PNAS RESEARCH ARTICLE | ECOLOGY

Predicting undetected native vascular plant diversity at a global scale

Barnabas H. Daru[a,1]

Edited by Douglas Soltis, University of Florida, Gainesville, FL; received November 13, 2023; accepted June 28, 2024

Vascular plants are diverse and a major component of terrestrial ecosystems, yet their geographic distributions remain incomplete. Here, I present a global database of vascular plant distributions by integrating species distribution models calibrated to species' dispersal ability and natural habitats to predict native range maps for 201,681 vascular plant species into unsurveyed areas. Using these maps, I uncover unique patterns of native vascular plant diversity, endemism, and phylogenetic diversity revealing hotspots in underdocumented biodiversity-rich regions. These hotspots, based on detailed species-level maps, show a pronounced latitudinal gradient, strongly supporting the theory of increasing diversity toward the equator. I trained random forest models to extrapolate diversity patterns under unbiased global sampling and identify overlaps with modeled estimations but unveiled cryptic hotspots that were not captured by modeled estimations. Only 29% to 36% of extrapolated plant hotspots are inside protected areas, leaving more than 60% outside and vulnerable. However, the unprotected hotspots harbor species with unique attributes that make them good candidates for conservation prioritization.

biodiversity hotspots | species distribution models | protected areas | vascular plant diversity | machine learning

Significance

To protect species, we need to know where they are, but high-resolution maps for vascular plants are lacking at a global scale. Here, I produce high-resolution native range maps for vascular plants at the species-level using informatics approaches that enable accurate assessment of biogeography patterns. My findings reveal important plant hotspots and provide a quantitative assessment of latitudinal diversity gradient for vascular plants at the species level. Machine learning identifies cryptic hotspots that can guide conservation priorities. However, over 60% of vascular plant diversity lies unprotected, emphasizing urgent conservation actions. This study explores fundamental questions about the distribution, evolution, and diversity of vascular plant life through unique approaches that are illuminating and unlocking ecological secrets hidden until now.

Vascular plants are a very diverse taxonomic group comprising about 340,000 species worldwide (1–3). They occur across all types of biomes, from rainforests to savannas, providing key ecosystem services upon which terrestrial life and human civilization depend (4). These services include provisioning (e.g., food and medicines), regulation of ecosystem processes (e.g., trophic regulation and water purification), cultural (firewood and ornamental), and supporting services (e.g., primary productivity), yet identifying concentrations of vascular plant species, endemism, and evolutionary diversity at a global scale rest largely on coarse-grained estimations (5, 6). Consequently, the underlying processes and principles governing vascular plant diversity at finer scales, which requires an accurate knowledge of the locations of species' geographic distributions, remain unknown at a global scale. Accurate knowledge of vascular plant geographic distributions is key for prioritizing conservation efforts and mitigating loss of species and their functions in the face of profound human impact on the planet (5).

The surge in the availability of vascular plant diversity data from heterogeneous sources has led to their compilation into major data hubs such as the Global Biodiversity Information Facility (7) that can facilitate macroecological analyses (8–12). These datasets are often available as point occurrences of where a species has been documented as present based on a voucher specimen in a herbarium or sighting in the field without linkage to tangible physical material (13). However, both voucher and observation records suffer from sampling biases and coverage gaps (14–16). Floristically rich regions like the Neotropics, Afrotropics, and Southeast Asia are particularly undersampled (14, 15, 17). These sampling biases and coverage gaps can lead to spurious ecological inferences (14, 15, 17) such as underestimation of true diversity (18) and potentially compromise effective conservation prioritization (19). Determining whether such coverage gaps and sampling biases reflect true absence or sampling artifact is challenging (20). Consequently, our understanding of fundamental biogeographic patterns such as latitudinal diversity gradients (21) or identifying global priority areas for conservation (such as biodiversity hotspots) relies largely on the distributions of well-studied animal groups (e.g., tetrapods).

Species distribution models (SDM) can predict vascular plant occurrences in unsampled areas (22) but often rely on biased occurrence records, leading to underestimates of true native diversity. The native range of a species is a fundamental unit of biological diversity that underpins our understanding of a species' natural habitat (23) but remains unknown for most vascular plant species across the globe. Nonetheless, vascular plants are known to associate with well-studied groups such as birds, mammals, amphibians, and reptiles (i.e., tetrapods), whose geographic sampling is less biased and may offer more accurate insights into native biodiversity. Furthermore, machine learning approaches can improve SDM predictions because

Author affiliations: [a]Department of Biology, Stanford University, Stanford, CA 94305

Author contributions: B.H.D. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

[1]Email: bdaru@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2319989121/-/DCSupplemental.

Published August 12, 2024.

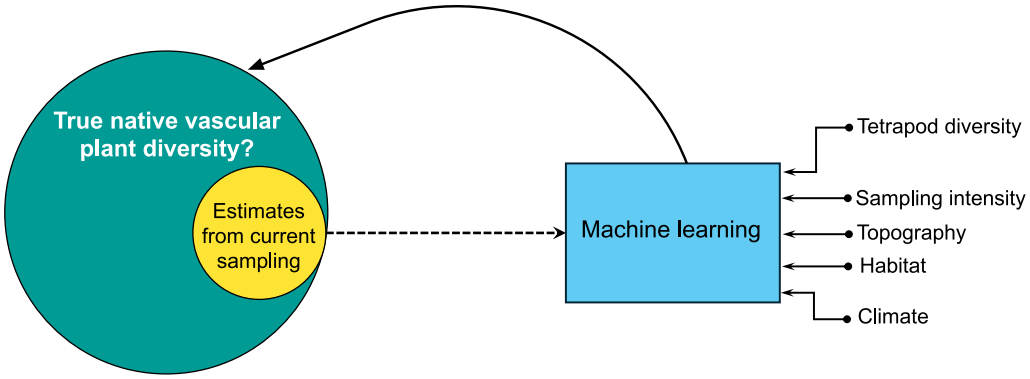

Fig. 1. Hypothesis for predicting native vascular plant diversity from modeled estimates based on current sampling. Species distribution models often rely on biased occurrence records, leading to underestimates of true native diversity. By incorporating the current sampling density alongside unbiased data (e.g., tetrapod distribution), habitat characteristics, and climate, it is possible to train machine learning models on the modeled estimates to improve predictions of true native vascular plant diversity despite current underestimates due to sampling biases and coverage gaps.

of their ability to deal with complex relationships between occurrences of poorly sampled groups like vascular plants, environmental factors, and evolutionary history that might present major challenges for conventional statistical models (24, 25). Therefore, I hypothesize that by training machine learning models on the modeled estimates as a function of plant sampling density, habitat characteristics, climate, alongside knowledge of tetrapod diversity whose geographic sampling is less biased (26, 27), it is possible to enhance the prediction of true native vascular plant diversity and potentially uncover hidden vascular plant diversity (Fig. 1). This information can guide future targeted biodiversity collecting (28). However, previous application of machine learning models to predict vascular plant diversity focused on specific taxonomic groups [e.g., Bromeliaceae; (29)] or are based on the compilation of regional checklists and floras aggregated to artificial and large administrative units such as countries (3, 30–32). While these approaches have provided insights into broad-scale patterns (30, 32–36), they assume that the ecological processes determining the native range of a vascular plant species within a given artificial administrative unit are similar across communities, precluding finer-scale ecological processes that determine a species' native range. A global analysis that addresses these limitations could reveal links between species' ecological preferences, shared evolutionary history, and potentially irreplaceable ecological and functional traits (37–39). Finally, integrating these approaches can be useful for identifying potential sites in need of conservation prioritization such as assessing the effectiveness of existing protected areas in capturing important vascular plant diversity hotspots.

Here, I address these knowledge gaps using an integrative workflow. I combine occurrence records, alpha hull polygons, species' dispersal capacity, natural habitat, and environmental variables within a framework of species distribution modeling (*SI Appendix*, Fig. S1) to generate estimates of species-level native distributions for 201,681 vascular plant species at a spatial grain of 5-arc min (~9 km at the equator). I stacked these modeled individual distributions to address three questions: 1) What are the key patterns of modeled vascular plant diversity? 2) How effectively can a random forest machine learning model validate these patterns and uncover potential refuges of hidden vascular plant diversity? 3) How effective are the existing protected areas in capturing important vascular plant diversity hotspots in need of conservation efforts?

Results and Discussion

Global Distributions of Native Vascular Plant Diversity and Endemism. Using predictive models and global environmental data layers (*SI Appendix*, Fig. S1), I generated species-level native range maps for 201,681 vascular plant species. The models showed high performance across the test statistics (*SI Appendix*, Fig. S2), underscoring the reliability of the predictions as indicators of potential vascular plant distributions. I found areas of high vascular plant species richness clustered across the globe including Southern Mexico, Mesoamerica, Amazon, Andes, Atlantic Forest, West and Central Africa, Eastern Arc, Cape Floristic Region, Madagascar, Hengduan-Himalaya, Indo-Malay, and Southeast and Southwest Australasia (Fig. 2). Importantly, the areas richest in native species richness also coincide with areas of high phylogenetic diversity (Fig. 2), which is in line with previous findings of strong correlation between phylogenetic diversity and species richness (40, 41). Additionally, I found a clear latitudinal gradient for both species richness and phylogenetic diversity, with higher richness and phylogenetic diversity near the equator, which gradually decreases at higher latitudes (Fig. 2). These patterns parallel those observed in tetrapods (Fisher's $z = 0.31$ for species richness and 0.32 vs. 0.35, for phylogenetic diversity) (Fig. 2 and *SI Appendix*, Fig. S3), albeit my maps show a slightly broader band of latitudinal diversity gradient compared with previous studies (21, 28) probably due to the limitations of my range polygon approach to modeling species distributions rather than biased point records which could result in underestimated range. While ecological theories explaining latitudinal diversity gradients have been empirically quantified for tetrapods (21) and recently for ants (28), comparable data for vascular plants were lacking until now. This finding for vascular plants based on detailed and quantitative species-level maps, strongly supports the classic theory of increasing diversity toward the tropics. An evaluation of the top areas of high species richness and phylogenetic diversity—hotspots, defined as the top 10% richest pixels of these metrics—reveals overlap with tetrapod hotspots (*SI Appendix*, Fig. S4) in some regions like Mesoamerica, Amazon, Central Africa, and Indo-Malay. However, correlations between vascular plants and tetrapods are slightly lower than those within tetrapod groups (Pearson's r: plants vs. tetrapods mean = 0.72 for species richness and 0.77 for phylogenetic diversity; tetrapods vs. tetrapods = 0.83 for both species richness and phylogenetic diversity; $P \ll 0.0001$, *SI Appendix*, Fig. S4 and Table S1). Nonetheless, the global overlap between vascular plants vs. tetrapods ($r = 0.72$) still indicates a high level of congruence that is consistent with the global spatial congruence of ants with tetrapods (28). These findings suggest that biogeographic patterns and conservation efforts focused on well sampled tetrapod groups can likely capture global vascular plant diversity but with some local variations. Notably, unique vascular plant hotspots are identified in Chaco

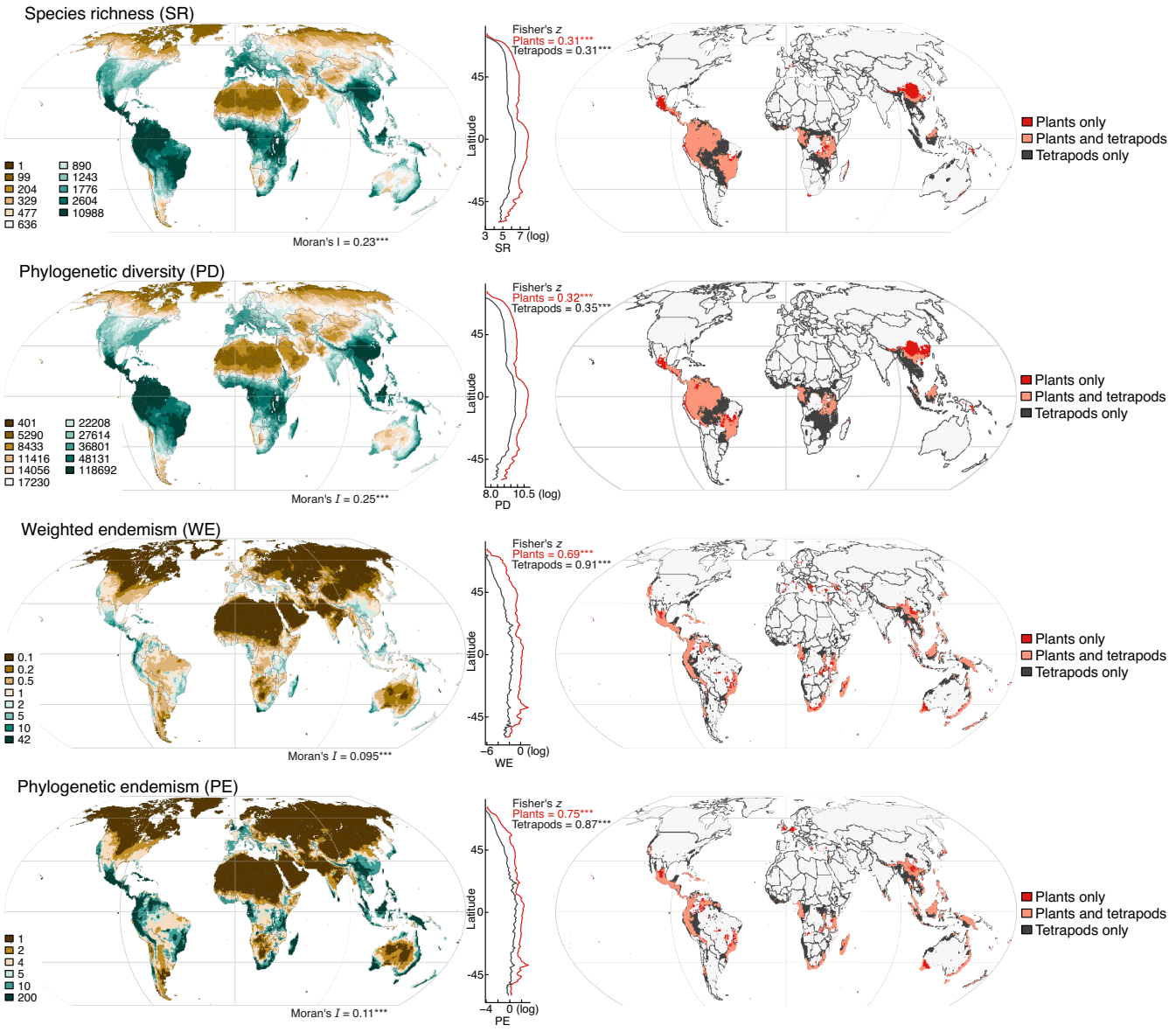

Fig. 2. Geographical patterns of native vascular plant diversity (n = 201,681 species) in relation to terrestrial tetrapods. left panel, Overall diversity patterns of species richness, phylogenetic diversity, weighted endemism, and phylogenetic endemism produced at 20 km × 20 km resolution and stacked into unique layers. Pixel values were binned into 10 quantiles to generate the color palette, noting that the scale differs between panels. Spatial evenness or clustering was calculated using Moran's *I* spatial autocorrelation with values of 1 indicating clustered patterns and 0 dispersed patterns. middle panel, Latitudinal plots of diversity patterns of each metric across 100 km latitudinal bins for plants (in red) and tetrapods (gray) along with effect size (Fisher's *z*) of the slope of latitudinal diversity gradient. right panel, Hotspots (diversity centers) defined as the top 10% of highest-ranking pixels for each metric, i.e., 90th percentile values, for species richness, phylogenetic diversity, weighted endemism, and phylogenetic endemism for plants (in red) in comparison to tetrapod hotspots (dark gray). All maps are projected under the Wagner IV projection. Significance codes: ***$P < 0.001$.

and the Cerrado savannas, Democratic Republic of Congo, and Yunnan, which do not align with tetrapod hotspots (Fig. 2). This pattern could be due to differences in relationships between vascular plants and climates from those of tetrapods.

Second, I analyzed patterns of species-weighted and phylogenetic endemism, which quantify the presence of rare species and geographically unique evolutionary lineages (38). I found that regions of high weighted and phylogenetic endemism are more dispersed (Moran's *I*: 0.095 and 0.11, both $P \ll 0.0001$) and distributed in several key areas: Americas, Afrotropics, Mediterranean, Himalaya and Southeast Asia, Australasia, and Oceania (Fig. 2). Areas of vascular plant endemism showed moderate correlation with endemism hotspots of tetrapods (Pearson's $r = 0.55$ to 0.57, $P \ll 0.0001$, *SI Appendix*, Fig. S4). However, this correlation is generally weaker than within tetrapod groups themselves (*SI Appendix*, Fig. S4 and Table S1). Notably, several of the regions harboring unique hotspots of vascular plant weighted and phylogenetic endemism do not coincide with endemism hotspots of tetrapods (Fig. 2). These findings indicate complex ecological processes and habitat diversity that may have evolved in these areas over time.

My modeled estimates incorporate dispersal limitation in a phylogenetic context, minimizing unrealistic predictions from unconstrained dispersal assumptions. Additionally, unlike standard workflows that rely on biased occurrence data, my approach defines a final training area that reflects the likely dispersal capabilities of the species and captures its natural habitat based on ecoregions. This avoids limitations associated with arbitrary spatial extents used in standard workflows, leading to more accurate distribution predictions (*SI Appendix*, Fig. S5).

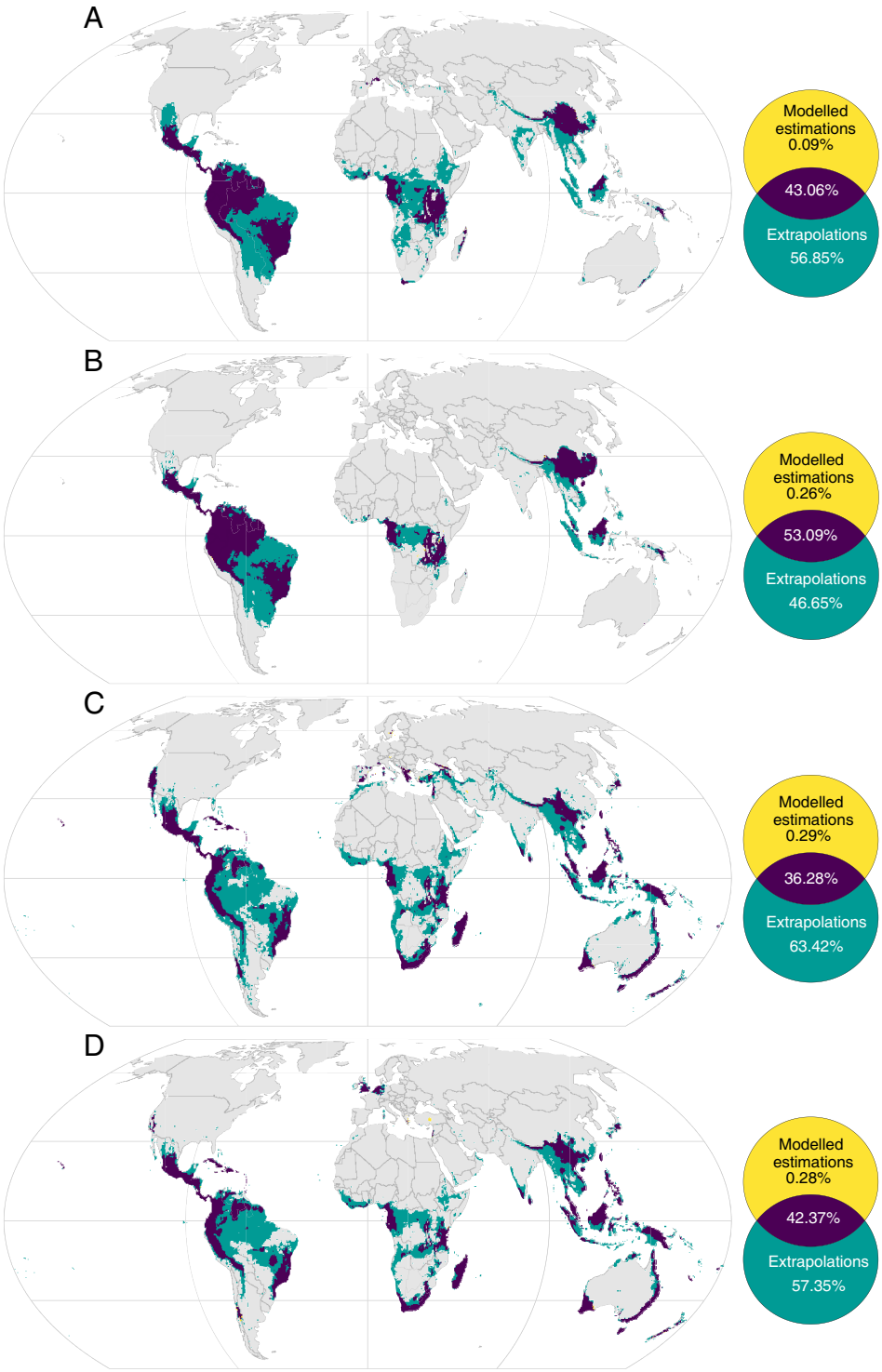

Fig. 3. Comparison of modeled estimates of vascular plant diversity hotspots based on current sampling vs. machine learning extrapolations assuming a globally unbiased sampling. Overlap of vascular plant diversity hotspots extrapolated by a machine learning random forest model under a scenario of globally unbiased sampling in comparison to modeled estimates based on current sampling for (*A*) species richness, (*B*) phylogenetic diversity, (*C*) weighted endemism, and (*D*) phylogenetic endemism. Refuge areas (in teal) represent unique hotspots predicted by random forest machine learning under a universally high global sampling but which do not fall into hotspots based on current sampling. The maps are projected under the Wagner IV projection.

Validation with Machine Learning Random Forest Model Extrapolations. I tested my modeled estimates for the effects of sampling biases common in vascular plant occurrence records (*SI Appendix*, Fig. S6) by training a random forest model to extrapolate vascular plant diversity patterns under a globally unbiased vascular plant sampling and as a function of tetrapod diversity alongside climatic variables and habitat characteristics. I assessed model accuracy using 5-fold spatial block cross-validation (42) by systematically dividing the data into spatial blocks using predefined floristic realms of the world (43). I identified optimal hyperparameters that minimize the root-mean-square error (RMSE) or "mtry," by tuning from 1 to 10, resulting in mtry = 4 for all diversity metrics except phylogenetic diversity with mtry = 5 (*SI Appendix*, Fig. S7). These settings were used to train the final models and generate extrapolated values per pixel. Assuming a globally high sampling in the future, I predict that 36% to 53% of diversity hotspots

will be robust to sampling effects. However, I uncover previously undetected hotspots of species richness and phylogenetic diversity that represent an increase of 47% to 63% assuming globally high sampling in the future. These undetected hotspots of species richness and phylogenetic diversity are located in southern United States, Sonora-Chihuahua and Yucatan Mexico, Mato Grosso Brazil, Paraguay, Central and Southeast Argentina, West Africa, Central Africa, Horn of Africa, Southern Angola, and Northern Namibia, Eastern Cape, India, Myanmar-Thailand-Malaysia, Sumatra, Kalimantan, and Southeast Australia (Fig. 3). Similarly, undetected hotspots of species and phylogenetic endemism are predicted in biodiversity-rich but currently undersampled regions including central South America, Central Africa, and Southeast Asia (Fig. 3). These findings are consistent with the correlation of continuous diversity values for each metric (*SI Appendix*, Table S1) where the correlation of modeled estimates versus extrapolated (without accounting for tetrapod diversity) are consistently higher than correlations of modeled estimates versus extrapolated (with high sampling and accounting for tetrapod diversity). These findings highlight the model's ability to capture potentially hidden vascular plant diversity, highlighting potential future diversity centers based on current sampling limitations.

Protection Levels of Hotspots Inside and Outside Protected Areas. Although protected areas are essential for conserving biodiversity, there are concerns that the success of the current network of protected areas may be biased toward certain locations rather than achieving conservation priorities (44). I tested this assumption by comparing predicted locations of plant hotspots (both modeled estimations and random forest extrapolations) within protected areas to their counterparts outside protected areas. This comparison provides insights into potential biases in the current protected area network and informs future conservation strategies. For hotspots based on modeled estimations, I found that species richness and phylogenetic diversity share equal protection levels with only 35% of hotspots cells intersecting with protected areas (Fig. 4). However, metrics of endemism including weighted and phylogenetic endemism are proportionally better protected, with 42% of these hotspots falling within protected areas (Fig. 4). These well-protected endemism hotspots tend to overlap larger reserves such as the Paríma Tapirapeco in Venezuela, Área De Proteção Ambiental Fernão Dias in Brazil, and Ngadju Indigenous Protected Area in Australia. This highlights the importance of these large protected areas for safeguarding unique evolutionary lineages. For extrapolated hotspots, protection levels yielded substantially fewer hotspot cells (29% to 36%) falling inside protected areas, leaving more than 60% outside and vulnerable. This means that the random forest machine learning extrapolations, while effective at uncovering hidden vascular plant diversity, can also reveal potential gaps in our protected area network. Although the random forest approach validates the modeled estimates in some regions, I found that large swathes of hidden vascular plant diversity in Bolivia, Brazil, Democratic Republic of Congo, Eastern China, South Africa, and Papua New Guinea are potentially located in places not covered by the current network of protected areas. These low protection levels may have arisen from ignorance of plant geographic distributions (17), supporting previous studies (28, 45–47) that protected areas do not maximize the protection of biodiversity. This finding suggests a critical need to expand protected areas or implement stricter conservation measures in these regions to safeguard these undiscovered hotspots of vascular plant diversity. These regions may represent previously overlooked centers of endemism or areas with unique ecological conditions that harbor a high diversity of vascular plant species. The unprotected hotspots highlighted in this study could guide efforts to expand the existing protected areas to achieve the United Nation's Kunming-Montreal Global Biodiversity Framework of expanding coverage of terrestrial protected areas and other conservation areas to 30% by 2030 (48).

The tendency of a species to be included inside or outside protected areas may depend on its evolutionary history (49, 50), intrinsic life history traits, or extinction risk (44). For example, reserves may be designed such that they may be biased toward organisms that are larger or appealing (51) or as response to biodiversity loss (44, 52). I tested these assumptions by assessing the impact of protected areas on common intrinsic functional traits (such as plant height and seed size), evolutionary history (evolutionary distinctiveness), and extinction risk (defined as degree of threat facing a species with data derived from published dataset of machine learning predictions of conservation status for over 150,000 land plant species) (53), for each diversity metric inside and outside protected areas. For modeled estimations, I found that protected hotspots predominantly harbor species with both higher extinction risk (Cohen's d = 0.40 to 0.70, $P < 0.01$) and larger stature (Cohen's d = 0.12 to 0.24, $P < 0.01$; Fig. 5). This means that protected hotspots may offer some conservation benefits by preserving vascular plants particularly those at elevated risk of extinction, consistent with the finding that protected areas slow down species declines (52). However, it could also indicate a bias in conservation decisions toward forested areas (54–56), as seen in certain parts of the USA (54). This suggests the need for future protected areas to better represent all ecosystems, not just those currently well protected, such as forests. Importantly, I identified greater evolutionary distinctiveness prevalent in unprotected hotspots of weighted endemism (Cohen's d = −0.15, $P < 0.01$). Evolutionary distinctiveness refers to species with fewer or no close living relatives (57). These species represent unique lineages on the tree of life and their loss would imply the disappearance of entire evolutionary branches. The fact that these irreplaceable species are found more prevalently in unprotected areas indicates the urgency for additional conservation efforts in these hotspots. Alternatively, this finding could also reflect a bias in habitat selection for protected areas, favoring places with plants commonly used for human sustenance and livelihoods (58) over those harboring evolutionarily distinct lineages including basal monocots like *Stylochaeton, Amorphophallus,* and *Pseudohydrosme* or the monotypic *Eremosyne* genus endemic to Western Australia. Similarly, for random forest extrapolations, while protected hotspots harbor larger statured and threatened species (Fig. 5), only evolutionary distinctiveness remained significant and effective in unprotected hotspots (Cohen's d = −0.27, $P < 0.01$; Fig. 5). This suggests that evolutionarily rare phylogenetic branches of the vascular plant tree of life are still outside and vulnerable and thus could be included in the global priority map for the expansion of protected areas.

Analyses of plant diversity and endemism patterns as well as their overlap with conservation areas are not new and have been explored at continental scales previously (59–61). However, this study applies these methods at a global scale with a high number of plant species (>200,000 species), allowing me to uncover unique patterns of native vascular plant diversity that would not be possible at a continental scale and how their relationships agree with tetrapods in some regions and differ in certain regions. By training a machine learning model constructed from random forest, I uncover previously unnoticed cryptic hotspots, providing a promising opportunity for future conservation efforts. These hidden refuges, especially in South America, tropical Africa, and Southeast Asia, could serve as targets for future vascular plant collecting and conservation of both species

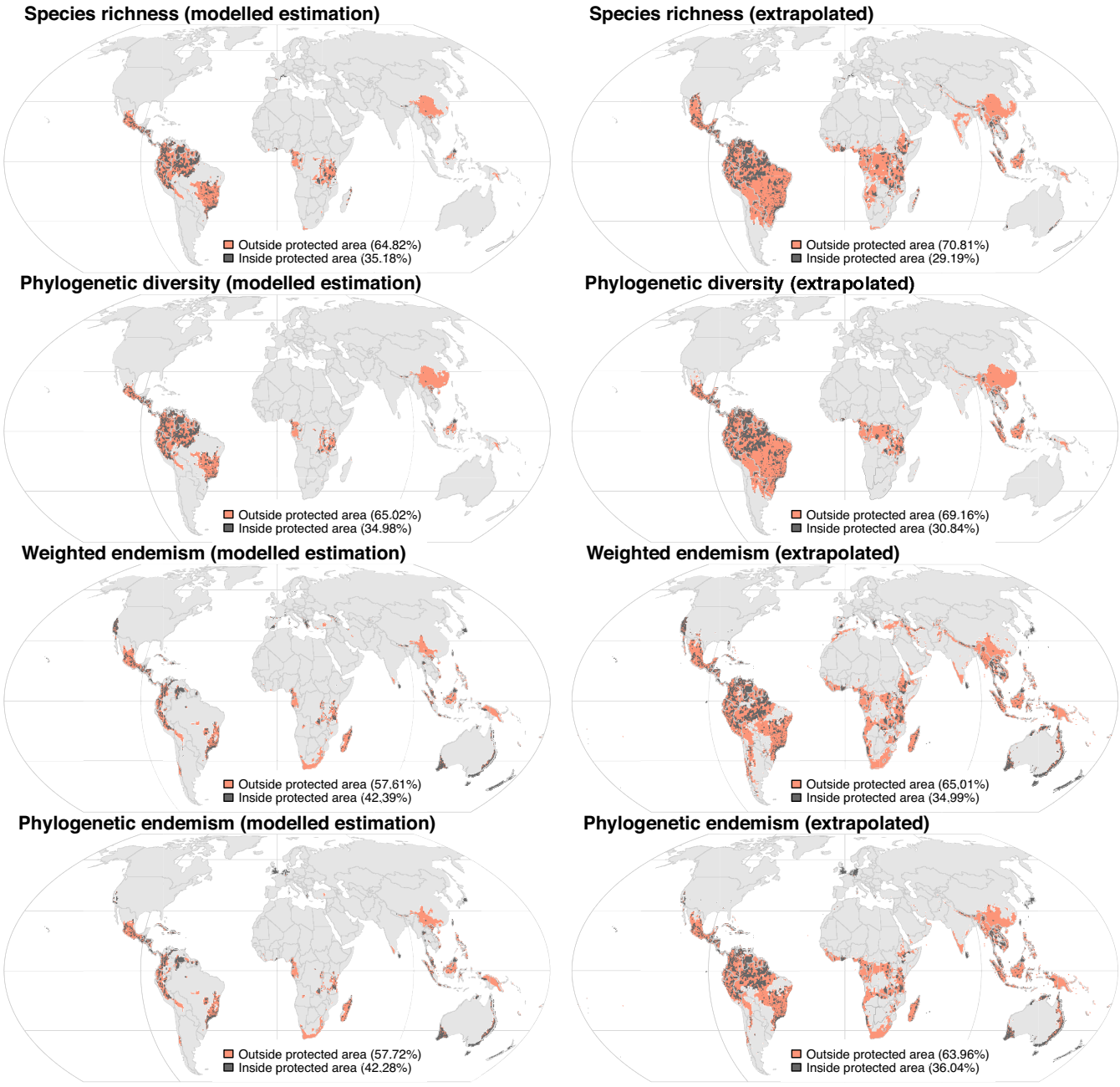

Fig. 4. Coverage of hotspots within and outside protected areas. Protection levels for hotspots inside and outside current networks of protected areas for species richness, phylogenetic diversity, weighted endemism, and phylogenetic endemism, based on modeled estimations (*Left panel*) and random forest extrapolations (*Right panel*). The maps are projected under Wagner IV projection (code "+proj=wag4").

and phylogenetic diversity. Furthermore, the evaluation of protection levels within and outside of protected areas raises concerns about the vulnerability of critical biodiversity centers in need of conservation, highlighting the need to expand the coverage of existing reserves and national parks, to align with global conservation goals, such as the Kunming-Montreal Global Biodiversity Framework. It is worth noting that majority of extrapolated hotspots lie outside protected areas, underscoring the effectiveness of random forest machine learning in uncovering hidden vascular plant diversity, while also revealing potential gaps in our protected area network. Finally, I highlight the importance of considering the evolutionary history and intrinsic traits of vascular plant species when designing protected areas. While some traits, like plant height and extinction risk, are better preserved within protected hotspots, evolutionary distinctiveness remains more prevalent outside protected areas, emphasizing the urgency of enhancing conservation efforts in these areas. The detailed species-level native range maps presented here could greatly enhance research into the mechanisms structuring and maintaining vascular plant diversity across ecological scales.

Materials and Methods

I obtained vascular plant occurrence data from Global Biodiversity Information Facility. These were thoroughly cleaned to remove errors and reflect species' known native ranges as defined by the World Checklist of Vascular Plants (3). The cleaned data were then used to generate alpha hull polygons which were cropped to land areas and finetuned by clipping them using the polygons of vascular plant families (62, 63). Using occurrence data derived from systematic sampling of the alpha hull polygons, I modeled species distributions as a function of environmental conditions and plant sampling intensity as background

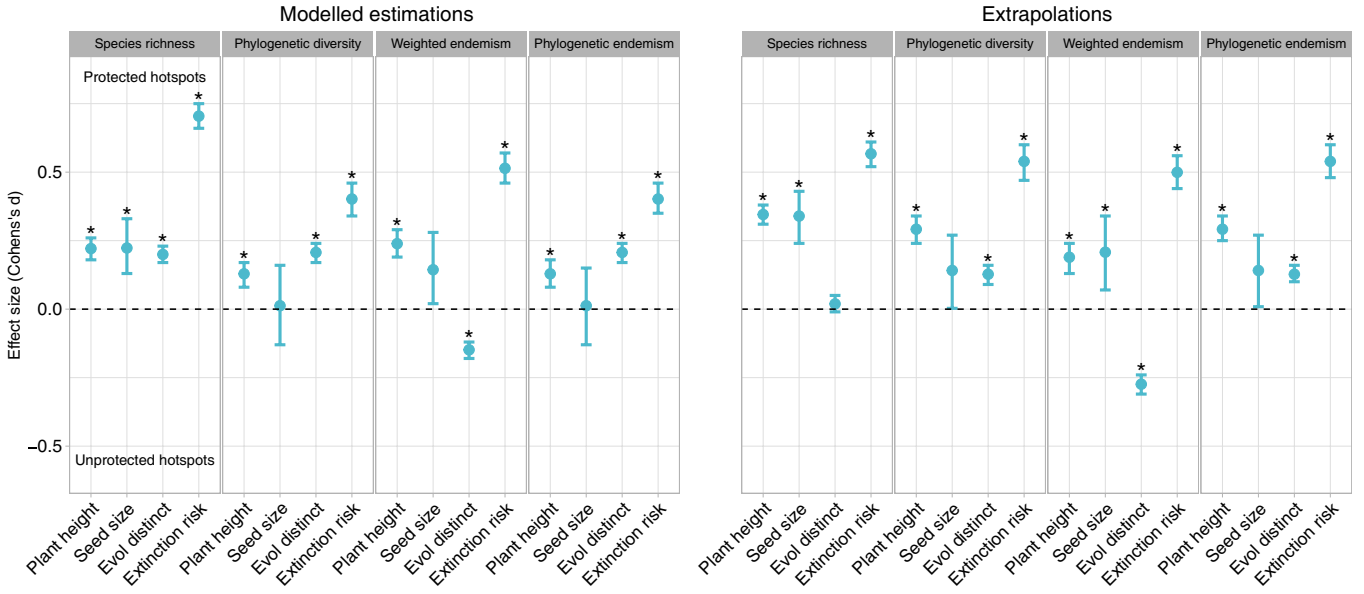

Fig. 5. Coverage of species-level attributes within protected and unprotected diversity hotspots of vascular plants. Differences in species attributes (functional traits, evolutionary distinctiveness, and extinction risk) for diversity hotspots were based on modeled estimations (left) and random forest extrapolations (right) computed using *t* test followed by Cohen's *d* with 1000 bootstrap replicates to estimate effect size. Values of Cohen's *d* range from 0 (no effect) to +1 or −1 (large effect), with positive values (above the dashed 0 line) indicating that the attribute is higher in protected hotspots, whereas negative values (below the zero line) indicate the opposite. The error bars indicate 95% CI, and the statistical significance of the *t* test is indicated with asterisks ($P < 0.01$).

points. The models were calibrated to species' realized niche defined using species-specific dispersal rates fitted using a phylogeographic Spherical Brownian Motion (64, 65). The data analysis was performed in four steps: 1) I calculated key vascular plant diversity patterns including species richness, phylogenetic diversity, weighted endemism, and phylogenetic endemism, 2) I tested my modeled estimates for the effects of sampling biases common in plant occurrence records by training a random forest model to extrapolate my modeled estimates under a globally unbiased plant sampling. 3) I conducted spatially corrected correlations between plant diversity patterns and those of the different tetrapod classes for each metric. 4) I compared predicted locations of plant hotspots (both modeled estimations and random forest extrapolations) within protected areas to their counterparts outside protected areas. A full discussion of *Materials and Methods* can be found in *SI Appendix, Materials and Methods*.

Data, Materials, and Software Availability. Geographic data, R package and codes have been deposited in DRYAD (https://doi.org/10.5061/dryad.5x69p8d9w) (66). I used the R codes from ref. 28 to conduct random forest extrapolations. All other analysis functions are described directly in the article and/or supporting information.

ACKNOWLEDGMENTS. I thank Stanford University for logistic support. Special thanks to M. W. Feldman, P. R. Ehrlich, C. P. Gross, and three anonymous reviewers for valuable comments and discussion during the early development of this paper. This study was supported by the US NSF (awards 2031928 and 2416314).

1. E. N. Lughadha *et al.*, Counting counts: Revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* **272**, 82–88 (2016).
2. K. J. Willis, *The State of the World's Plants 2017 Report*, R. B. Gardens, Ed. (Royal Botanic Gardens, Kew, UK, 2017).
3. R. Govaerts, E. Nic Lughadha, N. Black, R. Turner, A. Paton, The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Sci. Dat.* **8**, 215 (2021).
4. G. C. Daily, *Nature's Services: Societal Dependence on Natural Ecosystems* (Island Press, 1997).
5. A. Antonelli *et al.*, Why plant diversity and distribution matter. *New Phytol.* **240**, 1331–1336 (2023).
6. B. J. Enquist, R. Condit, R. K. Peet, M. Schildhauer, B. M. Thiers, Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints* **4**, e2615v2 (2016).
7. J. L. Edwards, M. A. Lane, E. S. Nielsen, Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* **289**, 2312–2314 (2000).
8. G. M. Moulatlet *et al.*, Global patterns of phylogenetic beta-diversity components in angiosperms. *J. Veg. Sci.* **34**, e13203 (2023).
9. S. Díaz *et al.*, The global spectrum of plant form and function. *Nature* **529**, 167–171 (2016).
10. B. H. Daru *et al.*, Widespread homogenization of plant communities in the Anthropocene. *Nat. Commun.* **12**, 6983 (2021).
11. A. Carta, L. Peruzzi, S. Ramírez-Barahona, A global phylogenetic regionalization of vascular plants reveals a deep split between Gondwanan and Laurasian biotas. *New Phytol.* **233**, 1494–1504 (2022).
12. Y. Liu *et al.*, An updated floristic map of the world. *Nat. Commun.* **14**, 2990 (2023).
13. S. Gaiji *et al.*, Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodivers. Inform.* **8**, 94–172 (2013).
14. C. Meyer, P. Weigelt, H. Kreft, Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
15. B. H. Daru *et al.*, Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytol.* **217**, 939–955 (2018).
16. C. König *et al.*, Biodiversity data integration–the significance of data resolution and domain. *PLoS Biol.* **17**, e3000183 (2019).
17. J. Stropp *et al.*, Mapping ignorance: 300 years of collecting flowering plants in Africa. *Glob. Ecol. Biogeogr.* **25**, 1085–1096 (2016).
18. B. H. Daru, J. Rodriguez, Mass production of unvouchered records fails to represent global biodiversity patterns. *Nat. Ecol. Evol.* **7**, 816–831 (2023).
19. S. Reddy, L. M. Dávalos, Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30**, 1719–1727 (2003).
20. R. P. Anderson, Real vs. artefactual absences in species distributions: Tests for Oryzomys albigularis (Rodentia: Muridae) in Venezuela. *J. Biogeogr.* **30**, 591–605 (2003).
21. H. Hillebrand, On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211 (2004).
22. A. Guisan, W. Thuiller, Predicting species distribution: Offering more than simple habitat models. *Ecol. Lett.* **8**, 993–1009 (2005).
23. T. A. Reydon, Why does the species problem still persist? *BioEssays* **26**, 300–305 (2004).
24. J. D. Olden, J. J. Lawler, N. L. Poff, Machine learning methods without tears: A primer for ecologists. *Q. Rev. Biol.* **83**, 171–193 (2008).
25. C. Crisci, B. Ghattas, G. Perera, A review of supervised machine learning algorithms and their applications to ecological data. *Ecol. Modell.* **240**, 113–122 (2012).
26. J. Troudet *et al.*, Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* **7**, 9132 (2017).
27. D. S. Park, Y. Xie, H. T. Thammavong, R. Tulaiha, X. Feng, Artificial hotspot occurrence inventory (AHOI). *J. Biogeogr.* **50**, 441–449 (2023).
28. J. M. Kass *et al.*, The global distribution of known and undiscovered ant biodiversity. *Sci. Adv.* **8**, eabp9908 (2022).
29. A. Zizka *et al.*, Biogeography and conservation status of the pineapple family (Bromeliaceae). *Divers. Distrib.* **26**, 183–195 (2020).
30. L. Cai *et al.*, Global models and predictions of plant diversity based on advanced machine learning techniques. *New Phytol.* **237**, 1432–1445 (2023).
31. R. K. Brummitt, World geographical scheme for recording plant distributions, 2nd ed (Biodiversity Information Standards (TDWG), 2001, http://www.tdwg.org/standards/109). Accessed 23 July 2022.
32. P. Weigelt, C. König, H. Kreft, GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography. *J. Biogeogr.* **47**, 16–43 (2020).
33. H. Kreft, W. Jetz, Global patterns and determinants of vascular plant diversity. *Proc. Natl Acad. Sci. U.S.A.* **104**, 5925–5930 (2007).
34. B. J. Enquist *et al.*, The commonness of rarity: Global and future distribution of rarity across land plants. *Sci. Adv.* **5**, eaaz0414 (2019).
35. P. Keil, J. M. Chase, Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nat. Ecol. Evol.* **3**, 390–399 (2019).
36. F. M. Sabatini *et al.*, Global patterns of vascular plant alpha diversity. *Nat. Commun.* **13**, 4683 (2022).

37. D. P. Faith, Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992).
38. D. Rosauer, S. W. Laffan, M. D. Crisp, S. C. Donnellan, L. G. Cook, Phylogenetic endemism: A new approach for identifying geographical concentrations of evolutionary history. *Mol. Ecol.* **18**, 4061–4072 (2009).
39. S. Veron *et al.*, High evolutionary and functional distinctiveness of endemic monocots in world islands. *Biodivers. Conserv.* **30**, 3697–3715 (2021).
40. W. D. Kissling *et al.*, Cenozoic imprints on the phylogenetic structure of palm species assemblages worldwide. *Proc. Natl Acad. Sci. U.S.A.* **109**, 7379–7384 (2012).
41. W. L. Eiserhardt *et al.*, Dispersal and niche evolution jointly shape the geographic turnover of phylogenetic clades across continents. *Sci. Rep.* **3**, 1164 (2013).
42. R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, J. Elith, Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecol. Monogr.* **92**, e01486 (2022).
43. D. M. Olson *et al.*, Terrestrial ecoregions of the world: A new map of life on Earth. *Bioscience* **51**, 933–938 (2001).
44. L. Naughton-Treves, M. B. Holland, K. Brandon, The role of protected areas in conserving biodiversity and sustaining local livelihoods. *Annu. Rev. Environ. Resour.* **30**, 219–252 (2005).
45. O. Venter *et al.*, Targeting global protected area expansion for imperiled biodiversity. *PLoS Biol.* **12**, e1001891 (2014).
46. J. E. M. Watson, N. Dudley, D. B. Segan, M. Hockings, The performance and potential of protected areas. *Nature* **515**, 67–73 (2014).
47. U. Roll *et al.*, The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nat. Ecol. Evol.* **1**, 1677–1682 (2017).
48. Convention on Biological Diversity, Conference of the Parties to the Convention on Biological Diversity (Fifteenth meeting – Part II, Montreal, Canada, 2022). https://www.cbd.int/doc/decisions/cop-15/cop-15-dec-04-en.pdf. Accessed 10 October 2023.
49. L. O. Frishkoff *et al.*, Loss of avian phylogenetic diversity in neotropical agricultural systems. *Science* **345**, 1343–1346 (2014).
50. A. J. Nowakowski, L. O. Frishkoff, M. E. Thompson, T. M. Smith, B. D. Todd, Phylogenetic homogenization of amphibian assemblages in human-altered habitats across the globe. *Proc. Natl Acad. Sci. U.S.A.* **115**, E3454–E3462 (2018).
51. C. M. Roberts *et al.*, Ecological criteria for evaluating candidate sites for marine reserves. *Ecol. Appl.* **13**, 199–214 (2003).
52. J. A. Nowakowski *et al.*, Protected areas slow declines unevenly across the tetrapod tree of life. *Nature* **622**, 101–106 (2023).
53. T. A. Pelletier *et al.*, Predicting plant conservation priorities on a global scale. *Proc. Natl Acad. Sci. U.S.A.* **115**, 13027–13032 (2018).
54. United States Forest Service, Forest Service Roadless Area Conservation Final Environmental Impact Statement (US Forest Service, Washington, D.C., 2000), vol. 1.
55. K. Bishop, N. Dudley, A. Phillips, S. Stolton, Speaking a Common Language: Uses and Performance of the IUCN System of Management Categories for Protected Areas (Cardiff University, The World Conservation Union & World Conservation Monitoring Centre, UK, 2004).
56. H. Locke, P. Dearden, Rethinking protected area categories and the new paradigm. *Environ. Conserv.* **32**, 1–10 (2005).
57. M. W. Cadotte, T. J. Davies, Rarest of the rare: Advances in combining evolutionary distinctiveness and scarcity to inform conservation at biogeographical scales. *Divers. Distrib.* **16**, 376–385 (2010).
58. S. Pironon *et al.*, The global distribution of plants used by humans. *Science* **383**, 293–297 (2024).
59. B. Mishler *et al.*, Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian Acacia. *Nat. Commun.* **5**, 4473 (2014).
60. B. H. Daru, M. Van der Bank, T. J. Davies, Spatial incongruence among hotspots and complementary areas of tree diversity in southern Africa. *Divers. Distrib.* **21**, 769–780 (2015).
61. L. M. Lu *et al.*, Evolutionary history of the angiosperm flora of China. *Nature* **554**, 234–238 (2018).
62. V. H. Heywood, *Flowering Plants of the World* (Batsford, 1993).
63. A. P. G. Iv, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
64. D. R. Brillinger, "A particle migrating randomly on a sphere" in *Selected works of David Brillinger*. (Springer, 2012), pp. 73–87.
65. S. Louca, Phylogeographic estimation and simulation of global diffusive dispersal. *Syst. Biol.* **70**, 340–359 (2021).
66. B. H. Daru, Predicting undetected native vascular plant diversity at a global scale [Dataset]. Dryad. https://doi.org/10.5061/dryad.5x69p8d9w. Deposited 16 July 2024.