

Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge

Laura J. Pollock^{1,2,17}✉, Justin Kitzes^{3,17}, Sara Beery⁴, Kaitlyn M. Gaynor^{5,6}, Marta A. Jarzyna^{7,8}, Oisín Mac Aodha⁹, Bernd Meyer¹⁰, David Rolnick^{2,11}, Graham W. Taylor^{12,13}, Devis Tuia¹⁴ & Tanya Berger-Wolf^{7,8,15,16}

Abstract

Large, well described gaps exist in both what we know and what we need to know to address the biodiversity crisis. Artificial intelligence (AI) offers new potential for filling these knowledge gaps, but where the biggest and most influential gains could be made remains unclear. To date, biodiversity-related uses of AI have largely focused on tracking and monitoring of wildlife populations. Rapid progress is being made in the use of AI to build phylogenetic trees and species distribution models. However, AI also has considerable unrealized potential in the re-evaluation of important ecological questions, especially those that require the integration of disparate and inherently complex data types, such as images, video, text, audio and DNA. This Review describes the current and potential future use of AI to address seven clearly defined shortfalls in biodiversity knowledge. Recommended steps for AI-based improvements include the re-use of existing image data and the development of novel paradigms, including the collaborative generation of new testable hypotheses. The resulting expansion of biodiversity knowledge could lead to science spanning from genes to ecosystems – advances that might represent our best hope for meeting the rapidly approaching 2030 targets of the Global Biodiversity Framework.

Sections

Introduction

Taxonomic descriptions

Estimates and patterns of abundance

Biogeographic species distribution

Abiotic tolerance and fundamental niche

Functional trait variation

Evolutionary relationships

Species interactions

Summary and future directions

A full list of affiliations appears at the end of the paper. ✉e-mail: laura.pollock@mcgill.ca

Introduction

Biodiversity is essential for human well-being, yet is increasingly threatened. Biodiversity is also complex, scale-dependent, hard to measure and full of surprises. Unlike the relatively simple causal link between atmospheric greenhouse gas concentration and climate change, the biodiversity story is more nuanced. The grand challenge in ecology and conservation is to be able to answer the following crucial questions. How many species do we have on Earth? Which populations are declining? Which areas are essential to protect? When will tipping points be reached? How should we best meet the 2030 global biodiversity targets set out in the Kunming–Montréal Global Biodiversity Framework (GBF; <https://www.cbd.int/gbf>)¹? Why do contemporary extinctions exceed background rates? Adequate answers to these and other questions are needed to capitalize on the current global momentum toward nature conservation.

Unfortunately, despite large volumes of data being collected, nearly all GBF targets and indicators are missing essential information, which is needed both to establish baselines and to monitor progress. Persistent biases dating from the 1980s have led to conservation efforts being focused repeatedly on the same taxa, which (counterintuitively) are not always those with the greatest levels of risk². Overall, biodiversity is well described in only a small fraction of the world³ and existing data are biased toward common species and populated areas in the Northern Hemisphere^{4,5}. Little is known for many species beyond their names and where they live. Information on how a particular species functions, how it speciated and how it interacts in communities is often absent, especially for species occurring in the ocean². These knowledge deficits span taxonomy to species interactions and have been organized into seven defined shortfalls in global biodiversity knowledge^{6,7} that capture the breadth and complexity of biodiversity in measurable ways. Overcoming these shortfalls is essential for calculating essential biodiversity variables⁸, meeting all biodiversity-based GBF indicators and addressing the most pressing challenges to biodiversity, which range from obtaining detailed on-the-ground knowledge to understanding national trends.

Ecologists and conservation researchers need to harness the current unprecedented levels of interest and global coordination in biodiversity protection generated by the GBF alongside emerging technologies, such as artificial intelligence (AI) and, more specifically, data-driven machine learning (ML), that can handle diverse and rapidly expanding datasets. An open question is how we can best leverage these technologies. To date, the rapid rise of AI technologies and methods in ecology and evolution has been mostly focused on a small set of conservation topics (reviewed elsewhere^{9–11}) and data-collection applications, such as bioacoustics¹², camera traps^{13,14}, satellite imagery and remote sensing¹⁵. Reviews of AI applications in biodiversity loss¹⁶ and AI methods for ecologists¹⁷ have already been published. The present Review considers how AI could address critical knowledge gaps in the broader fields of biodiversity science, which span spatial scales, genes, functions, phylogenies and species interactions. We note that owing to the extremely rapid development of AI, many of the publications cited in this Review are currently available only in non-peer-reviewed formats (conference proceedings or preprints).

In this Review, we delineate the current state of the seven biodiversity shortfalls^{7,18}, discuss how they are being addressed with AI and identify where AI offers the greatest potential for bridging the remaining gaps (Fig. 1). Where AI methods have not yet been used to address all seven biodiversity shortfalls, we recommend avenues of investigation that map the needs of each shortfall to specific AI solutions and place each solution in the context of the required steps, from

data imputation and analysis to conservation decisions. Finally, we provide a critical analysis of the realistic limitations of AI technologies; although the proliferation of foundation models (including molecular¹⁹, cellular^{20,21} and organismal²² models) coupled with generative AI holds promise for reducing all biodiversity shortfalls, AI is not the answer to every challenge.

Taxonomic descriptions

The Linnaean shortfall refers to the gap between the number of species on Earth and the number that have been formally described⁷. This shortfall is arguably the most foundational, as nearly every field in ecology, evolution and conservation relies on naming and cataloguing species to assess biodiversity. Nothing can be done to understand a species that is not known to exist.

Challenges

The size of the Linnaean shortfall is unknown, because estimating the size of this gap relies on counting the number of species that have been described and estimating the number of non-described species, both of which involve uncertainty. Approximately 2 million extant species have been described²³ of the estimated 8.7 million eukaryotic species thought to exist on Earth²⁴, although estimates vary widely²⁵. In general, the size of the Linnaean shortfall is thought to increase as the size of an organism and its complexity decreases²⁶ and to vary geographically along with a variety of other traits^{7,27,28} (see the Wallacean shortfall). Despite the fundamental nature of the Linnaean shortfall and the currently heightened global risk of species extinction, the number of newly named species has not increased since the 2000s². Indeed, taxonomy is a severely underfunded scientific discipline that is itself threatened with extinction²⁹.

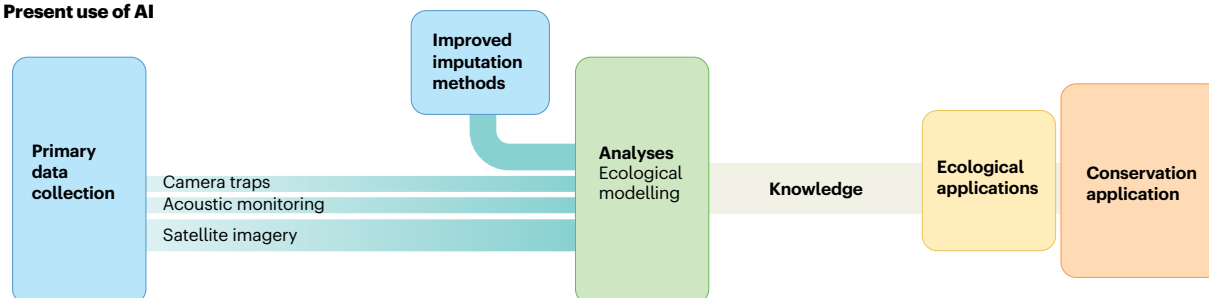
Past and future role of AI

AI has so far been used to mitigate the Linnaean shortfall in two major ways: to impute the estimated total number of taxa³⁰ and to identify new taxa in existing datasets (Fig. 1). Automated identification of previously undescribed taxa is a very promising task for AI that might be accomplished through identification of new taxa in existing images²², DNA samples³¹ or acoustic recordings and leveraging of modalities (such as DNA or acoustic analysis) that can indicate the presence of taxa as yet unseen³².

Novel taxa have been identified in raw sensor data, including via DNA barcodes^{31,33} and citizen science images^{22,34}. The image-classification models BioCLIP²² and BIOSCAN-CLIP³¹ are not designed to pinpoint new species, but can be used to label putative examples of new species by association with known template images or DNA sequences, respectively. WildCLIP focuses on the retrieval of images displaying certain attributes of an animal or its environment, which could be used to interrogate diverse datasets³⁵. Although such approaches are currently in the early stages of implementation, they offer considerable potential for incorporating techniques from the ML subareas of open world classification and category discovery³², which involve the identification of new categories (such as species) in unlabelled datasets (such as image libraries) that could contain both previously known and as-yet-undescribed categories.

Looking ahead, AI tools might contribute to tasks designed specifically to enable species discovery. For example, once a new species is identified (either manually or via AI-assisted discovery methods³²), AI vision–language models might be able to assist taxonomists in crafting a species description by picking out and describing its distinguishing features^{34,36,37}. Such methods could draw upon interpretable AI

a Present use of AI



b Future use of AI

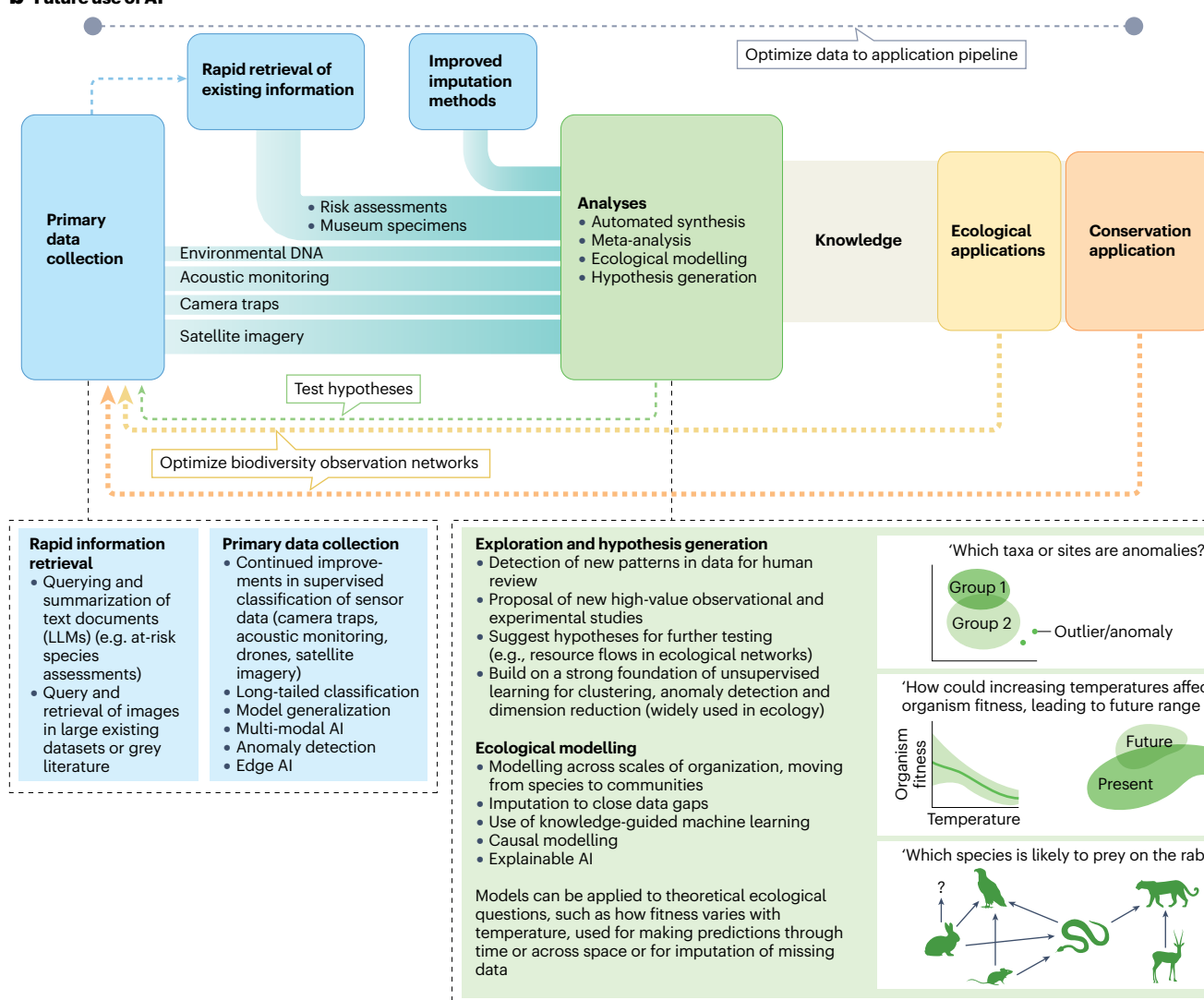


Fig. 1 | Potential roles of artificial intelligence in filling biodiversity knowledge gaps and downstream applications.

a, Artificial intelligence (AI) is widely implemented in the data–decision pipeline for conservation applications (those relating to species of conservation concern) but is less often integrated into the broader subfields of ecology. In consequence, most biodiversity shortfalls remain relatively unexplored. The increasing emphasis on satellite imagery and imputation methods, which generate large datasets and involve statistical

modelling, is likely to drive further uses of AI. **b**, The future development of AI could help to fill knowledge gaps in several areas. These improvements generally apply across multiple biodiversity shortfalls, although some tasks benefit more than others from specific improvements. The size of the boxes and the width of the connecting lines represent the relative importance of the contributions of AI to the data-generation and conservation-application processes. LLM, large language model.

techniques integrated with species-detection algorithms such as those used in BioCLIP or BIOSCAN-CLIP. Other algorithms might be used to recommend where, when and how to search for novel species: for example, approaches inspired by active learning could identify areas where there is high uncertainty in the diversity of certain taxa. This process results in an active learning feedback loop (Fig. 1), in which hypotheses proposed by both humans and algorithms are validated and the results are used to train further algorithms with improved performance. Such active learning processes are predicted to be instrumental in the design of future global monitoring networks³⁸. Contributions from AI might, therefore, not only increase the rate of species discovery and description but also the effectiveness of current and future generations of taxonomists.

Estimates and patterns of abundance

The Prestonian shortfall refers to the lack of knowledge of the abundance of a species and its trends in both space and time. Aside from the fundamental importance of accurate estimates of species abundance to population biology and evolution, knowledge of abundance is also critical to defining a species' conservation status and predicting its risk of extinction. However, addressing the Prestonian shortfall represents a considerable challenge because it involves counting (or estimating) the number of all individuals of a species present at a given point in space and time.

Challenges

From a data perspective, measuring the true abundance of a given population requires an exhaustive census of every individual of the relevant species present within a defined spatial and temporal window. This laborious exercise is very rarely completed even for relatively charismatic, easily detected and well studied taxa such as birds³⁹ or large mammals. As a notable exception, the Center for Tropical Forest Science (CTFS) Forest Global Earth Observatory (GEO) has established forest plots specifically for measuring the true abundance of tropical tree species⁴⁰. However, even for stationary and relatively well described tree species, conducting such censuses requires an enormous amount of effort.

A practical alternative to a complete census is to estimate both species abundance and its spatiotemporal trends from a statistically representative sample of individuals drawn from the entire population. The essential challenge of this approach is to accurately estimate the number of individuals that are present in the population but not sampled, which requires repeated sampling in either space or time. Two broad categories of population estimation model exist: marked, in which specific individuals are tagged or can otherwise be identified when re-sighted in repeated surveys; and unmarked, in which individuals are counted but cannot be identified as a specific (re-sighted) individual. The relative abundances of multiple species or trends in relative abundance among species might be sufficient for some applications and can often be inferred from count data alone^{18,39}.

Notably, although the explosion in citizen science data and the availability of digitized museum specimens have made important contributions to documenting the presence of species across their ranges (as discussed in the section on the Wallacean shortfall), these unmarked data sources usually do not measure abundance directly and might be too sparse in space and time to support robust estimates of abundance.

Past and future role of AI

To date, AI-based analysis of sensor data (such as camera trap images and acoustic recordings) has been applied to mitigating the Prestonian

shortfall by generating unmarked data for statistical abundance estimation^{41,42}. For example, occupancy estimates derived from repeated sampling⁴³ and time-to-detection models⁴⁴ that are based on the automated classification of bird song produce species abundance estimates similar to those derived from traditional human surveys⁴⁵. Automated classification can also be combined with grids of sensors, such as acoustic recorders, to produce high-resolution maps of sound sources that can be used to estimate the abundance of sound-producing species^{46,47}.

AI is already reducing the Prestonian shortfall for many species by increasing the efficiency of human experts in identifying specific individuals in collected images, which facilitates the non-invasive marked estimation of species abundance⁴⁸. This work started with the publication in 1990 of the first methods of re-identification based on computer vision⁴⁹. Early attempts involved the use of statistical pattern recognition derived from low-level features and geometry^{14,50,51}. Advances in person re-identification using deep learning⁵² have now been applied to images of animals of various species^{53–55} and have led to notable improvements in re-identification, particularly when high-quality, well focused images of single individuals taken by human experts were assessed.

AI methods that can identify individual organisms, not just their species identity, are poised to contribute even further to mitigating the Prestonian shortfall by enabling data obtained from passive sensors to be used in marked abundance models. However, much work remains to be done to increase the resilience of computer-vision-based re-identification to poor-quality images, for example those captured without a human photographer (such as in camera-trap data^{14,56}). In addition, these methods could be expanded to include many more data modalities, such as video^{57–59}, drone or unmanned aerial vehicle recordings⁶⁰ and audio files^{61,62}. AI applications for wildlife population monitoring (such as Wildbook; <https://www.wildme.org/wildbook.html>) would also benefit from progress on the 'open-set challenge', which involves not only identifying confidently that an individual has never before been seen⁶³ but also recognizing and matching sightings of new individuals over time⁶⁴. We anticipate that progress will also be made on the incorporation of experts into participatory and iterative human–AI systems to reduce the amount of expert input needed to derive abundance estimates⁶⁵ and to improve resistance to category errors, such as assigning images of multiple individuals to a single identification or splitting images of a single individual into multiple identifications.

In the long term, considerable development work is needed to enable computer vision systems to efficiently recognize individuals without clear biometric markings and to recognize the same individual as their markings change over time. Such improvements might involve the recognition of specific individuals from their behaviour, gait or vocalizations, or the integration of additional modalities, such as hyperspectral imaging or environmental DNA (eDNA) analyses. Additional biological attributes and contextual information derived from mechanistic studies of species behaviours, such as social interactions, demographic status and territories, could also be integrated into re-identification systems. Additionally, interdisciplinary research on statistical estimates of abundance is needed, for example, to develop estimation methods that take into account the continuous-value confidence scores of AI models⁴⁷ or make use of mixed-granularity species identification data, in which some sightings can be identified at the individual level, whereas others can only be confidently identified for a small subset of the population.

In contrast to the several decades of experience in the application of AI-based analytic methods to image databases, the application of these methods to acoustic⁶¹ and other data types collected in the field remains limited and emerging. Progress in computer vision could lead to the development of improved methods for the aerial census of large groups, which could build upon existing methods used to accurately count the number of individuals in crowds^{66,67}. The growing availability of eDNA data, particularly those obtained using field-deployable sequencing technologies, offers another potential avenue for species abundance estimation. However, the conversion of eDNA concentrations into reliable abundance estimates remains challenging, owing to the presence of complex environmental factors that affect DNA persistence and detection⁶⁸. AI approaches could make estimates derived from eDNA concentrations more robust by accounting for these environmental factors and by integrating eDNA data collected using other sensor types. Finally, we note that the AI-supported species abundance estimation methods described in this section have generally provided single snapshots in time. Additional progress towards the use of AI for forecasting time series and understanding drivers of population change, including through process-based and knowledge-guided ML models of population dynamics^{69,70}, is expected to contribute further towards reducing this shortfall.

Biogeographic species distribution

The Wallacean shortfall refers to the lack of detailed information on the biogeographic distribution of species. The documentation of species distributions dates from the early 1800s and, as such, is one of the oldest endeavours in biodiversity science. Moreover, the Wallacean shortfall affects nearly every subfield of ecology, including understanding the effects of climate change on biodiversity and the reconstruction of historical speciation events. Accurate data on species biogeographic distributions is essential for conservation because range size is one of the best predictors of extinction risk^{71–73} and is one of the main criteria for assignment of at-risk status⁷⁴. Knowledge of biogeographic species distributions is also essential for mapping species, biodiversity hotspots and ecosystem services. This information directly feeds several important biodiversity indicators, such as the Species Protection Index⁷⁵.

Challenges

Addressing the Wallacean shortfall is relatively simple in principle, in that it can be filled by simple occurrence data, which are increasingly drawn from crowdsourced initiatives such as iNaturalist⁷⁶. However, although the Global Biodiversity Information Facility (GBIF; <https://www.gbif.org/>) now contains over 3 billion records, these data are biased toward terrestrial areas², certain taxa (especially popular birds), the Northern Hemisphere and locations within 1.0 km of roads⁷⁷. Expert species-range maps derived from taxa-specific sources (including conservation guides and at-risk assessments) also provide distribution information for many taxa, but these often lack the granularity needed for use in conservation applications, such as the estimation of species–habitat relationships.

Past and future role of AI

One of the most promising ways in which AI can be used to fill the Wallacean shortfall is in the processing of incoming primary data collected from sensor arrays (Fig. 1). Technologies such as high-resolution satellite and aerial remote sensing in a wide range of spectra, stationary image capture, acoustic sensing and eDNA analysis are increasingly

being used to provide species data for sparsely covered and/or inaccessible locations. These innovations are already resulting in surprising discoveries, such as new colonies of emperor penguins⁷⁸ and advances such as the ability to monitor whales in remote locations⁷⁹. Species occurrences can also be extracted from non-target acoustic recordings, images on crowd science platforms (such as iNaturalist) and social media posts⁸⁰.

A second major and well integrated effort to fill the Wallacean shortfall has involved the use of species distribution models (SDMs) to impute missing data. SDMs used to predict a species' spatial distribution from environmental or habitat data have rapidly adopted ML techniques such as boosted regression trees⁸¹. Modern SDMs are now beginning to incorporate more-powerful ML techniques that can handle complex interactions with multiple data types, such as remotely sensed land-cover classes and continuous local climate measurements⁸². AI-based statistical model integration is an even more powerful tool⁸³ for the analysis of multimodal datasets that can handle multiple types of species occurrence data – for example, presence-only community science, presence–absence plot data and remote sensing imagery (Box 1). The development of standardized protocols and competitions such as GeoLifeCLEF (one of several challenges in ImageCLEF (<https://www.imageclef.org/>), the Conference and Labs of the Evaluation Forum (CLEF) cross language image retrieval track)⁸⁴ are extending the success of AI-assisted SDMs to other macroecological models used to predict biodiversity metrics, such as species richness⁸⁵.

However, many challenges remain to be addressed to enable meaningful gains to be made in primary data collection and data synthesis. AI-assisted methods offer the potential to decrease known data gaps by targeting severely under-sampled areas (such as the deep ocean) and taxa (such as fungi) via active learning methods that optimize future data acquisition from in situ sensor networks and community scientists^{86,87}. Extreme edge computational approaches increasingly move AI to the sensors themselves, in the form of smart camera traps and acoustic arrays that enable automated and adaptive data collection. Advanced techniques are also needed for spatial bias correction^{38,88}, to improve models of undersampled species that are based on well sampled species^{89,90}, and to model community assemblages and turnover⁹¹. The development of these methods requires cross-disciplinary collaborations of ecological statisticians and AI researchers⁹². Although multimodal datasets (for example, those that integrate camera trap and community science data) are already proving useful⁸³, future work could address the computational challenges associated with the analysis of large remote-sensing datasets by making use of models that efficiently encode spatially varying data representations, which are useful for many downstream tasks^{93,94}. Finally, where possible, the utility of AI methods for estimating spatial distributions of species should be rigorously quantified using best-in-class, expert-verified evaluation datasets. The creation of such 'gold standard' datasets is challenging but deserves attention from the biodiversity community^{95,96}.

Abiotic tolerance and fundamental niche

The Hutchinsonian shortfall refers to a gap in understanding of the tolerance of a species to abiotic conditions, including temperature, precipitation, soil, water and terrain. This suite of abiotic tolerances is often referred to as the Grinnellian fundamental niche, which is considered to be the multidimensional environmental space in which a species can persist^{97,98}. Knowledge of abiotic tolerances is particularly important in the context of climate change because rapidly changing abiotic

Box 1 | The use of imagery to fill biodiversity knowledge shortfalls

One of the biggest contributors to meeting the Wallacean and Hutchinsonian shortfalls (among others) has been the ever-increasing detail supplied by satellite imagery. The characterization of Earth's topography and terrain is relatively mature. National Aeronautics and Space Administration (NASA)- and European Space Agency (ESA)-based missions have provided planet-scale terrain models at resolutions of tens of metres. However, AI can still contribute to the assessment of other habitat variables. For example, global gridded soil data (SoilGrids250)¹⁹⁵ are based on traditional machine learning (ML) approaches such as random forests. These models were trained on publicly available soil profiles and spectral data in which reflectance relates to indices of vegetation, clouds or water vapour that populate a global datacube¹⁹⁶ and can be used, together with local observations, to characterize a habitat¹⁹⁷. Although historic indices were based on simple band ratios, the need for modelling of more-complex relationships or more-targeted variables has encouraged the AI community to provide these essential variables and thereby facilitate monitoring of the planet's climate¹⁹⁸ and biodiversity^{103,199}.

Advances in deep learning mean that new indices can be produced at high resolution and global scales. Tree height and above-ground biomass²⁰⁰, forest degradation²⁰¹, drivers of deforestation-related phenological factors such as plant functional types²⁰² and tree wood density²⁰³ are some important examples. However, most AI-based approaches require globally distributed ground observations to train the models, sometimes collected by missions such as the Global Ecosystem Dynamics Investigation (GEDI), which provided global tree-height ground points²⁰⁰. Thus, the development of AI-assisted models often requires extensive fieldwork collections and could potentially also benefit from citizen science²⁰⁴. Technical and economic challenges that remain to be solved to make the most of the immense observational power of satellite data include the integration of multiple sensor types (such as optical and radar^{205,206}), the limited spatial resolution of sensors, the high cost of large data volumes, and high computational costs. Multispectral or hyperspectral drone or airborne data can be used for smaller regions.

conditions are driving mismatches between species occurrences and tolerances that pose a threat to conservation efforts. Mitigation of the Hutchinsonian shortfall could improve predictions of population trajectories and range shifts under climate change⁹⁹.

Challenges

Information on species tolerances comes from two primary sources: physiological data derived from study of the organism and occurrence data from field observations. Physiological data collected about organisms in field or laboratory experimental settings can be used to generate models that generate performance curves for a given organism under different abiotic conditions. These data are challenging to obtain and exist only for a subset of species¹⁰⁰. Additionally, such studies often fail to account for environmental factors that affect performance under field conditions and rarely capture intraspecific variations

(see Raunkiaeran shortfall) or consider phenotypic plasticity¹⁰¹. Alternatively, occurrence data from field observations of an organism can be linked to data on prevalent abiotic conditions to characterize a species' realized niche.

The differentiation of fundamental niches (the potential for toleration of extant conditions) from realized niches (environments where an organism is actually found) remains a challenge, given that biotic factors and dispersal capabilities also shape species distributions¹⁰². A further challenge is the lack of data on the abiotic environment, particularly at the fine spatial scales needed to precisely quantify the abiotic tolerances of organisms.

Past and future role of AI

AI is an obvious candidate for inferring an organism's tolerances from field observations and occurrence data. Increasingly finer-resolution imagery from satellites (Box 1) and drones, thermal imaging data and precise light detection and ranging (LiDAR)-generated three-dimensional (3D) models of surface features can be the base ingredients for extracting a highly refined picture of species habitats or proxies that can be used to estimate such habitats. Although such raw data do not directly inform scientists about habitats, they can be used to extract the required information about phenology, soils and climate. Some sensors track organisms directly. Animal tracking, although particularly relevant to the Hutchinsonian shortfall, is also relevant to the Wallacean and other shortfalls.

Looking forward, alongside the advances in satellite-derived imagery used to identify habitat associations (Box 1), ground-to-space systems such as International Cooperation for Animal Research Using Space (ICARUS) could greatly enhance our understanding of abiotic tolerances and detailed behavioural responses to abiotic conditions (reviewed elsewhere¹⁰³). For plants, abiotic conditions could potentially be harvested from photographs (also discussed in relation to the Raunkiaeran shortfall). However, laboratory-based experimental work is needed to complement such observational data. AI could be helpful in building 3D models of wingbeats for birds using structured light and high-speed cameras¹⁰⁴, other detailed animal tracking movements^{105,106} or scent-based molecule detection¹⁰⁷ (which could be used to detect stress hormones in laboratory animals, for example). Looking to the future, the creation of accurate 'digital twin' ecological system models^{108–110} will require abiotic tolerance data derived from these AI methods as well as process-based AI approaches that combine abiotic tolerance information with additional observations to track and predict biodiversity change.

Functional trait variation

The Raunkiaeran shortfall highlights the lack of knowledge of both intraspecific and interspecific trait variations, the ecological functions arising from species traits, how these functions are influenced by interactions with other traits, and which traits act in tandem to provide specific ecosystem functions.

Challenges

Obtaining the true distributions of trait values within a given population requires a comprehensive assessment of every individual's traits to be conducted within a specified time frame, which is logistically impossible. For that reason, researchers have largely focused on compiling species mean values derived from measurements of museum specimens or individuals in the field as well as extraction of trait data from guidebooks and the scientific literature. These efforts led to the

development of large, publicly available trait databases^{111–115}. Efforts over the past 15 years have focused on intraspecific trait variability^{112,116} and the potential for some traits (such as diet and foraging characteristics) to show substantial spatiotemporal variation even within an individual¹¹⁷. However, considerable taxonomic gaps remain; most trait databases focus on plants¹¹⁶ or tetrapods^{111–113,115} and only a handful include invertebrates^{114,118}, fungi and microorganisms^{118,119}, for which comprehensive taxonomic coverage is often lacking. Geographic coverage is also uneven; regions in the Global South and ecosystems such as tropical forests and deep oceans typically lack comprehensive trait data. Finally, our knowledge of traits often centres on easily measurable morphological characteristics, which are only sometimes clearly linked to ecological functions^{120,121}. These factors limit our understanding of trait–environment relationships. Comprehensive data are urgently needed on behavioural, physiological and life-history traits, which are crucial for understanding species' ecological roles and interactions.

Past and future role of AI

Current AI techniques to address the Raunkiaeran shortfall typically focus on extracting trait information from digitized museum specimens¹²² and images provided by citizen scientists through the burgeoning field of imageomics¹²³. Most efforts have concentrated on easily measurable (often morphological) traits, but more-complex applications of AI are emerging. For example, in a study of birdwing butterflies, an AI model identified subtle differences in wing shape and colour – traits that are challenging for humans to discern¹²⁴. Likewise, AI has detected colour differences between genotypes of the polymorphic wood tiger moth (*Arctia plantaginis*) that are typically invisible to the human eye and has analysed pattern signatures of the mimetic eggs of the common cuckoo (*Cuculus canorus*)¹²⁵. ML techniques have also been used to impute missing trait data¹²⁶ and to infer trait combinations that are responsible for species interactions. Computer vision algorithms that learn from image attributes¹²⁷ could make important strides in addressing the Raunkiaeran shortfall (Box 1).

Several advances could be useful for quantifying ecosystem function. For example, the use of remotely sensed images to assess plant productivity could identify ecological interactions (as described for the Eltonian shortfall). The large strides made in using imagery to identify species and trait composition¹²⁸ are expected to facilitate the detailed measurement of ecosystem processes such as nutrient cycling, decomposition and food-web energy dynamics.

Looking ahead, AI is likely to continue to make surprising discoveries related to traits that are undetectable by human senses and to further develop this knowledge by adding further data modalities and identifying connections among them. AI could be used to analyse high-resolution 3D scans of collected natural history specimens, such as skulls, pollen¹²⁹ or fossil plants. Likewise, efforts are also advancing from 2D to 3D, static to video and visible-only to hyperspectral and other imaging techniques, which (for example) could capture the behaviour of flying insects around artificial light sources¹³⁰. AI could also play an important part in discovering and connecting traits to function within a single context in which multiple species exist.

Evolutionary relationships

The Darwinian shortfall highlights gaps in our understanding of the tree of life and the evolution of lineages, species and traits⁷. A related but distinct gap is the shortfall in understanding of evolutionary relationships at the population level, which we refer to here as the genetic diversity shortfall (described in Box 2).

The evolutionary history of species is particularly relevant to the study of the evolution of species traits (see the Raunkiaeran shortfall) and species niches (see the Hutchinsonian shortfall). However, beyond its fundamental importance to evolutionary biology, knowledge of evolutionary relationships is also crucial for conservation policy and planning, given that programmes such as EDGE (Evolutionarily Distinct, Globally Endangered) prioritize the conservation of evolutionarily distinct species¹³¹.

Challenges

The tree of life has been extensively revised with the advent of molecular biology, initially through the comparison of DNA fragments and now through whole-genome sequencing. Despite considerable advances in these methodologies, the use of molecular techniques for building comprehensive phylogenies remains constrained by data availability^{7,132}. As a result, only a few well known groups, such as birds, some plants and mammals, have comprehensive species-level phylogenies^{103,133} and many other clades (particularly highly diverse groups such as microorganisms, insects and fungi) lack comprehensive phylogenetic information. Limited data availability also hampers the estimation of time-calibrated phylogenies, which are crucial for the derivation of accurate evolutionary timelines. This process typically involves using the fossil record to set node age constraints, employing molecular clocks to estimate evolution rates and divergence times and using integrated models of fossil and phylogenetic data to concurrently estimate divergence times and diversification rates.

Past and future role of AI

All modern phylogenetic inference methods use computational (albeit mostly non-ML) algorithms¹³⁴. However, several advances in AI have shown promise for improving the understanding of evolutionary relationships. For example, frequently used ML algorithms such as random forests (which combine the output of multiple decision trees to reach a single result) have been used to improve the accuracy and efficiency of phylogenetic inference¹³⁵ and graph neural networks represent a potentially fruitful next step in this area. Moreover, guided by phylogeny and the increasing number of sequenced species, AI is starting to be used to extract trait information directly from images^{136–138} in studies of phenotype–genotype correlations¹³⁹ (see the Raunkiaeran shortfall). AI-harvested information can also be used to study trait evolution. For instance, Phylo-NN¹⁴⁰ integrates images with phylogenetic data in a hierarchical manner to generate imageomes (sequences of quantized feature vectors that capture evolutionary signals at varying ancestry levels). Other examples include phylogenetic Gaussian processes for reconstructing ancestral traits¹⁴¹, which provide detailed insights into evolutionary history and trait evolution across species.

Looking ahead, AI has the potential to revolutionize our understanding of evolutionary relationships. Advanced AI algorithms could accelerate phylogenetic reconstruction by leveraging parallel hardware and computational approximations that borrow core ideas from AI to address the combinatorial nature of phylogeny estimation. The power of modern AI methods to synthesize scientific knowledge, coupled with the increasing availability of sequenced genomes, could lead to the development of evolutionary foundation models similar to those emerging at the DNA sequence^{31,142}, cell²¹, organism and species²² levels. Future applications of AI might also enable trait dendrograms to be inferred directly from images of organisms, potentially including 3D fossil images, to generate testable hypotheses about their evolutionary relationships. AI could also be used to analyse high-resolution 3D scans

Box 2 | The shortfall in knowledge of genetic diversity

Genetic diversity is not included among the original seven global shortfalls in biodiversity knowledge, but it represents another critical knowledge gap. The genetic variants present in a (sub)population are important for maintaining its local environmental adaptation²⁰⁷. Intraspecific variation can be affected differently to interspecific variation by ecological processes, such as trophic cascades²⁰⁸, and is tied to ecosystem services such as pollination²⁰⁹. Genetic diversity is also one of the best indicators of the health of a population. Indeed, the adaptive potential of a population (which is reduced by factors such as undergoing a genetic bottleneck) has long been used in conservation assessments and as the basis for a headline indicator²⁰⁷ in the Global Biodiversity Framework. In many populations, genetic diversity is declining rapidly²¹⁰.

Two major challenges associated with this shortfall are that traditional methods of collecting genetic data across populations are extremely resource-intensive, and that genetic diversity cannot necessarily be inferred from conservation risk status²¹¹. In time, these challenges may be alleviated by the ever-increasing collection of environmental DNA²¹² and RNA²¹³. Other challenges include differing sensitivities to individual genetic markers and sampling procedures and the lack of coordinated collection efforts (with notable exceptions²¹⁴).

The resounding success of human genome-wide association studies and comparative population genomics offers considerable promise for the development of biodiversity applications. For example, one proposed framework leveraged large models for gene vectorization to predict habitat specificity from microbial whole-genome sequences²¹⁵. The model not only demonstrated strong predictive performance but also recapitulated known interaction networks and proposed new candidates for experimental follow-up²¹⁵. This work highlights the capacity of AI to process

large-scale genetic data and uncover genetic interactions within species. AI-assisted approaches are being used to correlate the similarity of individual biometric features, such as egg-spot patterns or facial features, with genetic distance. These advances underscore the potential of AI to increase understanding of genetic diversity within populations and aid conservation efforts.

Future AI technologies might transform our understanding of genetic diversity by overcoming current limitations in data collection and analysis. Emerging technologies such as nanopore-based sequencing could bring genetic analysis into the field by enabling real-time, inexpensive data collection²¹⁶. Megabarcoding techniques^{217,218} that facilitate large-scale genetic barcoding greatly enhance our ability to gather comprehensive population genetic data. In bioinformatics, the shift towards transformer-based AI models — such as Nucleotide Transformer²¹⁹ and DNA Bidirectional Encoder Representations from Transformers (DNABERT)²²⁰ — along with generalized state-space models, open up the ability to analyse long DNA sequences. These models can generate the detailed DNA embeddings that are crucial for capturing genetic diversity within and between populations. Further developments could connect DNA sequences to observable phenotypes, which might enable image and acoustic data to serve as a proxy for genetic variability.

How ecologists can best leverage these data and tools remains unclear. The role of AI in making genetic data accessible through natural language interactions, exemplified by transformer models such as ChatNT²²¹ and single-cell GPT²²², could greatly enhance our ability to interpret and use genetic information. Synergy between AI advances and innovative data collection techniques is key to resolving the genetic diversity shortfall.

of collected natural history specimens¹⁴³ to study trait evolution and reconstruct ancestral states. In acknowledgement of the receptiveness of the AI and/or ML communities to competitions and benchmarks, development of AI methods might be spurred by efforts to extend existing computational phylogenetic benchmarks to multimodal data for phylogeny inference. We note that such benchmarks must be carefully designed to ensure they are as representative as possible of the real-world challenge they pose. Other large-scale datasets such as Arboretum¹⁴⁴, TreeOfLife-10M²² and BIOSCAN-5M¹⁴⁵ already target the Linnaean shortfall by providing extensive multimodal biodiversity data. Similar phylogeny-focused datasets could challenge the AI community to address the Darwinian shortfall.

Species interactions

The Eltonian shortfall describes the gap in our knowledge of interspecific interactions (including competition, predation, herbivory, mutualism and parasitism) that are fundamental in shaping the ecological distribution and abundance of species. If such biotic interactions are not accounted for, predictions of (for example) population viability and species responses to disturbance are likely to be incorrect. An understanding of species interaction networks also enriches the understanding of ecosystem functioning and resilience to disturbance.

Challenges

The complexity of ecosystems and food webs makes it challenging to identify the presence and strength of pairwise interactions. Although some interactions (such as pollination and predation) involve direct contact between species, such events are often rare and difficult to observe. Other interactions remain invisible, such as competition for scarce resources. Species interactions are often not apparent until after the loss or removal of an individual species causes disruption that reverberates throughout ecosystems. Diet analysis of faecal or gut content — and, increasingly, DNA metabarcoding — has informed our understanding of the richness of trophic interactions (those that control energy flow through food webs via direct consumption of one species by another), although such methods remain costly and have low accuracy for some taxa¹⁴⁶. In relation to species interactions at biogeographic scales, field data on species co-occurrence can inform joint SDMs^{89,147,148}, although such correlative approaches can rarely tease apart the mechanistic interactions that underlie co-occurrence patterns¹⁴⁹.

Past and future role of AI

The Eltonian shortfall represents one of the most critical data gaps and many exciting opportunities exist for the use of AI to mitigate

this shortfall (Fig. 1). AI supports two broad ways to address the Eltonian shortfall: intelligent sensing in primary data collection and data imputation by interaction prediction.

AI is already being used for direct sensing of plant–pollinator¹⁵⁰ and plant–pest interactions¹⁵¹ in collections of visual^{152,153} and acoustic¹⁵⁴ data. These initiatives are partly motivated by smart agriculture (that is, the use of advanced technologies to improve agricultural productivity and sustainability). Exciting possibilities exist for using the soundtracks from recorded videos and images of non-focal species recorded in databases (such as iNaturalist) and social media posts to harvest species interactions from existing data. Similarly, large language models (LLMs) could generate summaries of existing detailed text descriptions from guidebooks, grey literature (produced at all levels of government, academia, business and industry outside commercial publishing channels) and conservation assessments. Predictive approaches are of great importance in data imputation because sampling of species interactions is much harder than sampling of species occurrence, even when supported by AI. Predictive tools could help to alleviate this data gap by providing both the basis for probabilistic analyses and effective guidance for additional observation efforts.

AI is now beginning to be used to adapt traditional models of species interactions based on plant–pollinator^{155–157}, predator–prey^{158,159} and host–parasite¹⁶⁰ networks by incorporating trait matching, eDNA¹⁶¹, species co-occurrence or even just network structures. Interactions can be predicted even in the absence of traits by extrapolating from partly known interaction networks¹⁶² or relying solely on co-occurrence¹⁶³ using graph dimensionality reduction techniques to generate a small latent interaction space that captures the essential features of the meta-web and even allows some transferability across taxa¹⁶⁴. Although these AI-assisted approaches remain limited by the need for more finely grained data sources (such as GPS tracking)¹⁶⁵, they are key to the alleviation of data sparsity. Some studies have demonstrated advantages of ML approaches (specifically deep neural networks and random forests) over classic statistical techniques¹⁵⁶ but much work remains to be done, particularly in addressing the mismatch between predictions of individual interactions and community properties¹⁶⁶ and in developing process-based ML models that realistically constrain network predictions¹⁶³.

In the long term, one of the obvious uses of AI will be to improve the analysis of ecological and other interaction networks, although such approaches are yet to be widely adopted. Graph neural networks¹⁶⁷, a form of deep neural network that performs inference on network structures rather than tensors, seem to be particularly promising. Graph neural networks represent a natural way to learn (multilayer) representations from graphs and to deal with a number of network-related problems¹⁶⁸ but have had slow uptake in ecology. Graph tokenization¹⁶⁹, a process similar to text tokenization, offers an alternative to graph-based learning with conventional deep neural networks. Ultimately, although graphs are likely to be a natural match for AI-assisted analysis of interaction and species networks, the extent of the improvement gained for ecological networks remains to be seen¹⁷⁰.

AI-assisted analysis of interaction and species networks has many potential applications, such as anticipation of extinction cascades. Once a metaweb has been characterized, algorithms can predict which extinctions and co-extinctions would have disproportionate effects on network structure¹⁷¹ (and, probably, on network function). Combined with Bayesian reasoning, this approach has delivered automated decision-making support that promises substantially improved conservation outcomes¹⁷² and could improve analyses of the effects of threats

on food webs¹⁷³. For example, the use of AI has led to considerable progress in understanding the patterns of historical mass extinctions by first inferring historical food webs from ecomorphological and phylogenetic traits and then evaluating their resilience¹⁷⁴.

Summary and future directions

The potential contributions of AI to filling the seven global biodiversity shortfalls extend far beyond those completed so far (Table 1). We believe that AI is poised to have a major effect on the seven shortfalls in three broad areas: continued improvements in data collection and processing; new approaches to ecological inference and prediction; and collaborative research design and hypothesis generation (Fig. 2). The first two areas have already received attention in ecology and further advances and applications are clearly on the way. The third area remains in its infancy.

The best established and most widespread applications of AI in filling biodiversity shortfalls relate to increasing the speed, scale and effectiveness of data collection and processing. To date, much of this work has centred on the use of automated sensors and ML methods, primarily from computer vision, to generate new records of species presence (that is, occurrence or abundance). These data are already helping to fill the Wallacean, Prestonian and Hutchinsonian shortfalls (Figs. 1 and 2; Box 1), mainly through their use in statistical models that relate species presence to abiotic conditions. Particularly promising are extensions of open-world classification and category discovery that could soon reinvigorate the naming of new species (that is, the Linnaean shortfall) via the automated discovery of taxonomic diversity. Below the species level, AI extensions to fine-scale features of organisms will help to fill the Raunkerian shortfall and the genetic diversity shortfall (Box 2). Finally, the generation of large-scale species co-occurrence data or even automated detection of direct species interactions, such as predation events or aggressive vocalizations, could further help to fill the Eltonian shortfall.

In the long term, the speed, scale and effectiveness of data processing must continue to improve as biodiversity datasets expand. A specific area for future growth is the application of emerging techniques for the detection and identification of rare events and anomalies. Addressing the challenges presented by hard-to-detect species, events and other kinds of unusual signal will help to fill all seven shortfalls. Another key role for AI that has largely gone unrealized in ecology is the ability of models to not only automate the detection, classification and analysis of data features that are already known to humans but also to discover novel features and patterns that had not previously been described or even imagined. Examples include traits that are not detectable by human senses, subtle features that distinguish individual organisms within a species from each other, and general patterns in community structure.

The assimilation capacity of AI also has untapped potential to gather and synthesize what we already know^{175,176}. Large amounts of (often publicly funded) high-quality ecological data are present in at-risk species assessments, consultant findings and government reports¹⁷⁶, many of which are available online and include more-detailed ecological information than is found in existing compiled datasets. Data-gathering and synthesis activities that could take many person-years could be done much more efficiently by using LLMs¹⁷⁷ for the initial stages. AI-assisted data gathering could be especially helpful for meta-analyses, in which data and findings from hundreds or thousands of papers need to be compiled to answer key unresolved ecological and evolutionary questions¹⁷⁸. Despite its speed and promise,

Table 1 | Open problems in AI of relevance to biodiversity shortfalls

Open problem	Description	Importance to biodiversity
Foundation models	Development and application of large-scale models, pretrained on extensive datasets, that can be adapted across diverse tasks without extensive retraining, thereby enabling existing knowledge to be leveraged for specific applications	Build on pre-existing knowledge rather than task-specific training and empower non-expert users to describe domain-specific observations while benefiting from the model’s broad knowledge base, which potentially accelerates and democratizes biodiversity research
Learning from long-tail distributions and fine-grained categories	Development of ML methods and evaluation metrics for highly imbalanced class distributions, especially data with many rare categories delineated by subtle differences in features and complex patterns	Essential for addressing the inherent complexity of ecological data, strong sampling bias, rare and hard-to-observe species and correlations between species. These methods could potentially improve the precision of biodiversity models and forecasts, as well as conservation impact assessments and conservation strategies
Out-of-distribution performance	Development of ML models that generalize effectively to data that differ markedly from the training data, owing to variations in sensors, geographical locations, environmental conditions or other factors	Biodiversity data collected for training models often come from limited regions and conditions, whereas deployment occurs in varied and unpredictable environments. Accurate out-of-distribution predictions are essential to identify general ecological relationships and enable reliable monitoring and assessment across different ecosystems
Continual learning	Development of ML models that continually learn and adapt from streaming data online, without the need for retraining on large, static datasets	Important for models deployed in dynamic and/or remote environments where traditional batch retraining is logistically challenging. Enables models to be updated with new data and changing conditions, which improves their accuracy and reliability for monitoring biodiversity in real time
Long-context learning	Development of ML models that effectively use and retain information from long sequences of data, essential for tasks such as natural language processing, time series forecasting, and whole-genome DNA-sequence analysis	Models increasingly integrate multimodal data and leverage advanced architectures such as transformers. Long-context learning enables the incorporation of extensive temporal and spatial information, which improves the understanding of complex ecological patterns and processes and ultimately leads to more-accurate and comprehensive biodiversity models
Core set selection	Identification and extraction of the most representative or informative subsets from large, noisy datasets, to optimize model-training efficiency and performance while minimizing computational resources	A crucial component of biodiversity studies, which often involve massive, unstructured datasets that require considerable expertise to curate. Effective core set selection can streamline the analysis of complex ecosystems, reduce computational demands, and potentially improve the accuracy of biodiversity models by focusing on the most relevant data points
Efficient parameter inference	Development of methods and toolkits for efficient parameter estimation and interpretation of those parameters	Essential for compatibility with statistical approaches commonly used in biodiversity analyses (such as understanding organism tolerances along environmental gradients) and estimation of statistical uncertainty and confidence
Process discovery and causal inference	Development of ML methods to identify underlying mechanistic relationships and causal structures from observational data, distinguish correlation from causation, and extract interpretable processes from complex ecological datasets	Required for understanding ecosystem dynamics, species interactions and responses to environmental change. Increases the reliability of predicted biodiversity responses to interventions and helps to identify key drivers of population decline, thereby supporting evidence-based conservation strategies

AI, artificial intelligence; ML, machine learning.

AI-assisted data synthesis still requires careful checks at all stages of data compilation, analysis and interpretation (Box 3).

Progress has also been made towards using AI-based deep learning to perform inference and prediction tasks that have traditionally been the remit of statistical methods such as regression. These tasks are ubiquitous across ecology and are a natural fit for AI. As the use of AI-based methods continues to grow, improved models are expected to support more-accurate SDMs to assist with the Wallacean shortfall as well as forecasting of time series (which has particular relevance to the Prestonian shortfall), phylogenetic reconstructions and food webs. Particularly important advances could come from the growing use of knowledge-guided ML, explainable AI¹⁷⁹ and causal inference^{180,181}, which could go beyond simply describing biodiversity patterns to uncovering their underlying ecological mechanisms. Such approaches have begun to make important contributions in related disciplines^{182,183} but have not yet been widely applied to addressing biodiversity shortfalls. Other improvements in bioinformatics tools and methods are

expected to help to fill the genetic diversity (Box 2) and Darwinian shortfalls. Across all shortfalls, targeting of geographically and taxonomically under-represented biodiversity should be a priority⁷⁶.

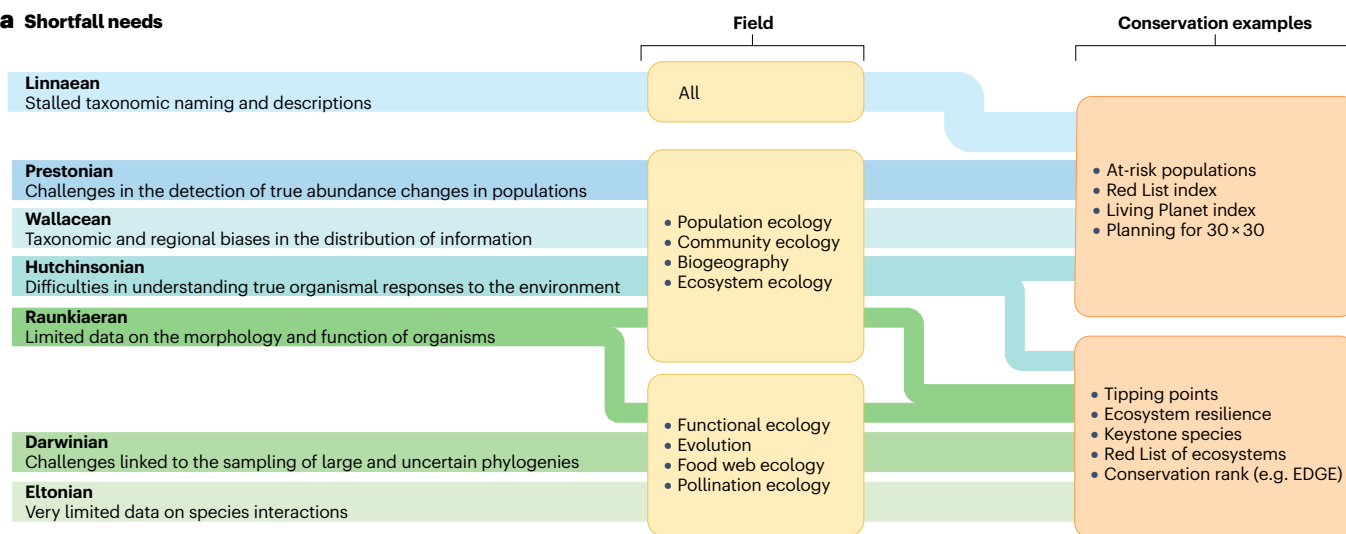
A specific and critical need in ecological inference is to understand connections and interactions across biological scales, especially at the community level. In general, AI is well-suited to the analysis of high-dimensionality data and systems and could be potentially more effective than existing methods at uncovering patterns and processes, both static and dynamic, in large interacting biological systems. The ability to use large, multimodal datasets and collections of models, together with nascent AI methods for data integration and synthesis, to extend knowledge from the species level to the community level will be essential for filling the Eltonian shortfall as well as other shortfalls that are driven by multi-species interactions. This process has begun with the creation of ML versions of joint SDMs¹⁸⁴, but moving from species to communities is expected to more accurately reflect ecological dynamics and provide more-direct links to ecosystem functions and services.

Review article

Finally, we propose that future generations of AI might fundamentally change the role of computers and computation in ecology and biodiversity research. At present, the contributions of AI remain restricted (like those of ecological models) to a largely top-down framework, in which a tool or model is used to conduct an analysis specified by a human scientist. The value and/or utility of AI in a particular context

is determined by how well different tools or models accomplish these predetermined tasks. This framework contrasts with collaborative research partnerships, in which participants iteratively discuss, challenge and build on each other's ideas to reach insights that could not have been reached by the individual participants alone. To date, such collaborations necessitate interactions between human colleagues.

a Shortfall needs



b Opportunities for AI

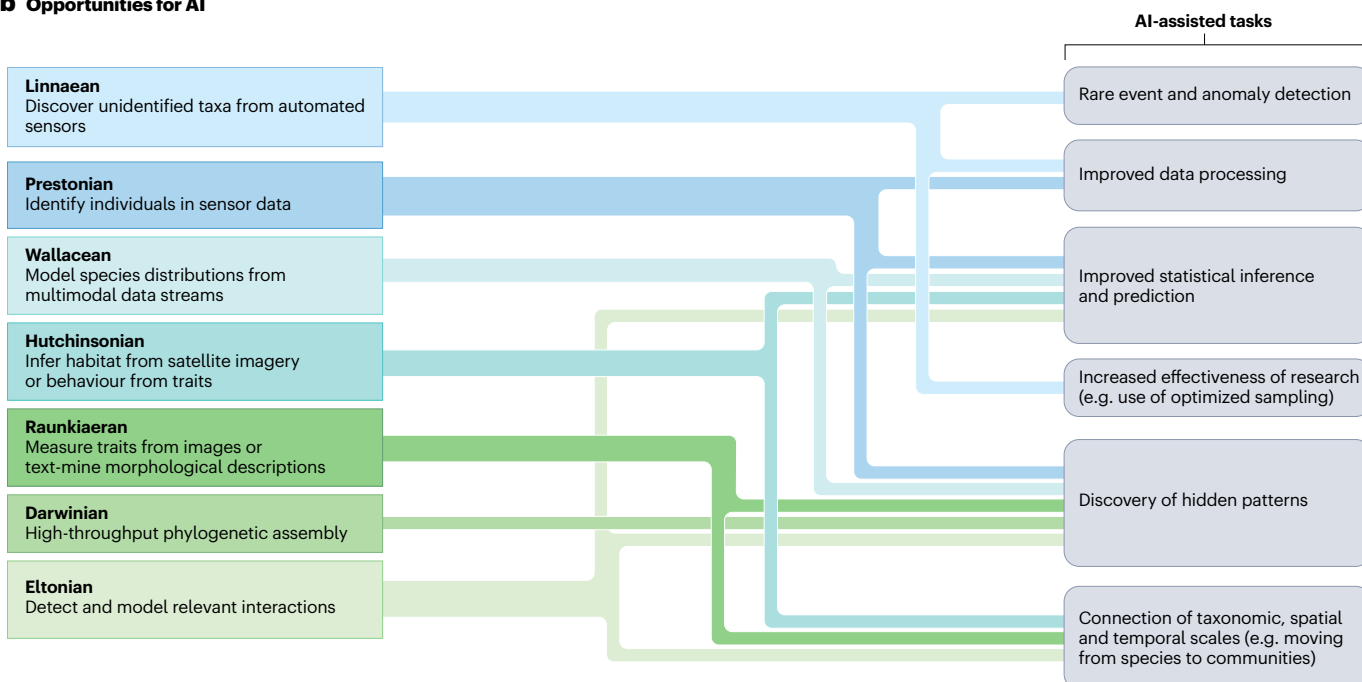


Fig. 2 | The seven shortfalls in biodiversity knowledge. **a**, Important challenges relating to each shortfall are presented along with summaries of the associated scientific fields and examples of conservation applications, including proposed or accepted indicators for the Kunming–Montréal Global Biodiversity Framework (GBF): the Red List Index (RLI), the Living Planet Index, the Red List of Ecosystems, Evolutionarily Distinct, Globally Endangered Species (EDGE) and

targets such as Target 3 of the GBF (protection of 30% of the land and sea by 2030, also known as 30 × 30). **b**, Connections between these shortfalls and applications of AI might provide the greatest potential for future advances beyond those already underway. The heights of the right-hand boxes indicate the relative potential for such gains.

Box 3 | Responsible use of artificial intelligence by ecologists

Artificial intelligence (AI) tools can help to address global biodiversity shortfalls but are not a panacea. Their effective application requires careful integration with scientific knowledge and application to real-world decision-making. Several broad areas are particularly relevant in the context of biodiversity shortfalls¹⁶:

- Errors or biases in data can be perpetuated in results, and poor algorithm design can mask unintended biases. For example, AI classifiers trained on abundant data from temperate regions could perform poorly in biodiverse tropical regions, and the development of AI models limited to well studied charismatic species might further bias research and management efforts towards those species. Explainable AI models can be helpful in exposing such biases.
- The performance expectations placed on AI systems can be unrealistically high, leading to disappointment and failures when these systems are used to perform roles for which they are unsuited, such as reasoning or prediction of future scenarios. We believe that, for the foreseeable future, AI should not replace human judgement in critical decision-making.
- The ability to easily and accurately predict the locations of species, particularly those that are threatened or endangered, could inadvertently increase threats to those species by supporting poaching and the wildlife trade.
- The analysis and sharing of datasets could potentially expose information about the individuals who collected such data or

such information could appear incidentally in images, recordings and other digital records. Although sensor networks and AI models are already used to detect individuals engaged in illegal activities (such as poaching or illegal logging), strong legal and regulatory frameworks might be needed to prevent the same networks from being used for broader human surveillance purposes.

- Widespread use of AI in biodiversity research could inadvertently increase inequality owing to its requirement for complex and costly computing resources, which represent a barrier to access by groups who would benefit from them the most, especially in the absence of effective communication. Accurate local knowledge could be overlooked in favour of deployment of large-scale, top-down models, which are often inaccurate at the local scales where biodiversity studies and conservation operate.
- The hardware running some AI systems, especially generative AI models such as ChatGPT, consumes substantial amounts of energy and materials²²³, the production and supply of which (ironically) pose considerable risks to global biodiversity. Such concerns might be alleviated by selection of more-energy-efficient transfer learning based on large foundation models, or smaller, more-focused AI models designed for specific tasks.

We speculate that several potential avenues exist through which AI might begin to join human scientists as a collaborative research partner. For example, we see potential for the growth of AI-assisted methods that iteratively and adaptively optimize experimental sampling schemes, in concert with changing input from human researchers, which will feed into coordinated monitoring efforts such as those of Group on Earth Observations Biodiversity Observation Network (GEO BON)³⁸. The use of AI in non-ecology fields, such as computer-aided drug discovery¹⁸⁵ and materials science¹⁸⁶, has shown the potential for such models to propose candidate research directions for experimental follow-up. Parallel tasks within biodiversity science could include AI-generated proposals of the existence of undetected species, resource flows, interactions, historical events, or intervention strategies that could be verified with additional research effort. Perhaps the broadest indications of such potential currently lie in LLM chatbots, which can be used by researchers to help them to think through research ideas and directions.

These potential applications of AI are largely in their infancy and breakthroughs are likely to come from the broad integration of ML with expert knowledge and models derived from first principles^{187,188}. The exploration of process-based or knowledge-guided ML models¹⁸⁹ that combine new information with existing scientific knowledge will be particularly important to derive ecological knowledge from newly obtained data. For example, AI systems trained on the extensive past climate record and based on transformer architecture¹⁹⁰ or graph neural networks¹⁹¹ have already contributed to climate modelling. Knowledge-guided ML^{69,70} offers a very promising future approach to this problem, in which the results of the AI model are constrained by boundary conditions dictated by physical knowledge of the climate system^{188,192} or phenological parameters^{188,192}.

Realization of these additional benefits of AI requires communication between biodiversity scientists and AI practitioners. A first useful step in this direction would be a concerted effort to translate biodiversity needs into tractable problems that attract attention from the AI community. Each biodiversity shortfall requires the identification of ecological use-cases along with evaluation data and benchmarks similar to GeoCLEF and BirdCLEF (tracks within CLEF that test and evaluate cross-language information retrieval of topics with a geographic specification and the ability to automate the identification of bird species from song, respectively), which are not usually prepared by ecologists. Development of such use-cases is already in progress for the Wallacean shortfall⁹⁵ and could be strategically set up for other biodiversity shortfalls. We stress that such collaborations also contribute to the AI community¹⁹³ by providing complex case studies that include multimodal data, long-tailed distributions, domain generalization, causal inference and other challenges. A long-term solution will be to integrate ecology into computer-science training programmes so that students trained in both specialties can speak the same language, clearly communicate issues and collectively arrive at solutions. The potential benefits of such joint training extend to related fields that face similar issues, such as Earth systems or climate sciences.

Beyond its contributions to fundamental ecology, an AI-inclusive research agenda is expected to have large follow-on benefits for policy and decision-making. Although AI is already making contributions to conservation in terms of data collection and analysis, its future contributions to fundamental ecology could greatly improve our understanding of conservation problems. As an immediate example, filling the Eltonian shortfall would provide a more-complete representation of species interactions and food webs in conservation that would help to

Glossary

Active learning

Sets of methodologies that aim to optimize data-collection strategies or to select the most informative samples among a large quantity of redundant items using iteration and the uncertainty of the AI model as guiding principles. Such models both increase confidence and reduce the amount of data required by adding examples to the training data in successive improvement cycles, in which the algorithm prompts the user for further information.

Benchmarks

Standardized, structured challenges posed to the AI community that often take the form of a fixed dataset, split into training, validation and testing subsets, alongside a carefully designed metric or set of metrics used to evaluate success.

Bioinformatics

A field of biology focused on the methods, tools, software and infrastructure needed to store, manage and analyse large, complex biological datasets.

Category discovery

A challenge in machine learning that aims to identify and group previously unknown or unlabelled categories within a dataset, thus allowing the model to autonomously discover and define new classes based on patterns or similarities in the data, in some cases alongside existing classes or domains.

Computer vision

A field of AI that enables machines to analyse and interpret information in images and videos to perform tasks such as image classification, object detection, semantic or instance segmentation, 3D image reconstruction, depth estimation, visual question answering, image retrieval and scene understanding.

Edge computational approaches

Computing systems that process the data at the device or sensor level and transmit only the desired results. Used when sensor behaviour is controlled by the results, in sensors that acquire a lot of unusable data, or under low-bandwidth conditions.

Embeddings

In machine learning, a representation of an object (such as an image, audio recording or word) as a numerical vector, such that some measure of discrepancy between vectors, generally called a distance, corresponds in a meaningful way to the relatedness of the objects they represent.

Foundation models

Machine learning models that are trained on a wide variety of data with the goal of being useful across a variety of different problems; broadly applicable foundation models require extremely large parameter spaces.

Fundamental niche

The role or position of an organism within an ecosystem, including its diet, behaviour and interactions with predator, prey or competitor species and its effect on its environment (habitat conditions and resources).

Gaussian processes

A type of statistical model based on the assumption that underlying random variables are normally distributed. Often used for continuous value prediction tasks that can naturally represent uncertainty in the modelled data.

Generative AI

Unlike discriminative AI, generative models are designed to generate novel content, often images or text, as opposed to providing information about existing data.

Imageomics

An emerging field, in which machine learning tools built around biological knowledge are used to analyse image data to characterize patterns and gain insights into traits and relationships at individual, population and species scales.

Machine learning

A subcategory of artificial intelligence in which models use an algorithm to pick out patterns in a training dataset that are relevant for solving the problem at hand.

Multimodal datasets

Datasets that observe the same entity with a variety of sensors. A modality could be an on-animal sensor, a drone image or a microphone, for example.

Natural language processing

A mechanism that enables human users to interact with artificial systems using natural language (that is via text or speech).

Open world classification

Also termed open set classification. A machine learning approach designed both to classify items into known classes and to recognize whether items belong to unknown classes that were absent during training.

Phenology

The study of how recurrent phenomena such as seasonal and climate variations affect events in the life cycle of an organism.

Realized niche

The subset of conditions within the fundamental niche actually used by a species, after interactions with other species (in particular predation and competition) and dispersal limitations have been taken into account.

Species distribution models

Also known as environmental niche models, these key tools predict species occurrence or abundance as a function of abiotic or biotic environmental variables; used for ecological inference of species responses to the environment and for mapping present and projected species distributions in response to climate or land-use change.

Tokenization

The process by which raw data (such as a text) are converted into smaller units (such as individual words) that can be used by models such as transformers.

Traits

Phenotypic attributes that affect an organism's fitness and/or influence its ecosystem functions and can also provide insights into the consequences of biodiversity loss for ecosystem functioning and human well-being.

Transformers

A neural network architecture that processes sequences of tokens (often text or images) in parallel rather than sequentially and is therefore highly effective at capturing long-range dependencies in data but computationally very expensive, requiring very large models and amounts of training data.

bridge the divide between biodiversity and ecosystem function^{172,173,194}. More-complete ecological knowledge is also predicted to greatly increase our ability to assess and monitor global indicators of the rapidly approaching 2030 GBF targets. Improved flow-through of ecological knowledge (and conservation-informed ecological hypothesis generation) to implementation of conservation strategies (Fig. 1) could generate

more opportunities to tailor analyses and scenarios to specific conservation questions from academia, government, non-governmental organizations and industry, thereby setting ecologists and conservationists on a direct path to data-informed solutions.

Published online: 20 February 2025

References

1. COP15: final text of Kunming–Montréal Global Biodiversity Framework. *Convention on Biological Diversity* <https://www.cbd.int/article/cop15-final-text-kunming-montreal-gbf-221222> (2022).
2. Caldwell, I. R. et al. Global trends and biases in biodiversity conservation research. *Cell Rep. Sustain.* **1**, 100082 (2024).
3. García-Roselló, E., González-Dacosta, J. & Lobo, J. M. The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. *Biol. Conserv.* **283**, 110118 (2023).
4. Daru, B. H. & Rodríguez, J. Mass production of unvouchered records fails to represent global biodiversity patterns. *Nat. Ecol. Evol.* **7**, 816–831 (2023).
5. Daru, B. H. et al. Widespread sampling biases in herbaria revealed from large-scale digitization. *N. Phytol.* **217**, 939–955 (2018).
6. Raven, P. H. & Wilson, E. O. A fifty-year plan for biodiversity surveys. *Science* **258**, 1099–1100 (1992).
7. Hortal, J. et al. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Syst.* **46**, 523–549 (2015).
8. Pereira, H. M. et al. Essential biodiversity variables. *Science* **339**, 277–278 (2013).
9. Berger-Tal, O. & Lahoz-Monfort, J. J. Conservation technology: the next generation. *Conserv. Lett.* **11**, e12458 (2018).
10. Tuia, D. et al. Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **13**, 792 (2022).
11. Barta, Z. Deep learning in terrestrial conservation biology. *Biol. Futura* **74**, 359–367 (2023).
12. Stowell, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**, e13152 (2022).
13. Weinstein, B. G. A computer vision for animal ecology. *J. Anim. Ecol.* **87**, 533–545 (2018).
14. Schneider, S., Taylor, G. W., Linquist, S. & Kremer, S. C. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Meth. Ecol. Evol.* **10**, 461–470 (2019).
15. Tuia, D. et al. Toward a collective agenda on AI for Earth science data analysis. *IEEE Geosci. Remote. Sens. Mag.* **9**, 88–104 (2021).
16. Biodiversity and Artificial Intelligence: Opportunities & Recommendations for Action. *The Global Partnership on AI* <https://gpa.ai/projects/responsible-ai/environment/biodiversity-and-ai-opportunities-recommendations-for-action.pdf> (2022).
17. Pichler, M. & Hartig, F. Machine learning and deep learning—a review for ecologists. *Meth. Ecol. Evol.* **14**, 994–1016 (2023).
18. Cardoso, P., Erwin, T. L., Borges, P. A. V. & New, T. R. The seven impediments in invertebrate conservation and how to overcome them. *Biol. Conserv.* **144**, 2647–2655 (2011).
19. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, 6723 (2024).
20. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Meth.* **21**, 1481–1491 (2024).
21. Rosen, Y. et al. Universal cell embeddings: a foundation model for cell biology. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.11.28.568918> (2023).
22. Stevens, S. et al. BioCLIP: a vision foundation model for the tree of life. In *2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* 19412–19424 (IEEE, 2024).
23. Bánki, O. Catalogue of Life Annual Release 2024. *Catalogue of Life* <https://www.catalogueoflife.org/2024/06/27/release> (2024).
24. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How many species are there on Earth and in the ocean? *PLoS Biol.* **9**, e1001127 (2011).
25. Caley, M. J., Fisher, R. & Mengersen, K. Global species richness estimates have not converged. *Trends Ecol. Evol.* **29**, 187–188 (2014).
26. Whittaker, R. J. et al. Conservation biogeography: assessment and prospect. *Divers. Distrib.* **11**, 3–23 (2005).
27. Winter, M. et al. Patterns and biases in climate change research on amphibians and reptiles: a systematic review. *R. Soc. Open Sci.* **3**, 160158 (2016).
28. Diniz-Filho, J. A. F., De Marco, P. Jr & Hawkins, B. A. Defying the curse of ignorance: perspectives in insect macroecology and conservation biogeography. *Insect Conserv. Divers.* **3**, 172–179 (2010).
29. Löbl, I., Klausnitzer, B., Hartmann, M. & Krell, F.-T. The silent extinction of species and taxonomists—an appeal to science policymakers and legislators. *Diversity* **15**, 1053 (2023).
30. Parsons, D. J., Pelletier, T. A., Wieringa, J. G., Duckett, D. J. & Carstens, B. C. Analysis of biodiversity data suggests that mammal species are hidden in predictable places. *Proc. Natl Acad. Sci. USA* **119**, e2103400119 (2022).
31. Gong, Z. et al. BIOSCAN-CLIP: bridging vision and genomics for biodiversity monitoring at scale. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2405.17537> (2024).
32. Zhao, B. & Mac Aodha, O. Incremental generalized category discovery. In *2023 Proc. IEEE/CVF Int. Conf. Comput. Vis.* 19080–19090 (IEEE, 2023).
33. Li, R., Ratnasingham, S., Zarubiieva, I., Somervuo, P. & Taylor, G. W. PROTAX-GPU: a scalable probabilistic taxonomic classification system for DNA barcodes. *Phil. Trans. R. Soc. B* **379**, 20230124 (2024).
34. Chen, Y. & Rolnick, D. Understanding insect range shifts with out-of-distribution detection. *climatechange.ai* <https://www.climatechange.ai/papers/neurips2023/130/paper.pdf> (Climate Change AI, 2023).
35. Gabeff, V., Rußwurm, M., Tuia, D. & Mathis, A. WildCLIP: scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *Int. J. Comput. Vis.* **132**, 3770–3786 (2024).
36. Paul, D. et al. A simple interpretable transformer for fine-grained image classification and analysis. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2311.04157> (2024).
37. Chiquier, M., Mall, U. & Vondrick, C. Evolving interpretable visual classifiers with large language models. In *Computer Vision—ECCV 2024: Lecture Notes in Computer Science* (eds Leonardi, A. et al.) 15122 (Springer, 2024).
38. Gonzalez, A. et al. A global biodiversity observing system to unite monitoring and guide action. *Nat. Ecol. Evol.* **7**, 1947–1952 (2023).
39. Lees, A. C. et al. A roadmap to identifying and filling shortfalls in Neotropical ornithology. *Auk* **137**, ukaa048 (2020).
40. Anderson-Teixeira, K. J. et al. CTFs-ForestGEO: a worldwide network monitoring forests in an era of global change. *Glob. Change Biol.* **21**, 528–549 (2015).
41. Moeller, A. K., Lukacs, P. M. & Horne, J. S. Three novel methods to estimate abundance of unmarked animals using remote cameras. *Ecosphere* **9**, e02331 (2018).
42. Gilbert, N. A., Clare, J. D. J., Stenglein, J. L. & Zuckerberg, B. Abundance estimation of unmarked animals based on camera-trap data. *Conserv. Biol.* **35**, 88–100 (2021).
43. Royle, J. A. & Nichols, J. D. Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**, 777–790 (2003).
44. Strebel, N. et al. Estimating abundance based on time-to-detection data. *Meth. Ecol. Evol.* **12**, 909–920 (2021).
45. Fiss, C. J. et al. Performance of unmarked abundance models with data from machine-learning classification of passive acoustic recordings. *Ecosphere* **15**, e4954 (2024).
46. Frommolt, K.-H. & Tauchert, K.-H. Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. *Ecol. Inform.* **21**, 4–12 (2014).
47. Rhinehart, T. A., Chronister, L. M., Devlin, T. & Kitzes, J. Acoustic localization of terrestrial wildlife: current practices and future opportunities. *Ecol. Evol.* **10**, 6794–6818 (2020).
48. Parham, J., Stewart, C., Berger-Wolf, T., Rubenstein, D. & Holmberg, J. The great Grevy's rally: a review on procedure. *cthlhu.dyn.wildme.io* https://cthlhu.dyn.wildme.io/public/papers/parham_jicai_aiwc_2018.pdf (2018).
49. Whitehead, H. Computer assisted individual identification of sperm whale flukes. *Rep. Int. Whal. Comm.* **12**, 71–77 (1990).
50. Crall, J. P., Stewart, C. V., Berger-Wolf, T. Y., Rubenstein, D. I. & Sundaresan, S. R. HotSpotter: patterned species instance recognition. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)* 230–237 (IEEE, 2013).
51. Arzoumanian, Z., Holmberg, J. & Norman, B. An astronomical pattern-matching algorithm for computer-aided identification of whale sharks, *Rhincodon typus*. *J. Appl. Ecol.* **42**, 999–1011 (2005).
52. Ye, M. et al. Deep learning for person re-identification: a survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 2872–2893 (2021).
53. Nepovniykh, E. et al. Species-agnostic patterned animal re-identification by aggregating deep local features. *Int. J. Comput. Vis.* **132**, 4003–4018 (2024).
54. Čermák, V., Pícek, L., Adam, L. & Papafitsoros, K. WildlifeDatasets: an open-source toolkit for animal re-identification. In *2024 IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)* 5941–5951 (IEEE, 2024).
55. Moskvayk, O., Maire, F., Armstrong, A. O., Dayoub, F. & Baktashmotlagh, M. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In *2021 Digital Image Computing: Techniques and Applications (DICTA)* <https://doi.org/10.1109/DICTA52665.2021.9647359> (IEEE, 2021).
56. Sundaresan, A. et al. Adapting the re-ID challenge for static sensors. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2412.00290> (2024).
57. Ravoor, P. C. & Sudarshan, T. S. B. Deep learning methods for multi-species animal re-identification and tracking—a survey. *Comput. Sci. Rev.* **38**, 100289 (2020).
58. Zuerl, M. et al. PolarBearVidD: a video-based re-identification benchmark dataset for polar bears. *Animals* **13**, 801 (2023).
59. Kuncheva, L. I., Garrido-Labrador, J. L., Ramos-Pérez, I., Hennessey, S. L. & Rodríguez, J. J. An experiment on animal re-identification from video. *Ecol. Inform.* **74**, 101994 (2023).
60. Koski, W. R. et al. Evaluation of UAS for photographic re-identification of bowhead whales, *Balaena mysticetus*. *J. Unman. Veh. Syst.* **3**, 22–29 (2015).
61. Knight, E. et al. Individual identification in acoustic recordings. *Trends Ecol. Evol.* **39**, 947–960 (2024).
62. Linhart, P., Mahamoud-Issa, M., Stowell, D. & Blumstein, D. T. The potential for acoustic individual identification in mammals. *Mamm. Biol.* **102**, 667–683 (2022).
63. Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: a survey. *Int. J. Comput. Vis.* **132**, 5635–5662 (2024).
64. Vaze, S., Han, K., Vedaldi, A. & Zisserman, A. Generalized category discovery. In *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* 7482–7491 (IEEE, 2022).
65. Kulits, P., Wall, J., Bedetti, A., Henley, M. & Beery, S. ElephantBook: a semi-automated human-in-the-loop system for elephant re-identification. In *ACM SIGCAS Conf. on Computing and Sustainable Societies (COMPASS)* 88–98 (Association for Computing Machinery, 2021).
66. Delplanque, A. et al. From crowd to herd counting: how to precisely detect and count African mammals using aerial imagery and deep learning? *ISPRS J. Photogramm. Remote. Sens.* **197**, 167–180 (2023).
67. Perez, G., Maji, S. & Sheldon, D. DISCount: counting in large image collections with detector-based importance sampling. *Proc. AAAI Conf. Artif. Intell.* **38**, 22294–22302 (2024).
68. Hebert, P. D. N., Floyd, R., Jafarpour, S. & Prosser, S. W. J. Barcode 100K specimens: in a single nanopore run. *Mol. Ecol. Resour.* **25**, e14028 (2025).

69. Reichstein, M. et al. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
70. Karpatne, A. et al. Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331 (2017).
71. Purvis, A., Gittleman, J. L., Cowlishaw, G. & Mace, G. M. Predicting extinction risk in declining species. *Proc. R. Soc. Lond. B* **267**, 1947–1952 (2000).
72. Newsome, T. M. et al. Constraints on vertebrate range size predict extinction risk. *Glob. Ecol. Biogeogr.* **29**, 76–86 (2020).
73. Staude, I. R., Navarro, L. M. & Pereira, H. M. Range size predicts the risk of local extinction from habitat loss. *Glob. Ecol. Biogeogr.* **29**, 16–25 (2020).
74. Mace, G. M. et al. Quantification of extinction risk: IUCN's system for classifying threatened species. *Conserv. Biol.* **22**, 1424–1442 (2008).
75. Jetz, W. et al. Include biodiversity representation indicators in area-based conservation targets. *Nat. Ecol. Evol.* **6**, 123–126 (2022).
76. Pimm, S. L. et al. Emerging technologies to conserve biodiversity. *Trends Ecol. Evol.* **30**, 685–696 (2015).
77. Geurts, E. M., Reynolds, J. D. & Starzowski, B. M. Turning observations into biodiversity data: broadscale spatial biases in community science. *Ecosphere* **14**, e4582 (2023).
78. Fretwell, P. T. & Trathan, P. N. Discovery of new colonies by Sentinel2 reveals good and bad news for emperor penguins. *Remote Sens. Ecol. Conserv.* **7**, 139–153 (2021).
79. Cubaynes, H. C. & Fretwell, P. T. Whales from space dataset, an annotated satellite image dataset of whales for training machine learning models. *Sci. Data* **9**, 245 (2022).
80. Manno, L. et al. Leveraging social media and deep learning to detect rare megafauna in video surveys. *Conserv. Biol.* **36**, e13798 (2022).
81. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **77**, 802–813 (2008).
82. Beery, S., Cole, E., Parker, J., Perona, P. & Winner, K. Species distribution modeling for machine learning practitioners: a review. In *Proc. 4th ACM SIGCAS Conf. on Computing and Sustainable Societies* 329–348 (Association for Computing Machinery, 2021).
83. Isaac, N. J. B. et al. Data integration for large-scale models of species distributions. *Trends Ecol. Evol.* **35**, 56–67 (2020).
84. Botella, C. et al. Overview of GeoLifeCLEF 2023: species composition prediction with high spatial resolution at continental scale using remote sensing. In *CLEF 2023—Working Notes of the Conference and Labs of the Evaluation Forum* article 3497, 1954–1971 (2023).
85. Mashiane, K., Ramoelo, A. & Adelabu, S. Prediction of species richness and diversity in sub-alpine grasslands using satellite remote sensing and random forest machine-learning algorithm. *Appl. Veg. Sci.* **27**, e12778 (2024).
86. Lange, C., Cole, E., Van Horn, G. & Mac Aodha, O. Active learning-based species range estimation. In *NIPS 2023: Proc. 37th Int. Conf. on Neural Information Processing Systems* article 1815, 41892–41913 (NeurIPS, 2024).
87. Mondain-Monval, T. et al. Adaptive sampling by citizen scientists improves species distribution model performance: a simulation study. *Meth. Ecol. Evol.* **15**, 1206–1220 (2024).
88. Seaton, F. M., Jarvis, S. G. & Henrys, P. A. Spatio-temporal data integration for species distribution modelling in R-INLA. *Meth. Ecol. Evol.* **15**, 1221–1232 (2024).
89. Pollock, L. J. et al. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Meth. Ecol. Evol.* **5**, 397–406 (2014).
90. Caradima, B., Schuwirth, N. & Reichert, P. From individual to joint species distribution models: a comparison of model complexity and predictive performance. *J. Biogeogr.* **46**, 2260–2274 (2019).
91. Talluto, M. V., Mokany, K., Pollock, L. J. & Thuiller, W. Multifaceted biodiversity modelling at macroecological scales using Gaussian processes. *Divers. Distrib.* **24**, 1492–1502 (2018).
92. Chen, D. & Gomes, C. P. Bias reduction via end-to-end shift learning: application to citizen science. In *Proc. Thirty-Third AAAI Conf. on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conf. and Ninth AAAI Symp. on Educational Advances in Artificial Intelligence* 493–500 (AAAI Press, 2019).
93. Rolf, E. et al. A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* **12**, 4392 (2021).
94. Klemmer, K., Rolf, E., Robinson, C., Mackey, L. & Rußwurm, M. SatCLIP: global, general-purpose location embeddings with satellite imagery. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2311.17179> (2024).
95. Cole, E. et al. Spatial implicit neural representations for global-scale species mapping. In *Proc. 40th Int. Conf. on Machine Learning* article 202, 6320–6342 (PMLR, 2023).
96. Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J. & Elith, J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecol. Monogr.* **92**, e01486 (2022).
97. Holt, R. D. Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc. Natl Acad. Sci. USA* **106**, 19659–19665 (2009).
98. Soberón, J. Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett.* **10**, 1115–1123 (2007).
99. Anderson, R. P. A framework for using niche models to estimate impacts of climate change on species distributions. *Ann. NY Acad. Sci.* **1297**, 8–28 (2013).
100. Bennett, J. M. et al. GlobTherm, a global database on thermal tolerances for aquatic and terrestrial organisms. *Sci. Data* **5**, 180022 (2018).
101. Childress, E. S. & Letcher, B. H. Estimating thermal performance curves from repeated field observations. *Ecology* **98**, 1377–1387 (2017).
102. Soberón, J. & Peterson, A. T. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiv. Inform.* <https://doi.org/10.17161/bi.v2i0.4> (2005).
103. Jetz, W. et al. Biological Earth observation with animal sensors. *Trends Ecol. Evol.* **37**, 293–298 (2022).
104. Deetjen, M. E., Biewener, A. A. & Lentink, D. High-speed surface reconstruction of a flying bird using structured light. *J. Exp. Biol.* **220**, 1956–1961 (2017).
105. Nath, T. et al. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* **14**, 2152–2176 (2019).
106. Pereira, T. D. et al. SLEAP: a deep learning system for multi-animal pose tracking. *Nat. Meth.* **19**, 486–495 (2022).
107. Wei, J. N. et al. A deep learning and digital archaeology approach for mosquito repellent discovery. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.09.01.504601> (2024).
108. Nativi, S., Mazzetti, P. & Craglia, M. Digital ecosystems for developing digital twins of the Earth: the Destination Earth case. *Remote Sens.* **13**, 2119 (2021).
109. Koning, Kde et al. Digital twins: dynamic model-data fusion for ecology. *Trends Ecol. Evol.* **38**, 916–926 (2023).
110. Trantas, A., Plug, R., Pileggi, P. & Lazovik, E. Digital twin challenges in biodiversity modelling. *Ecol. Inform.* **78**, 102357 (2023).
111. Wilman, H. et al. EltonTraits 1.0: species-level foraging attributes of the world's birds and mammals. *Ecology* **95**, 2027–2027 (2014).
112. Tobias, J. A. et al. AVONET: morphological, ecological and geographical data for all birds. *Ecol. Lett.* **25**, 581–597 (2022).
113. Oskyrko, O., Mi, C., Meiri, S. & Du, W. ReptTraits: a comprehensive dataset of ecological traits in reptiles. *Sci. Data* **11**, 243 (2024).
114. Shirey, V. et al. LepTraits 1.0: a globally comprehensive dataset of butterfly traits. *Sci. Data* **9**, 382 (2022).
115. Sheard, C. et al. Nest traits for the world's birds. *Glob. Ecol. Biogeogr.* **33**, 206–214 (2024).
116. Kattge, J. et al. TRY—a global database of plant traits. *Glob. Change Biol.* **17**, 2905–2935 (2011).
117. Murphy, S. J. et al. SAviTraits 1.0: seasonally varying dietary attributes for birds. *Glob. Ecol. Biogeogr.* **32**, 1690–1698 (2023).
118. Madin, J. S. et al. The coral trait database, a curated database of trait information for coral species from the global oceans. *Sci. Data* **3**, 160017 (2016).
119. Cébron, A. et al. BactoTraits—a functional trait database to evaluate how natural and man-induced changes influence the assembly of bacterial communities. *Ecol. Indic.* **130**, 108047 (2021).
120. Pigot, A. L. et al. Macroevolutionary convergence connects morphological form to ecological function in birds. *Nat. Ecol. Evol.* **4**, 230–239 (2020).
121. Augusto, L. & Boça, A. Tree functional traits, forest biomass, and tree species diversity interact with site properties to drive forest soil carbon. *Nat. Commun.* **13**, 1097 (2022).
122. Wilde, B. C., Bragg, J. G. & Cornwell, W. Analyzing trait–climate relationships within and among taxa using machine learning and herbarium specimens. *Am. J. Bot.* **110**, e16167 (2023).
123. Kline, J. et al. A framework for autonomic computing for in situ imageomics. In *2023 IEEE Int. Conf. on Autonomic Computing and Self-Organizing Systems (ACSOS)* 11–16 (IEEE, 2023).
124. Hoyal Cuthill, J. F., Guttenberg, N. & Huertas, B. Male and female contributions to diversity among birdwing butterfly images. *Commun. Biol.* **7**, 774 (2024).
125. Stoddard, M. C., Kilner, R. M. & Town, C. Pattern recognition algorithm reveals how birds evolve individual egg pattern signatures. *Nat. Commun.* **5**, 4117 (2014).
126. Johnson, T. F., Isaac, N. J. B., Paviolo, A. & González-Suárez, M. Handling missing values in trait data. *Glob. Ecol. Biogeogr.* **30**, 51–62 (2021).
127. Marcos, D., Potze, A., Xu, W., Tuia, D. & Akata, Z. Attribute prediction as multiple instance learning. *Trans. Mach. Learn. Res.* **8**, 253156463 (2022).
128. Crofts, A. L. et al. Linking aerial hyperspectral data to canopy tree biodiversity: an examination of the spectral variation hypothesis. *Ecol. Monogr.* **94**, e1605 (2024).
129. Wilson, O. J. The 3D pollen project: an open repository of three-dimensional data for outreach, education and research. *Rev. Palaeobot. Palynol.* **312**, 104860 (2023).
130. Fabian, S. T., Sondhi, Y., Allen, P. E., Theobald, J. C. & Lin, H.-T. Why flying insects gather at artificial light. *Nat. Commun.* **15**, 689 (2024).
131. Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C. & Baillie, J. E. M. Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE* **2**, e296 (2007).
132. Diniz Filho, J. A. F. et al. Macroecological links between the Linnean, Wallacean, and Darwinian shortfalls. *Front. Biogeogr.* **15**, e59566 (2023).
133. Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* **17**, e3000494 (2019).
134. Warnow, T. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation* (Cambridge Univ. Press, 2017).
135. Mo, Y. K., Hahn, M. W. & Smith, M. L. Applications of machine learning in phylogenetics. *Mol. Phylogenet. Evol.* **196**, 108066 (2024).
136. Lürig, M. D., Donoghue, S., Svensson, E. I., Porto, A. & Tsuboi, M. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Front. Ecol. Evol.* **9**, 642774 (2021).
137. Younis, S. et al. Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. *Bot. Lett.* **165**, 377–383 (2018).
138. Weaver, W. N., Ng, J. & Laport, R. G. LeafMachine: using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Appl. Plant. Sci.* **8**, e11367 (2020).
139. Stupp, D. et al. Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.* **12**, 6454 (2021).

140. Elhamod, M. et al. Discovering novel biological traits from images using phylogeny-guided neural networks. In *Proc. 29th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD 2023)* 3966–3978 (Association for Computing Machinery, 2023).
141. Meagher, J. P., Damoulas, T., Jones, K. E. & Girolami, M. in *Statistical Data Science* (ed. Adams, N.) Ch. 7, 111–124 (World Scientific, 2018).
142. Nguyen, T. Q., Ebnesajjad, C., Cole, S. R. & Stuart, E. A. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* **11**, 225–247 (2017).
143. Blackburn, D. P. et al. Increasing the impact of vertebrate scientific collections through 3D imaging: the openVertebrate (oVert) thematic collections network. *BioScience* **74**, 169–186 (2024).
144. Yang, C.-H. et al. Arboretum: a large multimodal dataset enabling AI for biodiversity. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2406.17720> (2024).
145. Gharaee, Z. et al. BIOSCAN-5M: a multimodal dataset for insect biodiversity. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2406.12723> (2024).
146. Alberdi, A. et al. Promises and pitfalls of using high-throughput sequencing for diet analysis. *Mol. Ecol. Resour.* **19**, 327–348 (2019).
147. Wilkinson, D. P., Golding, N., Guillera-Aroita, G., Tingley, R. & McCarthy, M. A. Defining and evaluating predictions of joint species distribution models. *Meth. Ecol. Evol.* **12**, 394–404 (2021).
148. Schliep, E. M. et al. Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data. *Glob. Ecol. Biogeogr.* **27**, 142–155 (2018).
149. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.* **23**, 1050–1063 (2020).
150. Ratnayake, M. N., Dyer, A. G. & Dorin, A. Tracking individual honeybees among wildflower clusters with computer vision-facilitated pollinator monitoring. *PLoS ONE* **16**, e0239504 (2021).
151. Liu, J. & Wang, X. Plant diseases and pests detection based on deep learning: a review. *Plant. Meth.* **17**, 22 (2021).
152. Hoye, T. T. et al. Deep learning and computer vision will transform entomology. *Proc. Natl Acad. Sci. USA* **118**, e2002545117 (2021).
153. Nawoya, S. et al. Computer vision and deep learning in insects for food and feed production: a review. *Comput. Electron. Agric.* **216**, 108503 (2024).
154. Folliot, A., Hupert, S., Ducretet, M., Sèbe, F. & Sueur, J. Using acoustics and artificial intelligence to monitor pollination by insects and tree use by woodpeckers. *Sci. Total. Environ.* **838**, 155883 (2022).
155. Ornai, A. & Keasar, T. Floral complexity traits as predictors of plant-bee interactions in a Mediterranean pollination web. *Plants* **9**, 1432 (2020).
156. Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M. & Hartig, F. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Meth. Ecol. Evol.* **11**, 281–293 (2020).
157. Sydenham, M. A. K. et al. MetaComNet: a random forest-based framework for making spatial predictions of plant–pollinator interactions. *Meth. Ecol. Evol.* **13**, 500–513 (2022).
158. Caron, D., Maiorano, L., Thuiller, W. & Pollock, L. J. Addressing the Eltonian shortfall with trait-based interaction models. *Ecol. Lett.* **25**, 889–899 (2022).
159. Llewellyn, J. et al. Predicting predator–prey interactions in terrestrial endotherms using random forest. *Ecography* **2023**, e06619 (2023).
160. Kotula, H. J., Peralta, G., Frost, C. M., Todd, J. H. & Tylanakis, J. M. Predicting direct and indirect non-target impacts of biocontrol agents using machine-learning approaches. *PLoS ONE* **16**, e0252448 (2021).
161. Roslin, T. & Majaneva, S. The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite! *Genome* **59**, 603–628 (2016).
162. Adhurya, S. & Park, Y.-S. A novel method for predicting ecological interactions with an unsupervised machine learning algorithm. *Methods Ecol. Evol.* **15**, 1247–1260 (2024).
163. Strydom, T. et al. A roadmap towards predicting species interaction networks (across space and time). *Phil. Trans. R. Soc. B* **376**, 20210063 (2021).
164. Strydom, T. et al. Food web reconstruction through phylogenetic transfer of low-rank network representation. *Meth. Ecol. Evol.* **13**, 2838–2849 (2022).
165. Suraci, J. P. et al. Beyond spatial overlap: harnessing new technologies to resolve the complexities of predator–prey interactions. *Oikos* **2022**, e090004 (2022).
166. Caron, D. et al. Trait-matching models predict pairwise interactions across regions, not food web properties. *Glob. Ecol. Biogeogr.* **33**, e13807 (2024).
167. Corso, G., Stark, H., Jegelka, S., Jaakkola, T. & Barzilay, R. Graph neural networks. *Nat. Rev. Meth. Primer* **4**, 17 (2024).
168. Hamilton, W. L. Graph representation learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning (SLAIML)* <https://doi.org/10.1007/978-3-031-01588-5> (Springer, 2020).
169. Kim, J. et al. Pure transformers are powerful graph learners. In *Proc. 36th Int. Conf. on Neural Information Processing Systems (NIPS 2022)* article 1060, 14582–14595 (NeurIPS, 2022).
170. Strydom, T. et al. Graph embedding and transfer learning can help predict potential species interaction networks despite data limitations. *Meth. Ecol. Evol.* **14**, 2917–2930 (2023).
171. Allesina, S. & Pascual, M. Googling food webs: can an eigenvector measure species’ importance for coextinctions? *PLoS Comput. Biol.* **5**, e1000494 (2009).
172. McDonald-Madden, E. et al. Using food-web theory to conserve ecosystems. *Nat. Commun.* **7**, 10245 (2016).
173. O’Connor, L. M. J. et al. Vulnerability of terrestrial vertebrate food webs to anthropogenic threats in Europe. *Glob. Change Biol.* **30**, e17253 (2024).
174. Fricke, E. et al. Collapse of terrestrial mammal food webs since the Late Pleistocene. *Science* **377**, 1008–1011 (2022).
175. Elliott, M. J. & Fortes, J. A. B. Toward reliable biodiversity information extraction from large language models. In *2024 IEEE 20th Int. Conf. on e-Science (IEEE, 2024)*.
176. Bledsoe, E. K. et al. Data rescue: saving environmental data from extinction. *Proc. R. Soc. B* **289**, 20220938 (2022).
177. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. 34th Int. Conf. on Neural Information Processing Systems (NIPS 2020)* article 793, 9459–9474 (NeurIPS, 2020).
178. Berger-Tal, O. et al. Leveraging AI to improve evidence synthesis in conservation. *Trends Ecol. Evol.* **39**, 548–557 (2024).
179. Ryo, M. et al. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography* **44**, 199–205 (2021).
180. Runge, J. et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **10**, 2553 (2019).
181. Camps-Valls, G. et al. Discovering causal relations and equations from data. *Phys. Rep.* **1044**, 1–68 (2023).
182. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **55**, 66 (2022).
183. Liu, L. et al. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nat. Commun.* **15**, 357 (2024).
184. Hartig, F. et al. Novel community data in ecology-properties and prospects. *Trends Ecol. Evol.* **39**, 280–293 (2024).
185. Niazi, S. K. & Mariani, Z. Computer-aided drug design and drug discovery: a prospective analysis. *Pharmaceuticals* **17**, 22 (2024).
186. Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.* **8**, 1153–1172 (2021).
187. Han, Z., Zhang, L., Jiang, Y., Wang, H. & Jiguet, F. Unravelling species co-occurrence in a steppe bird community of Inner Mongolia: insights for the conservation of the endangered Jankowski’s bunting. *Divers. Distrib.* **26**, 843–852 (2020).
188. Zérach, Y., Valero, S. & Inglada, J. Physics-driven probabilistic deep learning for the inversion of physical models with application to phenological parameter retrieval from satellite times series. *IEEE Trans. Geosci. Remote. Sens.* **61**, 4404723 (2023).
189. Karpatne, A., Kannan, R. & Kumar, V. (eds) *Knowledge Guided Machine Learning: Accelerating Discovery Using Scientific Knowledge and Data: Data Mining and Knowledge Discovery Series* (CRC Press, 2023).
190. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimateX: a foundation model for weather and climate. In *Proc. 40th Int. Conf. on Machine Learning (ICML 2023)* 25904–25938 (PMLR, 2023).
191. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1412 (2023).
192. de Bézenac, E., Pajot, A. & Gallinari, P. Deep learning for physical processes: incorporating prior scientific knowledge. *J. Stat. Mech.* **2019**, 124009 (2019).
193. Rolnick, D. et al. Position: application-driven innovation in machine learning. In *Proc. 41st Int. Conf. on Machine Learning* article 235, 42707–42718 (PMLR, 2024).
194. Harvey, E., Gounand, I., Ward, C. L. & Altermatt, F. Bridging ecology and conservation: from ecological networks to ecosystem function. *J. Appl. Ecol.* **54**, 371–379 (2017).
195. Hengl, T. et al. SoilGrids250m: global gridded soil information based on machine learning. *PLoS ONE* **12**, e0169748 (2017).
196. Mahecha, M. D. et al. Earth system data cubes unravel global multivariate dynamics. *Earth Syst. Dyn.* **11**, 201–234 (2020).
197. Bush, A. et al. Connecting Earth observation to high-throughput biodiversity data. *Nat. Ecol. Evol.* **1**, 0176 (2017).
198. Bojinski, S. et al. The concept of essential climate variables in support of climate research, applications, and policy. *Bull. Am. Meteor. Soc.* **95**, 1431–1443 (2014).
199. Skidmore, A. K. & Pettorelli, N. Agree on biodiversity metrics to track from space: ecologists and space agencies must forge a global monitoring strategy. *Nature* **523**, 403–406 (2015).
200. Lang, N., Jetz, W., Schindler, K. & Wegner, J. D. A high-resolution canopy height model of the Earth. *Nat. Ecol. Evol.* **7**, 1778–1789 (2023).
201. Reiche, J. et al. Combining satellite data for better tropical forest monitoring. *Nat. Clim. Change* **6**, 120–122 (2016).
202. Moreno-Martinez, Á. et al. A methodology to derive global maps of leaf traits using remote sensing and climate data. *Remote. Sens. Environ.* **218**, 69–88 (2018).
203. Yang, H. et al. Global patterns of tree wood density. *Glob. Change Biol.* **30**, e17224 (2024).
204. Wolf, S. et al. Citizen science plant observations encode global trait patterns. *Nat. Ecol. Evol.* **6**, 1850–1859 (2022).
205. Gómez-Chova, L., Tuia, D., Moser, G. & Camps-Valls, G. Multimodal classification of remote sensing images: a review and future directions. *Proc. IEEE* **103**, 1560–1584 (2015).
206. Bachmann, R., Mizrahi, D., Atanov, A. & Zamir, A. MultiMAE: multi-modal multi-task masked autoencoders. *Computer Vision—ECCV 2022 17th Eur. Conf. Proc. XXXVII* 348–367 (Springer, 2022).
207. Hoban, S. et al. Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biol. Conserv.* **248**, 108654 (2020).
208. Des Roches, S. et al. The ecological importance of intraspecific variation. *Nat. Ecol. Evol.* **2**, 57–64 (2018).

209. Des Roches, S., Pendleton, L. H., Shapiro, B. & Palkovacs, E. P. Conserving intraspecific variation for nature's contributions to people. *Nat. Ecol. Evol.* **5**, 574–582 (2021).
210. Exposito-Alonso, M. et al. Genetic diversity loss in the Anthropocene. *Science* **377**, 1431–1435 (2022).
211. Schmidt, C., Hoban, S., Hunter, M., Paz-Vinas, I. & Garroway, C. J. Genetic diversity and IUCN Red List status. *Conserv. Biol.* **37**, e14064 (2023).
212. Yates, M. C., Derry, A. M. & Cristescu, M. E. Environmental RNA: a revolution in ecological resolution? *Trends Ecol. Evol.* **36**, 601–609 (2021).
213. Cristescu, M. E. Can environmental RNA revolutionize biodiversity science? *Trends Ecol. Evol.* **34**, 694–697 (2019).
214. Hobern, D. BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. *Genome* **64**, 161–164 (2021).
215. Li, Z., Cranganore, S. S., Youngblut, N. & Kilbertus, N. Whole genome transformer for gene interaction effects in microbiome habitat specificity. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2405.05998> (2024).
216. Srivathsan, A. & Meier, R. Scalable, cost-effective, and decentralized DNA barcoding with Oxford nanopore sequencing. *Meth. Mol. Biol.* **2744**, 223–238 (2024).
217. Meier, R., Hartop, E., Pylatiuk, C. & Srivathsan, A. Towards holistic insect monitoring: species discovery, description, identification and traits for all insects. *Phil. Trans. R. Soc. B* **379**, 20230120 (2024).
218. Pei, W. et al. Megabarcoding reveals a tale of two very different dark taxa along the same elevational gradient. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.04.29.591578> (2024).
219. Dalla-Torre, H. et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat. Meth.* <https://doi.org/10.1038/s41592-024-02523-z> (2024).
220. Zhou, Z. et al. DNABERT-S: learning species-aware DNA embedding with genome foundation models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2402.08777> (2024).
221. Richard, G. et al. ChatNT: a multimodal conversational agent for DNA, RNA and protein tasks. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.04.30.591835> (2024).
222. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Meth.* **21**, 1470–1480 (2024).
223. Luccioni, S., Jernite, Y. & Strubell, E. Power hungry processing: watts driving the cost of AI deployment? In *Proc. 2024 ACM Conf. on Fairness, Accountability, and Transparency (FAccT)* 85–99 (Association for Computing Machinery, 2024).

Acknowledgements

The authors' research work is supported by the AI and Biodiversity Change (ABC) Global Center, which is funded by the US National Science Foundation under award 2330423 (to J.K., S.B., M.A.J. and T.B.-W.) and the Natural Sciences and Engineering Research Council of Canada under award 585136 (to L.J.P., K.G., D.R. and G.W.T.). This paper also draws on research supported in part by the Social Sciences and Humanities Research Council of Canada.

Author contributions

All authors researched data for the article and wrote the first draft of the manuscript. L.J.P., J.K., S.B., K.G., O.M.A., B.M., D.R., G.W.T., D.T. and T.B.-W. reviewed or edited the manuscript before submission. L.J.P., J.K., K.G., M.A.J., O.M.A. and T.B.-W. made substantial contributions to discussions of its content. In addition, T.B.-W. initiated the review paper, L.J.P., J.K., M.A.J. and K.G. conceived the original idea, and L.J.P. and J.K. led the organization of paper sections and refined the drafts.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Biodiversity* thanks Sarab Sethi, who co-reviewed with Peggy Bevan, Russell Dinnage and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2025

¹Department of Biology, McGill University, Montréal, Quebec, Canada. ²Mila—Quebec AI Institute, Montréal, Quebec, Canada. ³Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA. ⁴Faculty of AI and Decision Making, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁵Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. ⁶Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada. ⁷Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, USA. ⁸Translational Data Analytics Institute, The Ohio State University, Columbus, OH, USA. ⁹School of Informatics, University of Edinburgh, Edinburgh, UK. ¹⁰Department of Data Science and AI, Monash University, Clayton, Victoria, Australia. ¹¹School of Computer Science, McGill University, Montréal, Quebec, Canada. ¹²School of Engineering, University of Guelph, Guelph, Ontario, Canada. ¹³Vector Institute, Toronto, Ontario, Canada. ¹⁴School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ¹⁵Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. ¹⁶Department of Electrical & Computer Engineering, The Ohio State University, Columbus, OH, USA. ¹⁷These authors contributed equally: Laura J. Pollock, Justin Kitzes.