Trends in
Ecology & Evolution

CellPress
OPEN ACCESS

Review

Species delimitation 4.0: integrative taxonomy meets artificial intelligence☆

Kevin Karbstein [1,*], Lara Kösters [1], Ladislav Hodač [1], Martin Hofmann [2], Elvira Hörandl [3], Salvatore Tomasello [3], Natascha D. Wagner [3], Brent C. Emerson [4], Dirk C. Albach [5], Stefan Scheu [6,7], Sven Bradler [6], Jan de Vries [8,9,10], Iker Irisarri [11], He Li [12], Pamela Soltis [13], Patrick Mäder [2,14,15], and Jana Wäldchen [1,14]

Although species are central units for biological research, recent findings in genomics are raising awareness that what we call species can be ill-founded entities due to solely morphology-based, regional species descriptions. This particularly applies to groups characterized by intricate evolutionary processes such as hybridization, polyploidy, or asexuality. Here, challenges of current integrative taxonomy (genetics/genomics + morphology + ecology, etc.) become apparent: different favored species concepts, lack of universal characters/markers, missing appropriate analytical tools for intricate evolutionary processes, and highly subjective ranking and fusion of datasets. Now, integrative taxonomy combined with artificial intelligence under a unified species concept can enable automated feature learning and data integration, and thus reduce subjectivity in species delimitation. This approach will likely accelerate revising and unraveling eukaryotic biodiversity.

Highlights

Modern species delimitation is challenged by past morphological descriptions, a mix of applied species concepts, missing tools for complex evolutionary processes and large multi-approach datasets, and non-standardized data integration.

The vision is to have less subjective and standardized species delimitation approaches based on modern integrative taxon-omics under a unified species concept, integrating the discovery of genetic entities with the fusion of automatically extracted information from multi-approach data to find natural taxonomic units.

Artificial Intelligence (AI) approaches have been launched to tackle delimitation issues using classification/clustering methods based on supervised/unsupervised learning, although data fusion and problematic and unknown species represent active fields of research.

AI can help accelerate the revision and unraveling of eukaryotic biodiversity on a scale not seen before.

The species challenge

Globally, >2 million eukaryotic **species** (see Glossary) have been recognized, comprising approximately 1 600 000 animals, 420 000 land plants, and 140 000 fungi and protists [1][i]. Estimates even range from 2 million to 1 trillion existing species [2][i], leaving most of Earth's biodiversity to be described, particularly in animals and protists [1–3]. The description of **taxa** and their recognition as species is a long-standing debate in evolutionary biology. Scientists have been pondering about what a species is now for more than 2000 years [4–7]. Species are the basic units we use to understand evolutionary biology and biodiversity, and the cornerstone of nature conservation and policy-making [5,8,9][i]. However, the assessment of biodiversity has been hampered so far by intricate evolutionary processes, as well as missing objective and reproducible concepts or methods.

The diversity of evolutionary processes asks for integrative concepts

Evolution is driven by mutation, gene flow, genetic drift, and natural selection, leading to adaptation and speciation. However, many eukaryotic groups are characterized by further, partly intricate evolutionary processes. These processes can overlap and interact, resulting in groups that are difficult to distinguish both morphologically and genetically, known as **taxonomically complex groups (TCGs)** [10]. Intricate evolutionary processes in eukaryotes predominantly comprise **hybridization**, **polyploidy**, and **asexuality** (other factors reviewed in [11–13]).

Hybridization

Speciation represents a protracted continuum [14], and emerging species are often incompletely reproductively isolated for several million years, allowing gene flow [15]. Hybridization is thus

[1] Max Planck Institute for Biogeochemistry, Department of Biogeochemical Integration, 07745 Jena, Germany
[2] Technical University of Ilmenau, Institute for Computer and Systems Engineering, 98693 Ilmenau, Germany
[3] University of Göttingen, Albrecht-von-Haller Institute for Plant Sciences, Department of Systematics, Biodiversity and Evolution of Plants (with Herbarium), 37073 Göttingen, Germany

© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

considered a key factor in eukaryotic speciation and diversification [13,16–18]. At least 25% of vascular plants, but only 10% of animals (including hominids) and only a few percent of fungi are known to be involved in hybridization events [13,15,19,20]. In plants with a more frequent biparental organellar inheritance, hybridization is less likely to lead to disruption of nuclear-organellar coadaptations that are essential for core energy production, resulting in more fertile hybrid offspring [21–23]. Increased hybrid numbers in plants may also be attributed to less active, restrictive mating compared to animals (behavior) or fungi (genetic mating types/loci) [5,13]. In general, hybridization has multiple potential outcomes, ranging from infertile/inviable offspring, introgression of adaptive traits, or even extinction of progenitors [18,20–22]. Nevertheless, due to mutation buffering, increased heterozygosity, and hybrid novelty/vigor, hybrids are also able to establish and persist successfully over evolutionary time scales [13,18,20,24–27]. Hybridization leads to network-like evolutionary patterns (**reticulate evolution**), which often violate assumptions of model-based phylogenetic and species delimitation approaches (Tables 1 and 2).

Polyploidy and asexuality

Modern **genomics** has revealed episodes of ancient whole-genome duplication that preceded key innovations in several eukaryotic lineages, especially in flowering plants, all of which share a polyploid common ancestor [13,28,29]. In recent polyploids, multiple gene copies allow for higher physiological and phenotypic flexibility in response to environmental conditions [30,31]. For example, polyploids can better perform in past glaciated areas [25,31–33]. Allopolyploidy (hybridization + polyploidy) is particularly likely to generate novel genomic and phenotypic features, and new species can saltatorially emerge in <100–200 years, or few generations [18,24–26,31,34,35]. Fungi are considered to be largely haploid, and most animals are diploid, with neopolyploids scattered across fishes or insects [13,36]. Animals usually have sex chromosomes [36,37], resulting in distorted chromosome ratios after polyploidization and thus intersterility/infertility [17,26,38]. In contrast, at least 35% of flowering plants are known to be neopolyploid [39], the majority of them having no sex chromosomes [17]. In addition, polyploidy in hermaphroditic plants often leads to self-fertilization and/or asexual reproduction, ensuring reproductive success [38]. Species delimitation approaches are mainly based on diploid models (Tables 1 and 2), and thus cannot satisfactorily handle high intragenomic variability or complex origins of polyploids.

Asexuality is closely linked to hybridization and/or polyploidy in eukaryotes, and represents a modification of the sexual pathway [12,26,36,40–42]. Eukaryotes reproduce asexually, for example, via unfertilized egg cells (parthenogenesis, animals), clonal seeds (apomixis, plants), or haploid spores (sporulation, fungi) [43]. In fungi, exclusive asexual reproduction is known for 20% of species [13]. In plants and animals, asexual reproduction occurs in <1% of species, although asexual species are important in many ecosystems (e.g., soil mites or dandelions). Without sex, each individual can evolve to a distinct uniparentally reproducing lineage, form populations, and might be theoretically considered a species. However, facultative asexuality allows for a return to sexual recombination, which can result in highly reticulate complexes with hundreds of morphotypes/taxa [26,43]. In evolutionary time scales, asexuals can be successful and may re-evolve to a functional diploid, sexual state [16,30,36,44]. Species delimitation approaches must consequently consider genetic variation, differentiation, and stability of asexual lineages to properly reconstruct their relationships.

Species concepts

Since Darwin's and Wallace's articulation of a theory of evolution, many species concepts have been controversially discussed for eukaryotic groups [6,13,45,46], leading to different **species delimitation** results and thus species *per se* (Box 1). The confusion surrounding this topic was recognized by De Queiroz [6], who revolutionized **taxonomy** by strictly separating the species

[4]Institute of Natural Products and Agrobiology (IPNA-CSIC), Island Ecology and Evolution Research Group, 38206 La Laguna, Tenerife, Canary Islands, Spain
[5]Carl von Ossietzky-Universität Oldenburg, Institute of Biology and Environmental Science, 26129 Oldenburg, Germany
[6]University of Göttingen, Johann-Friedrich-Blumenbach Institute of Zoology and Anthropology, 37073 Göttingen, Germany
[7]University of Göttingen, Centre of Biodiversity and Sustainable Land Use (CBL), 37073 Göttingen, Germany
[8]University of Göttingen, Institute for Microbiology and Genetics, Department of Applied Bioinformatics, 37077 Göttingen, Germany
[9]University of Göttingen, Campus Institute Data Science (CIDAS), 37077 Göttingen, Germany
[10]University of Göttingen, Göttingen Center for Molecular Biosciences (GZMB), Department of Applied Bioinformatics, 37077 Göttingen, Germany
[11]Leibniz Institute for the Analysis of Biodiversity Change (LIB), Centre for Molecular Biodiversity Research, Phylogenomics Section, Museum of Nature, 20146 Hamburg, Germany
[12]Eastern China Conservation Centre for Wild Endangered Plant Resources, Chenshan Botanical Garden, 201602 Shanghai, China
[13]University of Florida, Florida Museum of Natural History, 32611 Gainesville, USA
[14]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Puschstrasse 4, 04103 Leipzig, Germany
[15]Friedrich Schiller University Jena, Faculty of Biological Sciences, Institute of Ecology and Evolution, Philosophenweg 16, 07743 Jena, Germany

☆ **Twitter Accounts**
Consortium at Max Planck Institute for Biogeochemistry and TU Ilmenau working on AI in ecological and evolutionary research (authors belonging to this consortium: K.K., L.H., L.K., M.H., P.M., J.W.): https://twitter.com/Flora_Incognita

*Correspondence:
kkarb@bgc-jena.mpg.de (K. Karbstein).

concept from delimitation. Many modern authors agree that species represent distinct genetic ancestor-descendant lineages, interconnected by populations throughout time and/or space [6,8,26,47–50]. Previous concepts are now treated as operational criteria, allowing the application of De Queiroz's **unified species concept (USC)** [6,45,51]. Genetically distinct eukaryotic lineages characterized by persistence in time and space, with its individuals sharing a common phenotype, ecological niche, or behavior, would thus be accepted as species. Unlike other concepts, the USC universally applies to (a)sexual di- and polyploids, and/or hybrids [8,45,47,52,53]. Nevertheless, past species descriptions, data collection, and bioinformatic implementations of intricate evolutionary processes remain a persistent challenge for biologists. Species delimitation has thus been a mixture of several approaches, which are applied until now and partly in parallel, and which we classify as species delimitation 1.0–4.0 (Figure 1A–D).

Species delimitation 1.0

Traditionally, the most commonly applied delimitation criterion is morphology, which has been occasionally supplemented by other datatypes (Figure 1A; [5,52]). For many **morphospecies** outside TCGs, the classification is still valid and could be confirmed by modern methods [19,54,55]. However, there are several challenges when using morphology-based delimitation and also identification. Morphospecies do not necessarily represent natural evolutionary entities, for example, due to the high subjectivity of selected diagnostic characters or potentially overinterpreted, (epi)genetically based plasticity in relation to environmental influences [5,25,52], or because specific morphotypes can originate multiple times, or underrepresent genome-level diversification **(cryptic species)** [24,25,47,48,56–58]. Species diversity may therefore be over- or underestimated (e.g., 5 lineages instead of 12 buttercup morphospecies [47]; 7 lineages instead of 1 beetle morphospecies [56]). In addition, the existence of cryptic species is likely to underpin a vast underestimation of biodiversity (e.g., probably >80% of insects are undescribed [2,59]). Different taxonomic treatments also can substantially impact species numbers of regional floras (e.g., increase by 50–100%, when taxa of a few large European asexual TCGs are considered as species, [60,61]). All of this is highly problematic for biodiversity assessment, conservation biology, and ecosystem modeling.

Species delimitation 2.0

Beyond traditional taxonomy, new species concepts proposed between the 1960s and 1990s improved the theoretical background of delimitation (Box 1). Within this period, greater emphasis on genetic data has provided an increasingly reliable view of species evolution [5,46,65]. Approaches relying on single- or multi-**locus** datasets [e.g., mtDNA (*COI*), cpDNA (*rbcL*, *matK*), or nuclear sequences (*ITS*, microsatellites) [5,13,58]] became the primary approach for reconstructing **population genetic** and **phylogenetic** relationships with **bioinformatic** tools since the 1990s (Figure 1B). In parallel, morphological descriptions were enhanced with (geometric) morphometric methods to better define levels of (dis)similarity; cytological data (chromosome counts) were improved by modern techniques for estimating ploidy and reproductive modes; ecological and behavioral data were quantified by multivariate statistics to better circumscribe niche separation among species [5].

Approaches have typically been applied in parallel and manually weighted by authors to make taxonomic decisions, but rarely integrated with reproducible analytical tools (Table 1). Awareness substantially rose that multiple, complementary datasets are needed to describe or revise species, and to reduce approach-specific failure rates [24,25,47,51,52,66]. Species delimitation solely based on genetic data (also including eDNA [59]), as exemplified by the widely applied **multispecies coalescent (MSC)** model, tends to infer populations or subspecies as independent evolutionary lineages and overestimates species numbers [8,67,68]. Morphological criteria also have little discriminatory power in cases of cryptic diversity. Ecological, chemical, or behavioral criteria often provide insufficient resolution in groups with highly overlapping, hybridogenous diversity. Knowledge

Glossary

Artificial intelligence (AI): technology that aims to simulate animal/human intelligence.

Artificial neural network (ANN): within ML, ANNs are interconnected neurons organized into input, hidden, and output layers.

Asexuality: type of reproduction in which meiotic recombination and merging of genomes from two parents is absent, resulting in offspring that are genetically identical to a single parent.

Bioinformatics: interdisciplinary field that develops methods and tools for analyzing large biological datasets.

Convolutional neural network (CNN): within DL, CNNs represent a type of neural network that extracts information through convolutions for feature learning. Convolutions are specialized neurons, with a small fixed-sized receptive filter sliding over an input tensor (matrix), summing all multiplied values at every slide position.

Cryptic species: different genetic lineages showing the same/very similar morphology.

Deep learning (DL): within ML, DL represents ANNs with feature learning.

Gene/locus: specific coding/general region on a DNA strand.

Genomics: study of the structure, function, and evolution of genomes.

High-throughput sequencing (HTS): DNA sequencing technologies in a massively parallelized manner, providing fast and cost-effective methods.

Hybridization: fusion of previously diverged genomes.

Machine learning (ML): within AI, ML aims to recognize patterns in data and learn from them in order to make predictions.

Morphospecies: species described exclusively or predominantly based on morphological characters.

Multispecies coalescent (MSC): stochastic process model that describes the genealogical relationships of DNA sequences (or alleles) sampled from multiple species.

Polyploidy: presence of more than two sets of chromosomes in a nucleus, often referred to as whole-genome duplication.

Population genetics/-omics and phylogenetics/-omics: research fields that study the evolutionary relationships among groups of populations/species using genetic/(sub)genomic data.

about genetics, ploidy levels, morphology, and geography (e.g., isolation by mountains/oceans) can indicate whether lineages represent reproductively isolated, stable geno- and phenotypes with specific environmental adaptations. Consequently, **integrative taxonomy** approaches that systematically combine genetics with other sources of evidence allow for greater confidence and less subjectivity in species delimitation [13,25,52,53,66,69].

Species delimitation 3.0

Despite integrative taxonomy, the speciation process still challenges taxonomists in deciding whether observed entities represent subspecies or species. Increasing knowledge about intricate evolutionary processes requires highly informative multi-approach datasets evaluated by efficient bioinformatic tools. Information-rich **high-throughput sequencing (HTS)** data and new or improved bioinformatic tools are now available. This astonishing progress has resulted in a plethora of taxonomic descriptions and revisions [47,48,64,70], and has led to a significant increase in awareness that morphospecies particularly in TCGs often represent ill-founded entities [24,47,48,56,57]. **Integrative taxon-omics** approaches that innovatively combine taxonomy with 21st century large-scale -omics and other complementary data sources therefore started to become the new gold standard for species delimitation in TCGs (Figure 1C) [24,25,53,56,70].

Two main HTS techniques are currently applied: **reduced-representation sequencing (RRS)** and to a lesser extent **whole genome (re)sequencing (WGS/WGR)** [24,26,71]. These (sub) genomic approaches result in a large number of informative markers that enable us to address complex evolutionary questions (Table 1), which was not feasible with a more limited number of loci before. HTS typically provides genetic information from tens to hundreds of thousands of **SNPs**, or hundreds of locus/**gene** sequences [24,71–76]. In delimitation 2.0–3.0 studies, both SNPs and genes are used to study phylogenetic tree conflicts, which can deliver the first evidence for reticulate evolution [11,47,50,56,76]. Analyses of SNPs via genetic structure and network approaches proved to be efficient in detecting interspecific gene flow, post-origin evolution, and lineage stability, while allelic information from nuclear genes also helps to clarify hybrid origins and delimitation of a/sexual di- and polyploids under the MSC [19,24,71–73,77,78]. In addition, HTS often codelivers mitochondrial or plastid (organellar) DNA, which can be helpful in delimiting species in certain hybridogenous eukaryotic TCGs, as these markers allow us to detect nuclear-organellar tree conflicts and maternal/paternal or even extinct progenitors [19,24,56,79].

Species delimitation 4.0

The models used so far in taxonomic approaches are based on predefined human model assumptions simplifying complexity to make biological processes assessable or predictable. For example, the MSC requires no hybridization, non-saltatory evolutionary rates, and random mating within species [14,91], which is only the case in tree-like speciation of diploid, sexually reproducing organisms. For taxonomic treatments, other challenges relate to the ranking of results according to their importance and dataset disagreement (e.g., genotype-morphotype mismatches due to intricate evolution [25,66]), or the favored species concepts (e.g., USC, BSC, or clustering concepts; Box 1). The theory for a lineage-species concept and integrative taxonomy is widely accepted [8,13,25,45,51,64,66], but in practice this still depends on implementability and author preferences yielding a mix of applied species concepts and thus delimitation results. New approaches would be highly desirable for standardized data evaluation that is independent of the focal species group, and for automating the processing of large datasets, including feature extraction, learning, and data integration (fusion) in feasible time frames.

Combining taxonomy with **artificial intelligence (AI)** may help delimit species in a less subjective and more integrative and rapid way. The vision is an integrative USC (iUSC) that uses

Reduced-representation sequencing (RRS): techniques that generate a subset of the genome using either random or targeted approaches: RAD-Seq/GBS methods use restriction enzymes and result in genome-wide SNP datasets; target enrichment (TEG) methods are based on a collection of hundreds of nuclear genes selected from a target reference.
Reticulate evolution: origin of a new lineage by the (partial) fusion of two or more ancestral lineages. It represents evolutionary processes that cannot be described as bifurcating trees.
SNP: single nucleotide polymorphism. Specific site that shows base variation among aligned genetic sequences.
Species: basic unit of systematic biology, and the result of the speciation process. Current research is focused on finding universal definitions (e.g., the USC).
(Unified) species concept (USC): a theoretical concept for defining a species, or species in general across all organism groups.
Species delimitation (1.0-4.0): process of inferring boundaries among sampled individuals and determining whether they belong to different species. We use numbers 1.0–4.0 to classify species delimitation developments.
Taxon: rankless taxonomic unit of organisms.
(Integrative) taxonomy/-omics: branch of systematics concerned with the documentation, classification, and naming of biodiversity. 'Integrative', when based on multi-approach data, and 'omics', when (sub)genomic data are used. Systematics is the study of the evolutionary history and relationships of organisms.
Taxonomically complex groups (TCGs): group of related individuals characterized by intricate evolutionary processes that complicate species delimitation.
Whole genome (re)sequencing (WGS/WGR): HTS techniques to sequence entire genomes, either *de novo* or using a reference genome.

Table 1. Genetic datasets evaluated by appropriate bioinformatic tools and supplemented by other data sources, to delimit eukaryotic species based on modern integrative taxonomy (species delimitation 2.0 and 3.0)[a]

Genetic dataset	Phylogenetic/-omic analysis	Examples of bioinformatic tools	Evolutionary processes	Can be integrated with	Evolutionary stage	Examples of recent studies
Genome-wide SNPs (e.g., from RAD-Seq, WGS/WGR)	Tree (tree-like evolution) Structure Network SNP origin (reticulate evolution)	Astral-III [80], IQ-Tree2 [81], RAxML_NG [82] SNPs: PhyloNetworks [83], RADpainter [84], SNiPloid [85], sNMF [86], Structure [87] Genes under MSC: BPP, iBPP [88,89] BFD*, DISSECT [90,91] PhyloNet incl. MPAllopp [92], Stacey [93]	Evolution of diploids (progenitors), sexuals, (obligate asexual) autopolyploids, and detection of incongruences hinting at reticulate evolution For example, SNPs: interspecific gene flow (introgression), hybrid origins, hybrid/polyploid postorigin evolution (e.g., lineage composition or stability) For example, genes under MSC: exact hybrid origins (parental contributions), and delimitation of a/sexual di- and polyploid lineages	Morphology (all groups) Biogeography and distribution (all groups) Ecology (all groups) Ploidy (predominantly plants) Reproduction and recombination (all groups) Physiology and chemistry [predominantly plants (algae) and fungi]	Early to middle species divergence	[19,24,48,56,64,76,94,95]
(Phased) nuclear genes (e.g., from TEG, WGS/WGR)					Middle to late species divergence	[25,47,48,50,57,94]
Mitochondrial regions (e.g., from TEG, WGS/WGR) Plastid regions (e.g., from TEG, WGS/WGR)	Tree Network (verification of progenitors, further evidence for reticulate evolution)	RAxML_NG [82], TCS [96]	Detection of extinct species, progenitors, or lineages Nuclear-plastid discordance (reticulate evolution)		Early to late species divergence	[19,24,48,56,77,79]

[a] Cited studies usually perform a mix of 2.0 and 3.0 approaches, which is a combination of single-/multi-locus and subgenomic datasets. The bioinformatic tools listed are related to species delimitation in various ways: only a few approaches are strict delimitation approaches (i.e., assigning individuals into groups/species, such as RADpainter, sNMF, Structure, (i)BPP, BFD*, DISSECT, and Stacey), whereas many others provide rather indirect evidence for delimitation (e.g., support statistics of tree-building approaches). More bioinformatic tools and recent studies can be found on FigShare upon publication (**https://doi.org/10.6084/m9.figshare.23815407**).

Table 2. Summary of the most popular MSC, and classical and current ML approaches developed for species classification[a–d]

Approach[a]	Analytical framework[b]	Testing datasets[d]	Approach suitable for?				Refs
			Diploids/ polyploids	Sexuals/ asexuals	Tree-like/ reticulate evolution	Data integration tested?	
e.g., BFD*, BPP, DISSECT, iBPP, SPEEDEMON, Stacey	MSC	genetic (partly simulated, single to several genes/loci) + partly morphometric animal datasets	yes/ yes (only autopolyploid)	yes / - (only autopolyploid)	yes / - (only Stacey)	no (only iBPP, SPEEDEMON potentially)	[88–91,93,118]
DELINEATE	MSC tree + PBD	genetic (simulated)	yes / -	yes / -	yes / yes (only within the speciation process)	no	[14]
mix of ML approaches (unsupervised)	classical ML (RF, VAE, t-SNE)	genetic (up to 1k SNPs) animal dataset	yes / yes (potentially)	yes / yes (potentially)	yes / yes (potentially)	no	[55]
CLADES (supervised)	classical ML (SVM)[c]	genetic - subgenomic (few to hundreds of genes) animal datasets	yes / -	yes / - (potentially if diploid)	yes / yes (only within speciation process)	no	[106]
delimitR (supervised)	classical ML (SFS + RF)[c]	genetic - subgenomic (up to 20k SNPs) animal datasets	yes / -	yes / - (potentially if diploid)	yes / yes (only within speciation process)	no	[105]
MMNet (supervised)	DL (CNNs)	genetic - subgenomic (few genes, up to 10k SNPs) + image datasets with several animal and a single plant group/s	yes / yes (potentially)	yes / yes (potentially)	yes / yes (potentially)	yes (fusion)	[109]

[a] 'Supervised' or 'unsupervised' indicates that the ML classifier is trained with or without labeled species data, respectively.
[b] Abbreviations: CNNs, convolutional neural networks; DL, deep learning; PBD, protracted birth death model; RF, random forest classifier; SFS + RF, site frequency spectrum + random forest classifier; SVM, support vector machine; t-SNE, distributed stochastic neighbor embedding; VAE, variational autoencoder.
[c] Analytical frameworks: trained classifier based on species evolve (CLADES)/classifier-based model selection (delimitR) under standard diploid population genetic models (coalescent theory, F-statistics).
[d] Testing datasets: genetic, species delimitation 2.0; subgenomic, species delimitation 3.0

AI to integrate the genetic lineage concept with operational criteria and species hypothesis testing (Figure 1D) within the following steps: (i) examine genetic diversity, stability, and differentiation of lineages (observed as phyla/clusters/discontinuities); (ii) describe the most likely lineage-species scenarios as hypotheses; (iii) add information from the taxonomically most important criteria to reduce criterion-dependent failures and clarify the evolutionary role; and (iv) select the most likely hypothesis as species scenario. All datatypes can be used in AI systems as long as they are transformable into numeric values (Figure 2A,B) [97], and can be efficiently supplemented by collection data (e.g., molecular [98], morphological [25], or spectral ploidy [99]) and information from critically examined online databases for genetics (e.g., NCBI or BOLD), images (e.g., BOLD, GBIF, iDigBio, or citizen science such as iNaturalist, Flora Incognita, etc.), ploidy (e.g., Plant DNA C-value Database or ploiDB), or biogeography/ecology (e.g., WorldClim).

Machine learning for biology

Alongside popular applications such as text generation (e.g., ChatGPT), medical diagnostics, or self-driving cars, **machine learning (ML)** has gained increasing attention in biological research. In contrast to previous model-based approaches (e.g., MSC), ML models aim to recognize patterns in data and learn from them to make predictions [97,101]. Pioneering ML applications include automated species identification using images or sound, DNA variant calling, or ploidy

Box 1. Species concepts, limitations, and modern integrative taxonomy

Species delimitation depends on the applied concept, that is, the theoretical framework to define a species. Most concepts agree to treat species as separately evolving metapopulation lineages, but differ in the criteria to define lineage characteristics [6]. More than 30 concepts have been applied to different eukaryotic groups, developed around phenomena such as reproductive isolation, shared origin (monophyly), morphological/genetic cohesion, or phenotypic similarity, acting as criteria for delimitation. Consequently, each concept has its limitations, and different criteria to define or delimit species often deliver contradictory species delimitation results [6,13,45].

For example, the biological species concept (BSC) [4] defines species as reproductively isolated, interbreeding populations. Sexual reproduction keeps lineages together and separates lineages by successive establishment through selective pressures for mating compatibility. The BSC lacks suitability when, for example, incomplete crossing barriers exist with related but morphologically clearly distinct species, or in obligate asexual or self-compatible taxa. Phylogenetic species concepts (PSCs) [6,46] define species by shared ancestry (monophyly), and can be applied to both sexual and obligate asexual taxa. However, the PSC cannot be applied, for example, to diploid groups with recently originated non-monophyletic, auto- and allopolyploid species [62]. Genetic [63] and morphological/phenetic [6,46] cluster concepts treat distinguishable clusters without intermediates as species, and are particularly appropriate for TCGs.

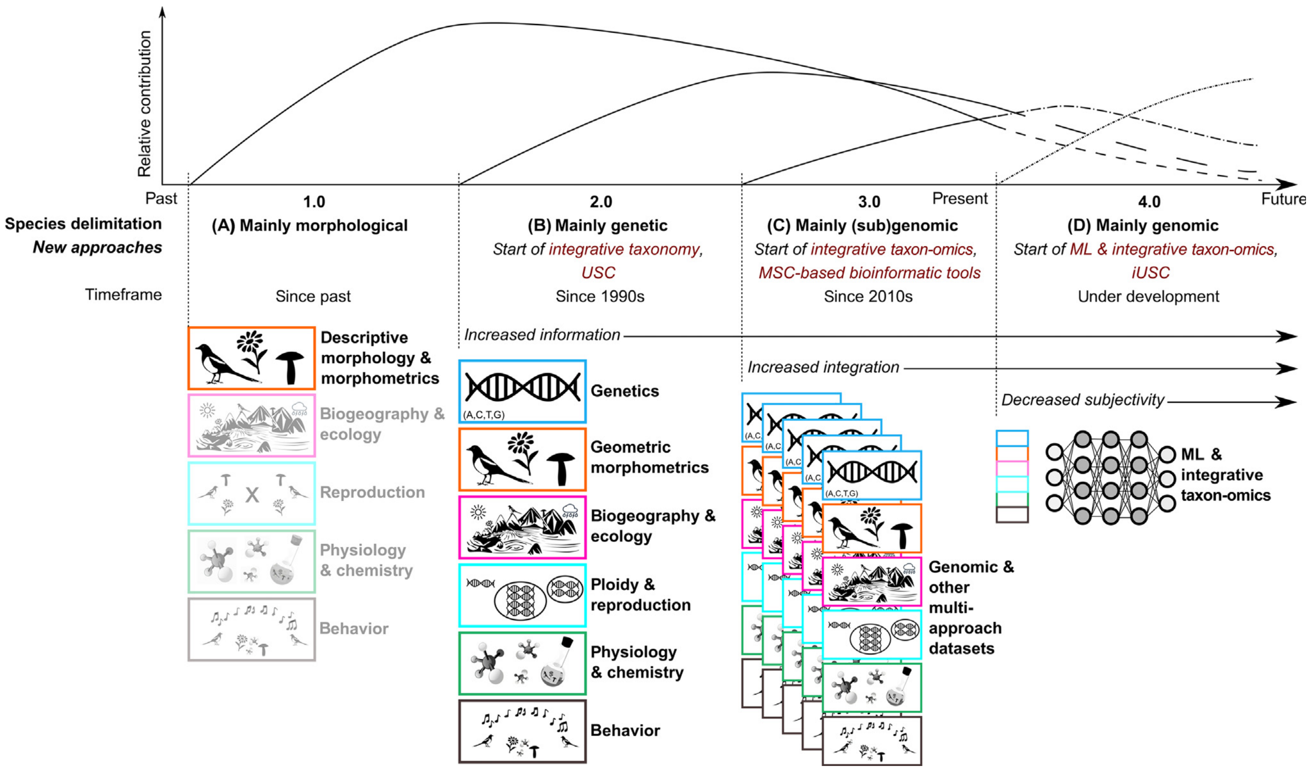
Recent species delimitation studies have often used a combination of different criteria for different datasets to delimit a specific group of interest and to follow an integrative taxonomic approach. For example, genetic-based criteria (e.g., in PSCs) are well suited for animals and fungi because species become reproductively isolated efficiently or are asexual leading to tree-like evolutionary patterns, or are only genetically recognizable in the case of cryptic species [13,46,49,50,56,58]. Particularly in flowering plant groups, a combination of criteria is needed because reproductive isolation is often incomplete, leading to phenotypically diverse, but macroscopically recognizable species evolving in a reticulate manner [5,26,45,47,64].

In the past, and still today, species have been described and distinguished primarily on the basis of morphological differences (species delimitation 1.0). Using modern integrative taxonomic approaches, we know that morphospecies, especially in TCGs, often lack distinctive genetic, ecological, and even morphological features. They are the main target of current taxonomic research (species delimitation 2.0/3.0). However, the interpretation of results from multiple approaches and datasets for final taxonomic treatments remains highly author-dependent, but AI can help to enable less subjective and more integrative species delimitation (species delimitation 4.0).

estimation [59,97,99,100]. Recent ML developments predominantly rely on **deep learning (DL)**, which represents automated feature extraction and learning based on **artificial neural networks (ANNs)** [102]. DL needs no prior expert knowledge, which is particularly advantageous for delimiting species in TCGs where discriminative features are difficult to identify or taxonomic expertise is not available. ANNs are inspired by animal brains: the neuron as the basic unit represents the sum of all inputs multiplied by their trainable weight and bias factors that is activated by a nonlinear function (neuron firing), with all neurons organized into multiple, interconnected layers (Figure 2C) [97,103]. Conceptually speaking, ANNs map an input to the (desired) output based on different learning and optimization strategies. Species delimitation, treated as a classification or clustering task, can therefore be performed using features learned from labeled data (supervised), unlabeled data using only the inherent structure of the data (unsupervised), or a mix of both (semi-supervised; Figure 2C) [103,104].

ML for species delimitation

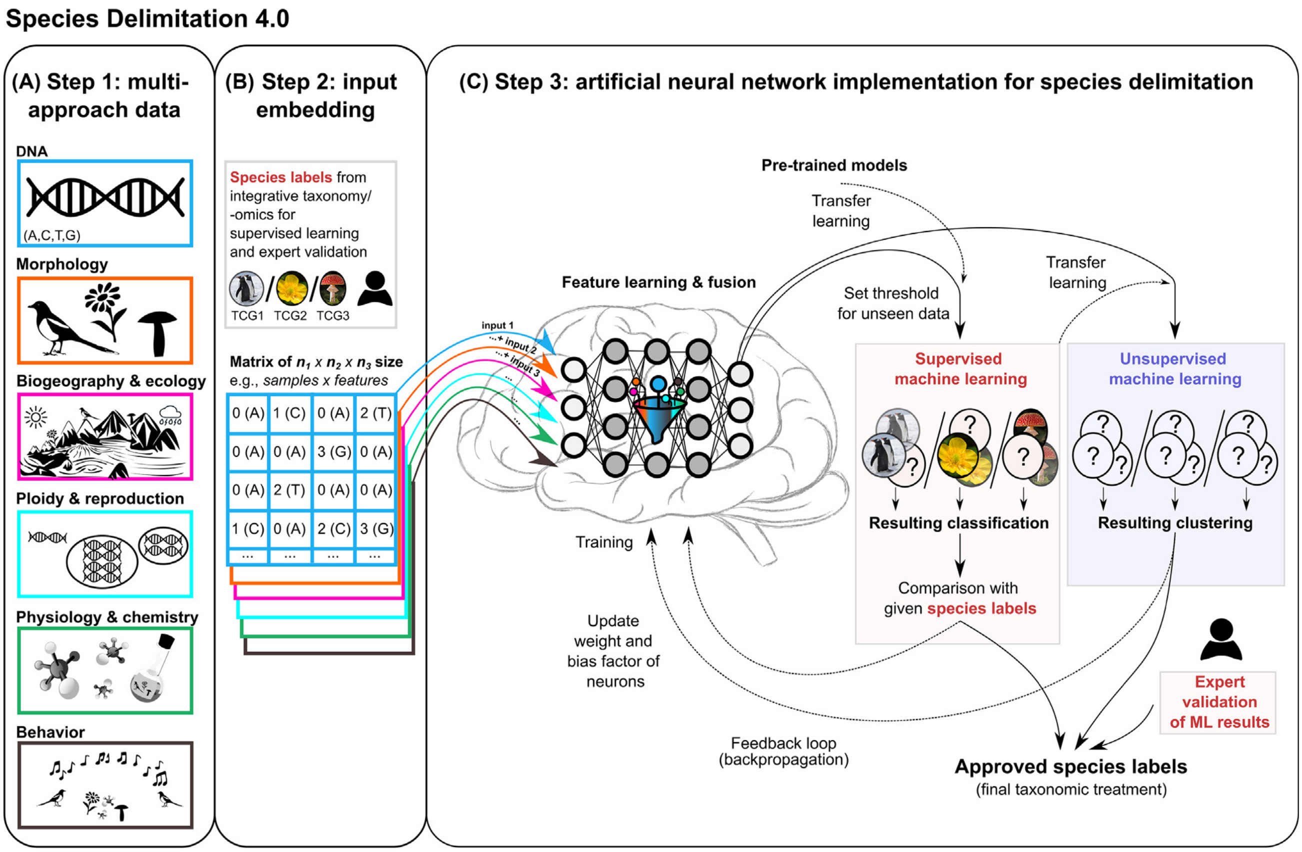
Tools such as (i)BPP, DISSECT, or Stacey [88,89,91,93] are currently the most reliable and trusted approaches for species delimitation 2.0/3.0, but their applicability suffers from high computational effort for locus- or species-rich datasets and mentioned biological limitations of the MSC. The first promising, supervised (e.g., delimitR or CLADES) and unsupervised (e.g., RF or t-SNE) attempts have been made using predominantly classical ML building on animal genetic data, which were partly supplemented by phylogenetics and morphology [55,69,105,106]. However, these methods are often not suitable for integrative taxonomy or TCGs, for example, due to disregard of gene flow, implementation for few-locus-based, diploid animal genetic data only, and/or lack of strategies for dataset fusion.



Trends in Ecology & Evolution

Figure 1. Past to present to future developments in species delimitation approaches, including dataset descriptions and integration. (A) For >2000 years, people have been describing species mainly based on their morphology (1.0). (B) Since the 1990s, genetic data have started to become the primary source for species delimitation, including the boosted development of integrative taxonomy and a unified species concept (USC; 2.0). (C) Since the 2010s, genetics turned into genomics, including the development of sophisticated, multispecies coalescent (MSC)-based bioinformatics and integrative taxon-omics for inferring eukaryotic species delimitation (3.0). (D) Currently, work is underway to develop machine learning (ML)-based strategies for species delimitation, and in this study, ML-based integrative taxon-omics under a unified species concept (iUSC) for species delimitation (4.0). Approaches 1.0–4.0 have been used to date (see details in the main text), and their relative application for species description and delimitation over time is approximately estimated. Images, except for the artificial neural network scheme, are from pixabay.com (free to use under the Content License).

ANN architectures such as **convolutional neural networks (CNNs)** are currently most suitable for molecular, image, video, or environmental classification tasks [97,100,102,109]. Convolutions represent specialized neurons, with a small fixed-sized receptive filter sliding over input matrices [97,102]. CNNs thus enable the extraction of spatially autocorrelated, hierarchical, and subtle biological patterns with low optimization effort; they efficiently learn the importance of extracted features via unique character combinations and under intraspecific variation [97,102,107,108]. Other network architectures like recurrent neural networks (RNNs) [102] are applied to sequential prediction tasks (e.g., DNA base calling or songbird classification) [97,102]. However, with long inputs, RNNs usually forget initially learned features. Recent transformer networks outperform specific RNN/CNN tasks [97], but their computational cost grows with input length. RNNs and transformers (to date) are therefore impractical for large integrative taxonomic datasets. CNNs have recently been shown to efficiently and accurately integrate genetic and morphological data to identify closely related, diploid animal species [109]. However, the fusion of large genomic datasets with other data sources and the ability to analyze unseen species, as is usually the case in species delimitation, are missing so far. Consequently, new ML approaches need to be developed for species discovery and delimitation.


Species Delimitation 4.0
(A) Step 1: multi-approach data
DNA
(A,C,T,G)
Morphology
Biogeography & ecology
Ploidy & reproduction
Physiology & chemistry
Behavior
(B) Step 2: input embedding
Species labels from integrative taxonomy/-omics for supervised learning and expert validation
TCG1 TCG2 TCG3
Matrix of n_1 x n_2 x n_3 size
e.g., samples x features
0 (A) 1 (C) 0 (A) 2 (T)
0 (A) 0 (A) 3 (G) 0 (A)
0 (A) 2 (T) 0 (A) 0 (A)
1 (C) 0 (A) 2 (C) 3 (G)
... ...
(C) Step 3: artificial neural network implementation for species delimitation
input 1
...+ input 2
...+ input 3
Pre-trained models
Transfer learning
Feature learning & fusion
Set threshold for unseen data
Transfer learning
Supervised machine learning
Unsupervised machine learning
Resulting classification
Resulting clustering
Comparison with given species labels
Training
Update weight and bias factor of neurons
Feedback loop (backpropagation)
Expert validation of ML results
Approved species labels
(final taxonomic treatment)

Trends in Ecology & Evolution

Figure 2. Scheme of ML-based implementation for integrative taxonomy under the USC (iUSC). (A) Step 1: collecting multi-approach datasets. (B) Step 2: transforming/embedding the data as numeric matrices to use as input for the first ANN layer. (C) Step 3: running the ANN for feature extraction using supervised learning with labeled training data from integrative taxonomy/-omics results or unsupervised learning without labeled training data (or both, semi-supervised) [97,109], including data ranking and fusion (first DNA, then merging with other datasets), and final species classification/clustering as output of the last ANN layer. ANNs with supervised learning are trained by adjusting layer weights and bias factors so that the predicted label matches the expected label ('ground truth'). This feedback loop is called backpropagation, which also can be applied in unsupervised ANNs using different optimization strategies [102,103]. Finally, outputs should be validated by taxonomic experts. As examples, we illustrate datasets of intricate eukaryotic species complexes (cryptic speciation in gentoo penguins or fly agaric complex, and young speciation in the goldilocks buttercup complex): *Pygoscelis papua* (gentoo penguins), *Ranunculus auricomus* (goldilocks buttercup plants), and *Amanita muscaria* (fly agaric mushrooms), with taxonomically unrevised/problematic, or unknown species within groups highlighted as question marks. ANNs based on supervised learning are applicable to datasets with low levels of problematic or unknown species, while ANNs based on unsupervised learning are more suitable for datasets without reliable species labels. Images of symbols (A) and of species (B, C) from pixabay.com (free to use under the Content License), *Ranunculus* images from Kevin Karbstein. Abbreviations: ANN, artificial neural network; iUSC, integrative unified species concept; ML, machine learning.

In general, a basic ML model would be highly desirable for eukaryotic species delimitation (Figure 2C). To achieve this along with optimal model accuracy, selecting biological and phylogenetically diverse, coarse- to fine-grained (e.g., sampling within different families or family/genus of the examined group, and TCGs) datasets is highly recommended for learning processes. The idea is that this pretrained ML model ascertains basic ancestor-descendant lineage (evolutionary) relationships but also biological features of hybridizing, polyploid, or asexual species within TCGs important for the delimitation process. Via transfer learning [103,110], pretrained ML models can be used as starting points to fine-tune other supervised ML models or for further unsupervised delimitation in specific TCGs. Nevertheless, species labels and features not included in training processes represent an ongoing challenge for ML approaches. In supervised learning, generating data-specific learned thresholds for identifying out-of-distribution observations that substantially deviate from trained taxa can be applied, among other techniques, to delimit groups with low levels of unknown taxa [107,109,111]. Emerging unsupervised ANNs that operate without potentially biased training labels may be more powerful, less subjective extensions to existing approaches in discovering unknown species via clusters and gaps (Figure 2C), such as, similarity learning or deep clustering [112,113]. These approaches do not necessarily need intensive (phylo)genomic work and taxonomically revised treatments before running the model.

In addition, ranking of datasets and timing of data fusion are also critical for reliable ML approaches, and represent active fields of research. Data fusion is known to improve model accuracy, and is especially relevant for species groups characterized by high intraspecific variation but low interspecific differentiation [109,111]. To follow an iUSC, genomic data should be evaluated first, and the most likely species scenarios can then be fused with other datasets (e.g., genomics, genomics + morphology, genomics + morphology + ecology) (Figure 2C). As this process is highly complex, future ML delimitation approaches particularly need to focus on incremental weighting and ranking of genomic data with further multi-approach data for reliable species delimitation. To quantify uncertainty of ML findings, specific scores (e.g., accuracy score) and confusion matrices (e.g., predicted vs. true species labels) for supervised learning [107,111], or bootstrap sampling with replacement for unsupervised learning (or similar techniques) can be applied [104]. To identify highly important features for the ML classification or clustering process (e.g., specific DNA or image regions), and to recognize similar or hitherto taxonomically unrecognized features, explainable AI approaches such as Grad-CAM [114,117] are increasingly available and can be supported by multivariate statistics based on feature correlation with previous phylogenomic, morphometric, or ecological results.

However, the quality of ML-model predictions strongly relates to the quality of provided datasets. Although ML models continuously improve due to richer datasets and more efficient ML architectures, there are still some pitfalls. For example, training an ML model with a few DNA sequences that happen to be duplicates or that do not capture species variability can lead to overfitting or loss of separability between closely related species. Rare species entries or strict dataset filtering also often result in uneven or generally low sample sizes, negatively impacting model performance. Statistically meaningful training, validation, and testing of ML models require at least a few samples each (>5–10 samples/species recommended, depending on intraspecific variation and species number [109,115]), which has often not been the case in taxonomic studies. Another problem relates to artificial gaps or false overlaps in DNA alignments or images with different colors or backgrounds leading to spurious features being recognized by the ML model [115,116].

There are some optimization strategies to handle dataset bias (e.g., training with taxonomically revised species groups, or loci selection), uneven sample size (e.g., data augmentation techniques), or dataset noise (e.g., standardized image backgrounds or DNA alignment workflows) [97,109,115], but these issues cannot be eliminated entirely. Moreover, there is probably no universal approach that could perform optimally for all given species delimitation tasks ('no-free lunch' theorem, [103]), and thus no one-size-fits-all solution. For example, new delimitation approaches based on unsupervised ML models may suffer from missing biological theory or information content, and semi-supervised ML models trained with highly accurate integrative taxonomic data could outperform them in scenarios with low levels of unknown species. In addition, the learning process of ML models is still a 'black box' for human observers. Although new explainable

AI (XAI) approaches are emerging to visualize and detect biological features learned by the ML model [117], previous delimitation results from integrative taxonomy are needed, as well as experienced taxonomists to control and validate ML-based species classification or clustering, and to find biological explanations in the background of group-related evolutionary and ecological hypotheses.

Concluding remarks

Several unresolved tasks remain (see Outstanding questions). Taxonomic workflows are still challenging due to missing integrative tools to support decisions as to whether lineages represent populations, subspecies, or species. High sampling effort for multi-approach datasets, bioinformatic skills, and the cost of DNA sequencing and laboratory equipment are particularly problematic for traditional alpha-taxonomists, the number of which is steadily decreasing, with a concomitant global decline in taxonomic expertise. A new generation of biodiversity scientists should be trained to combine expertise in methods from species delimitation 1.0–4.0, to meet the need of documenting species in a world of declining biodiversity. Increased standardization, automatization, and public support will be required for this task. We also need widely accepted integrative ways of dealing with a unified species concept (e.g., iUSC) to (re)define species as objectively and naturally as possible for biodiversity research. This should be followed by rapid automatic integration into biodiversity databases. Now is the time for evolutionary biologists to face current challenges in taxonomy, for example, resolving hundreds of synonyms or doubtful names per described species, or covering the need of naming the global biodiversity.

Integrative taxonomy based on ML may help to delimit species less subjectively as well as more reliably and rapidly than traditional methods do and, may therefore help to revise and unravel the eukaryotic diversity on a global scale. Basic ML networks pave the way for broader applicability across eukaryotes and act as a starting point for delimitation at lower taxonomic levels. Because no single universal genetic marker for species delimitation among all eukaryotes exists, multigenomic (genomic, nuclear, and organellar) data will be needed and should be combined with nonmolecular data. It is not clear yet how machines will best tackle species delimitation (semi-/supervised vs. unsupervised learning), handle disagreement among datasets, or rank and fuse data. However, these approaches are under current development and require further research.

Acknowledgments

K.K., L.K., P.M., and J.W.'s research is supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS20062, M.H. by the German Ministry of Agriculture and Food (BMEL-BLE) grant 2819NA106, and L.H. by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety, and Consumer Protection grant 67KI2086. K.K. also received funding from the German Research Foundation (DFG) Priority Program Taxon-Omics SPP1991 funding line Seed Money for Post-Doc Grants. E.H., S.T., and N.W. are supported by the DFG priority program Taxon-Omics SPP1991 grants Ho4395/10-2, To1400/1-1, and Wa3684/2-1, respectively. J.d.V. is supported by the DFG priority program SPP 2237 MAdLand grant VR 132/4-1. J.d.V. further thanks the European Research Council for funding under the EU Horizon 2020 Research and Innovation Program (Grant Agreement No. 852725; ERC-StG 'TerreStriAL'). H.L. is supported by the National Natural Science Foundation of China grant 32171813.

Declaration of interests

No interests are declared.

Resources

[i] www.iucnredlist.org

Outstanding questions

Will ML approaches be able to delimit lineages/species, or just classes of objects? Will they be able to reliably integrate different datasets? Is there a universal ML approach, or will approaches always need training for specific groups? Can ML flexibilize model assumptions of current (phylo)genomic approaches mainly developed for tree-like evolving diploids? Who should judge ML results? First ML approaches allow for less subjective, integrative species delimitation. Research should focus on dataset fusion, the combination of ML with current phylogenomics, and a deeper understanding of the ML decision-making process. Current ML results must be judged by trained taxonomists.

How often do past descriptions overlap with modern classifications? What impact do revised treatments have on local, regional, and global assessments of biodiversity? Can we infer species objectively as natural entities, or will there always be subjectivity? Species are the cornerstone of biodiversity research. To define species as naturally as possible, and to confidentially differentiate them from subspecies, varieties, or forms (which can be treated as infraspecific taxa before more research is done), an iUSC should be universally applied in research, including peer-reviewed, open-access publication of taxonomic treatments.

How can the collection of multi-approach data be simplified, automatized, and made less expensive? Can we automatically revise a substantial number of species? Many eukaryotic groups have not yet been studied using even simple genetic approaches. This is mainly due to a lack of resources, specifically in developing countries where most of the present biodiversity is found. The current rate of biodiversity loss urges us to simplify lab workflows with universal methods, and to automatize data collection from online databases to accelerate species revision and naming. Yet unknown lineages are likely to be discovered by AI.

References

1. Burki, F. *et al.* (2020) The new tree of eukaryotes. *Trends Ecol. Evol.* 35, 43–55
2. Larsen, B.B. *et al.* (2017) Inordinate fondness multiplied and redistributed: the number of species on Earth and the new Pie of Life. *Q. Rev. Biol.* 92, 229–265
3. Cicconardi, F. *et al.* (2013) Collembola, the biological species concept and the underestimation of global species richness. *Mol. Ecol.* 22, 5382–5396
4. Mayr, E. (1942) *Systematics and the origin of species*, Columbia University Press
5. Stuessy, T.F. (2009) *Plant Taxonomy: The Systematic Evaluation of Comparative Data* (2nd ed), Columbia University Press
6. De Queiroz, K. (2007) Species concepts and species delimitation. *Syst. Biol.* 56, 879–886
7. Mishler, B.D. (2009) Three centuries of paradigm changes in biological classification: is the end in sight? *Taxon* 58, 61–67
8. Sukumaran, J. and Knowles, L.L. (2017) Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci.* 114, 1607–1612
9. Qiu, L. *et al.* (2023) Defining honeybee subspecies in an evolutionary context warrants strategized conservation. *Zool. Res.* 44, 483–493
10. Ennos, R. *et al.* (2005) Conserving taxonomic complexity. *Trends Ecol. Evol.* 20, 164–168
11. Mirarab, S. *et al.* (2021) Multispecies coalescent: theory and applications in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 52, 247–268
12. Jaron, K.S. *et al.* (2021) Genomic features of parthenogenetic animals. *J. Hered.* 112, 19–33
13. Maharachchikumbura, S.S.N. *et al.* (2021) Integrative approaches for species delimitation in Ascomycota. *Fungal Divers.* 109, 155–179
14. Sukumaran, J. *et al.* (2021) Incorporating the speciation process into species delimitation. *PLoS Comput. Biol.* 17, e1008924
15. Mallet, J. (2005) Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20, 229–237
16. Soltis, P.S. *et al.* (2015) Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* 35, 119–125
17. He, L. and Hörandl, E. (2022) Does polyploidy inhibit sex chromosome evolution in angiosperms? *Front. Plant Sci.* 13, 1–9
18. Abbott, R. *et al.* (2013) Hybridization and speciation. *J. Evol. Biol.* 26, 229–246
19. Keuler, R. *et al.* (2020) Genome-scale data reveal the role of hybridization in lichen-forming fungi. *Sci. Rep.* 10, 1–14
20. Zinner, D. *et al.* (2011) The strange blood: natural hybridization in primates. *Evol. Anthropol. Issues News Rev.* 20, 96–103
21. Giordano, L. *et al.* (2018) Mitonuclear interactions may contribute to fitness of fungal hybrids. *Sci. Rep.* 8, 1706
22. Postel, Z. and Touzet, P. (2020) Cytonuclear genetic incompatibilities in plant speciation. *Plants* 9, 487
23. Weaver, R.J. *et al.* (2022) Genomic signatures of mitonuclear coevolution in mammals. *Mol. Biol. Evol.* 39, 1–14
24. Karbstein, K. *et al.* (2022) Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large *Ranunculus auricomus* species complex. *New Phytol.* 235, 2081–2098
25. Hodač, L. *et al.* (2023) Geometric morphometric versus genomic patterns in a large polyploid plant species complex. *Biology (Basel)* 12, 418
26. Hörandl, E. (2022) Novel approaches for species concepts and delimitation in polyploids and hybrids. *Plants* 11, 204
27. Selz, O.M. and Seehausen, O. (2019) Interspecific hybridization can generate functional novelty in cichlid fish. *Proc. R. Soc. B Biol. Sci.* 286, 20191621
28. Van De Peer, Y. *et al.* (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424
29. Leebens-Mack, J.H. *et al.* (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685
30. Blischak, P.D. *et al.* (2018) Integrating networks, phylogenomics, and population genomics for the study of polyploidy. *Annu. Rev. Ecol. Evol. Syst.* 49, 253–278
31. Van de Peer, Y. *et al.* (2021) Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* 33, 11–26
32. Karbstein, K. *et al.* (2021) Moving beyond assumptions: polyploidy and environmental effects explain a geographical parthenogenesis scenario in European plants. *Mol. Ecol.* 30, 2659–2675
33. David, K.T. (2022) Global gradients in the distribution of animal polyploids. *Proc. Natl. Acad. Sci.* 119, 2017
34. Braasch, I. (2020) Genome evolution: domestication of the allopolyploid goldfish. *Curr. Biol.* 30, R812–R815
35. Soltis, D.E. *et al.* (2004) Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biol. J. Linn. Soc.* 84, 458–502
36. Campbell, M.A. *et al.* (2016) The case of the missing ancient fungal polyploids. *Am. Nat.* 188, 602–614
37. Renner, S.S. (2014) The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am. J. Bot.* 101, 1588–1596
38. Spoelhof, J.P. *et al.* (2020) Does reproductive assurance explain the incidence of polyploidy in plants and animals? *New Phytol.* 227, 14–21
39. Wood, T.E. *et al.* (2009) The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci.* 106, 13875–13879
40. Hojsgaard, D. and Hörandl, E. (2019) The rise of apomixis in natural plant populations. *Front. Plant Sci.* 10, 1–13
41. Lamelza, P. *et al.* (2019) Hybridization promotes asexual reproduction in *Caenorhabditis* nematodes. *PLoS Genet.* 15, e1008520
42. Kondrashov, A.S. (1994) The asexual ploidy cycle and the origin of sex. *Nature* 370, 213–216
43. Hörandl, E. *et al.* (2020) Genome evolution of asexual organisms and the paradox of sex in eukaryotes. In *Evolutionary biology — a transdisciplinary approach* (Pontarotti, P., ed.), pp. 133–167, Springer
44. Brandt, A. *et al.* (2017) Effective purifying selection in ancient asexual oribatid mites. *Nat. Commun.* 8, 873
45. Hörandl, E. (2018) The classification of asexual organisms: old myths, new facts, and a novel pluralistic approach. *Taxon* 67, 1066–1081
46. Coyne, J.A. and Orr, H.A. (2004) *Speciation*, Sinauer
47. Karbstein, K. *et al.* (2020) Phylogenomics supported by geometric morphometrics reveals delimitation of sexual species within the polyploid apomictic *Ranunculus auricomus* complex (Ranunculaceae). *Taxon* 69, 1191–1220
48. Irisarri, I. *et al.* (2021) Unexpected cryptic species among streptophyte algae most distant to land plants. *Proc. R. Soc. B Biol. Sci.* 288
49. Wang, P.M. *et al.* (2018) Phylogeny and species delimitation of *Flammulina*: taxonomic status of winter mushroom in East Asia and a new European species identified using an integrated approach. *Mycol. Prog.* 17, 1013–1030
50. Bank, S. *et al.* (2021) Reconstructing the nonadaptive radiation of an ancient lineage of ground-dwelling stick insects (Phasmatodea: Heteropterygidae). *Syst. Entomol.* 46, 487–507
51. Freudenstein, J.V. *et al.* (2017) Biodiversity and the species concept — lineages are not enough. *Syst. Biol.* 66, 644–656
52. Dayrat, B. (2005) Towards integrative taxonomy. *Biol. J. Linn. Soc.* 85, 407–415
53. Oberprieler, C. (2023) The Wettstein tesseract: a tool for conceptualising species-rank decisions and illustrating speciation trajectories. *Taxon* 72, 1–7
54. Wagner, N.D. *et al.* (2018) RAD sequencing resolved phylogenetic relationships in European shrub willows (*Salix* L. subg. *Chamaetia* and subg. *Vetrix*) and revealed multiple evolution of dwarf shrubs. *Ecol. Evol.* 17, 8243–8255
55. Derkarabetian, S. *et al.* (2019) A demonstration of unsupervised machine learning in species delimitation. *Mol. Phylogenet. Evol.* 139, 106562
56. Pérez-Delgado, A.J. *et al.* (2022) Hidden island endemic species and their implications for cryptic speciation within soil arthropods. *J. Biogeogr.* 49, 1367–1380

57. Boluda, C.G. *et al.* (2019) Evaluating methodologies for species delimitation: the mismatch between phenotypes and genotypes in lichenized fungi (*Bryoria* sect. *Implexae*, Parmeliaceae). *Persoonia - Mol. Phylogeny Evol. Fungi* 42, 75–100
58. Dietz, L. *et al.* (2023) Standardized nuclear markers improve and homogenize species delimitation in Metazoa. *Methods Ecol. Evol.* 14, 543–555
59. van Klink, R. *et al.* (2022) Emerging technologies revolutionise insect ecology and monitoring. *Trends Ecol. Evol.* 37, 872–885
60. Haveman, R. *et al.* (2002) Apomicten: het belang van een genuanceerde taxonomie voor plantensociologisch onderzoek en natuurbeheer. *Stratiotes* 25, 3–25
61. Richards, A.J. (2003) Apomixis in flowering plants: an overview. *Philos. Trans. R. Soc. B Biol. Sci.* 358, 1085–1093
62. Hörandl, E. (2006) Paraphyletic versus monophyletic taxa-evolutionary versus cladistic classifications. *Taxon* 55, 564–570
63. Mallet, J. (1995) A species definition for the modern synthesis. *Trends Ecol. Evol.* 10, 294–299
64. Ott, T. *et al.* (2022) The warps and wefts of a polyploidy complex: Integrative species delimitation of the diploid *Leucanthemum* (Compositae, Anthemideae) representatives. *Plants* 11, 1878
65. Dawkins, R. (2008) *The selfish gene*, Springer Spektrum (in German)
66. Schlick-Steiner, B.C. *et al.* (2010) Integrative taxonomy: a multi-source approach to exploring biodiversity. *Annu. Rev. Entomol.* 55, 421–438
67. Chambers, E.A. and Hillis, D.M. (2020) The multispecies coalescent over-splits species in the case of geographically widespread taxa. *Syst. Biol.* 69, 184–193
68. Caron, D.A. and Hu, S.K. (2019) Are we overestimating protistan diversity in nature? *Trends Microbiol.* 27, 197–205
69. Derkarabetian, S. *et al.* (2022) Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data. *Front. Zool.* 19, 8
70. Wibberg, D. *et al.* (2021) High quality genome sequences of thirteen Hypoxylaceae (Ascomycota) strengthen the phylogenetic family backbone and enable the discovery of new taxa. *Fungal Divers.* 106, 7–28
71. McKain, M.R. *et al.* (2018) Practical considerations for plant phylogenomics. *Appl. Plant Sci.* 6, e1038
72. Eriksson, J.S. *et al.* (2018) Allele phasing is critical to revealing a shared allopolyploid origin of *Medicago arborea* and *M. strasseri* (Fabaceae). *BMC Evol. Biol.* 18, 9
73. Andermann, T. *et al.* (2018) Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Syst. Biol.* 68, 32–46
74. Ribeiro, P.G. *et al.* (2021) A bioinformatic platform to integrate target capture and whole genome sequences of various read depths for phylogenomics. *Mol. Ecol.* 30, 6021–6035
75. Gulyaev, S. *et al.* (2022) The phylogeny of *Salix* revealed by whole genome re-sequencing suggests different sex-determination systems in major groups of the genus. *Ann. Bot.* 129, 485–498
76. He, L. *et al.* (2023) Evolutionary origin and establishment of a dioecious diploid-tetraploid complex. *Mol. Ecol.* 32, 2732–2749
77. Bank, S. *et al.* (2021) A tree of leaves: phylogeny and historical biogeography of the leaf insects (Phasmatodea: Phylliidae). *Commun. Biol.* 4, 932
78. Tomasello, S. (2018) How many names for a beloved genus? – coalescent-based species delimitation in *Xanthium* L. (Ambrosiinae, Asteraceae). *Mol. Phylogenet. Evol.* 127, 135–145
79. Šlenker, M. *et al.* (2021) Allele sorting as a novel approach to resolving the origin of allotetraploids using Hyb-Seq data: a case study of the Balkan Mountain endemic *Cardamine barbaraeoides*. *Front. Plant Sci.* 12, 1–22
80. Zhang, C. *et al.* (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinforma.* 19, 153
81. Minh, B.Q. *et al.* (2020) New methods to calculate concordance factors for phylogenomic datasets. *Mol. Biol. Evol.* 37, 2727–2733
82. Kozlov, A.M. *et al.* (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455
83. Solís-Lemus, C. *et al.* (2017) PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34, 3292–3298
84. Malinsky, M. *et al.* (2018) RADpainter and fineRADstructure: population inference from RADseq data. *Mol. Biol. Evol.* 35, 1284–1290
85. Peralta, M. *et al.* (2013) SNiPloid: a utility to exploit high-throughput SNP data derived from RNA-Seq in allopolyploid species. *Int. J. Plant Genomics* 2013, 1–6
86. Frichot, E. *et al.* (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196, 973–983
87. Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959
88. Solís-Lemus, C. *et al.* (2015) Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution* 69, 492–507
89. Yang, Z. and Rannala, B. (2010) Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci.* 107, 9264–9269
90. Leaché, A.D. *et al.* (2014) Species delimitation using genome-wide SNP data. *Syst. Biol.* 63, 534–542
91. Jones, G. *et al.* (2015) DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics* 31, 991–998
92. Yan, Z. *et al.* (2022) Maximum parsimony inference of phylogenetic networks in the presence of polyploid complexes. *Syst. Biol.* 71, 706–720
93. Jones, G. (2017) Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *J. Math. Biol.* 74, 447–467
94. Wagner, F. *et al.* (2017) “Fix me another Marguerite!”: species delimitation in a group of intensively hybridizing lineages of ox-eye daisies (*Leucanthemum* Mill., Compositae-Anthemideae). *Mol. Ecol.* 26, 4260–4283
95. Wagner, N.D. *et al.* (2020) Phylogenomic relationships and evolution of polyploid *Salix* species revealed by RAD Sequencing data. *Front. Plant Sci.* 11, 36–41
96. Múrias dos Santos, A. *et al.* (2016) tcsBU: a tool to extend TCS network layout and visualization. *Bioinformatics* 32, 627–628
97. Borowiec, M.L. *et al.* (2022) Deep learning as a tool for ecology and evolution. *Methods Ecol. Evol.* 13, 1640–1660
98. Raxworthy, C.J. and Smith, B.T. (2021) Mining museums for historical DNA: advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060
99. Buono, D. and Albach, D.C. (2023) Infrared spectroscopy for ploidy estimation: an example in two species of *Veronica* using fresh and herbarium specimens. *Appl. Plant Sci.* 11, e11516
100. Mäder, P. *et al.* (2021) The Flora Incognita app – interactive plant species identification. *Methods Ecol. Evol.* 12, 1335–1342
101. Yang, Z. and Rannala, B. (2014) Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.* 31, 3125–3135
102. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
103. Goodfellow, I. *et al.* (2016) *Deep Learning*, MIT Press
104. Liu, T. *et al.* (2022) Stability estimation for unsupervised clustering: a review. *WIREs Comput. Stat.* 14, 1–17
105. Smith, M.L. and Carstens, B.C. (2020) Process-based species delimitation leads to identification of more biologically relevant species. *Evolution (N. Y)* 74, 216–229
106. Pei, J. *et al.* (2018) CLADES: a classification-based machine learning method for species delimitation from population genetic data. *Mol. Ecol. Resour.* 18, 1144–1156
107. Seeland, M. *et al.* (2019) Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinforma.* 20, 4
108. Wäldchen, J. and Mäder, P. (2018) Machine learning for image based species identification. *Methods Ecol. Evol.* 9, 2216–2225
109. Yang, B. *et al.* (2022) Identification of species by combining molecular and morphological data using convolutional neural networks. *Syst. Biol.* 71, 690–705
110. Tan, C. *et al.* (2018) A survey on deep transfer learning. In *Lecture Notes in Computer Science 11141 LNCS*, pp. 270–279, Springer

111. Seeland, M. and Mäder, P. (2021) Multi-view classification with convolutional neural networks. *PLoS One* 16, e0245230
112. Schneider, S. *et al.* (2022) Similarity learning networks for animal individual re-identification: an ecological perspective. *Mamm. Biol.* 102, 899–914
113. Millán Arias, P. *et al.* (2022) DeLUCS: deep learning for unsupervised clustering of DNA sequences. *PLoS One* 17, e0261531
114. Selvaraju, R.R. *et al.* (2020) Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128, 336–359
115. Badirli, S. *et al.* (2020) Classifying the unknown: insect identification with deep hierarchical Bayesian learning. *Methods Ecol. Evol.* 14, 1515–1530
116. Rzanny, M. *et al.* (2017) Acquiring and preprocessing leaf images for automated plant identification: understanding the tradeoff between effort and information gain. *Plant Methods* 13, 97
117. Samek, W. *et al.* (2021) Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* 109, 247–278
118. Douglas, J. and Bouckaert, R. (2022) Quantitatively defining species boundaries with more efficiency and more biological realism. *Commun. Biol.* 5, 755