It contained all georeferenced plots from Europe and adjacent areas (i.e., 1,731,055 vegetation plots and 36,670,535 observations from 34,643 different taxa).

These vegetation plots were first split into two sets, depending on the presence or absence of a habitat type label:

1. a dataset containing unlabeled data, i.e., vegetation plots with a missing indication of EUNIS habitat type. This dataset (henceforth "fill-mask dataset") containing 572,231 vegetation plots could only be used for training the masked language model.

2. a dataset containing labeled data, i.e., vegetation plots with an indication of EUNIS habitat type. This dataset (henceforth "text classification dataset") containing 850,933 vegetation plots could be used for training both the masked language model and the text classification model.

To ensure a clean dataset representing vegetation patterns well, some additional pre-processing steps were conducted. We removed the few species with a given cover percentage of 0, assuming these were errors or scientists reporting absent species (which resulted in 31,813,043 observations remaining). We merged duplicated species in the same vegetation plots (i.e., species that appeared twice or more in one vegetation plot because they were in different layers) and their percentage covers were summed[97] (which resulted in 31,036,661 observations remaining). The taxon names were then standardized[98] using the API of the Global Biodiversity Information Facility (GBIF). It relies on the GBIF Backbone Taxonomy as its nomenclatural source for species taxon names and integrates and harmonizes taxonomic data from multiple authoritative sources (e.g., Catalogue of Life[99], International Plant Names Index[100], World Flora Online)[101]. As EVA is an aggregator of national and regional vegetation-plot databases, this step ensured that the same species collected in two very distant areas still shared the same name[102]. If no direct match was found for the species name (e.g., the GBIF Backbone Taxonomy is not able to provide a scientific name for the EVA species "Carex cuprina"), then it was dropped. As we focused on the species taxonomic rank, taxa identified only to the genus level were dropped, and taxa identified at the subspecies level were lumped together at the species level (e.g., *Hedera* was dropped but both *Hedera helix* subsp. *helix* and *Hedera helix* subsp. *poetarum* were merged into *Hedera helix*). This resulted in 29,859,407 observations remaining. We removed hybrid species and very rare species (i.e., species that appeared less than ten times in the whole dataset), which resulted in 29,836,079 observations remaining. Vegetation plots that lost more than 25% of their taxa or their most abundant taxon after the species names matching were removed from the dataset, to ensure that the remaining plots still provided reliable representations of vegetation patterns (which resulted in the final number of 29,149,022 observations remaining). Finally, vegetation plots belonging to very rare habitat types (i.e., habitat types that appeared less than ten times in the whole dataset) were considered unlabeled data and added to the fill-mask dataset.

The set of labeled vegetation plots was then strategically split. Indeed, to avoid overfitting, ideally part of the available labeled data must be held out as a test set. However, the quantity of available full lists of plant species with estimates of cover-abundance of each species and habitat type assignment is not very high (i.e., less than 1M vegetation plots for all of Europe, a relatively low number compared to the vast amount of biodiversity data available). Partitioning the available data into a training set and a test set would reduce the number of training samples to a level too low for effective model training. As a result, it is possible to instead used k-fold cross-validation (CV)[103] to split the dataset into k subsets. Then, for each of the splits, the model can be trained using $k-1$ of the subsets for training and the latter one for validation. However, cross-validation scores for the classification of vegetation plots can be biased if the data is randomly split, because they are commonly spatially autocorrelated (spatially closer data points have similar values). One strategy to reduce the bias is splitting data along spatial blocks[104]. This procedure avoids fitting structural patterns and allows the separation of near-duplicates. Such vegetation plots differ from each other in a very small portion of species (e.g., if they are close in space, two vegetation plots may exhibit identical plant composition but feature species with slightly contrasting abundances). The data set was thus first split into spatial blocks of 6 arc-minutes (0.1 degree on the World Geodetic System 1984, or WGS 84, spheroid). Then, the blocks were split into folds. Since the geographic distribution of vegetation plots across Europe is unequal, each block can have a different number of data points. The folds were thus balanced to have approximately equal number of plots instead of assigning the same number of blocks to each fold (which could have led to folds with very different numbers of data points). This process was facilitated by the use of the research software Verde[105].

With over 1.4M vegetation plots, 29M observations and 14K species, the dataset used in this paper is one of the most extensive datasets of vegetation plots ever analysed[106]. The entire description of the dataset can be found in Supplementary Table S2, and a visualization of the data can be found in Supplementary Figure S32. An overview of the long tail distribution of species (i.e., there is a strong class imbalance, meaning that a few species are present in many of the vegetation plots) can be found in Supplementary Figure S14, and more taxonomic information of the species (e.g., class, order, and family), mostly vascular plants with some bryophytes and lichens, can be found in Supplementary Table S16.

The EUNIS habitat types[107] are referred by their codes instead of their names, as they better reflect the classification hierarchy. The coding system is structured so that each broad habitat group is represented by one letter (except the broad habitat group *Littoral biogenic habitats*, which is designated by the code *MA2*). Then, a new alphanumeric character is added for each subsequent level. For instance, the habitat type *Mediterranean, Macaronesian and Black Sea shifting coastal dune* is identified by the code *N14*, indicating its belonging to the habitat group *N1* (i.e., *Coastal dunes and sandy shores*), and more generally to the broad habitat group *N* (i.e., *Coastal habitats*). The entire list of the 227 habitat types used in this work can be found in Supplementary Table S24, but to exemplify the habitat types included, we list eight broad habitat groups used in this paper below:

- ***Littoral biogenic habitats*** (code: *MA2*) - 11 habitat types belonging to littoral habitats formed by animals such as worms and mussels or plants (salt marshes)

- ***Coastal habitats*** (code: *N*) - 25 habitat types belonging to habitats above spring high tide limit (or above mean water level in non-tidal waters) occupying coastal features and characterised by their proximity to the sea, including coastal

dunes and wooded coastal dunes, beaches and cliffs

- ***Wetlands*** (code: *Q*) - 17 habitat types belonging to wetlands, with the water table at or above ground level for at least half of the year, dominated by herbaceous or ericoid vegetation
- ***Grasslands and lands dominated by forbs, mosses or lichens*** (code: *R*) - 52 habitat types belonging to non-coastal land which is dry or only seasonally wet (with the water table at or above ground level for less than half of the year) with greater than 30% vegetation cover
- ***Heathlands, scrub and tundra*** (code: *S*) - 42 habitat types belonging to non-coastal land which is dry or only seasonally inundated (with the water table at or above ground level for less than half of the year) usually with greater than 30% vegetation cover and with the development of soil
- ***Forests and other wooded land*** (code: *T*) - 45 habitat types belonging to land where the dominant vegetation is, or was until very recently, trees with a canopy cover of at least 10%
- ***Inland habitats with no or little soil and mostly with sparse vegetation*** (code: *U*) - 23 habitat types belonging to non-coastal habitats on substrates with no or little development of soil, mostly with less than 30% vegetation cover which are dry or only seasonally wet (with the water table at or above ground level for less than half of the year)
- ***Vegetated man-made habitats*** (code: *V*) - 12 habitat types belonging to anthropogenic habitats which are dominated by vegetation and usually subject to regular management but also arising from recent abandonment of previously cultivated ground

The final dataset created solely for the fill-mask task, i.e., fill-mask dataset, contained a total of 572 231 vegetation plots covering 14 069 different species. This dataset of 10 853 856 species observations (on average 19 species per plot) was only used for fine-tuning the masked language model, as each sample was unlabeled (the vegetation plots in this set were not classified to a habitat type). Each sample was used for the fill-mask task during each split in the training set, along with around 90% of the text classification dataset.

The text classification dataset, which was created both for the fill-mask task and the text classification task, contained a total of 850 933 vegetation plots covering 13 727 different species. This dataset of 18 295 166 species observations (on average around 22 species per plot) was used for fine-tuning the masked language model and for training the classifier head (i.e., the module added on top of the masked language model to transform its outputs into predictions for assigning habitat types to vegetation plots), as each sample was labeled (the vegetation plots in this set were classified to a habitat type). Each sample was used nine times in the training set and once in the test set.

Pl@nBERT fill-mask model training

Every plant species has specific environmental preferences that shape its presence. Therefore, the task of masking some of the species in a vegetation plot and predicting which species should replace those masks can help get a good contextual understanding of an entire ecosystem. This process is known as fill-mask. A detailed description of the hardware used to train the models can be found in Supplementary Text S3.

Pl@ntBERT is based on the vanilla Transformer model BERT[46]. Hence, to predict a masked species in a vegetation plot, the model can consider (i.e., focus on and process information using the attention mechanism in the Transformer architecture) all species bidirectionally. This means the model, when looking at a specific species, has full access to the species on the left (i.e., more abundant species) and right (i.e., less abundant species). The two original BERT models (i.e., base and large) were leveraged for this study. BERT-base has 12 Transformer layers (i.e., Transformer blocks) and 110M parameters (i.e., number of learnable variables) and BERT-large has 24 Transformer layers and 340M parameters. A detailed description of the architecture of the two sizes can be found in Supplementary Table S1. Moreover, the uncased version of BERT was leveraged to train Pl@ntBERT. This version does not distinguish between "*hedera*" and "*Hedera*". Hence, as all outputs from Pl@ntBERT would be in lowercase, all inputs (abundance-ordered plant species sequences) were also lowercased to ensure consistency. For these two reasons, each sentence fed into the model was formed by listing all the species by descending abundance order, in lowercase, and separated by commas. In case of species having the same cover (which is frequent as most EVA data come from ordinal scales with a few steps only), they were randomly ordered.

For many NLP applications involving Transformer models, it is possible to simply take a pre-trained model and fine-tune it directly on some data for the task at hand. Provided that the dataset used for pre-training is not too different from the dataset used for fine-tuning, transfer learning will usually produce good results. The predictions depend on the dataset the model was trained on, since it learns to pick up the statistical patterns present in the data. However, our dataset contains binomial names (i.e., the scientific names given to species and used in biological classification, which consist of a genus name followed by a

species epithet). Because it has been pre-trained on the English Wikipedia and BookCorpus datasets, the predictions of the vanilla Transformer model BERT for the masked tokens will reflect these domains. BERT will typically treat the species names in the dataset as rare tokens, and the resulting performance will be less than satisfactory. By fine-tuning the language model on in-domain data, we can boost the performance of the downstream task[108]. This process of fine-tuning a pre-trained language model on in-domain data is called domain adaptation. Vegetation-plot records from EVA that were not assigned to a habitat type were used for this task. The sentences were created by ordering each species within a plot in their descending order of abundance, separating them by commas. Two different ways were used to tokenize (i.e., prepare the inputs for the models) the names of the species:

1. the "term" way: a species name is divided into two tokens, one for the genus name and one for the species epithet.
2. the "species" way: a whole binomial name is equivalent to a token.

More information about the versions of Pl@ntBERT can be found in Supplementary Table S7. For each approach, two model sizes were leveraged: base and large.

Unlike other NLP tasks, such as token classification or question answering, where a labeled dataset to train on is given, there is not any explicit labels in masked language modeling. A good language model is one that assigns high probabilities to sentences that are grammatically correct, and low probabilities to nonsense sentences. Assuming our test dataset consists of sentences that are coherent plant assemblages, then one way to measure the quality of our language model is to calculate the probabilities it assigns to the masked species in all the sequences of the test set. High probabilities indicate that the model is not "surprised" or "perplexed" by the unseen examples (i.e., describing the model's uncertainty or difficulty in predicting masked elements, hence reflecting how well it has learned the underlying structure of the data), and suggests it has learned the basic patterns of grammar in the language (in the case of Pl@ntBERT, the language being "floristic composition"). As a result, the perplexity, which is defined as the exponential of the cross-entropy loss, is one of the most common metrics to measure the performance of language models (the smaller its value, the better its performance). It was used in our experiments to evaluate the model in addition to the species masking accuracy.

Except for commas, the classify tokens [CLS], which represent entire input sequences, and the separate tokens [SEP], which mark the separation between different input sequences), 15% of the tokens were "masked" during the experiments. These tokens consisted of full species names in the case of Pl@ntBERT-species and of genus names or species epithets in the case of Pl@ntBERT-term. We followed the same procedure used in the original BERT paper[46]: each selected token was replaced by (i) the [MASK] token 80% of the time, (ii) a random species 10% of the time, or (iii) the same species 10% of the time. Each model was trained for five epochs (i.e., five complete pass of the training dataset through the model). This process was facilitated by the use of the deep learning package Pytorch[109] and the open-source library HuggingFace[110].

To compare how Pl@ntBERT models species assemblages compared to traditional approaches, we also implemented three alternative baseline methods solely based on species co-occurrence information. The first one is a version of Pl@ntBERT for which species are given as input in random order rather than abundance-ordered. This makes it possible to remove the information linked to the order of species so that most of the syntax rules cannot be learned anymore apart from co-occurrence patterns. The second baseline method is a naive Bayes predictor based on the species co-occurrence matrix. Ten different co-occurrence matrices were built, each time leveraging all the dataset minus one fold (to always keep the ground truth hidden). As a result, each matrix indicates how many times species of each pair co-occur in the same vegetation plots in the nine training folds. From the co-occurrence matrix, we can derive the probability of each species conditionally to an observed species assemblage. More details about how this naive Bayes predictor is constructed can be found in Supplementary Equation S25. The other baseline method is a neural network optimizing the log-loss function using stochastic gradient descent[111]. It was trained on incomplete species assemblages (i.e., for every vegetation plot of the training set, a species was randomly masked and the goal of the model was to retrieve it). More details about how the multilayer perceptron is implemented can be found in Supplementary Table S21.

Identifying habitat types

The classification of vegetation provides a useful way of summarizing our knowledge of vegetation patterns. Therefore, the task of assigning a habitat type to sentences describing floristic compositions serves to describe many facets of ecological processes. This process is called text classification.

Pl@ntBERT is based on the fine-tuned version of BERT, meaning it has already adapted its weights to predict species that are more strongly associated with the plants from the sentence. It provides a better foundation for learning task-specific models, such as a text classification model. To create a state-of-the-art model for vegetation classification, we added one additional output layer (i.e., a fully connected layer that matched the number of habitat types) on top of the pooled output.

Vegetation-plot records from EVA that were assigned to a habitat type were used for this task. The habitat labels were generated using the expert system EUNIS-ESy version v2021-06-01[60] directly by the coordinators of the EVA database using

the JUICE program[112]. This means that using the EUNIS-ESy to identify the habitat types of the raw data from EVA (without the pre-processing steps such as harmonizing the taxon names) should lead to an accuracy of 100%. Each model was trained for five epochs.

To evaluate the classification performance, we computed accuracy, precision, recall, and F1-score on the test set. Given the class imbalance in habitat labels (e.g., the habitat type R22, i.e., Low and medium altitude hay meadow, is present 69,533 times in the text classification dataset, and the habitat type U35, i.e., Boreal and arctic base-rich inland cliff, is present 12 times in the text classification dataset), the F1-score was particularly useful in assessing how well the model performed across different habitat types. We also compared Pl@ntBERT's performance against a standard BERT model trained from scratch on the same dataset to assess the benefits of domain adaptation. Finally, we compared the results with EUNIS-ESy and hdm-framework, respectively a classification expert system and a deep-learning framework.

Code availability

The generic, free, and open-source framework that supports the findings of this study is available in GitHub at `https://github.com/cesar-leblanc/plantbert`. See Figure 5 for an overview of the list of tasks that Pl@ntBERT can achieve.

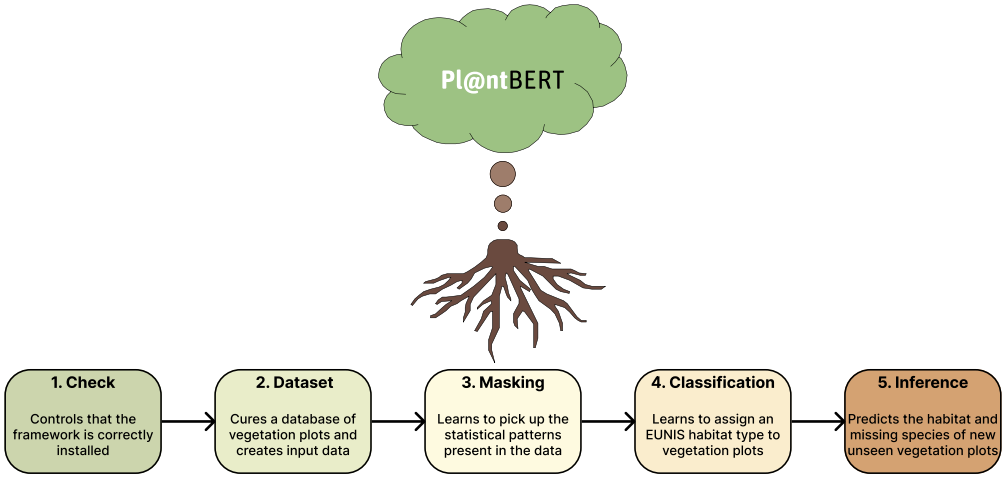


Figure 5. Overview of the framework. The panels display the sequence of tasks performed during each of the five main stages (installation check, dataset curation, masking training, classification training, and outcomes prediction).

References

1. Tilman, D., Isbell, F. & Cowles, J. M. Biodiversity and ecosystem functioning. *Annu. review ecology, evolution, systematics* **45**, 471–493 (2014).
2. Callaway, R. M. *et al. Positive interactions and interdependence in plant communities*, vol. 415 (Springer, 2007).
3. Diamond, J. M. Assembly of species communities. *Ecol. evolution communities* 342–444 (1975).
4. Keddy, P. A. Assembly and response rules: two goals for predictive community ecology. *J. Veg. Sci.* **3**, 157–164, DOI: https://doi.org/10.2307/3235676 (1992).
5. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. review Ecol. Syst.* **31**, 343–366 (2000).
6. Cody, M. L. & Diamond, J. M. *Ecology and evolution of communities* (Harvard University Press, 1975).
7. Lavergne, S., Mouquet, N., Thuiller, W. & Ronce, O. Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annu. review ecology, evolution, systematics* **41**, 321–350 (2010).
8. Thuiller, W., Pollock, L. J., Gueguen, M. & Münkemüller, T. From species distributions to meta-communities. *Ecol. Lett.* **18**, 1321–1328, DOI: https://doi.org/10.1111/ele.12526 (2015). https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12526.
9. Münkemüller, T. *et al.* Dos and don'ts when inferring assembly rules from diversity patterns. *Glob. Ecol. Biogeogr.* **29**, 1212–1229 (2020).
10. Veech, J. A. A probabilistic model for analysing species co-occurrence. *Glob. Ecol. Biogeogr.* **22**, 252–260 (2013).

11. Gotelli, N. J. & Ulrich, W. The empirical bayes approach as a tool to identify non-random species associations. *Oecologia* **162**, 463–477 (2010).

12. Siefert, A., Laughlin, D. C. & Sabatini, F. M. You shall know a species by the company it keeps: Leveraging co-occurrence data to improve ecological prediction. *J. Veg. Sci.* **35**, e13314 (2024).

13. Meynard, C. N. *et al.* Disentangling the drivers of metacommunity structure across spatial scales. *J. Biogeogr.* **40**, 1560–1571, DOI: https://doi.org/10.1111/jbi.12116 (2013). https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbi.12116.

14. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.* **23**, 1050–1063 (2020).

15. Bruelheide, H. *et al.* A checklist for using beals' index with incomplete floristic monitoring data: Reply to christensen et al.(2021): Problems in using beals' index to detect species trends in incomplete floristic monitoring data. *Divers. Distributions* **27**, 1328–1333 (2021).

16. Garbolino, E., De Ruffray, P., Brisse, H. & Grandjouan, G. Probable flora: An expression mean of ecological gradients in france. *Comptes rendus biologies* **336**, 73–81 (2013).

17. Whittaker, R. H. Dominance and diversity in land plant communities: Numerical relations of species express the importance of competition in community function and evolution. *Science* **147**, 250–260 (1965).

18. Keddy, P. A. & Shipley, B. Competitive hierarchies in herbaceous plant communities. *Oikos* 234–241 (1989).

19. Kinlock, N. L. A meta-analysis of plant interaction networks reveals competitive hierarchies as well as facilitation and intransitivity. *The Am. Nat.* **194**, 640–653 (2019).

20. Lin, Z. *et al.* A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

21. Hall, L. S., Krausman, P. R. & Morrison, M. L. The habitat concept and a plea for standard terminology. *Wildl. society bulletin* 173–182 (1997).

22. Janssen, J. *et al. European Red List of Habitats* (Publications Office of the European Union Luxembourg, 2016).

23. Directive, H. Council directive 92/43/eec of 21 may 1992 on the conservation of natural habitats and of wild fauna and flora. *Off. J. Eur. Union* **206**, 50 (1992).

24. Union, E. Regulation (eu) 2024/1991 of the european parliament and of the council of 24 june 2024 on nature restoration and amending regulation (eu) 2022/869. *Off. J. Eur. Union* **1991**, 1 (2024).

25. Davies, C. & Moss, D. Eunis habitat classification. final report to the european topic centre on nature conservation. *Eur. Environ. Agency* **256** (1999).

26. Chytr, M. *et al.* Eunis habitat classification: Expert system, characteristic species combinations and distribution maps of european habitats. *Appl. Veg. Sci.* **23**, 648–675 (2020).

27. Evans, D. The eunis habitats classification–past, present & future. *Revista de investigación marina* **19**, 28–29 (2012).

28. Bohn, U. *et al.* Karte der natürlichen vegetation europas, maßstab 1: 2 500 000.[map of the natural vegetation of europe. scale 1: 2 500 000]. *Bundesamt für Naturschutz, Bonn* (2000).

29. Dinerstein, E. *et al.* An ecoregion-based approach to protecting half the terrestrial realm. *BioScience* **67**, 534–545 (2017).

30. Chytrỳ, M. Formalized approaches to phytosociological vegetation classification. (2000).

31. Noble, I. The role of expert systems in vegetation science. *Vegetatio* **69**, 115–121 (1987).

32. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, vol. 2 (Springer, 2009).

33. Bruelheide, H. Using formal logic to classify vegetation. *Folia Geobot.* **32**, 41–46 (1997).

34. Tich, L., Chytr, M. & Landucci, F. Grimp: A machine-learning method for improving groups of discriminating species in expert systems for vegetation classification. *J. Veg. Sci.* **30**, 5–17 (2019).

35. De Cáceres, M. *et al.* A comparative framework for broad-scale plot-based vegetation classification. *Appl. Veg. Sci.* **18**, 543–560 (2015).

36. Deneu, B. *et al.* Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology* **17**, e1008856 (2021).

37. Černá, L. & Chytr, M. Supervised classification of plant communities with artificial neural networks. *J. Veg. Sci.* **16**, 407–414 (2005).

38. Olden, J. D., Lawler, J. J. & Poff, N. L. Machine learning methods without tears: a primer for ecologists. *The Q. Rev. Biol.* **83**, 171–193 (2008).

39. Wisz, M. S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. reviews* **88**, 15–30 (2013).

40. Xu, J. *et al.* Cascading predictions from common to rare species improves species distribution models. .

41. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).

42. Marcos, D. *et al.* Fully automatic extraction of morphological traits from the web: utopia or reality? *arXiv preprint arXiv:2409.17179* (2024).

43. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *nature* **596**, 583–589 (2021).

44. Lam, H. Y. I., Ong, X. E. & Mutwil, M. Large language models in plant biology. *Trends Plant Sci.* (2024).

45. Leblanc, C., Bonnet, P., Servajean, M. & Joly, A. Pl@ ntbert: leveraging large language models to enhance vegetation classification through species composition analysis (Università di Bologna, 2024).

46. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

47. Chytr, M. *et al.* European vegetation archive (eva): an integrated database of european vegetation plots. *Appl. vegetation science* **19**, 173–180 (2016).

48. Pärtel, M., Szava-Kovats, R. & Zobel, M. Dark diversity: shedding light on absent species. *Trends ecology & evolution* **26**, 124–128 (2011).

49. Bayes, T. Naive bayes classifier. *Article Sources Contributors* 1–9 (1968).

50. Bebis, G. & Georgiopoulos, M. Feed-forward neural networks. *Ieee Potentials* **13**, 27–31 (1994).

51. Veech, J. A. The pairwise approach to analysing species co-occurrence (2014).

52. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).

53. Nilsson, I. N. & Nilsson, S. G. Turnover of vascular plant species on small islands in lake möckeln, south sweden 1976–1980. *Oecologia* **53**, 128–133 (1982).

54. Klimeš, L., Dančák, M., Hájek, M., Jongepierová, I. & Kučera, T. Scale-dependent biases in species counts in a grassland. *J. Veg. Sci.* **12**, 699–704 (2001).

55. Morrison, L. W. Nonsampling error in vegetation surveys: understanding error types and recommendations for reducing their occurrence. *Plant Ecol.* **222**, 577–586 (2021).

56. Bengio, Y., Goodfellow, I. & Courville, A. *Deep learning*, vol. 1 (MIT press Cambridge, MA, USA, 2017).

57. Mucina, L. *et al.* Vegetation of europe: hierarchical floristic classification system of vascular plant, bryophyte, lichen, and algal communities. *Appl. vegetation science* **19**, 3–264 (2016).

58. Chytrỳ, M. *et al.* Floraveg. eu—an online database of european vegetation, habitats and flora. *Appl. vegetation science* **27**, e12798 (2024).

59. Beltagy, I., Lo, K. & Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

60. Chytrỳ, M. *et al.* Eunis-esy: Expert system for automatic classification of european vegetation plots to eunis habitats. (2021).

61. Leblanc, C. *et al.* A deep-learning framework for enhancing habitat identification based on species composition. *Appl. Veg. Sci.* **27**, e12802 (2024).

62. Bruelheide, H., Tichỳ, L., Chytrỳ, M. & Jansen, F. Implementing the formal language of the vegetation classification expert systems (esy) in the statistical computing environment r. *Appl. Veg. Sci.* **24**, e12562 (2021).

63. Haykin, S. *Neural networks: a comprehensive foundation* (Prentice Hall PTR, 1998).

64. Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, 278–282 (IEEE, 1995).

65. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

66. Arik, S. Ö. & Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 6679–6687 (2021).

67. Gorishniy, Y., Rubachev, I., Khrulkov, V. & Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **34**, 18932–18943 (2021).

68. Van Rossum, G., Drake, F. L. *et al.* *Python reference manual*, vol. 111 (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

69. Nickolls, J., Buck, I., Garland, M. & Skadron, K. Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for? *Queue* **6**, 40–53 (2008).

70. Greig-Smith, P. Pattern in vegetation. *J. Ecol.* **67**, 755–779 (1979).

71. Keith, D. A. *et al.* The iucn red list of ecosystems: Motivations, challenges, and applications. *Conserv. Lett.* **8**, 214–226 (2015).

72. Bruelheide, H. *et al.* splot–a new tool for global vegetation analyses. *J. vegetation science* **30**, 161–186 (2019).

73. Carmona, C. P. & Pärtel, M. Estimating probabilistic site-specific species pools and dark diversity from co-occurrence data. *Glob. Ecol. Biogeogr.* **30**, 316–326 (2021).

74. Bruelheide, H. *et al.* Using incomplete floristic monitoring data from habitat mapping programmes to detect species trends. *Divers. Distributions* **26**, 782–794 (2020).

75. Klimeš, L. Scale-dependent variation in visual estimates of grassland plant cover. *J. Veg. Sci.* **14**, 815–821 (2003).

76. Morin, P. J. *Community ecology* (John Wiley & Sons, 2009).

77. Morales-Castilla, I., Matias, M. G., Gravel, D. & Araújo, M. B. Inferring biotic interactions from proxies. *Trends ecology & evolution* **30**, 347–356 (2015).

78. Kissling, W. D. *et al.* Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *J. Biogeogr.* **39**, 2163–2178 (2012).

79. Wootton, J. T. Indirect effects in complex ecosystems: recent progress and future challenges. *J. Sea Res.* **48**, 157–172 (2002).

80. Ryo, M. *et al.* Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography* **44**, 199–205 (2021).

81. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328 (PMLR, 2017).

82. Tichỳ, L. & Chytrỳ, M. Probabilistic key for identifying vegetation types in the field: A new method and android application. *J. Veg. Sci.* **30**, 1035–1038 (2019).

83. Chao, A. *et al.* An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis) similarity measures. *Ecol. monographs* **89**, e01343 (2019).

84. Guisan, A. & Zimmermann, N. E. Predictive habitat distribution models in ecology. *Ecol. modelling* **135**, 147–186 (2000).

85. Guisan, A., Thuiller, W. & Zimmermann, N. E. *Habitat suitability and distribution models: with applications in R* (Cambridge University Press, 2017).

86. Joly, A. *et al.* Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of birds, plants, snakes and fungi. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 416–439 (Springer, 2023).

87. Joly, A. *et al.* Overview of lifeclef 2024: Challenges on species distribution prediction and identification. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 183–207 (Springer, 2024).

88. Botella, C. *et al.* Overview of geolifeclef 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing. In *CLEF 2023: Conference and Labs of the Evaluation Forum* (2023).

89. Picek, L. *et al.* Overview of geolifeclef 2024: Species presence prediction based on occurrence data and high-resolution remote sensing images. *Work. Notes CLEF* (2024).

90. Bonnet, P. *et al.* Synergizing digital, biological, and participatory sciences for global plant species identification: Enabling access to a worldwide identification service. *Biodivers. Inf. Sci. Standards* **7** (2023).

91. Contini, M. *et al.* Seatizen atlas: a collaborative dataset of underwater and aerial marine imagery. *Sci. Data* **12**, 67 (2025).

92. Botella, C. *et al.* The geolifeclef 2023 dataset to evaluate plant species distribution models at high spatial resolution across europe. *arXiv preprint arXiv:2308.05121* (2023).

93. Joly, A. *et al.* Lifeclef 2024 teaser: Challenges on species distribution prediction and identification. In *European Conference on Information Retrieval*, 19–27 (Springer, 2024).

94. Marcenò, C. *et al.* Facebook groups as citizen science tools for plant species monitoring. *J. Appl. Ecol.* **58**, 2018–2028 (2021).

95. Leblanc, C., Joly, A., Lorieul, T., Servajean, M. & Bonnet, P. Species distribution modeling based on aerial images and environmental features with convolutional neural networks. In *CLEF (Working Notes)*, 2123–2150 (2022).

96. Yang, X., Gao, J., Xue, W. & Alexandersson, E. Pllama: An open-source large language model for plant science. *arXiv preprint arXiv:2401.01600* (2024).

97. Fischer, H. S. On the combination of species cover values from different vegetation layers. *Appl. Veg. Sci.* **18**, 169–170 (2015).

98. Grenié, M. *et al.* Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods Ecol. Evol.* **14**, 12–25 (2023).

99. Bánki, O. Catalogue of life checklist. *(No Title)* (2023).

100. Gilmour, R. The international plant names index. *Electron. Resour. Rev.* **4**, 60–61 (2000).

101. Borsch, T. *et al.* World flora online: Placing taxonomists at the heart of a definitive and comprehensive global resource on the world's plants. *Taxon* **69**, 1311–1341 (2020).

102. Jansen, F. & Dengler, J. Plant names in vegetation databases–a neglected source of bias. *J. Veg. Sci.* **21**, 1179–1186 (2010).

103. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. royal statistical society: Ser. B (Methodological)* **36**, 111–133 (1974).

104. Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).

105. Uieda, L. Verde: Processing and gridding spatial data using green's functions. *J. Open Source Softw.* **3**, 957 (2018).

106. Picek, L. *et al.* Geoplant: Spatial plant species prediction dataset. *arXiv preprint arXiv:2408.13928* (2024).

107. Davies, C. E. *et al.* The eunis habitat classification. (2000).

108. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

109. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).

110. Wolf, T. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

111. Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

112. Tichỳ, L. Juice, software for vegetation classification. *J. vegetation science* **13**, 451–453 (2002).

Acknowledgements

The research described in this paper was funded by the European Commission through the GUARDEN (safeGUARDing biodivErsity aNd critical ecosystem services across sectors and scales) and MAMBO (Modern Approaches to the Monitoring of BiOdiversity) projects. These projects received funding from the European Union's Horizon Europe research and innovation programme under grant agreements 101060693 (start date: 01/11/2022; end date: 31/10/2025) and 101060639 (start date: 01/09/2022; end date: 31/08/2026), respectively. W.T. also acknowledge support from the HorizonEurope OBSGESSION project (N°101134954). The content of this paper reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. Our major thanks go to thousands of European vegetation scientists of several generations who collected the original vegetation-plot data in the field, published them or made their unpublished data available to others, and to those who spent myriad hours digitizing data and managing the contributing databases. Vegetation-plot data for this study were provided by Sylvain Abdulhak, Alicia Acosta,

Emiliano Agrillo, Pierangela Angelini, Iva Apostolova, Olivier Argagnon, Fabio Attorre, Svetlana Aćić, Christian Berg, Ariel Bergamini, Erwin Bergmeier, Idoia Biurrun, Maxim Bobrovsky, Steffen Boch, Gianmaria Bonari, Anne Bonis, Zoltán Botta-Dukát, Jan-Bernard Bouzillé, Helge Bruelheide, Vanessa Bruzzaniti, Juan Antonio Campos, Maria Laura Carranza, Laura Casella, Alessandro Chiarucci, Andrei Chuvashov, Milan Chytrý, János Csiky, Olga Demina, Jürgen Dengler, Panayotis Dimopoulos, Dmytro Dubyna, Tetiana Dziuba, Alexei Egorov, Rasmus Ejrnæs, Franz Essl, Jörg Ewald, Giuliano Fanelli, Federico Fernández-González, Úna FitzPatrick, Xavier Font, Gianpietro Giusso del Galdo, Emmanuel Garbolino, Itziar García-Mijangos, Rosario G. Gavilán, Jean-Michel Genis, Michael Glaser, Valentin Golub, Friedemann Goral, Jean-Claude Gégout, Behlül Güler, Rense Haveman, Stephan Hennekens, Adrian Indreica, Maike Isermann, Ute Jandt, Florian Jansen, Jan Jansen, John Janssen, Anni Kanerva Jašková, Borja Jiménez-Alfaro, Martin Jiroušek, Veronika Kalníková, Ali Kavgacı, Larisa Khanina, Ilona Knollová, Vitaliy Kolomiychuk, Łukasz Kozub, Daniel Krstonošić, Helmut Kudrnovsky, Anna Kuzemko, Filip Küzmič, Zygmunt Kącki, Flavia Landucci, Igor Lavrinenko, Mariya Lebedeva, Jonathan Lenoir, Armin Macanović, Corrado Marcenò, Aleksander Marinšek, Marco Massimi, Ruth Mitchell, Jesper Erenskjold Moeslund, Pavel Novák, Vladimir Onipchenko, Viktor Onyshchenko, Robin Pakeman, Hristo Pedashenko, Tomáš Peterka, Remigiusz Pielech, Vadim Prokhorov, Ricarda Pätsch, Aaron Pérez-Haase, Valerijus Rašomavičius, Maria Pilar Rodríguez-Rojo, John S. Rodwell, Iris de Ronde, Eszter Ruprecht, Solvita Rūsiņa, Michele De Sanctis, Joop Schaminée, Joachim Schrautzer, Ingrid Seynave, Desislava Sopotlieva, Angela Stanisci, Milica Stanišić-Vujačić, Zvjezdana Stančić, Zora Dajić Stevanović, Danijela Stešević, Jens-Christian Svenning, Grzegorz Swacha, Irina Tatarenko, Ioannis Tsiripidis, Ruslan Tsvirko, Pavel Dan Turtureanu, Domas Uogintas, Emin Uğurlu, Milan Valachovič, Kiril Vassilev, Roberto Venanzoni, Sophie Vermeersch, Risto Virtanen, Denys Vynokurov, Lynda Weekes, Wolfgang Willner, Thomas Wohlgemuth, Sergey Yamalov, Svitlana Yemelianova, Dominik Zukal, Mirjana Krstivojević Ćuk, Renata Ćušterevska, Andraž Čarni, Jozef Šibík, Urban Šilc, and Željko Škvorc. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

Author contributions

All authors conceived the experiments, C.L. conducted the experiments, all authors interpreted the results. C.L. wrote the main manuscript text and prepared all figures and tables. All authors reviewed the manuscript.

Data Availability

The data that support the findings of this study are available from EVA but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available from the authors or EVA custodians upon reasonable request and with permission of EVA. The DOI of the EVA data selection for this project is https://doi.org/10.58060/QR4B-G979.

Additional information

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf