



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/168439/>

Version: Accepted Version

Article:

Draper, FC, Baker, TR, Baraloto, C et al. (2020) Quantifying Tropical Plant Diversity Requires an Integrated Technological Approach. *Trends in Ecology & Evolution*, 35 (12). pp. 1100-1109. ISSN: 0169-5347

<https://doi.org/10.1016/j.tree.2020.08.003>

© 2020, Elsevier. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 **Quantifying tropical plant diversity requires an integrated technological**
2 **approach**

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

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

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

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

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

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

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

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

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

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

19 **Key words**

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

21 **Abstract**

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

32 **The challenge of tropical plant diversity**

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

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

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

65 **Limitations with current process of quantifying tropical plant diversity**

66 Currently, almost all studies seeking to quantify tropical plant diversity are underpinned by
67 morphological botanical approaches to species identification (Box 1). However, attempts to
68 quantify taxonomic uncertainty in large synthetic datasets have revealed substantial errors [8–
69 10]. We suggest that these uncertainties arise from limitations in both underlying taxonomic
70 frameworks (point 1) and the process of species identification (points 2-7):

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

- 75 2. Local herbaria are often relied upon to identify species, but these collections are often
76 incomplete and specimen identifications may not be reliable [13]. Furthermore, specimen
77 identifications are rarely standardised among herbaria, but see [14,15].
- 78 3. Species level identifications in diverse tropical forests often require samples of fruits or
79 flowers. Given the often short and unpredictable phenologies of many tropical species
80 [16], short field research visits can easily miss the reproductive period of species,
81 meaning species level identifications are made on vegetative samples, thereby decreasing
82 their accuracy.
- 83 4. **Voucher samples** (see glossary) from ecological inventories, when collected, frequently
84 lack reproductive structures (flowers and fruit) and are rarely accepted by herbaria.
85 Therefore, ecological inventories typically contribute little to species delimitation and
86 developing taxonomies, despite considerable potential to do so [17].
- 87 5. In practice, identifying species based on morphological characters is, at least to some
88 extent, subjective if it cannot be done by the taxonomic specialist for a given group,
89 which is seldom the case. Identifications by non-specialists vary and depend on previous
90 experience and resources available (i.e., taxonomic monographs, flora accounts and
91 specimens identified by taxonomic specialists).
- 92 6. In many cases, vouchers are not collected for every individual plant within inventory
93 plots. Instead, individuals from the same plot that are deemed to be the same species are
94 grouped together and one or more vouchers are collected to represent that group. This
95 effectively means that the initial judgment of the field botanist introduces uncertainty
96 which is difficult to quantify post-hoc.

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

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

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

114 **Technological approaches to quantifying tropical plant diversity**

115 **DNA approaches**

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

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

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

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

155 **Spectroscopy approaches**

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

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

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

180 Spectroscopic approaches for species identification share many of the advantages that DNA
181 barcoding has over traditional approaches; for example, only vegetative material (which can be
182 older and dried) is required to make identifications which are quantitative and reproducible.
183 While correct use of spectrometers and the analysis of spectral data also requires time and
184 dedication, training in these approaches can be undertaken in weeks to months rather than years.
185 Importantly, spectroscopy holds several key advantages for species identification in addition to
186 those shared with DNA-based approaches. First, spectra reflect not only the taxonomic identity
187 but also several functional traits (e.g. foliar nitrogen and water content) [43–45], which can
188 improve our understanding of the interaction between taxonomic diversity and ecosystem
189 functioning. Second, imaging spectroscopy provides a method for scaling up biodiversity
190 estimates to far greater areas than will ever be possible with field work alone (Box 2). Third,
191 while the initial expense of a precise lab-based spectrometer is not insignificant, many thousands
192 of samples can be processed with relatively modest maintenance and operation costs and can be
193 operated in the field or herbarium without the need of a wet lab.

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

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

202 **Artificial intelligence (AI) approaches**

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

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

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

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

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

239 **Developing a framework for progress**

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

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

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

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

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

270 **Concluding remarks**

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

278 There have been several previous calls to leverage different forms of technology to revolutionize
279 species identification [53,54]. We are now at a stage where the technology has come of age and
280 necessary tools for identification are available, affordable, and tested. It is time to move beyond
281 demonstrating the capabilities of these tools through small scale comparisons, and instead begin
282 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).

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

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

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

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

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

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

318

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

320 A major advantage of spectroscopic approaches is that imaging spectrometers can be mounted on
321 airborne and satellite platforms, and therefore can be used to scale-up biodiversity estimates
322 across vast spatial scales (e.g. 10^6 km²) [57,58]. This is important because most tropical forests
323 occur in vast inaccessible areas of wilderness, and accumulating field data over such large scales
324 would be impossible. Furthermore, existing approaches for scaling up ground-based biodiversity
325 estimates across large areas of tropical forests have had limited success. For example, species
326 distribution modelling approaches perform poorly in tropical forest regions because climate and
327 edaphic gradients are either poorly characterized at relevant spatial scales (e.g., soil fertility) or
328 represent relatively narrow breadth across large areas (e.g., precipitation). Indeed, equivalent
329 performance to describe species distributions can be obtained through simple spatial
330 extrapolation [59].

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

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

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

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

348

349 **Box 3 Open data and analytical tools [400 words]**

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

360 Fortunately, existing databasing tools provide several of the key elements required to overcome
361 these challenges. GenBank – an online publicly available database of DNA data for more than
362 420,000 species – has transformed genetic analyses since its inception [67]. GenBank is already
363 used to store thousands of tropical plant DNA barcodes and full plastomes, and well-developed
364 data pipelines exist for inputting and extracting future collections. In addition, the volume of
365 online voucher specimens with images is increasing all the time, delivered by individual herbaria
366 and aggregated internationally (e.g. [7]), with some exemplar national programmes that have
367 mobilized many small, local collection (e.g. [14,15]. Forestplots.net is an online resource for
368 storing and sharing tree biodiversity and biomass data from tropical regions [68]. Crucially, this
369 online repository now links individual trees to relevant voucher samples and their images,
370 thereby providing a pathway for standardizing and revising identifications across locations.
371 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.

387 **Glossary Box (500 words)**

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

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

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

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

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

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

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

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

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

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

422

423 **Acknowledgements**

424 FCD is funded by an EU MSC global fellowship 794973 ‘E-FUNDIA’. The authors thank the
425 “Investissement d’avenir” grant from the Agence Nationale de la Recherche (CEBA, ref. ANR-
426 10- LABX-25-01)

427 **References**

- 428 1 Barlow, J. *et al.* (2018) The future of hyperdiverse tropical ecosystems. *Nature* 559, 517–
429 526
- 430 2 Feeley, K. (2015) Are We Filling the Data Void? An Assessment of the Amount and
431 Extent of Plant Collection Records and Census Data Available for Tropical South
432 America. *PLOS ONE* 10, e0125629
- 433 3 Díaz, S. *et al.* (2019) IPBES. 2019. Summary for policymakers of the global assessment
434 report on biodiversity and ecosystem services of the Intergovernmental Science-Policy
435 Platform on Biodiversity and Ecosystem Services., IPBES secretariat.
- 436 4 Wilson, E.O. (2017) Biodiversity research requires more boots on the ground. *Nature*
437 *Ecol. Evol.* 1, 1590–1591
- 438 5 Ríos-Saldaña, C.A. *et al.* (2018) Are fieldwork studies being relegated to second place in
439 conservation science? *Glob. Ecol. Conserv.* 14, e00389
- 440 6 Enquist, B.J. *et al.* (2016) Cyberinfrastructure for an integrated botanical information
441 network to investigate the ecological impacts of global climate change on plant
442 biodiversity. *PeerJ Prepr.* 4:e2615v2,
- 443 7 GBIF: The Global Biodiversity Information Facility (2020) What is GBIF?. Available
444 from <https://www.gbif.org/what-is-gbif> [13 January 2020]

- 445 8 Gomes, A.C.S. *et al.* (2013) Local plant species delimitation in a highly diverse
446 Amazonian forest: Do we all see the same species? *J Veg. Sci.* 24, 70–79
- 447 9 Dexter, K.G. *et al.* (2010) Using DNA to assess errors in tropical tree identifications:
448 How often are ecologists wrong and when does it matter? *Ecol. Monogr.* 80, 267–286
- 449 10 Cardoso, D. *et al.* (2017) Amazon plant diversity revealed by a taxonomically verified
450 species list. *Proc. Natl. Acad. Sci. U.S.A* 114, 10695–10700
- 451 11 Pennington, T.D. (1997) *Genus Inga: Botany*, Royal Botanic Gardens, Kew.
- 452 12 Prance, G.T. (1989) Chrysobalanaceae. *Flora Neotrop.* 9, 1–267
- 453 13 Goodwin, Z.A. *et al.* (2015) Widespread mistaken identity in tropical plant collections.
454 *Curr. Biol.* 25, R1066–R1067
- 455 14 Canteiro, C. *et al.* (2019) Enhancement of conservation knowledge through increased
456 access to botanical information. *Conserv. Biol.* 33, 523–533
- 457 15 The speciesLink network (2006). Available from <http://splink.cria.org.br/> [17/07/2020]
- 458 16 Martinez, R.V. and Phillips, O.L. (2000) Allpahuayo: Floristics, Structure, and Dynamics
459 of a High-Diversity Forest in Amazonian Peru. *Ann. Missouri Bot.* 87, 499
- 460 17 Baker, T.R. *et al.* (2017) Maximising Synergy among Tropical Plant Systematists,
461 Ecologists, and Evolutionary Biologists. *Trends Ecol. & Evol.* 32, 258–267
- 462 18 Arellano, G. *et al.* (2014) Commonness patterns and the size of the species pool along a
463 tropical elevational gradient: insights using a new quantitative tool. *Ecography* 37, 536–
464 543
- 465 19 Pos, E. *et al.* (2014) Are all species necessary to reveal ecologically important patterns?
466 *Ecol. Evol.* 4, 4626–4636

- 467 20 Maitner, B.S. *et al.* (2018) The bien r package: A tool to access the Botanical
468 Information and Ecology Network (BIEN) database. *Methods Ecol. Evol.* 9, 373–379
- 469 21 ter Steege, H. *et al.* (2019) Towards a dynamic list of Amazonian tree species. *Sci. Rep.*
470 9, 3501
- 471 22 De Queiroz, K. (2007) Species Concepts and Species Delimitation. *Syst. Biol.* 56, 879–
472 886
- 473 23 Hart, M.L. *et al.* (2016) Retrieval of hundreds of nuclear loci from herbarium specimens.
474 *Taxon* 65, 1081–1092
- 475 24 Kress, W.J. and Erickson, D.L. (2008) DNA Barcoding-a Windfall for Tropical Biology?
476 *Biotropica* 40, 405–408
- 477 25 Dick, C.W. and Kress, W.J. (2009) Dissecting Tropical Plant Diversity with Forest Plots
478 and a Molecular Toolkit. *BioScience* 59, 745–755
- 479 26 Hollingsworth, P.M. *et al.* (2011) Choosing and using a plant DNA barcode. *PLoS ONE*
480 6, e19254
- 481 27 Hajibabaei, M. *et al.* (2006) DNA barcodes distinguish species of tropical Lepidoptera.
482 *Proc. Natl. Acad. Sci. U.S.A* 103, 968–971
- 483 28 Gonzalez, M.A. *et al.* (2009) Identification of Amazonian Trees with DNA Barcodes.
484 *PLoS ONE* 4, e7483
- 485 29 Parmentier, I. *et al.* (2013) How Effective Are DNA Barcodes in the Identification of
486 African Rainforest Trees? *PLoS ONE* 8, e54921
- 487 30 Kress, W.J. *et al.* (2009) Plant DNA barcodes and a community phylogeny of a tropical
488 forest dynamics plot in Panama. *Proc. Natl. Acad. Sci. U.S.A* 106, 18621–18626

- 489 31 Hollingsworth, P.M. *et al.* (2016) Telling plant species apart with DNA: From barcodes
490 to genomes. *Philos Trans R Soc Lond B Biol Sci.* 371, 20150338
- 491 32 Johnson, M.G. *et al.* (2018) A Universal Probe Set for Targeted Sequencing of 353
492 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Syst.*
493 *Biol.* 68, 594–606
- 494 33 Padial, J.M. *et al.* (2010) The integrative future of taxonomy. *Front. Zool.* 7, 16
- 495 34 Prata, E.M.B. *et al.* (2018) Towards integrative taxonomy in Neotropical botany:
496 disentangling the *Pagamea guianensis* species complex (Rubiaceae). *Bot. J. Linn. Soc.*
497 188, 213–231
- 498 35 Damasco, G. *et al.* (2019) Reestablishment of *Protium cordatum* (Burseraceae) based on
499 integrative taxonomy. *Taxon* 68, 34–46
- 500 36 Antonelli, A. *et al.* (2018) Conceptual and empirical advances in Neotropical biodiversity
501 research. *PeerJ* 2018, 6, e5644
- 502 37 Asner, G.P. *et al.* (2014) Functional and biological diversity of foliar spectra in tree
503 canopies throughout the Andes to Amazon region. *New Phytol.* 204, 127–139
- 504 38 Endara, M.J. *et al.* (2018) Chemocoding as an identification tool where morphological-
505 and DNA-based methods fall short: *Inga* as a case study. *New Phytol.* 218, 847–858
- 506 39 Lang, C. *et al.* (2017) Discrimination of taxonomic identity at species, genus and family
507 levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). *For. Ecol.*
508 *Manag.* 406, 219–227
- 509 40 Lang, C. *et al.* (2015) Near Infrared Spectroscopy Facilitates Rapid Identification of Both
510 Young and Mature Amazonian Tree Species. *PLOS ONE* 10, e0134521

- 511 41 Hadlich, H.L. *et al.* (2018) Recognizing Amazonian tree species in the field using bark
512 tissues spectra. *For. Ecol. Manag.* 427, 296–304
- 513 42 Durgante, F.M. *et al.* (2013) Species Spectral Signature: Discriminating closely related
514 plant species in the Amazon with Near-Infrared Leaf-Spectroscopy. *For. Ecol. Manag.*
515 291, 240–248
- 516 43 Costa, F.R.C. *et al.* (2018) Near-infrared spectrometry allows fast and extensive
517 predictions of functional traits from dry leaves and branches. *Ecol. Appl.* 28, 1157–1167
- 518 44 Asner, G.P. *et al.* (2011) Spectroscopy of canopy chemicals in humid tropical forests.
519 *Remote Sens. Environ.* 115, 3587–3598
- 520 45 Asner, G.P. and Martin, R.E. (2008) Spectral and chemical analysis of tropical forests:
521 Scaling from leaf to canopy levels. *Remote Sens. Environ.* 112, 3958–3970
- 522 46 Wäldchen, J. and Mäder, P. (2018) Machine learning for image based species
523 identification. *Methods Ecol. Evol.* 9, 2216–2225
- 524 47 Brodrick, P.G. *et al.* (2019) Uncovering Ecological Patterns with Convolutional Neural
525 Networks. *Trends Ecol. & Evol.* 34, 734–745
- 526 48 Christin, S. *et al.* (2019) Applications for deep learning in ecology. *Methods Ecol. Evol.*
527 10, 1632–1644
- 528 49 Wäldchen, J. *et al.* (2018) Automated plant species identification—Trends and future
529 directions. *PLOS Comput. Biol.* 14, e1005993.
- 530 50 Joly, A. *et al.* (2019) , Overview of LifeCLEF 2019: Identification of Amazonian Plants,
531 South & North American Birds, and Niche Prediction. *Lect. Notes Comput. Sci.* 11696,
532 387–401

- 533 51 ter Steege, H. *et al.* (2013) Hyperdominance in the Amazonian tree flora. *Science* 342,
534 1243092
- 535 52 Fauset, S. *et al.* (2015) Hyperdominance in Amazonian forest carbon cycling. *Nat.*
536 *Commun.* 6, 6857
- 537 53 Janzen, D.H. (2004) Now is the time. *Philos Trans R Soc Lond B Biol Sci.* 359 731–732
- 538 54 Gaston, K.J. and O’Neill, M.A. (2004) Automated species identification: Why not?
539 *Philos Trans R Soc Lond B Biol Sci.* 359, 655–667
- 540 55 Esquivel-Muelbert, A. *et al.* (2019) Compositional response of Amazon forests to climate
541 change. *Glob. Chang. Biol.* 25, 39–56
- 542 56 Enquist, B.J. *et al.* (2019) The commonness of rarity: Global and future distribution of
543 rarity across land plants. *Sci. Adv.* 5, eaaz0414
- 544 57 Jetz, W. *et al.* (2016) Monitoring plant functional diversity from space. *Nat. Plants* 2,
545 16024
- 546 58 Asner, G.P. *et al.* (2012) Carnegie Airborne Observatory-2: Increasing science data
547 dimensionality via high-fidelity multi-sensor fusion. *Remote Sens. Environ.* 124, 454–
548 465
- 549 59 Gomes, V.H.F. *et al.* (2018) Species Distribution Modelling: Contrasting presence-only
550 models with plot abundance data. *Sci. Rep.* 8, 1003
- 551 60 Féret, J.-B. and Asner, G.P. (2014) Mapping tropical forest canopy diversity using high-
552 fidelity imaging spectroscopy. *Ecol. Appl.* 24, 1289–1296
- 553 61 Draper, F.C. *et al.* (2019) Imaging spectroscopy predicts variable distance decay across
554 contrasting Amazonian tree communities. *J. Ecol.* 107,

- 555 62 Asner, G.P. *et al.* (2015) Landscape biogeochemistry reflected in shifting distributions of
556 chemical traits in the Amazon forest canopy. *Nat. Geosci.* 8, 567–575
- 557 63 Asner, G.P. *et al.* (2017) Airborne laser-guided imaging spectroscopy to map forest trait
558 diversity and guide conservation. *Science* 355, 385–389
- 559 64 Baldeck, C.A. *et al.* (2015) Operational Tree Species Mapping in a Diverse Tropical
560 Forest with Airborne Imaging Spectroscopy. *PLOS ONE* 10, e0118403
- 561 65 Asner, G.P. (1998) Biophysical and biochemical sources of variability in canopy
562 reflectance. *Remote Sens. Environ.* 64, 234–253
- 563 66 Tuomisto, H. *et al.* (2019) Discovering floristic and geocological gradients across
564 Amazonia. *J. Biogeogr.* 46, 1734–1748
- 565 67 Sayers, E.W. *et al.* (2018) GenBank. *Nucleic Acids Res.* 47, D94–D99
- 566 68 Lopez-Gonzalez, G. *et al.* (2011) ForestPlots.net: A web application and research tool to
567 manage and analyse tropical forest plot data. *J. Veg. Sci.* 22, 610–613
- 568 69 Asner, G.P. and Martin, R.E. (2016) Spectranomics: Emerging science and conservation
569 opportunities at the interface of biodiversity and remote sensing. *Glob. Ecol. Conserv.* 8,
570 212-219
- 571 70 Asner, G.P. and Martin, R.E. (2009) Airborne spectranomics: Mapping canopy chemical
572 and taxonomic diversity in tropical forests. *Front. Ecol. Environ.* 7, 269–276
- 573 71 Baraloto, C. *et al.* (2010) Functional trait variation and sampling strategies in species-
574 rich plant communities. *Funct. Ecol.* 24, 208–216

575

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