Quantifying tropical plant diversity requires an integrated technological approach

Frederick C. Draper, Center for Global Discovery and Conservation Science, Arizona State University, Tempe, USA and School of Geography, University of Leeds, UK

Timothy R. Baker, School of Geography, University of Leeds, UK

Christopher Baraloto, Institute of Environment, Department of Biological Sciences, Florida International University, Miami, USA

Jerome Chave, Laboratoire Evolution et Diversité Biologique (EDB) CNRS/UPS, Toulouse, France

Flavia Costa, Instituto Nacional de Pesquisas da Amazônia - INPA, Manaus, Brazil

Roberta E. Martin, Center for Global Discovery and Conservation Science, Arizona State University, Tempe, USA

R. Toby Pennington, Geography Department, University of Exeter, UK and Royal Botanic Garden Edinburgh, UK

Alberto Vicentini, Instituto Nacional de Pesquisas da Amazônia - INPA, Manaus, Brazil

Gregory P. Asner, Center for Global Discovery and Conservation Science, Arizona State University, Tempe, USA

Corresponding author: F.C. Draper (freddie.draper@gmail.com)

Key words

Tropical botany, plant biodiversity, technology, spectroscopy, DNA, Artificial intelligence

Abstract

Tropical biomes are the most diverse plant communities on Earth, and quantifying this diversity at large spatial scales is vital for many purposes. As macroecological approaches proliferate, the taxonomic uncertainties in species occurrence data are easily neglected and can lead to spurious findings in downstream analyses. Here, we argue that technological approaches offer potential solutions, but there is no single silver bullet to resolve uncertainty in plant biodiversity quantification. Instead, we propose the use of AI approaches to build a data-driven framework that integrates several data sources - including spectroscopy, DNA sequences, image recognition and morphological data. Such a framework would provide a foundation for improving species identification in macroecological analyses while simultaneously improving the taxonomic process of species delimitation.

The challenge of tropical plant diversity

Much of global biodiversity is concentrated in tropical biomes [1]. Yet, the tropics face the twin challenges of being among the most data-deficient regions on Earth in terms of occurrence records [2], while also being among the most threatened by rapid human development and climate change [3]. As a result, describing, measuring, monitoring, and conserving tropical biodiversity is now recognized as a priority by relevant intergovernmental panels [3]. Despite three centuries of biodiversity research, we remain unable to quantify tropical plant diversity, i.e. to provide the fundamental spatially explicit information required to effectively monitor and conserve tropical ecosystems; and to answer vital questions such as how many species exist in tropical forests, which areas are the most species rich, and which areas house the most unique (endemic) species.

Prominent voices have recently called for a Linnaean renaissance, arguing that an increase in field biologists cataloguing and describing this diversity is urgently required [4]. Despite this call to arms, the number of biologists collecting field data in the tropics continues to decline [5]. Although an increase in field collections is essential, quantifying biodiversity in the highly diverse tropics is not only an issue of boots on the ground. Each year field biologists continue to collect large amounts of species occurrence and abundance data, but taxonomic uncertainty surrounding these data persist. Furthermore, vast quantities of data are increasingly being combined to develop large synthetic databases [6,7]. While such datasets are an essential tool for assessing large-scale vegetation responses to global change, the accessibility of such huge datasets makes it easy to overlook two issues associated with these data: (i) many areas in the tropics remain unexplored and lack collections of museum specimens and ecological inventories; and (ii) significant underlying uncertainties in tropical plant taxonomy persist.

One of the main innovations in biodiversity research over the last decades is the increasing appreciation for different dimensions of diversity beyond taxonomic species diversity, including functional and phylogenetic diversity, as well as more abstract proxies such as remotely sensed spectral diversity and environmental DNA. Although these approaches can provide insights into broad biodiversity patterns and the ecological mechanisms underlying them at landscape or community scales, many of the fundamental processes underpinning biodiversity patterns (e.g., extinction, speciation, competition) occur at the species or population level. While the huge task of identifying species remains daunting, monitoring species-level changes in tropical forests, which requires accurate species identifications, will be essential to understanding and mitigating the impacts of global change.

Limitations with current process of quantifying tropical plant diversity

Currently, almost all studies seeking to quantify tropical plant diversity are underpinned by morphological botanical approaches to species identification (Box 1). However, attempts to quantify taxonomic uncertainty in large synthetic datasets have revealed substantial errors [8–10]. We suggest that these uncertainties arise from limitations in both underlying taxonomic frameworks (point 1) and the process of species identification (points 2-7):

1. The taxonomy of many tropical plant lineages is out of date or incomplete. For example, up to 40% of the species described in neotropical plant monographs are new to science, while in other cases re-circumscribed species can 'sink' as synonyms multiple species (sometimes >10) previously considered distinct [11,12].

2. Local herbaria are often relied upon to identify species, but these collections are often incomplete and specimen identifications may not be reliable [13]. Furthermore, specimen identifications are rarely standardised among herbaria, but see [14,15].
3. Species level identifications in diverse tropical forests often require samples of fruits or flowers. Given the often short and unpredictable phenologies of many tropical species [16], short field research visits can easily miss the reproductive period of species, meaning species level identifications are made on vegetative samples, thereby decreasing their accuracy.
4. **Voucher samples** (see glossary) from ecological inventories, when collected, frequently lack reproductive structures (flowers and fruit) and are rarely accepted by herbaria. Therefore, ecological inventories typically contribute little to species delimitation and developing taxonomies, despite considerable potential to do so [17].
5. In practice, identifying species based on morphological characters is, at least to some extent, subjective if it cannot be done by the taxonomic specialist for a given group, which is seldom the case. Identifications by non-specialists vary and depend on previous experience and resources available (i.e., taxonomic monographs, flora accounts and specimens identified by taxonomic specialists).
6. In many cases, vouchers are not collected for every individual plant within inventory plots. Instead, individuals from the same plot that are deemed to be the same species are grouped together and one or more vouchers are collected to represent that group. This effectively means that the initial judgment of the field botanist introduces uncertainty which is difficult to quantify post-hoc.

7. In ecological inventories, there is a lack of taxonomic standardization amongst plots and surveys, hampering the use of these data both within and among different tropical regions. This is true of both named species and especially the unnamed "**morphotypes**"(see glossary). Morphotypes are often standardized within a plot or dataset because identifications are done by the same individual or team; but they are rarely standardized among datasets (but see [18,19]).

Together these uncertainties lead to many individual tropical plants remaining unidentified or incorrectly identified, despite being collected or observed in inventory plots. Because much of this uncertainty remains unquantified, it is propagated through to downstream data products such as large-scale biodiversity databases. While removing taxonomic synonyms and flagging erroneous coordinates are crucially important steps in cleaning botanical data [20], this is not the same as standardizing taxonomy because it still assumes that underlying species identifications are correct.

Recent initiatives have addressed some of these issues by promoting closer collaboration between taxonomic specialists and ecologists [17], digitizing and standardizing voucher specimens among plot networks and herbaria, as well as providing taxonomically verified and expertly curated regional scale species lists [10,21].

Technological approaches to quantifying tropical plant diversity

DNA approaches

The best-known technological solution for addressing issues with species delimitation and species identification is DNA sequencing. DNA sequences are ideal for estimating evolutionary relationships among individuals, populations and species and therefore now form the basis for lineage-based species concepts [22]. Furthermore, DNA sequencing can be applied both to vegetative samples, and now, using next-generation approaches such as target capture, even to two-century old herbarium collections [23]. Because of these advantages, DNA-based approaches were predicted to revolutionize biodiversity research in the tropics [24,25]. Although DNA-based approaches are used in both the delimitation and identification of tropical plants, neither of these tasks have been transformed by DNA-based techniques, and both are still most frequently based on traditional morphological methods.

One approach to aid species identification is **DNA barcoding** [26] (see glossary). Although it has been highly successful in some taxonomic groups (e.g. moths [27]), DNA barcoding has had less impact on tropical plant biodiversity surveys [28,29]. This lack of success can be attributed in part to the incomplete reference library that is required for identification by barcoding, which requires existing sequences from authoritatively identified specimens. In addition, while standard barcodes can distinguish a high percentage of species at some local sites (e.g., 97% tree species on Barro Colorado Island [30]), they are less accurate at other sites [28]; and at a global scale, at least 30% of tropical plant species cannot be differentiated using these barcodes, because they are insufficiently variable both in lineages with slow mutation rates relative to speciation rate, and in groups showing recent and rapid divergence [31].

Species discrimination can be improved by adding additional, more variable, DNA loci. The advent of **next-generation sequencing technologies** (see glossary) has made the sequencing of high numbers of such additional loci feasible over the past decade. For example, rather than just two standard plastid barcodes (*rbcL* and *matk*; 1400 bp in total), whole plastome sequences can provide 150,000 bp of sequence. Hybrid capture techniques work well with degraded DNA from herbarium specimens and can simultaneously offer sequence from thousands of individual nuclear genes, some of which may work in combination as barcodes angiosperm-wide [32]. Genome skimming offers access to loci from plastid, mitochondrial and repeated nuclear regions, though low copy nuclear genes are more difficult to assemble [31]. The costs of these approaches are decreasing, but at this time they remain a limiting factor to allow use at massive scales.

In some cases, these large datasets cannot solve fundamental conceptual issues such as the failure of plastid genomes to track species boundaries because of interspecific gene flow [31], though this can be mitigated by using multiple, unlinked nuclear loci. In addition, there is a practical problem that new loci will require the construction of new sequence reference libraries to allow them to be used as identification tools, and the reference libraries themselves pre-suppose a stable and accurate underlying taxonomy. Yet, these issues may be overcome rapidly by using next-generation DNA sequence data as a foundation for species delimitation as part of an **integrative taxonomic approach** (see glossary [33–35]). Such an approach will simultaneously improve taxonomy and build a barcode reference library for the loci used.

Spectroscopy approaches

Other technological approaches that could aid species delimitation and identification include lab-based **spectroscopy** and remotely sensed **imaging spectroscopy** (see glossary). Although spectroscopy is a well-established discipline, it is rarely considered for quantifying biodiversity in the tropics [36]. Spectroscopy dramatically expands the dimensionality of a vegetative plant sample, effectively providing several hundred characters that reflect different chemical and physical properties of an individual's leaves or wood. As variation in foliar chemistry and physical properties is greater among species than within species [37], spectroscopy can differentiate among species in a manner similar to "chemocoding" [38] but with considerably lower running costs.

The few studies that have tested the accuracy of spectroscopy in determining species identifications have produced promising results, often surpassing the accuracy typically obtained by DNA barcoding in tropical plant lineages [39–42]. For example, trees in two families for which classical DNA barcodes provide less resolution [28], the Burseraceae and Lecythidaceae, were identified by spectroscopy to species level with an accuracy of 97-98 % [40-41]. In a wide-ranging study, 1449 canopy species in the Andes – Amazon region were classified to species level with an accuracy of >85% [37]. Other research has demonstrated the utility of using bark and branch tissue in addition to leaf tissue for spectroscopic identifications [39,41]. One recent study across a number of Amazonian taxa found that species level identifications made with branch samples had an accuracy > 90%, which could be increased to 94% if leaf tissue was included[39].

Spectroscopy approaches have also been used effectively in the species delimitation process. For example, spectroscopy was recently used alongside DNA data to delimit the species complexes *Protium heptaphyllum* (Burseraceae) and *Pagamea guianensis* (Rubiaceae) into two and fourteen distinct species, respectively [34,35].

Spectroscopic approaches for species identification share many of the advantages that DNA barcoding has over traditional approaches; for example, only vegetative material (which can be older and dried) is required to make identifications which are quantitative and reproducible. While correct use of spectrometers and the analysis of spectral data also requires time and dedication, training in these approaches can be undertaken in weeks to months rather than years.

Importantly, spectroscopy holds several key advantages for species identification in addition to those shared with DNA-based approaches. First, spectra reflect not only the taxonomic identity but also several functional traits (e.g. foliar nitrogen and water content) [43–45], which can improve our understanding of the interaction between taxonomic diversity and ecosystem functioning. Second, imaging spectroscopy provides a method for scaling up biodiversity estimates to far greater areas than will ever be possible with field work alone (Box 2). Third, while the initial expense of a precise lab-based spectrometer is not insignificant, many thousands of samples can be processed with relatively modest maintenance and operation costs and can be operated in the field or herbarium without the need of a wet lab.

Like DNA-based approaches, spectroscopy will not solve all identification problems. Several factors including: leaf ontogeny, leaf light environment and leaf sample preparation are known to increase variation within species; therefore, a standardized protocol will be essential. Furthermore, though initial results are promising, spectroscopy for botanical identification has

not been widely tested across lineages and locations, so we do not yet know the limits to these approaches. Finally, because spectroscopy provides a phenotypic measurement, it does not represent an alternative for lineage-based species delimitation methods for which DNA sequences are required [22].

Artificial intelligence (AI) approaches

Together, traditional morphological botanical approaches alongside genetic and spectroscopic technologies provide huge potential for identifying plant individuals by expanding the data dimensionality of vegetative samples. However, like traditional identification approaches, genetic and spectroscopic techniques are still dependent on comparisons with a reference library, which is currently lacking for many tropical species. Once we have started to develop a unified reference library using a combination of DNA and spectroscopic approaches alongside morphological characteristics, how can we make robust, repeatable and objective comparisons with this reference library across the tropics?

Artificial intelligence (AI, see glossary) presents a suite of robust and objective computational methods with huge potential for taxonomic identification. In recent years there has been an explosion in the use of AI approaches to a range of ecological questions including species identification [46–48]. This increase is due largely to the accessibility of high-performance algorithms and the availability of high-performance GPU -accelerated distributed computing systems.

Recent efforts have used **deep learning approaches** (see glossary) to successfully identify plant species from images taken both in the field and in herbaria [46,49]. However, such efforts have proven more challenging in tropical ecosystems [50], where identifications made by expert

botanists are more accurate. This may be because in species-rich tropical regions many species can appear extremely similar, and image-based approaches cannot detect the subtle features such as texture that expert botanists use to distinguish samples. Alternatively, the poor performance of image-based classifiers may be due to insufficient or inaccurate image training data across taxa. Further testing of the limits of image-based classification is required with expanded image libraries. Nevertheless, while image-based approaches are likely an effective tool to classify samples to the family or genus level, we suggest that AI approaches will be more successful at species level classification if they are expanded to include more feature rich data such as foliar spectra and DNA barcodes.

An important limitation of AI approaches, particularly deep learning, is that they require extensive training data. Large online image libraries can be rapidly developed; for example, several hundred thousand images of 10,000 Amazonian plant species [50] have been collected. However, these libraries are based on online image search engine results that have not been authoritatively identified and therefore will contain significant error. Libraries of DNA barcodes, spectra and well identified herbarium specimens are smaller, but better curated. Initial collections of standard DNA barcodes and foliar spectra have been made for many thousands of tropical species, providing a solid foundation for future training data [37]. Furthermore, DNA, images and potentially spectra can be readily extracted from herbarium vouchers, so building large databases is just a matter of funding and will.

Developing a framework for progress

The framework we outline here will require an integrative multidisciplinary approach (figure 1), building upon existing collaborations (e.g. among systematists and ecologists) as well as forging entirely new ones (e.g. with data scientists). The greatest challenge to our proposed framework is that it relies on an underlying reference library that must be dynamic to future changes in plant systematics and available to the many thousands of tropical biodiversity scientists. How can such a reference library be built for the many thousands of plant species that exist in tropical forests?

A first step is to reduce the scope of the task. Skilled field botanists can often assign individuals to family or genus with little error. Therefore, following the current paradigm of developing family or genus-level reference collections, presents the most tractable pathway that builds on current knowledge and resources. Additionally, concentrating on those lineages that contain many 'hyperdominant' species [51], would reduce the taxonomic uncertainty surrounding those species that dominate ecosystem functioning [52]. Several lineages containing hyperdominant species already have well developed molecular phylogenies (e.g. *Inga* (Fabaceae), *Protium* (Burseraceae)). By prioritizing these dominant lineages, we can build a modular reference library which can be expanded, thereby balancing near term practicality with long term potential. As complete lineage specific modules are populated with relevant DNA and spectral reference libraries, deep-learning classification models can be developed and published in publicly available online repositories [Box 3].

The next step will be to apply these approaches broadly across existing datasets including herbarium collections and permanent plot networks. Working with herbaria across the tropics, it will be possible to transform these vast collections into unified identifications for potentially

thousands of species. There are significant costs associated with meeting this challenge at scale; in this respect, spectral approaches are likely the most cost-effective option, and developing standardized protocols to take uniform spectral measurements represents a priority.

Not all individuals will be identified with a high degree of confidence by deep learning classification models. Unidentified individuals should be highlighted as either taxonomically described species missing from the reference collection, or putative novel species that remain undescribed. Therefore, although the primary focus of the workflow we outline is to improve species identification, this process will simultaneously accelerate the process of species discovery.

Concluding remarks

Although the idea of scanning a tropical forest plant specimen with a handheld device and instantly obtaining a correct species-level identification [53] remains science fiction for now, the technological approaches we outline have significant potential for revolutionizing our ability to quantify plant diversity in tropical forests at global scales in coming decades. The limitations we describe could be overcome by integrating these new technologies to generate a dynamic, data-driven framework for biodiversity research, while simultaneously strengthening the link between ecological and taxonomic practices.

There have been several previous calls to leverage different forms of technology to revolutionize species identification [53,54]. We are now at a stage where the technology has come of age and necessary tools for identification are available, affordable, and tested. It is time to move beyond demonstrating the capabilities of these tools through small scale comparisons, and instead begin to develop a unified, objective and scalable framework from which we can quantify tropical

283 plant diversity globally and answer some of the most pressing issues in tropical plant ecology

284 (see Outstanding Questions).

Box 1 Current approaches for quantifying plant diversity [400 words]

The quantification of plant diversity consists of two distinct elements, hereafter labelled ‘species delimitation’ and ‘species identification’. Species delimitation is the process of delimiting plant species based on characters that generally come from macro-morphology, but may also include micro-morphology and genetic data. Species delimitation is typically carried out by taxonomists, who are concerned with producing taxonomies for specific lineages and describing new species. This species delimitation process therefore develops the underlying taxonomy that underpins all subsequent biodiversity analyses. Recent approaches that integrate data sources and especially DNA sequence data have proven powerful in delimiting tropical species, for example revealing cryptic variation in widespread Amazonian species [34,35].

Species identification is the process of assigning individual specimens to known plant species using pre-existing taxonomy. In tropical forests this process is often carried out by ecologists who establish vegetation survey plots where individuals are identified to the finest possible taxonomic level and often measured for diameter, height and other plant traits. Collections of survey plots can then be grouped into plot networks, which can be used to ask ecological questions at local, landscape, regional or even global scales.

Identifying an individual plant sample can take many forms. A skilled botanist may be able to make a genus or species level identification in the field if the individual belongs to a species that is particularly easy to identify or is locally or regionally common. More commonly, though, this process requires a representative voucher sample for each species found in the field subsequently to be compared with reference collections in local herbaria as well as increasingly available digital herbaria, published taxonomic treatments and keys. Using a range of morphological

characters, botanists are then able to assign an individual to a species. Of course, many individuals in forest inventory plots cannot be identified to species level. In these instances, unidentified individuals are assigned to ‘morphospecies’. These morphospecies may be abundant and well-known locally but awaiting scientific description, or existing species that have not been previously collected in that locality, or errant discriminations that ultimately will be integrated into existing species.

Vouchers are not always collected for every individual within a forest census plot, but more often only a representative voucher for every species or morphospecies encountered within the plot is collected. Implicit in this process is the assumption that the collecting teams are able to accurately delimit different species at the plot scale even if they are not able to assign an identification.

Box 2 Scaling up biodiversity estimates with imaging spectroscopy [400 words]

A major advantage of spectroscopic approaches is that imaging spectrometers can be mounted on airborne and satellite platforms, and therefore can be used to scale-up biodiversity estimates across vast spatial scales (e.g. 10^6 km^2) [57,58]. This is important because most tropical forests occur in vast inaccessible areas of wilderness, and accumulating field data over such large scales would be impossible. Furthermore, existing approaches for scaling up ground-based biodiversity estimates across large areas of tropical forests have had limited success. For example, species distribution modelling approaches perform poorly in tropical forest regions because climate and edaphic gradients are either poorly characterized at relevant spatial scales (e.g., soil fertility) or represent relatively narrow breadth across large areas (e.g., precipitation). Indeed, equivalent performance to describe species distributions can be obtained through simple spatial extrapolation [59].

Imaging spectroscopy has now been used successfully to map different dimensions of tropical plant biodiversity at a range of scales, including landscape scale spectral alpha and beta diversity which are shown to be effective proxies of taxonomic alpha and beta diversity [60,61], as well as landscape and regional scale functional beta diversity [62,63] from foliar traits and species distributions [64].

Top-of-canopy reflectance spectra obtained from airborne or spaceborne platforms do not form a one-to-one relationship with leaf spectra collected in-situ due to variation in leaf orientation, canopy structure, soil reflectance, illumination conditions and viewing geometry [65]. This disconnect is increased when leaves are dried, making it difficult to scale directly from

herbarium specimens to the landscape. Nevertheless, species-specific mapping can be achieved across the landscape if training data are collected as canopy spectra in the field.

A major limitation to imaging spectroscopy of tropical forests is that only the uppermost sunlit canopies are detected by sensors, therefore excluding the many thousands of species that never make it to the forest canopy. While understory species will remain hidden from imaging spectrometers, patterns of canopy composition correlate strongly with composition and diversity patterns in lower forest strata [61,66]. Therefore, canopy biodiversity may offer an effective proxy for understanding broader community level patterns.

Box 3 Open data and analytical tools [400 words]

If the technological approaches that we advocate for here are to have widespread impact on biodiversity quantification in the coming decades, then the data produced need to be open and accessible to the many researchers working across tropical regions. Equally, the reference libraries necessary to form the taxonomic foundations on which **machine learning models** (see glossary) are based must be carefully curated and validated by expert systematists. Additionally, the computational approaches needed to build classification models require both significant computational expertise and resources, neither of which are possessed by most plant ecologists or systematists working in the tropics. Finally, plant taxonomy and systematics is a dynamic process, and classification algorithms must be flexible to revision if they are to be 'future proof'. Working to reconcile these various requirements presents a major challenge.

Fortunately, existing databasing tools provide several of the key elements required to overcome these challenges. GenBank – an online publicly available database of DNA data for more than 420,000 species – has transformed genetic analyses since its inception [67]. GenBank is already used to store thousands of tropical plant DNA barcodes and full plastomes, and well-developed data pipelines exist for inputting and extracting future collections. In addition, the volume of online voucher specimens with images is increasing all the time, delivered by individual herbaria and aggregated internationally (e.g. [7]), with some exemplar national programmes that have mobilized many small, local collection (e.g. [14,15]. Forestplots.net is an online resource for storing and sharing tree biodiversity and biomass data from tropical regions [68]. Crucially, this online repository now links individual trees to relevant voucher samples and their images, thereby providing a pathway for standardizing and revising identifications across locations. Linking vouchers to associated spectral or DNA reference material would provide the

372 infrastructure that is required to develop the approach we advocate. The Spectranomics and
373 BRIDGE databases provide important examples of how to link voucher samples from tropical
374 trees to coupled spectra and chemical measurements from the same individuals [69–71]. In
375 summary, much of the core databasing infrastructure required to build reference libraries for
376 multidimensional datasets have been developed, but these tools have existed in isolation from
377 one another and are now ripe for integration. Building upon these foundations, and crucially
378 making any future databases publicly accessible, will be essential.

379 Computational literacy among biodiversity researchers has grown enormously in recent decades,
380 particularly within the R environment, but building and training deep-learning classification
381 models still need to be developed by specialist groups. Applying such models to newly collected
382 data will be within the capabilities of many biodiversity researchers, particularly if a companion
383 R package is developed as has been done successfully for the BIEN database [20]. As taxon-
384 specific reference libraries are developed and machine learning models are constructed, they can
385 be rapidly published online (e.g. through GitHub) and seamlessly integrated into existing
386 workflows.

Glossary Box (500 words)

Voucher sample: A dried and pressed plant sample representative of an individual specimen that is used for species identification. Samples can be vegetative (consisting of leaves and small branches) or fertile (including flowers and/or fruits).

Morphotype: A voucher sample that cannot be identified to species, and is therefore given an individual morphospecies code.

Integrative taxonomy: The process of delimiting species by integration of different data types (e.g., morphological characters, chemical characters, DNA sequences), generally in a lineage-based, phylogenetic framework.

Spectroscopy: The study of the interaction between matter (in this case plant leaves or wood) and electromagnetic radiation (in this case frequently infrared radiation). By measuring the radiation that is reflected and absorbed from a sample across a range of wavelengths a spectrum of radiation is produced. This spectrum reflects the chemical and physical properties of the substance (leaf or wood sample) being measured.

Imaging spectroscopy: A branch of remote sensing where, for each pixel of the acquired image, reflected solar radiation is measured across a range of wavelengths, producing a spectrum for each pixel.

DNA barcoding: The process of sequencing short sequences of DNA (400 – 800 base pairs), which can then be used to identify the species of an individual plant. For plants there are four established standard barcodes *rbcL, matK, trnH-psbA*, and ITS2.

Next-generation sequencing: Also called high-throughput sequencing, encompasses a range of modern DNA sequencing approaches that allow for rapid sequencing of far greater quantities of DNA than was possible with traditional Sanger sequencing approaches.

Artificial intelligence (AI): A suite of computational approaches that are able to perform tasks that require intelligent behaviour such as learning and problem solving. Here we include machine learning and deep learning approaches as subfields of AI.

Machine learning: A branch of AI that includes a range of computational algorithms that are able to use training data to make predictions without being programmed explicitly to do so. In this context, machine learning approaches can be used to learn the differences among plant species and then use this learning to classify unknown individuals based on specified features.

Deep learning: Deep learning can be considered a subset of machine learning. Unlike machine learning where relevant features are specified, in deep learning features are not specified, instead the entire dataset and relevant features are identified and used independently. Convolutional Neural Networks (CNNs) are a set of deep learning approaches that are increasingly being used in ecology.

Acknowledgements

FCD is funded by an EU MSC global fellowship 794973 ‘E-FUNDIA’. The authors thank the “Investissement d'avenir” grant from the Agence Nationale de la Recherche (CEBA, ref. ANR-10- LABX-25-01)

References

1 Barlow, J. *et al.* (2018) The future of hyperdiverse tropical ecosystems. *Nature* 559, 517–526

2 Feeley, K. (2015) Are We Filling the Data Void? An Assessment of the Amount and Extent of Plant Collection Records and Census Data Available for Tropical South America. *PLOS ONE* 10, e0125629

3 Díaz, S. *et al.* (2019) IPBES. 2019. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services., IPBES secretariat.

4 Wilson, E.O. (2017) Biodiversity research requires more boots on the ground. *Nature Ecol. Evol.* 1, 1590–1591

5 Ríos-Saldaña, C.A. *et al.* (2018) Are fieldwork studies being relegated to second place in conservation science? *Glob. Ecol. Conserv.* 14, e00389

6 Enquist, B.J. *et al.* (2016) Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Prepr.* 4:e2615v2,

7 GBIF: The Global Biodiversity Information Facility (2020) What is GBIF?. Available from https://www.gbif.org/what-is-gbif [13 January 2020]

8 Gomes, A.C.S. *et al.* (2013) Local plant species delimitation in a highly diverse Amazonian forest: Do we all see the same species? *J Veg. Sci.* 24, 70–79

9 Dexter, K.G. *et al.* (2010) Using DNA to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter? *Ecol. Monogr.* 80, 267–286

10 Cardoso, D. *et al.* (2017) Amazon plant diversity revealed by a taxonomically verified species list. *Proc. Natl. Acad. Sci. U.S.A* 114, 10695–10700

11 Pennington, T.D. (1997) *Genus Inga: Botany*, Royal Botanic Gardens, Kew.

12 Prance, G.T. (1989) Chrysobalanaceae. *Flora Neotrop.* 9, 1–267

13 Goodwin, Z.A. *et al.* (2015) Widespread mistaken identity in tropical plant collections. *Curr. Biol.* 25, R1066–R1067

14 Canteiro, C. *et al.* (2019) Enhancement of conservation knowledge through increased access to botanical information. *Conserv. Biol*. 33, 523–533

15 The speciesLink network (2006). Available from http://splink.cria.org.br/ [17/07/2020]

16 Martinez, R.V. and Phillips, O.L. (2000) Allpahuayo: Floristics, Structure, and Dynamics of a High-Diversity Forest in Amazonian Peru. *Ann. Missouri Bot.* 87, 499

17 Baker, T.R. *et al.* (2017) Maximising Synergy among Tropical Plant Systematists, Ecologists, and Evolutionary Biologists. *Trends Ecol. & Evol.* 32, 258–267

18 Arellano, G. *et al.* (2014) Commonness patterns and the size of the species pool along a tropical elevational gradient: insights using a new quantitative tool. *Ecography* 37, 536–543

19 Pos, E. *et al.* (2014) Are all species necessary to reveal ecologically important patterns? *Ecol. Evol.* 4, 4626–4636

20 Maitner, B.S. *et al.* (2018) The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods Ecol. Evol.* 9, 373–379

21 ter Steege, H. *et al.* (2019) Towards a dynamic list of Amazonian tree species. *Sci. Rep.* 9, 3501

22 De Queiroz, K. (2007) Species Concepts and Species Delimitation. *Syst. Biol.* 56, 879–886

23 Hart, M.L. *et al.* (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65, 1081–1092

24 Kress, W.J. and Erickson, D.L. (2008) DNA Barcoding-a Windfall for Tropical Biology? *Biotropica* 40, 405–408

25 Dick, C.W. and Kress, W.J. (2009) Dissecting Tropical Plant Diversity with Forest Plots and a Molecular Toolkit. *BioScience* 59, 745–755

26 Hollingsworth, P.M. *et al.* (2011) Choosing and using a plant DNA barcode. *PLoS ONE* 6, e19254

27 Hajibabaei, M. *et al.* (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proc. Natl. Acad. Sci. U.S.A* 103, 968–971

28 Gonzalez, M.A. *et al.* (2009) Identification of Amazonian Trees with DNA Barcodes. *PLoS ONE* 4, e7483

29 Parmentier, I. *et al.* (2013) How Effective Are DNA Barcodes in the Identification of African Rainforest Trees? *PLoS ONE* 8, e54921

30 Kress, W.J. *et al.* (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proc. Natl. Acad. Sci. U.S.A* 106, 18621–18626

31 Hollingsworth, P.M. *et al.* (2016) Telling plant species apart with DNA: From barcodes to genomes. *Philos Trans R Soc Lond B Biol Sci*. 371, 20150338

32 Johnson, M.G. *et al.* (2018) A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Syst. Biol.* 68, 594–606

33 Padial, J.M. *et al.* (2010) The integrative future of taxonomy. *Front. Zool.* 7, 16

34 Prata, E.M.B. *et al.* (2018) Towards integrative taxonomy in Neotropical botany: disentangling the Pagamea guianensis species complex (Rubiaceae). *Bot. J. Linn. Soc.* 188, 213–231

35 Damasco, G. *et al.* (2019) Reestablishment of Protium cordatum (Burseraceae) based on integrative taxonomy. *Taxon* 68, 34–46

36 Antonelli, A. *et al.* (2018) Conceptual and empirical advances in Neotropical biodiversity research. *PeerJ* 2018, 6, e5644

37 Asner, G.P. *et al.* (2014) Functional and biological diversity of foliar spectra in tree canopies throughout the Andes to Amazon region. *New Phytol.* 204, 127–139

38 Endara, M.J. *et al.* (2018) Chemocoding as an identification tool where morphological- and DNA-based methods fall short: Inga as a case study. *New Phytol.* 218, 847–858

39 Lang, C. *et al.* (2017) Discrimination of taxonomic identity at species, genus and family levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). *For. Ecol. Manag.* 406, 219–227

40 Lang, C. *et al.* (2015) Near Infrared Spectroscopy Facilitates Rapid Identification of Both Young and Mature Amazonian Tree Species. *PLOS ONE* 10, e0134521

41 Hadlich, H.L. *et al.* (2018) Recognizing Amazonian tree species in the field using bark tissues spectra. *For. Ecol. Manag.* 427, 296–304

42 Durgante, F.M. *et al.* (2013) Species Spectral Signature: Discriminating closely related plant species in the Amazon with Near-Infrared Leaf-Spectroscopy. *For. Ecol. Manag.* 291, 240–248

43 Costa, F.R.C. *et al.* (2018) Near-infrared spectrometry allows fast and extensive predictions of functional traits from dry leaves and branches. *Ecol. Appl.* 28, 1157–1167

44 Asner, G.P. *et al.* (2011) Spectroscopy of canopy chemicals in humid tropical forests. *Remote Sens. Environ.* 115, 3587–3598

45 Asner, G.P. and Martin, R.E. (2008) Spectral and chemical analysis of tropical forests: Scaling from leaf to canopy levels. *Remote Sens. Environ.* 112, 3958–3970

46 Wäldchen, J. and Mäder, P. (2018) Machine learning for image based species identification. *Methods Ecol. Evol.* 9, 2216–2225

47 Brodrick, P.G. *et al.* (2019) Uncovering Ecological Patterns with Convolutional Neural Networks. *Trends Ecol. & Evol.* 34, 734–745

48 Christin, S. *et al.* (2019) Applications for deep learning in ecology. *Methods Ecol. Evol.* 10, 1632–1644

49 Wäldchen, J. *et al.* (2018) Automated plant species identification—Trends and future directions. *PLOS Comput. Biol.* 14, e1005993.

50 Joly, A. *et al.* (2019) , Overview of LifeCLEF 2019: Identification of Amazonian Plants, South & North American Birds, and Niche Prediction. *Lect. Notes Comput. Sci.* 11696, 387–401

51 ter Steege, H. *et al.* (2013) Hyperdominance in the Amazonian tree flora. *Science* 342, 1243092

52 Fauset, S. *et al.* (2015) Hyperdominance in Amazonian forest carbon cycling. *Nat. Commun.* 6, 6857

53 Janzen, D.H. (2004) Now is the time. *Philos Trans R Soc Lond B Biol Sci*. 359 731–732

54 Gaston, K.J. and O'Neill, M.A. (2004) Automated species identification: Why not? *Philos Trans R Soc Lond B Biol Sci*. 359, 655–667

55 Esquivel-Muelbert, A. *et al.* (2019) Compositional response of Amazon forests to climate change. *Glob. Chang. Biol.* 25, 39–56

56 Enquist, B.J. *et al.* (2019) The commonness of rarity: Global and future distribution of rarity across land plants. *Sci. Adv.* 5, eaaz0414

57 Jetz, W. *et al.* (2016) Monitoring plant functional diversity from space. *Nat. Plants* 2, 16024

58 Asner, G.P. *et al.* (2012) Carnegie Airborne Observatory-2: Increasing science data dimensionality via high-fidelity multi-sensor fusion. *Remote Sens. Environ.* 124, 454–465

59 Gomes, V.H.F. *et al.* (2018) Species Distribution Modelling: Contrasting presence-only models with plot abundance data. *Sci. Rep.* 8, 1003

60 Féret, J.-B. and Asner, G.P. (2014) Mapping tropical forest canopy diversity using high-fidelity imaging spectroscopy. *Ecol. Appl.* 24, 1289–1296

61 Draper, F.C. *et al.* (2019) Imaging spectroscopy predicts variable distance decay across contrasting Amazonian tree communities. *J. Ecol.* 107,

62 Asner, G.P. *et al.* (2015) Landscape biogeochemistry reflected in shifting distributions of chemical traits in the Amazon forest canopy. *Nat. Geosci*. 8, 567–575

63 Asner, G.P. *et al.* (2017) Airborne laser-guided imaging spectroscopy to map forest trait diversity and guide conservation. *Science* 355, 385–389

64 Baldeck, C.A. *et al.* (2015) Operational Tree Species Mapping in a Diverse Tropical Forest with Airborne Imaging Spectroscopy. *PLOS ONE* 10, e0118403

65 Asner, G.P. (1998) Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sens. Environ.* 64, 234–253

66 Tuomisto, H. *et al.* (2019) Discovering floristic and geoecological gradients across Amazonia. *J. Biogeogr.* 46, 1734–1748

67 Sayers, E.W. *et al.* (2018) GenBank. *Nucleic Acids Res*.47, D94–D99

68 Lopez-Gonzalez, G. *et al.* (2011) ForestPlots.net: A web application and research tool to manage and analyse tropical forest plot data. *J. Veg. Sci.* 22, 610–613

69 Asner, G.P. and Martin, R.E. (2016) Spectranomics: Emerging science and conservation opportunities at the interface of biodiversity and remote sensing. *Glob. Ecol. Conserv.* 8, 212-219

70 Asner, G.P. and Martin, R.E. (2009) Airborne spectranomics: Mapping canopy chemical and taxonomic diversity in tropical forests. *Front. Ecol. Environ.* 7, 269–276

71 Baraloto, C. *et al.* (2010) Functional trait variation and sampling strategies in species-rich plant communities. *Funct. Ecol.* 24, 208–216

Figure 1: Schematic of possible framework for unifying different approaches and data sources to make high confidence species level identifications using a range of data sources and AI classifications. The Green shaded box represents the start point of specimen collection. Yellow boxes represent different input data types that can be used for species identification or species delimitation. Purple boxes represent different species classification processes, including both human decision-making (hierarchical family classification) and AI approaches. Blue boxes represent different forms of reference material or training data required for the classification approaches. Classification models can be applied to different data types independently, therefore not all types of data are necessary for species identification, although combining different data types (e.g. DNA-barcodes and spectroscopy data) will increase accuracy. Red boxes represent possible incomplete identifications, while red shading indicate the ultimate end point of the framework.