



**HAL**  
open science

## The dimensionality and structure of species trait spaces

David Mouillot, Nicolas Loiseau, Matthias Grenié, Adam Algar, Michele Allegra, Marc Cadotte, Nicolas Casajus, Pierre Denelle, Maya Guéguen, Anthony Maire, et al.

► **To cite this version:**

David Mouillot, Nicolas Loiseau, Matthias Grenié, Adam Algar, Michele Allegra, et al.. The dimensionality and structure of species trait spaces. *Ecology Letters*, 2021, 24 (9), pp.1988-2009. 10.1111/ele.13778 . hal-03405358

**HAL Id: hal-03405358**

**<https://hal.science/hal-03405358v1>**

Submitted on 6 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The dimensionality and structure of species trait spaces

David Mouillot<sup>1,2,\*</sup>, Nicolas Loiseau<sup>1,\*</sup>, Matthias Grenié<sup>3</sup>, Adam C. Algar<sup>4</sup>, Michele Allegra<sup>5</sup>, Marc W. Cadotte<sup>6</sup>, Nicolas Casajus<sup>7</sup>, Pierre Denelle<sup>8</sup>, Maya Guéguen<sup>9</sup>, Anthony Maire<sup>10</sup>, Brian Maitner<sup>11</sup>, Brian McGill<sup>12</sup>, Matthew McLean<sup>13</sup>, Nicolas Mouquet<sup>1,7</sup>, François Munoz<sup>14</sup>, Wilfried Thuiller<sup>9</sup>, Sébastien Villéger<sup>1</sup>, Cyrille Violle<sup>3</sup> & Arnaud Auber<sup>15</sup>

## Affiliations

<sup>1</sup> MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France

<sup>2</sup> Institut Universitaire de France, IUF, Paris 75231, France

<sup>3</sup> Centre d'Ecologie Fonctionnelle et Evolutive—UMR 5175 CEFÉ, Univ Montpellier, CNRS, IRD, EPHE, Univ Paul Valéry, Montpellier, France

<sup>4</sup> Department of Biology, Lakehead University, Thunder Bay, Ontario, Canada

<sup>5</sup> Institut de Neurosciences de la Timone, Aix Marseille Université, UMR 7289 CNRS, 13005, Marseille, France

<sup>6</sup> Department of Biological Sciences, University of Toronto-Scarborough, Toronto, ON, Canada

<sup>7</sup> FRB—CESAB, Institut Bouisson Bertrand. 5, rue de l'École de médecine, 34000, Montpellier, France

<sup>8</sup> Biodiversity, Macroecology & Biogeography, University of Goettingen, Göttingen, Germany

<sup>9</sup> Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France.

<sup>10</sup> EDF R&D, LNHE (Laboratoire National d'Hydraulique et Environnement), Chatou, France

<sup>11</sup> Department of Ecology and Evolutionary Biology, University of Connecticut, Connecticut, USA

<sup>12</sup> School of Biology and Ecology and Mitchell Center for Sustainability Solutions, University of Maine, Orono, Maine, USA

<sup>13</sup> Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

<sup>14</sup> LiPhy (Laboratoire Interdisciplinaire de Physique), Université Grenoble-Alpes, Grenoble, France

<sup>15</sup> IFREMER, Unité Halieutique Manche Mer du Nord, Laboratoire Ressources Halieutiques Boulogne-sur-Mer, France.

\*Co-first authors

**Running title:** The structure of species trait spaces

**Keywords:** Functional ecology, species clustering, species uniqueness, hypervolume, complexity

37 **Statement of authorship:** All authors conceived the ideas and designed the methodology;  
38 NL, MG and AA collected the data; DM, NL, MG, NC, MG and AA analyzed the data and  
39 participated in script development; DM, NL and AA led the writing of the manuscript. All  
40 authors contributed critically to the drafts and gave final approval for submission.

41 **Data accessibility statement:** All relevant R codes are available from the GitHub Repository:  
42 <https://github.com/LoiseauN/dimensionality>. Code for the Elbow approach is also available  
43 from the GitHub Repository: <https://github.com/ahasverus/elbow>. All data are available  
44 (Table 1).

45 **Type of article:** Reviews and Syntheses

46 **Number of words in the main text:** 7788

47 **Number of words in the abstract:** 199

48 **Number of references:** 111

49 **Number of figures:** 9

50 **Number of tables:** 1

51

52 **Corresponding author details:** mailing address: University of Montpellier, 93 Place Eugène  
53 Bataillon, 34090 Montpellier, France; email address: david.mouillot@umontpellier.fr; phone  
54 number: + 33 6 09 47 21 47

55

56

57 **Abstract**

58

59 Trait-based ecology aims to understand the processes that generate the overarching diversity  
60 of organismal traits and their influence on ecosystem functioning. Achieving this goal  
61 requires simplifying this complexity in synthetic axes defining a trait space and to cluster  
62 species based on their traits while identifying those with unique combinations of traits.  
63 However, so far, we know little about the dimensionality, the robustness to trait omission, and  
64 the structure of these trait spaces. Here, we propose a unified framework and a synthesis  
65 across 30 trait datasets representing a broad variety of taxa, ecosystems and spatial scales to  
66 show that a common trade-off between trait space quality and operability appears between  
67 3 and 6 dimensions. The robustness to trait omission is generally low but highly variable  
68 among datasets. We also highlight invariant scaling relationships, whatever organismal  
69 complexity, between the number of clusters, the number of species in the dominant cluster  
70 and the number of unique species with total species richness. When species richness  
71 increases, the number of unique species saturates, while species tend to disproportionately  
72 pack in the richest cluster. Based on these results, we propose some rules of thumb to build  
73 species trait spaces and estimate subsequent functional diversity indices.

74

75

76

77

## 78 **Introduction**

79

80 Biodiversity comprises a great variety of organismal forms, functions, diets, physiologies and  
81 life histories — hereafter called traits — that have been shaped by large-scale evolutionary  
82 and ecological processes (Schluter 1993; Reich *et al.* 1999) and that have important  
83 implications for ecosystem functioning (Hector *et al.* 1999; Duffy *et al.* 2001). Thus,  
84 quantifying and characterizing trait variation among species is key to understand species  
85 assembly rules (Bruehlheide *et al.* 2018; Jarzyna *et al.* 2020), evolutionary dynamics (Deline *et al.*  
86 *et al.* 2018; Pigot *et al.* 2020), and ecosystem functioning (Gagic *et al.* 2015; Cadotte 2017) but  
87 also to predict biodiversity responses to global changes (McLean *et al.* 2019; R uger *et al.*  
88 2020) and to guide conservation efforts (Pollock *et al.* 2017; Sala *et al.* 2021). For instance,  
89 experiments show that plant communities with higher levels of trait diversity are more  
90 productive and have a higher resource use efficiency by intercepting more light, taking up  
91 more nitrogen, and occupying more of the available space (Spehn *et al.* 2005) but can also  
92 limit plant disease risks (Le Bagousse-Pinguet *et al.* 2021).

93

94 Yet, owing to the increasing availability of widespread — but also incomplete and  
95 heterogeneous — information on multiple traits collected with various methods across most  
96 kingdoms of life (Jones *et al.* 2009; Schneider *et al.* 2017; Perez *et al.* 2019; Kattge *et al.*  
97 2020), the characterization of species ecological strategies and relationships with  
98 environmental conditions is becoming more complex and multidimensional than ever  
99 (Villeger *et al.* 2011; Bruehlheide *et al.* 2018). Reducing this complexity has both theoretical  
100 and practical benefits. First, clustering thousands of species into a limited number of entities  
101 sharing similar trait values can reveal the amount of functional vulnerability within  
102 assemblages (Mouillot *et al.* 2014) or a functional backbone common to separate geographic  
103 realms (McLean *et al.* 2021). Second, many traits are strongly correlated owing to life-history  
104 trade-offs or adaptive constraints, suggesting that trait diversity within a clade is more limited  
105 than expected (Winemiller *et al.* 2015; D iaz *et al.* 2016; Pigot *et al.* 2020). Birds with  
106 relatively long, narrow wings, pointed tips, and strong sweep back (such as those of a  
107 swallow) fly at high speeds but are energetically inefficient and cannot fly over long distances  
108 (Savile 1957). Third, the hyper-dimensionality of trait spaces, where species are placed  
109 according to their combinations of traits, prevents the computation of hypervolume-based  
110 functional diversity indices or null models to test community assembly hypotheses (Blonder  
111 *et al.* 2014; Maire *et al.* 2015). Fourth, predicting biodiversity and ecosystem trajectories

112 under various environmental scenarios needs parsimonious trait-based models (Barros *et al.*  
113 2017; Cooke *et al.* 2019b; Ruger *et al.* 2020) since the use of too many traits may induce  
114 overfitting (Bernhardt-Romerer *et al.* 2008).

115

116 However, we still lack a unified methodological framework to assess the different aspects of a  
117 species trait space. The dimensionality and the structure of a species trait space are indeed two  
118 sides of the same coin since they both refer to its complexity, i.e. the way species and their  
119 traits are organized in this space. We also lack a synthesis on the main factors shaping the  
120 different aspects of species trait spaces. The degree of organismal complexity, which is  
121 related to the diversity of cell types (Valentine *et al.* 1994), can indeed influence the  
122 complexity of species trait space following key functional innovation in multicellular clades  
123 (Knoll 2011; Cox *et al.* 2021; Sosiak & Barden 2021). The environment can also be crucial in  
124 determining the course of multicellular evolution and organismal complexity, with  
125 aggregative multicellularity evolving more frequently on land while clonal multicellularity is  
126 more frequent in water (Fisher *et al.* 2020). On the other hand, the number of species and trait  
127 characteristics are likely to influence the complexity of species trait spaces beyond the type of  
128 organism and the environment (Zhu *et al.* 2017; Kohli & Jarzyna 2021). Yet, the relative  
129 importance of these different potential drivers has never been tested across kingdoms and  
130 realms for a vast number and diversity of traits and taxa.

131

132 A first critical aspect of a species trait space refers to the well-known dimensionality issue  
133 (Laughlin 2014; Maire *et al.* 2015). While dimension reduction is appealing, the devil lies in  
134 the details. Indeed, going from a large number of traits to a reduced trait space (Figure 1a-d),  
135 that represents meaningful ecological dimensions or axes, is conceptually and  
136 methodologically difficult (Maire *et al.* 2015; Winemiller *et al.* 2015; Pigot *et al.* 2020;  
137 Sosiak & Barden 2021). High-dimensional spaces might indeed be required to fully capture  
138 trait variation among species (Carscadden *et al.* 2017) or clades (Cooney *et al.* 2017).  
139 Moreover, the extent to which collected traits, some being potentially uninformative,  
140 redundant or incomplete, can be summarized with a few dimensions to reliably represent the  
141 diversity of organism forms and functions has not been quantitatively tested across a large set  
142 of taxa, ecosystems and traits.

143

144 A second key aspect of any species trait space is its robustness to the choice or the omission  
145 of traits so its capacity to consistently position species relative to each other whatever the sub-

146 selection of traits for a given goal (environmental filtering, competitive interactions, etc.).  
147 This capacity ultimately determines the confidence by which we can estimate metrics like  
148 species trait dissimilarity or functional diversity (Carscadden *et al.* 2017; Zhu *et al.* 2017;  
149 Kohli & Jarzyna 2021). However, this robustness has been largely overlooked and deserves a  
150 dedicated analysis across multiple datasets where the number, completeness, correlation, and  
151 type of traits cover a broad range of options.

152

153 A third key aspect of any species trait space relates to its structure, and particularly how  
154 species are distributed and clustered in that space. Species with very similar traits are likely to  
155 play comparable roles in ecosystems (Dehling *et al.* 2016; Pigot *et al.* 2020; Sosiak & Barden  
156 2021), and are packed within a trait space into clusters (Figure 1d). The size (i.e. species  
157 richness) of these clusters relates to functional redundancy (Walker 1992; Fonseca & Ganade  
158 2001), which could act as an insurance against the loss of certain combinations of traits and  
159 the disruption of ecosystem functioning under disturbance (Sanders *et al.* 2018; McLean *et al.*  
160 2019). The other side of the same coin is functional uniqueness represented by species having  
161 no neighbors in the trait space owing to their unique combinations of traits (species B and D  
162 in Figure 1d). Several studies suggested that, beyond the positive influence of species trait  
163 diversity on ecosystem functioning (Gross *et al.* 2017; Craven *et al.* 2018), these unique  
164 species can play key and irreplaceable functional roles (O'Gorman *et al.* 2011; Pigot *et al.*  
165 2016a; Maire *et al.* 2018; Le Bagousse-Pinguet *et al.* 2021). The filling of this trait space  
166 through evolutionary history, and more particularly the emergence of species with unique  
167 traits, has also motivated numerous studies investigating specialization in clades or  
168 competition footprint across the Tree of life (Ricklefs 2010; Cornwell *et al.* 2014; Stubbs &  
169 Benton 2016; Phillips *et al.* 2018; Jarzyna *et al.* 2020; Cox *et al.* 2021). Yet, we still lack a  
170 flexible framework in which the number and composition of species clusters but also unique  
171 species are automatically detected regardless of the shape, the density in terms of species  
172 richness and the dimensionality of the trait space in which they are embedded.

173

174 Here, we propose a unified and flexible framework to assess (i) the optimal number of axes  
175 representing species trait diversity (dimensionality), (ii) the consistency of the trait space in  
176 species placement when sub-setting a limited number of traits (robustness), and (iii) the  
177 distribution of species among clusters including the proportion of unique species (structure).  
178 To better understand the drivers of these three key aspects, we apply our framework on 30  
179 trait datasets spreading across most kingdoms of life (e.g. bacteria, plants, vertebrates) and

180 biomes (terrestrial and marine) at different scales (local to global), and spanning two orders of  
181 magnitude in species richness and one order of magnitude in the number of traits with  
182 different types (e.g. continuous, categorical, etc..) and varying proportions of missing values  
183 (Table 1). To disentangle the drivers of trait space complexity, we then model the  
184 dimensionality, the robustness and the structure of these 30 trait spaces as a function of the  
185 type of species, the type of ecosystem, the number of species, the number of traits, the type of  
186 traits, the correlation between traits and the proportion of missing values. Ultimately, we  
187 provide guidance to deal with the heterogeneity and incompleteness of species trait databases  
188 when building species trait spaces and assessing trait-based metrics in community ecology,  
189 evolution and biogeography.

190

## 191 **Materials and Methods**

### 192 **Building species trait space**

193 Among the myriad of methods proposed to reduce the dimensionality of data (Laughlin 2014;  
194 Kraemer *et al.* 2018; Nguyen & Holmes 2019), we chose one that is commonly used in ecology,  
195 based on well-established ordination techniques, and flexible enough to be adapted to any kind  
196 of trait data. Our goal is not to review or compare existing methods but rather to assemble a  
197 suite of methods able to extract the main features of any species trait space and test their drivers.

198

199 First, we calculated trait dissimilarity between species pairs using the Gower pairwise distance  
200 (Gower & Legendre 1986). This metric can handle multiple types of data (e.g., categorical,  
201 ordinal and continuous traits) and is also less sensitive to missing values than other distance  
202 estimation methods (Podani & Schmera 2006; Pavoine *et al.* 2009). The dissimilarity between  
203 two species is only evaluated on traits with known values for both species but this dissimilarity  
204 is standardized across all pairs whatever the number of traits considered. This step (Figure 1b)  
205 was carried out with the *daisy()* function in the *cluster* R package.

206

207 Second, we performed ordination of species in a space of reduced dimensionality by mean of  
208 Principal Coordinates Analysis (PCoA), which identifies orthogonal axes along which trait  
209 dissimilarity is decomposed (Legendre & Legendre 1998). For this step (Figure 1c) we used  
210 the *pcoa()* function in the *ape* R package.

211

### 212 **Quality of species trait space**



213 To assess the dimensionality and robustness of species trait spaces, we needed a metric  
214 measuring the degree of distortion between the initial trait distance matrix between species pairs  
215 (Gower distance on all traits) and the distance matrix after dimensionality reduction (Euclidean  
216 distance on PCoA axes) or after removing traits (Gower distance on the sub-selection of traits),  
217 respectively. We assumed that a trait space is a high-quality representation of the full dataset if  
218 distances between species in that space are close to the initial distances computed with all traits  
219 (Maire *et al.* 2015). The approach of comparing the similarity of two distance matrices has  
220 precedent in Mantel tests (Legendre & Legendre 1998), although the end goal here is quite  
221 different – producing a metric of robustness for low-dimensional trait space. Indeed, Mantel  
222 tests only correlate values or ranks between two distance matrices, ignoring the global co-  
223 ranking between species and their neighborhood which are key features of species trait space  
224 when the ultimate goal is to cluster species and identify functionally unique ones (Pimiento *et*  
225 *al.* 2020a).

226

227 Several measures of trait space quality have been proposed (Mérigot *et al.* 2010; Maire *et al.*  
228 2015), but we chose a new one in the field of ecology with five key properties that overcome  
229 classical limitations: (i) being unitless so independent of the number, range or value of traits,  
230 (ii) being standardized between 0 and 1 with a clear and intuitive interpretation of these extreme  
231 values, (iii) avoiding the dilemma of whether or not to square the error, which arises in distance-  
232 based quality metrics, (iv) being asymmetric by construction so only considering that the lower-  
233 dimensional distance matrix is a poorer representation of species distribution in trait space  
234 compared to the initial distance matrix, and (v) proposing a common, albeit arbitrary, threshold  
235 to define quality.

236

237 This method is based on the co-ranking matrix  $Q$  which compares the ranking of distance  
238 between objects in the initial distance matrix and in a lower-dimensional space (Lee &  
239 Verleysen 2009). In our case, let us denote by  $\delta_{i,j}$  the distance between species  $i$  and  $j$  in the  
240 initial trait matrix (Figure 1a) and  $d_{i,j}$  their distance in the lower-dimension matrix (Figure 1c).  
241 Then, for any fixed species  $i$ , we assessed the ranks of the distances between this species  $i$  and  
242 all other  $S-1$  species  $j$  in both the initial and the lower-dimensional matrix denoted as  $\rho_{i,j}$  and  
243  $r_{i,j}$ , respectively. These ranks varied between 1 and  $(S-1)$  with  $S$  being the total number of  
244 species. The co-ranking matrix  $Q$  is of size  $(S-1)$  by  $(S-1)$  and has for elements the number of  
245 species pairs that have the rank  $k$  in the initial (all traits) Gower distance matrix and the rank  $l$   
246 in the lower-dimensional (PCoA axes) Euclidean distance matrix (Figure 1e). Since the roles

247 played by species  $i$  and species  $j$  are asymmetric, matrix  $Q$  sums at  $S(S-1)$ , so the total number  
248 of pairs ( $S-1$ ) made by each of the  $S$  species.

249

250 Then, we defined the rank error to be the difference  $\rho_{i,j} - r_{i,j}$ . If there is no error, i.e. a perfect  
251 match in species neighbors between the initial and the lower-dimensional distance matrices,  
252 then  $Q$  is a diagonal matrix, i.e. ranks  $k$  and  $l$  will be similar so  $\rho_{i,j} - r_{i,j}=0$  for all species pairs.  
253 At the opposite, rank mismatches or errors, due to dimensionality reduction or trait omission,  
254 induce off-diagonal species pairs in this co-ranking matrix (Figure 1e). These off-diagonal  
255 species pairs represent pairs that come at a lower distance rank (intrusion) or at a higher distance  
256 rank (extrusion) in the lower-dimensional space compared to the initial space (Lee & Verleyesen  
257 2009).

258

259 To assess whether the lower-dimensional space was a good representation of the initial space,  
260 we needed an asymmetric measure. In other words, a measure that compares the ranks of  
261 species pairs in the lower-dimensional matrix to those of the initial matrix and not the way  
262 around. A spearman-rank correlation is symmetric (the correlation between A and B equals the  
263 correlation between B and A) since it compares the ranks without any primary structure like in  
264 Mantel tests. We thus chose the Area Under the Curve (AUC) criteria, which is based on the  
265 Somer's  $D$  statistic, as an asymmetric rank measure (Somers 1962). AUC is unitless and varies  
266 between 0 and 1. A value of 1 represents the best-case scenario where the ranking of species  
267 pairs would be perfectly preserved between the initial and the lower-dimensional distance  
268 matrix (Kraemer *et al.* 2018). A rule of thumb to interpret this metric is that above 0.7  
269 dimensionality reduction can be considered as good or acceptable and above 0.8 as excellent.  
270 Below 0.5 the lower-dimensional space is a poor representation of the initial trait space while  
271 0 means as good as random. It corresponds to the null or independence hypothesis in Mantel  
272 tests (Legendre & Legendre 1998). More details can be found in Kraemer *et al.* (2018) who  
273 developed the *dimRed* and *coRanking* R packages for computing the co-ranking matrix  $Q$  with  
274 the function *coranking* and then the AUC metric with the function *AUC\_InK\_R\_NX*.

275

276 Complementary to the AUC metric, which is only based on ranks so potentially weakly  
277 influenced by some extreme distortion values, we also compared the initial and lower-  
278 dimensional distances between species pairs by using the Euclidean distance for  
279 multidimensional spaces, also known as the Mean Absolute Deviation (MAD) (Maire *et al.*  
280 2015).

281

## 282 **Dimensionality of species trait space**

283 To determine how many dimensions are needed to build a trait space of enough quality that  
284 correctly positions species between each other, we used two approaches: a parsimonious one  
285 based on the elbow inflection point for the AUC metric and the other one based on a quality  
286 threshold for the AUC metric, both tested on 1 to 20 PCoA axes. The idea behind the elbow  
287 method is to maximize a given benefit (AUC gain in our case) while reducing the cost (number  
288 of dimensions in our case) (Thorndike 1953). Consequently, the inflection point corresponds to  
289 the additional PCoA axis above which the benefit becomes lower than the cost (Supplementary  
290 Figure 2). This elbow method is classically used in dimensionality analyses (Nguyen & Holmes  
291 2019) but never in combination with AUC.

292

293 As a complementary method, we used the AUC quality threshold of 0.7 to determine the  
294 dimensionality of the trait space so here the cumulated number of PCoA axes needed to obtain  
295 a good or acceptable positioning of species in the lower-dimensional space compared to the  
296 initial one based on all traits. This dimensionality assessment is more subjective than the elbow  
297 one since based on an arbitrary threshold. However, it has the merit of providing a standardized,  
298 so comparable, quality value across datasets for the low-dimensional representations.

299

300 The amount of variance explained by the PCoA axes could also be considered as a quality  
301 metric of species trait space (Pimiento *et al.* 2020b) like with Principal Components Analyses  
302 (PCA) (Pigot *et al.* 2020; R uger *et al.* 2020). Yet, for non-Euclidean distances like Gower,  
303 PCoA axes may obtain negative eigenvalues corresponding to imaginary dimensions (Legendre  
304 & Legendre 1998). In that case, the sum of all positive eigenvalues (real axes) is higher than  
305 the total variance of data. This intuitive additional piece of information was nonetheless  
306 included in our study through the examination of the relationship between the AUC-based  
307 dimensionality and the number of axes necessary to explain 50% of trait variation. The  
308 proportion of explained variance by PCoA axes was extracted using the `ape::pcoa()` R function.

309

## 310 **Robustness to trait omission**

311 To test the robustness, or the lack of sensitivity, of the trait space to trait omission or sub-  
312 selection, we randomly removed between 10% and 80% (increments of 10%) of the total  
313 number of traits, and then estimated a new Gower distance between all species pairs for each  
314 removal percentage; we did not use PCoA axes in this robustness analysis, only traits. Then,

315 we assessed the level of congruence between the initial distance matrix and the lower-  
316 dimensional distance matrix by computing the AUC and MAD metrics. These simulations were  
317 performed 100 times for each removal percentage. We then extracted an index of robustness  
318 defined as the opposite of sensitivity so the mean loss of AUC when 50% of the traits are  
319 removed.

320

### 321 **Species clustering and uniqueness**

322 To cluster species in the trait space and potentially identify unique species we used the  
323 “clustering by fast search and find of density peaks” algorithm which is based on initial pairwise  
324 distances and does not require dimensionality reduction (Rodriguez & Laio 2014). Yet, the  
325 robustness of the clustering critically depends on the robustness of pairwise species distances  
326 to trait omission. Among the many clustering algorithms that have been proposed (Jain & Dubes  
327 1988; Xu & Tian 2015; Condon *et al.* 2016), this one combines the advantages of (i) clustering  
328 objects regardless of the shape and dimensionality of the space in which they are embedded,  
329 (ii) detecting isolated objects automatically independently of their number, and (iii) making the  
330 number and size of clusters emerge with no a priori expectation or arbitrary choice.

331

332 In our case, this algorithm first computed the density of neighbors for each species, defined as  
333 the number of species that are within a given small distance  $d_0$  (Figure 1f). Given this density,  
334 the algorithm then relied on two basic principles: (1) cluster centers were species characterized  
335 by a higher density of neighbors than their own neighbors and by a relatively large distance  
336 from other species with a higher density of neighbors, and (2) isolated or unique species had no  
337 neighbors at maximum  $d_0$  (zero density or redundancy). Once cluster centers and unique species  
338 were identified, all remaining species were assigned to a cluster corresponding to the nearest  
339 neighbor of higher density (Rodriguez & Laio 2014). We adopted two modifications to reduce  
340 arbitrary choices. First, the identification of cluster centers was fully automated: all species with  
341 higher neighbor density than their own neighbors and at a distance of at least  $d_0$  from species  
342 with higher density were considered as cluster centers. Second, if two clusters were not  
343 separated by a “low density valley”, i.e. a region of radius  $d_0$  where densities were lower than  
344 those of the cluster centers, they were merged.

345

346 The whole clustering process thus required only a single free parameter, the threshold  $d_0$ , fixed  
347 by a rule of thumb by which the minimum distance to the nearest neighbor defining isolation,  
348 i.e. species uniqueness in trait space, is the average number of neighbors around each object

349 corresponding to 1 or 2% of the total number of species in the dataset (Rodriguez & Laio 2014).  
350 This procedure has the advantage of not fixing a  $d_0$  value a priori for all datasets but instead to  
351 define a  $d_0$  value for each dataset only depending on species number. Unique species can thus  
352 be considered as relative isolates in the trait space. We chose 1% as a conservative rule to not  
353 cluster species being too different in traits so keeping  $d_0$  small. We provide an R implementation  
354 of this algorithm along with the code to reproduce all the analyses of this paper (R Core Team,  
355 2021; see section Data and Code availability).

356

### 357 **Influence of trait dataset characteristics**

358 To test whether the characteristics of species, ecosystems and traits can influence the  
359 dimensionality, robustness, and structure of species trait space we performed General Linear  
360 Models (GLMs) with a Gaussian distribution for all response variables, i.e., the elbow-based  
361 dimensionality, the threshold-based dimensionality, the robustness to 50% trait removal, the  
362 log-transformed number of species clusters, the percentage of species packed in the first  
363 cluster and the percentage of unique species (distributions are shown in Supplementary Figure  
364 3). As explanatory factors, we used the type of species life form (plant, invertebrate and  
365 vertebrate) and the type of ecosystem (aquatic and terrestrial) to test the potential effects of  
366 organismal complexity. We also used the log-transformed number of species and number of  
367 traits as the dimensions of the initial species trait matrix. Trait characteristics were then used  
368 as potential drivers like the percentage of missing values, the percentage of quantitative traits  
369 and the mean pairwise correlation between traits, expressed as the rank-based Kendall index  
370 able to mix continuous and categorical traits. Pairwise correlations between quantitative trait  
371 dataset characteristics are rather low ( $-0.19 < r < 0.45$ ) and mainly non-significant  
372 (Supplementary Figure 4).

373

374 We then used partial regression plots to highlight the effect of each factor while controlling  
375 for the others (set at their mean). Statistical analyses were carried out using the function *glm*  
376 from the *stats* R package while partial plots were drawn using the function *visreg* from the  
377 *visreg* R package.

378

379 In addition to the analyses performed on empirical datasets, we also built three simulated  
380 datasets to test the effect of species and trait number on the dimensionality of species trait  
381 space without changing the type of traits as a controlled experiment. Continuous traits for  
382 1,000 species were generated following a uniform distribution (0-1) with no missing value. In

383 the first dataset we simulated 10 uncorrelated traits, in the second 10 correlated traits ( $r = 0.5$ )  
384 and in the third 20 uncorrelated traits. We then estimated the trait space dimensionality for  
385 each level of species number and each dataset using the AUC threshold of 0.7.

386

## 387 **Results**

### 388 **Trait space dimensionality**

389 Over the 30 datasets we obtained an optimal reduced dimensionality ranging between 2 and 8  
390 axes (Median=4) using the elbow method and between 2 to 17 axes (Median=6) using the  
391 AUC threshold of 0.7 when attained. For all datasets, we could reach the AUC threshold of  
392 0.7 with less than 20 dimensions or PCoA axes, except for plants of the French Alps for  
393 which AUC remained low ( $<0.6$ ) even with many axes (Figure 2). For the remaining 29  
394 datasets, the correlation between the elbow-based and threshold-based dimensionality was  
395 positive but weak ( $r = 0.3$ ) and non-significant ( $p\text{-value}=0.10$ ) highlighting their  
396 complementarity (Supplementary Figure 5). With a more demanding threshold of AUC=0.8  
397 (high quality trait space), up to 24 datasets could reach this value with a maximum of 20  
398 dimensions (Figure 2).

399

400 Two first GLMs, including all explanatory factors but only 29 datasets out of 30 (bacteria  
401 were excluded since they are the only representative of a kingdom), showed that the type of  
402 life form (plant, invertebrate and vertebrate) and the type of ecosystem (aquatic and  
403 terrestrial) did not significantly explain the elbow-based and the threshold-based  
404 dimensionality (Supplementary Table 1). The partial regression plots illustrate these weak  
405 influences while controlling for the other factors (Figure 3). We thus retained only  
406 quantitative variables related to the characteristics of the species trait datasets in the following  
407 analyses.

408

409 The elbow-based dimensionality was weakly explained by the five quantitative characteristics  
410 of the datasets ( $R^2=0.15$ ) but the correlation between traits had by far the main effect, albeit  
411 non-significant ( $p\text{-value}=0.09$ ) (Supplementary Table 2), with a lower optimal number of axes  
412 when the correlation between traits increased (Figure 4). The threshold-based dimensionality  
413 was well explained by characteristics of the datasets ( $R^2=0.61$ ) with the log-number of traits  
414 and the correlation between traits having the strongest and only significant effects  
415 (Supplementary Table 2). The partial regression plots showed that the threshold-based  
416 dimensionality strongly increased with the log-number of traits while it decreased with the

417 correlation between traits (Figure 4). As a complementary analysis, our simulated trait  
418 datasets confirmed the main influence of the number and the correlation of traits on species  
419 trait space dimensionality while the number of species had only an effect for less than 100  
420 species and no effect above 200 species (Supplementary Figure 6).

421  
422 The number of axes necessary to explain 50% of trait variation was a weak predictor of the  
423 elbow-based dimensionality ( $R^2=0.18$ ) but was a strong predictor of the threshold-based  
424 dimensionality ( $R^2=0.82$ ), albeit underestimated (Supplementary Figure 7).

### 425 426 **Robustness to trait omission**

427 The robustness to trait omission was generally low over the 30 datasets with a mean AUC loss  
428 of 0.54 (SD=0.12) when 50% of the traits were deleted. In these cases, most low-dimensional  
429 trait spaces were poor representations of the initial distances between species. Yet, this  
430 robustness was highly heterogeneous among datasets ranging from 0.33 to 0.85 of AUC loss  
431 (Figure 5). To stay above the AUC threshold of 0.7, trait omission should not exceed 20% on  
432 average when we ignored the five datasets for which even removing 10% of traits induced an  
433 AUC loss of more than 0.3 (i.e.  $AUC < 0.7$ ).

434  
435 Like for the dimensionality, the robustness to trait omission was not significantly influenced  
436 by either the type of species life form or the type of ecosystem (Figure 3, Supplementary  
437 Table 1) so these factors were ignored in the following analyses focused on quantitative  
438 factors. The robustness to trait omission was strongly dependent on the dataset characteristics  
439 ( $R^2=0.84$ ) with the log-number of traits, the percentage of missing values and the correlation  
440 between traits having the strongest and only significant effects (Supplementary Table 2). The  
441 partial regression plots revealed quite logically that the robustness to trait omission (opposite  
442 to AUC loss) increased with the number of traits but also with the correlation between traits  
443 (Figure 4). In contrast, robustness was negatively related to the percentage of missing values,  
444 which again makes sense. With many missing values, the trait space is likely to be unstable  
445 under trait omission so dimensionality reduction may distort the representation of the initial  
446 distances between species.

### 447 448 **Species clustering in trait space**

449 Over the 30 datasets, the number of species clusters, delineated by the “fast search and find of  
450 density peaks” algorithm, varied between 4 and 434 and was moderately explained by the

451 dataset characteristics ( $R^2=0.57$ ). The number of clusters was not significantly influenced by  
452 either the type of species life form or the type of ecosystem (Figure 6, Supplementary Table  
453 3) so these factors were ignored in the following analyses. The main and only significant  
454 drivers were the log-number of species and percentage of missing values (Supplementary  
455 Table 4). The number of clusters logically decreased with the percentage of missing values  
456 since less trait combinations can be realized but increased with the number of species (Figure  
457 7, Supplementary Figure 8). Yet, the number of clusters increased as a saturating power-law  
458 with the number of species owing to a slope much lower than 1 (0.41) in the log-log  
459 relationship when we controlled for other effects (Figure 8a).

460

461 The proportion of species belonging to the first or dominant cluster was not significantly  
462 driven by either the type of species life form or ecosystem (Figure 6, Supplementary Table 3)  
463 so these factors were ignored in the following analyses. This species packing into the  
464 dominant cluster was mainly driven by the log-number of species with a predictive power of  
465  $R^2=0.58$  while all the other dataset characteristics had non-significant influences  
466 (Supplementary Table 4). The slope of the relationship between the proportion of species  
467 clustered within the first group and the log-number of species was positive (Figure 7),  
468 highlighting that species tended to pack in the richest trait cluster when species richness  
469 increased, regardless of the other dataset characteristics. Yet, the log-log relationship between  
470 the total species richness and the richness of the first cluster revealed a power law with a  
471 slope higher than 1 (1.38) when we controlled for other effects (Figure 8b), suggesting that  
472 species packing disproportionately increased with species richness.

473

#### 474 **Unique species in trait space**

475 The number of unique species, i.e. species that did not belong to any cluster so isolated in the  
476 trait space, varied between 27 and 1750 among datasets with a percentage ranging from 2% to  
477 74% (Median=42%). These unique species were widespread in trait space and not just located  
478 on the edges, suggesting openings scattered throughout species trait spaces (Figure 9). Yet,  
479 well-known unique species appeared clearly far on the edge such as the whale shark  
480 (*Rhincodon typus*) which is the largest shark (20 meters long and body mass of 34 tonnes)  
481 while being a planktivore, so an ecological outlier among Chondrichthyes.

482

483 The proportion of unique species was not significantly influenced by either the type of species  
484 life form or ecosystem (Figure 6, Supplementary Table 3) so these categorical factors were



485 ignored in the following analyses only based on quantitative factors. The proportion of unique  
486 species was strongly explained by dataset characteristics ( $R^2=0.82$ ) with the log-number of  
487 species and, to a less extent, the percentage of missing values, being the main drivers  
488 (Supplementary Table 4).

489

490 The partial regression plots revealed that the proportion of unique species had a marked  
491 negative relationship with the log-number of species while controlling for other effects  
492 (Figure 7), suggesting that species-rich assemblages left less space for ecological uniqueness  
493 or that species tended to disproportionately pack into the richest cluster when diversity  
494 increased (Figure 8b). This saturating relationship was highlighted by the partial plot linking  
495 the total number of species and the number of unique species with a power log-log slope of  
496 0.47 (Figure 8c). The proportion of unique species also decreased with the proportion of  
497 missing values since it mechanically reduced the diversity of trait combinations and increased  
498 species similarity (Figure 7).

499

## 500 **Discussion**

### 501 **The necessary trade-off between trait space quality and operationality**

502 Trait-based approaches have a long tradition in life science since the development of the two-  
503 strategy life-history framework from ‘fast’ (r) to ‘slow’ (K) organisms (MacArthur & Wilson  
504 1967; Pianka 1972). This oversimplified view was later extended to triangular continuums of  
505 plant life-history strategies with the well-known competitive ability - physiological tolerance  
506 to stress - adaptation to disturbance (C-S-R) schema introduced by Grime (1977) and the Leaf-  
507 Height-Seed (LHS) framework by Westoby (1998). Such meaningful simplifications of trait  
508 variability among species have revolutionized functional ecology and inspired similar  
509 successful approaches for insects (Greenslade 1983), freshwater fishes (Winemiller & Rose  
510 1992), corals (Darling *et al.* 2012) and microbes (Malik *et al.* 2020). In the case of well-  
511 established or experimentally tested causal relationships between traits and environments or  
512 functions, the dimensionality issue is of marginal importance when building species spaces with  
513 few relevant traits delineating clearly defined ecological strategies. By contrast, when such  
514 knowledge is lacking, so when many traits are available with low evidence of particular causal  
515 relevance, when big data analyses are performed with many missing values, or when species  
516 strategies cannot be summarized by a limited set of traits, ecologists face the challenge of trait  
517 space hyper-dimensionality (Blonder *et al.* 2014).

518

519 Dimensionality reduction can then be a necessary step since some widely used functional  
520 diversity indices (e.g. functional richness) are based on the volume of trait space (convex hull  
521 volume) occupied by species of a given ecosystem (Villegger *et al.* 2008; Laliberte & Legendre  
522 2010; Trindade-Santos *et al.* 2020) that can be hardly calculated beyond 6 dimensions, even  
523 less (4-5) if null models are required or when pair-wise site measures like  $\beta$ -diversity have to  
524 be estimated (Villegger *et al.* 2011; Loiseau *et al.* 2017; Pimienta *et al.* 2020b; Su *et al.* 2021).  
525 Since most common functional diversity indices are sensitive to the degree of correlation among  
526 traits (Zhu *et al.* 2017), we also suggest to compute these indices from a reduced number of  
527 independent PCoA axes to improve the capacity to distinguish between communities along  
528 gradients of stress (Trindade-Santos *et al.* 2020).

529

530 Beyond practical reasons, this dimensionality value also informs about the extent to which  
531 species traits can be reduced to a limited number of ecologically meaningful axes (Díaz *et al.*  
532 2016; Pigot *et al.* 2020). This quest for ecological syndromes or strategies is not new (Westoby  
533 1998; Reich *et al.* 2003) and some previous studies have investigated the intrinsic  
534 dimensionality of species traits using various linear and non-linear methods (Westoby 1998;  
535 Laughlin 2014; Maire *et al.* 2015). Here, we proposed two complementary ways to estimate  
536 linear dimensionality and we applied them to 30 datasets to ultimately identify their main  
537 drivers, if any.

538

539 Using the parsimonious elbow-based AUC method, we found a median dimensionality of 4  
540 axes which is a rather low value given that we only considered datasets with at least 10 traits  
541 in our study (Table 1). Interestingly, for most datasets (25 out of 30) the elbow-based  
542 dimensionality is lower than 6 axes (2-5) (Figure 2) suggesting that the calculation of most  
543 volume-based functional diversity indices can be performed even with null models. Using the  
544 AUC-threshold criteria of 0.7, the dimensionality is higher (median of 6 axes) and generally  
545 out of the operational range for calculations of hypervolume-based metrics like functional  
546 richness (Villegger *et al.* 2008) or functional  $\beta$ -diversity (Loiseau *et al.* 2017). It reinforces the  
547 idea that the diversity of organism forms and functions has a larger dimensionality than  
548 previously thought (Pigot *et al.* 2016b; Messier *et al.* 2017) whatever the kingdom and  
549 ecosystem. Only poor assemblages (<30 species) can be accurately described with low-  
550 dimensionality (<4 axes) as shown in our simulations (Supplementary Figure 6).

551

552 This can be partly due to the coexistence of different syndromes related to different sets of  
553 traits, corresponding to different ecological strategies, under a given environment (Reich *et al.*  
554 2003; Sosiak & Barden 2021). For instance, landscape filters can shape trait community  
555 composition with species sharing some traits (trait syndromes) responding in a similar way  
556 under the same environmental conditions (e.g. agricultural intensification) (Gámez-Virués *et*  
557 *al.* 2015). When using large species datasets mixing various environments and many traits  
558 like in most our cases (Table 1), the potential multiplication of trait syndromes could explain  
559 the relatively high dimensionality in the trait space we have observed, particularly for the  
560 plants in the French Alps or stream macroinvertebrates (Figure 2). We may expect lower  
561 dimensionality in species trait space built from local communities under severe filters owing  
562 to the predominance of a few but highly constrained trait syndromes. We may also expect  
563 lower dimensionality when using effect vs. response traits in a more coherent and systematic  
564 manner with a clear defined goal (Luck *et al.* 2012).

565

566 The most surprising result is the weak positive correlation between the elbow-based and  
567 threshold-based dimensionality values showing that a low elbow-based AUC value does not  
568 imply passing the 0.7 AUC threshold and vice-versa (Figure 2). This is because the elbow-  
569 based method imposes a compromise between the quantity of axes and the quality of the trait  
570 space to avoid selecting more poorly informative axes (over-dimensionality) while the  
571 threshold-based method only considers quality whatever the quantity of axes. Given this  
572 constraint, the elbow-based method provides lower dimensionality values (2-8 axes against 2-  
573 17 axes for the threshold method; Supplementary Figure 5) which are also less influenced by  
574 dataset characteristics. As a practical guide, we suggest to use the elbow-based method as a first  
575 estimate of dimensionality on a given trait dataset and then to increase the number of  
576 dimensions to be considered until passing the 0.7 threshold if necessary. With this rule of  
577 thumb, we should end-up with an optimal dimensionality comprising between 3 and 6 axes for  
578 most datasets, as a trade-off between operationality and quality. Obviously, the operational  
579 constraint depends on species number, diversity indices being used and power facilities.

580

581 In case a value of AUC=0.5 cannot be reached with a reasonable number of dimensions (<10  
582 axes) like on the French Alps plants (Figure 2) we suggest either to carefully select the most  
583 relevant traits given the question being addressed (Thuiller *et al.* 2014) or to avoid indices based  
584 on trait space reduction (like functional richness) but instead to use distance-based indices (Rao)  
585 only (Laliberte & Legendre 2010; Mouillot *et al.* 2013; Chao *et al.* 2019). For representation

586 purposes, which are classically drawn in 2 or 3 dimensions with PCoA axes 1 to 4 (Stubbs &  
587 Benton 2016; Bruelheide *et al.* 2018; Loiseau *et al.* 2020; Pimiento *et al.* 2020b), we suggest  
588 to provide the corresponding AUC value as a key information on trait space quality along with  
589 the percentage of trait variation explained by axes. Since dimensionality is weakly influenced  
590 by dataset characteristics, except trait correlations that decrease dimensionality for both elbow-  
591 based and threshold-based criteria (Figure 4), we suggest to pay particular attention to  
592 unnecessary or meaningless traits that are strongly independent from the others and would  
593 inflate dimensionality potentially biasing biodiversity metrics. Conversely, considering  
594 redundant or correlated traits, even if meaningless, has no expected impact on dimensionality  
595 so can be very neutral in the building of species trait space and the computation of indices. Yet,  
596 using surrogate traits or traits with a coarse resolution to describe a given dimension of  
597 ecological strategy can substantially affect the results (Loranger *et al.* 2016; Kohli & Jarzyna  
598 2021).

599

#### 600 **The low but predictable robustness to trait omissions or choices**

601 Choosing a set of traits always means ignoring some, while important traits can be missed  
602 because they are unavailable or unknown. Often traits are ignored for non-biological reasons  
603 such as the difficulty of measuring them or the lack of standardization in the research  
604 community. The consequences of this sub-selection have been poorly investigated, despite its  
605 potential to modify the perceived dissimilarity between species (Carscadden *et al.* 2017) and  
606 profoundly affect the estimates of functional diversity (Zhu *et al.* 2017). Here, we randomly  
607 reduced our trait datasets to assess the impact of trait omission on AUC loss between the  
608 initial distance matrix (all traits) and that based on 90% to 20% of the traits only (Figure 5).  
609 When only 10% of traits are removed, AUC is still higher than 0.7 on average across  
610 simulations in 21 datasets out of 30, suggesting overall high robustness to low rate of trait  
611 omission except for some taxa like palm trees, sharks, thermal fauna and corals which belong  
612 to different kingdoms and ecosystems. At 50% of trait removal, AUC severely drops below  
613 the 0.7 threshold for all datasets except fishes of the Jakarta Bay (Figure 5).

614

615 This overall low but highly variable robustness of species distances to trait omission is very  
616 well explained by datasets characteristics (Figure 4). Unsurprisingly, AUC loss at 50%  
617 omission rate is negatively related to the number of traits, so that trait-poor datasets (corals,  
618 sharks or freshwater fishes) are more sensitive to the removal of traits than their trait-rich  
619 counterparts (macro-invertebrates or bacteria). Our statistical model also shows an expected

620 negative relationship between AUC loss and trait correlation, so with more redundant traits  
621 the distances between species pairs in a low dimensional space are more strongly preserved.  
622 This might explain why dimensionality reduction has been successful for some research fields  
623 in functional ecology (e.g. leaf traits and the leaf economic spectrum (Wright *et al.* 2004;  
624 Díaz *et al.* 2016)), while other studies such as those spanning many organs of plants have  
625 failed to find meaningful reduction in trait dimensionality (Carscadden *et al.* 2017; Messier *et*  
626 *al.* 2017). We also point out that the number of missing values strongly impacts robustness to  
627 trait omission so including traits with many missing values (>10%) can be a  
628 counterproductive effort, especially with Gower-like metrics which only consider traits with  
629 no missing values to assess the distance between two species. We also show a high variability  
630 in robustness for a given level of trait omission (Figure 5) suggesting that robustness to trait  
631 omission depends on traits being removed, some being more critical than others,  
632 independently of their ecological relevance. This reinforces the advice to carefully select traits  
633 prior to analyses and pay a particular attention to those being uncorrelated to the others given  
634 their disproportionate importance in the structuring of species trait spaces and subsequent  
635 analyses.

636

637 Taken together these results point out that the robustness of species space to trait omissions or  
638 choices is on average lower than previously thought (Douma *et al.* 2012) and that dataset  
639 characteristics, not the species life form or ecosystem type, explain this robustness, notably  
640 the presence of too many missing values. As a precautionary principle, we suggest to perform  
641 sensitivity analyses where traits are removed one by one or until a certain percentage of  
642 removal to assess the robustness of the results (Mouillot *et al.* 2014; Pollock *et al.* 2017;  
643 McLean *et al.* 2018; Cooke *et al.* 2019a; Loiseau *et al.* 2020). Trait-gap filling through  
644 automatic imputation might also be an interesting perspective (Penone *et al.* 2014; Schrodte *et*  
645 *al.* 2015; Goberna & Verdú 2016; Johnson *et al.* 2020). However, given the way most of these  
646 approaches work, this is likely that trait imputations will follow the main trends and the main  
647 syndromes and will unlikely generate unique species artificially hidden in the space.

648

### 649 **Species packing in trait space disproportionately increases with species richness**

650 The species packing in trait space, or so-called over-redundancy (Mouillot *et al.* 2014),  
651 provides functional insurance and resilience to ecosystems under disturbances (McLean *et al.*  
652 2019). This packing can be easily assessed with categorical traits since each unique  
653 combination of traits, also called functional entity, is a cluster so the clusters with a high

654 number of species, or higher than expected under a null model, are considered as over-packed  
655 or over-redundant while those with few species are vulnerable to biodiversity loss (Mouillot  
656 *et al.* 2014). With continuous traits or a large mix of traits as in our study, the clustering of  
657 species remains an arbitrary decision depending on the methods and thresholds used. We  
658 chose a clustering method with the lowest number of arbitrary decisions as possible  
659 independently of the shape and structure of species distribution in trait space (Rodriguez &  
660 Laio 2014). Surprisingly, this method, despite its attractiveness in other fields (medical and  
661 social sciences) and its parsimony (one parameter), has never been applied in ecology and  
662 evolution so far.

663  
664 Using a “fast search and find of density peaks” algorithm (Rodriguez & Laio 2014), we show  
665 that the number of clusters increases with the number of species when we control for the other  
666 factors (Figure 7) but with a strongly saturating relationship (Figure 8a) suggesting that  
667 species tend to over-pack into some clusters instead of creating new clusters in species-rich  
668 assemblages as shown for reef fishes (Mouillot *et al.* 2014) or passerine birds (Pigot *et al.*  
669 2016b). With a slope of 0.41 on the log-log scale it means that when species richness doubles,  
670 the number of clusters only increases by 30%. As a corollary, the richness of the dominant  
671 cluster increases with total species richness on a log-log scale with a slope higher than 1  
672 (Figure 8b) suggesting that additional species disproportionately pack into the most speciose  
673 cluster. More precisely two times more species in a given assemblage induces the packing of  
674 2.6 times more species in the dominant cluster. So, biodiversity only reinforces the  
675 redundancy of the most common traits instead of providing the level of insurance we should  
676 expect from species richness only under a random or proportional distribution of species  
677 among clusters (Mazel *et al.* 2014; Mouillot *et al.* 2014). This remarkable trend is observed  
678 for all taxa and ecosystem types.

679

### 680 **The saturating scaling of uniqueness with species richness**

681 The identification of ecological disparity, gaps, distinctiveness or uniqueness in trait spaces is  
682 a long-standing issue in ecology and evolution (Foote 1990; Winemiller 1991; Ricklefs 2005;  
683 Bapst *et al.* 2012; Violle *et al.* 2017; Gauzere *et al.* 2020). It contributes, for instance, to  
684 estimate the level of functional insurance and vulnerability to species extinction (Mouillot *et*  
685 *al.* 2014) but also to better understand the influence of trait rarity on ecosystem functioning  
686 (Maire *et al.* 2018), to set conservation priorities targeting unique species (Loiseau *et al.*  
687 2020), and to illuminate the capacity for innovation in clades (Cornwell *et al.* 2014; Deline *et*

688 *al.* 2018; Reeves *et al.* 2020). Yet, there is no consensus on the way to determine which  
689 species are isolated enough in trait spaces to be considered as unique species. Among the  
690 myriad of clustering algorithms (Xu & Tian 2015), the method based on fast search and find  
691 of density peaks was able to extract unique species in a very intuitive, standard, biodiversity-  
692 independent and distribution-free way. We show that the proportion of unique species  
693 decreases with species richness (Figure 7) while the number of unique species saturates  
694 rapidly with species richness (Figure 8c) suggesting that ecological novelty does not scale  
695 proportionally with taxonomic diversity but at a much lower rate whatever the kingdom or  
696 ecosystem. With a slope of 0.47 on the log-log scale it means that when species richness  
697 doubles, the number of unique species increases by 38%. This result resonates with the  
698 saturating link between ecological disparity and species richness across geological periods  
699 (Bapst *et al.* 2012) contrary to predictions from theory on adaptive radiations and ecological  
700 speciation (Rundell & Price 2009). More precisely, some entire lineages remained  
701 ecologically conservative throughout the Mesozoic without exploring vacant portions of trait  
702 space and then trait bursts occurred owing to changing abiotic conditions during the Late  
703 Jurassic (Reeves *et al.* 2020). Both adaptive radiations due to species interactions and  
704 innovative solutions to face new environments are certainly at play to explain the invariant  
705 saturating scaling of ecological uniqueness with species richness.

706

## 707 **Conclusions**

708 Four take-home messages can be extracted from this analysis. First of all, when no prior  
709 selection of traits can be carried out, the minimum dimensionality of trait space is rather large  
710 with around 3-6 dimensions. The success of identifying axes of variation, especially when trait  
711 correlations are strong, suggests that the research program of finding major trade-off axes  
712 grounded in ecological principles shows more promise than the arbitrary selection and removal  
713 of traits. Second, most trait spaces are highly sensitive to trait omission, which thus requires  
714 careful thinking about which traits might be overlooked, missed and targeted into the future.  
715 Third, there are plenty of unique species and the success of the clustering approach suggests  
716 that we need to pay more attention to how species pack relative to each other in trait space and  
717 not only focus on dimensionality reduction of trait spaces. Fourth, the complexity of  
718 multicellular organisms from plants to vertebrates or from aquatic to terrestrial species has little  
719 influence on the dimensionality, robustness and structure of trait space. Instead our synthesis  
720 suggests that the rate of key functional innovations and the subsequent complexity of trait space  
721 are consistent across multicellular clades with multicellularity evolution in plants sharing many

722 features with that leading to animals. Yet, these results are based on only 30 datasets and may  
723 lack statistical power to detect some effects. Moreover, these results are only valid for the range  
724 of dataset characteristics that we used in our analyses so more than 40 species and ten traits.  
725 We obtained different patterns for species-poor assemblages in our simulations but we are  
726 confident that our empirical assessment may embrace most species richness conditions  
727 encountered in temperate or tropical assemblages for most taxa when building regional or  
728 global species trait space.

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743



744 **Acknowledgements**

745 This research is supported by the Fondation pour la Recherche sur la Biodiversité (FRB) and  
746 Electricité de France (EDF) in the context of the CESAB project ‘Causes and consequences of  
747 functional rarity from local to global scales’ (FREE)

748 **Supplementary materials**

749 Supplementary Table 1: Statistics of GLM.

750 Supplementary Table 2: Statistics of GLM.

751 Supplementary Table 3: Statistics of GLM.

752 Supplementary Table 4: Statistics of GLM.

753

754 Supplementary Figure 1: Number of species and traits in the CESTES database.

755 Supplementary Figure 2: Theoretical illustration of the Elbow method.

756 Supplementary Figure 3: Distribution of the six metrics characterizing species trait spaces.

757 Supplementary Figure 4: Pairwise Pearson correlations between all dataset characteristics.

758 Supplementary Figure 5: Relation between the dimensionality found with the elbow vs.

759 threshold method.

760 Supplementary Figure 6: Simulated relationships between dimensionality and species-trait

761 numbers.

762 Supplementary Figure 7: Relation between dimensionality and the number of PCoA axes

763 explaining 50% of trait variation.

764 Supplementary Figure 8: Relationships between the log number of species and the log number

765 of clusters, the log number of species in the most dominant clusters and the log number of

766 unique species.

767 Supplementary Figure 9: Relation between the Area Under the Curve (AUC) metric and the

768 mean absolute deviation (MAD).

769

770

771

772

773

774

775 **References**

776

777 Bapst, D.W., Bullock, P.C., Melchin, M.J., Sheets, H.D. & Mitchell, C.E. (2012). Graptoloid  
778 diversity and disparity became decoupled during the Ordovician mass extinction.  
779 *Proceedings of the National Academy of Sciences of the United States of America*,  
780 109, 3428-3433.

781

782 Barros, C., Guéguen, M., Douzet, R., Carboni, M., Boulangeat, I., Zimmermann, N.E. *et al.*  
783 (2017). Extreme climate events counteract the effects of climate and land-use changes  
784 in Alpine tree lines. *Journal of Applied Ecology*, 54, 39-50.

785

786 Bernhardt-Römermann, M., Römermann, C., Nuske, R., Parth, A., Klotz, S., Schmidt, W. *et*  
787 *al.* (2008). On the identification of the most suitable traits for plant functional trait  
788 analyses. *Oikos*, 117, 1533-1541.

789

790 Blonder, B., Lamanna, C., Violle, C. & Enquist, B.J. (2014). The n-dimensional  
791 hypervolume. *Global Ecology and Biogeography*, 23, 595-609.

792

793 Bruehlheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S.M. *et*  
794 *al.* (2018). Global trait–environment relationships of plant communities. *Nature*  
795 *Ecology & Evolution*, 2, 1906-1917.

796

797 Cadotte, M.W. (2017). Functional traits explain ecosystem function through opposing  
798 mechanisms. *Ecology Letters*, 20, 989-996.

799

800 Carscadden, K.A., Cadotte, M.W. & Gilbert, B. (2017). Trait dimensionality and population  
801 choice alter estimates of phenotypic dissimilarity. *Ecology and Evolution*, 7, 2273-  
802 2285.

803

804 Chao, A., Chiu, C.-H., Villéger, S., Sun, I.F., Thorn, S., Lin, Y.-C. *et al.* (2019). An attribute-  
805 diversity approach to functional diversity, functional beta diversity, and related  
806 (dis)similarity measures. *Ecological Monographs*, 89, e01343.

807

808 Condon, A., Ding, J. & Shah, S. (2016). densityCut: an efficient and versatile topological  
809 approach for automatic clustering of biological data. *Bioinformatics*, 32, 2567-2576.

810

811 Cooke, R.S.C., Bates, A.E. & Eigenbrod, F. (2019a). Global trade-offs of functional  
812 redundancy and functional dispersion for birds and mammals. *Global Ecology and*  
813 *Biogeography*, 28, 484-495.

814

815 Cooke, R.S.C., Eigenbrod, F. & Bates, A.E. (2019b). Projected losses of global mammal and  
816 bird ecological strategies. *Nature Communications*, 10, 2279.

817

818 Cooney, C.R., Bright, J.A., Capp, E.J.R., Chira, A.M., Hughes, E.C., Moody, C.J.A. *et al.*  
819 (2017). Mega-evolutionary dynamics of the adaptive radiation of birds. *Nature*, 542,  
820 344-347.

821

822 Cornwell, W.K., Westoby, M., Falster, D.S., FitzJohn, R.G., O'Meara, B.C., Pennell, M.W. *et*  
823 *al.* (2014). Functional distinctiveness of major plant lineages. *Journal of Ecology*, 102,  
824 345-356.

825  
826 Cox, D.T.C., Gardner, A.S. & Gaston, K.J. (2021). Diel niche variation in mammals  
827 associated with expanded trait space. *Nature Communications*, 12, 1753.  
828  
829 Craven, D., Eisenhauer, N., Pearse, W.D., Hautier, Y., Isbell, F., Roscher, C. *et al.* (2018).  
830 Multiple facets of biodiversity drive the diversity–stability relationship. *Nature*  
831 *Ecology & Evolution*, 2, 1579-1587.  
832  
833 Darling, E.S., Alvarez-Filip, L., Oliver, T.A., McClanahan, T.R. & Côté, I.M. (2012).  
834 Evaluating life-history strategies of reef corals from species traits. *Ecology Letters*, 15,  
835 1378-1386.  
836  
837 Dehling, D.M., Jordano, P., Schaefer, H.M., Böhning-Gaese, K. & Schleuning, M. (2016).  
838 Morphology predicts species' functional roles and their degree of specialization in  
839 plant–frugivore interactions. *Proceedings of the Royal Society of London B:*  
840 *Biological Sciences*, 283.  
841  
842 Deline, B., Greenwood, J.M., Clark, J.W., Puttick, M.N., Peterson, K.J. & Donoghue, P.C.J.  
843 (2018). Evolution of metazoan morphological disparity. *Proceedings of the National*  
844 *Academy of Sciences*, 115, E8909.  
845  
846 Díaz, S., Kattge, J., Cornelissen, J.H.C., Wright, I.J., Lavorel, S., Dray, S. *et al.* (2016). The  
847 global spectrum of plant form and function. *Nature*, 529, 167-171.  
848  
849 Douma, J.C., Shipley, B., Witte, J.P.M., Aerts, R. & van Bodegom, P.M. (2012). Disturbance  
850 and resource availability act differently on the same suite of plant traits: revisiting  
851 assembly hypotheses. *Ecology*, 93, 825-835.  
852  
853 Duffy, J.E., Macdonald, K.S., Rhode, J.M. & Parker, J.D. (2001). Grazer diversity, functional  
854 redundancy, and productivity in seagrass beds: An experimental test. *Ecology*, 82,  
855 2417-2434.  
856  
857 Fisher, R.M., Shik, J.Z. & Boomsma, J.J. (2020). The evolution of multicellular complexity:  
858 the role of relatedness and environmental constraints. *Proceedings of the Royal*  
859 *Society B: Biological Sciences*, 287, 20192963.  
860  
861 Fonseca, C.R. & Ganade, G. (2001). Species functional redundancy, random extinctions and  
862 the stability of ecosystems. *Journal of Ecology*, 89, 118-125.  
863  
864 Foote, M. (1990). Nearest-Neighbor Analysis of Trilobite Morphospace. *Systematic Biology*,  
865 39, 371-382.  
866  
867 Gagic, V., Bartomeus, I., Jonsson, T., Taylor, A., Winqvist, C., Fischer, C. *et al.* (2015).  
868 Functional identity and diversity of animals predict ecosystem functioning better than  
869 species-based indices. *Proceedings of the Royal Society B: Biological Sciences*, 282,  
870 20142620.  
871  
872 Gámez-Virués, S., Perović, D.J., Gossner, M.M., Börschig, C., Blüthgen, N., de Jong, H. *et*  
873 *al.* (2015). Landscape simplification filters species traits and drives biotic  
874 homogenization. *Nature Communications*, 6, 8568.

875  
876 Gauzere, P., Morin, X., Violle, C., Caspeta, I., Ray, C. & Blonder, B. (2020). Vacant yet  
877       invasible niches in forest community assembly. *Functional Ecology*, 34, 1945-1955.  
878  
879 Goberna, M. & Verdú, M. (2016). Predicting microbial traits with phylogenies. *The ISME*  
880       *Journal*, 10, 959-967.  
881  
882 Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarities  
883       coefficients. *Journal of Classification*, 3, 5-48.  
884  
885 Greenslade, P.J.M. (1983). Adversity Selection and the Habitat Templet. *The American*  
886       *Naturalist*, 122, 352-365.  
887  
888 Grime, J.P. (1977). Evidence for the Existence of Three Primary Strategies in Plants and Its  
889       Relevance to Ecological and Evolutionary Theory. *The American Naturalist*, 111,  
890       1169-1194.  
891  
892 Gross, N., Bagousse-Pinguet, Y.L., Liencourt, P., Berdugo, M., Gotelli, N.J. & Maestre, F.T.  
893       (2017). Functional trait diversity maximizes ecosystem multifunctionality. *Nature*  
894       *Ecology & Evolution*, 1, 0132.  
895  
896 Hector, A., Schmid, B., Beierkuhnlein, C., Caldeira, M.C., Diemer, M., Dimitrakopoulos,  
897       P.G. *et al.* (1999). Plant diversity and productivity experiments in European  
898       grasslands. *Science*, 286, 1123-1127.  
899  
900 Jain, A.K. & Dubes, R.C. (1988). *Algorithms for clustering data*. Prentice Hall, Englewood  
901       Cliffs, New Jersey.  
902  
903 Jarzyna, M.A., Quintero, I. & Jetz, W. (2020). Global functional and phylogenetic structure of  
904       avian assemblages across elevation and latitude. *Ecology Letters*, n/a.  
905  
906 Jeliaskov, A., Mijatovic, D., Chantepie, S., Andrew, N., Arlettaz, R., Barbaro, L. *et al.*  
907       (2020). A global database for metacommunity ecology, integrating species, traits,  
908       environment and space (vol 7, 6, 2020). *Scientific Data*, 7.  
909  
910 Johnson, T.F., Isaac, N.J.B., Paviolo, A. & González-Suárez, M. (2020). Handling missing  
911       values in trait data. *Global Ecology and Biogeography*, n/a.  
912  
913 Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L. *et al.* (2009).  
914       PanTHERIA: a species-level database of life history, ecology, and geography of  
915       extant and recently extinct mammals. *Ecology*, 90, 2648-2648.  
916  
917 Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I.C., Leadley, P. *et al.* (2020). TRY  
918       plant trait database – enhanced coverage and open access. *Global Change Biology*, 26,  
919       119-188.  
920  
921 Knoll, A.H. (2011). The Multiple Origins of Complex Multicellularity. *Annual Review of*  
922       *Earth and Planetary Sciences*, 39, 217-239.  
923

- 924 Kohli, B.A. & Jarzyna, M.A. (2021). Pitfalls of ignoring trait resolution when drawing  
925 conclusions about ecological processes. *Global Ecology and Biogeography*, n/a.  
926
- 927 Kraemer, G., Reichstein, M. & Mahecha, M.D. (2018). dimRed and coRanking-Unifying  
928 Dimensionality Reduction in R. *R Journal*, 10, 342-358.  
929
- 930 Laliberte, E. & Legendre, P. (2010). A distance-based framework for measuring functional  
931 diversity from multiple traits. *Ecology*, 91, 299-305.  
932
- 933 Laughlin, D.C. (2014). The intrinsic dimensionality of plant traits and its relevance to  
934 community assembly. *Journal of Ecology*, 102, 186-193.  
935
- 936 Le Bagousse-Pinguet, Y., Gross, N., Saiz, H., Maestre, F.T., Ruiz, S., Dacal, M. *et al.* (2021).  
937 Functional rarity and evenness are key facets of biodiversity to boost  
938 multifunctionality. *Proceedings of the National Academy of Sciences*, 118,  
939 e2019355118.  
940
- 941 Lee, J.A. & Verleysen, M. (2009). Quality assessment of dimensionality reduction: Rank-  
942 based criteria. *Neurocomputing*, 72, 1431-1443.  
943
- 944 Legendre, P. & Legendre, L. (1998). *Numerical Ecology*. Second edn. Elsevier, Amsterdam.  
945
- 946 Loiseau, N., Legras, G., Gaertner, J.C., Verley, P., Chabanet, P. & Merigot, B. (2017).  
947 Performance of partitioning functional beta-diversity indices: Influence of functional  
948 representation and partitioning methods. *Global Ecology and Biogeography*, 26, 753-  
949 762.  
950
- 951 Loiseau, N., Mouquet, N., Casajus, N., Grenié, M., Guéguen, M., Maitner, B. *et al.* (2020).  
952 Global distribution and conservation status of ecologically rare mammal and bird  
953 species. *Nature Communications*, 11, 5071.  
954
- 955 Loranger, J., Blonder, B., Garnier, É., Shipley, B., Vile, D. & Violle, C. (2016). Occupancy  
956 and overlap in trait space along a successional gradient in Mediterranean old fields.  
957 *Am. J. Bot.*, 103, 1050-1060.  
958
- 959 Luck, G.W., Lavorel, S., McIntyre, S. & Lumb, K. (2012). Improving the application of  
960 vertebrate trait-based frameworks to the study of ecosystem services. *Journal of*  
961 *Animal Ecology*, 81, 1065-1076.  
962
- 963 MacArthur, R.H. & Wilson, E.O. (1967). *The Theory of Island Biogeography*. Princeton  
964 University Press.  
965
- 966 Maire, E., Grenouillet, G., Brosse, S. & Villéger, S. (2015). How many dimensions are  
967 needed to accurately assess functional diversity? A pragmatic approach for assessing  
968 the quality of functional spaces. *Global Ecology and Biogeography*, 24, 728-740.  
969
- 970 Maire, E., Villéger, S., Graham, N.A.J., Hoey, A.S., Cinner, J., Ferse, S.C.A. *et al.* (2018).  
971 Community-wide scan identifies fish species associated with coral reef services across  
972 the Indo-Pacific. *Proceedings of the Royal Society B: Biological Sciences*, 285,  
973 20181167.

974  
975 Malik, A.A., Martiny, J.B.H., Brodie, E.L., Martiny, A.C., Treseder, K.K. & Allison, S.D.  
976 (2020). Defining trait-based microbial strategies with consequences for soil carbon  
977 cycling under climate change. *The ISME Journal*, 14, 1-9.  
978  
979 Mazel, F., Guilhaumon, F., Mouquet, N., Devictor, V., Gravel, D., Renaud, J. *et al.* (2014).  
980 Multifaceted diversity-area relationships reveal global hotspots of mammalian species,  
981 trait and lineage diversity. *Global Ecology and Biogeography*, 23, 836-847.  
982  
983 McLean, M., Auber, A., Graham, N.A.J., Houk, P., Villéger, S., Violle, C. *et al.* (2019). Trait  
984 structure and redundancy determine sensitivity to disturbance in marine fish  
985 communities. *Global Change Biology*, 25, 3424-3437.  
986  
987 McLean, M., Mouillot, D., Lindegren, M., Engelhard, G., Villéger, S., Marchal, P. *et al.*  
988 (2018). A Climate-Driven Functional Inversion of Connected Marine Ecosystems.  
989 *Current Biology*, 28, 3654-3660.e3653.  
990  
991 McLean, M., Stuart-Smith, R.D., Villéger, S., Auber, A., Edgar, G.J., MacNeil, M.A. *et al.*  
992 (2021). Trait similarity in reef fish faunas across the world's oceans. *Proceedings of*  
993 *the National Academy of Sciences*, 118, e2012318118.  
994  
995 Mérigot, B., Durbec, J.-P. & Gaertner, J.-C. (2010). On goodness-of-fit measure for  
996 dendrogram-based analyses. *Ecology*, 91, 1850-1859.  
997  
998 Messier, J., Lechowicz, M.J., McGill, B.J., Violle, C. & Enquist, B.J. (2017). Interspecific  
999 integration of trait dimensions at local scales: the plant phenotype as an integrated  
1000 network. *Journal of Ecology*, 105, 1775-1790.  
1001  
1002 Mouillot, D., Graham, N.A.J., Vileger, S., Mason, N.W.H. & Bellwood, D.R. (2013). A  
1003 functional approach reveals community responses to disturbances. *Trends in ecology*  
1004 *& evolution*, 28, 167-177.  
1005  
1006 Mouillot, D., Vileger, S., Parravicini, V., Kulbicki, M., Ernesto Arias-Gonzalez, J., Bender,  
1007 M. *et al.* (2014). Functional over-redundancy and high functional vulnerability in  
1008 global fish faunas on tropical reefs. *Proceedings of the National Academy of Sciences*  
1009 *of the United States of America*, 111, 13757-13762.  
1010  
1011 Nguyen, L.H. & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction.  
1012 *PLOS Computational Biology*, 15, e1006907.  
1013  
1014 O'Gorman, E.J., Yearsley, J.M., Crowe, T.P., Emmerson, M.C., Jacob, U. & Petchey, O.L.  
1015 (2011). Loss of functionally unique species may gradually undermine ecosystems.  
1016 *Proceedings of the Royal Society B: Biological Sciences*, 278, 1886-1893.  
1017  
1018 Pavoine, S., Vallet, J., Dufour, A.B., Gachet, S. & Daniel, H. (2009). On the challenge of  
1019 treating various types of variables: application for improving the measurement of  
1020 functional diversity. *Oikos*, 118, 391-402.  
1021

- 1022 Penone, C., Davidson, A.D., Shoemaker, K.T., Di Marco, M., Rondinini, C., Brooks, T.M. *et al.* (2014). Imputation of missing data in life-history trait datasets: which approach  
1023 performs the best? *Methods in Ecology and Evolution*, 5, 961-970.  
1024  
1025
- 1026 Perez, T.M., Valverde-Barrantes, O., Bravo, C., Taylor, T.C., Fadrique, B., Hogan, J.A. *et al.*  
1027 (2019). Botanic gardens are an untapped resource for studying the functional ecology  
1028 of tropical plants. *Philosophical Transactions of the Royal Society B: Biological*  
1029 *Sciences*, 374, 20170390.  
1030
- 1031 Phillips, A.G., Töpfer, T., Rahbek, C., Böhning-Gaese, K. & Fritz, S.A. (2018). Effects of  
1032 phylogeny and geography on ecomorphological traits in passerine bird clades. *Journal*  
1033 *of Biogeography*, 45, 2337-2347.  
1034
- 1035 Pianka, E.R. (1972). r and K Selection or b and d Selection? *The American Naturalist*, 106,  
1036 581-588.  
1037
- 1038 Pigot, A.L., Bregman, T., Sheard, C., Daly, B., Etienne, R.S. & Tobias, J.A. (2016a).  
1039 Quantifying species contributions to ecosystem processes: a global assessment of  
1040 functional trait and phylogenetic metrics across avian seed-dispersal networks.  
1041 *Proceedings of the Royal Society B: Biological Sciences*, 283, 20161597.  
1042
- 1043 Pigot, A.L., Sheard, C., Miller, E.T., Bregman, T.P., Freeman, B.G., Roll, U. *et al.* (2020).  
1044 Macroevolutionary convergence connects morphological form to ecological function  
1045 in birds. *Nature Ecology & Evolution*, 4, 230-239.  
1046
- 1047 Pigot, A.L., Trisos, C.H. & Tobias, J.A. (2016b). Functional traits reveal the expansion and  
1048 packing of ecological niche space underlying an elevational diversity gradient in  
1049 passerine birds. *Proceedings of the Royal Society B: Biological Sciences*, 283,  
1050 20152013.  
1051
- 1052 Pimiento, C., Bacon, C.D., Silvestro, D., Hendy, A., Jaramillo, C., Zizka, A. *et al.* (2020a).  
1053 Selective extinction against redundant species buffers functional diversity.  
1054 *Proceedings of the Royal Society B: Biological Sciences*, 287, 20201162.  
1055
- 1056 Pimiento, C., Leprieur, F., Silvestro, D., Lefcheck, J.S., Albouy, C., Rasher, D.B. *et al.*  
1057 (2020b). Functional diversity of marine megafauna in the Anthropocene. *Science*  
1058 *Advances*, 6, eaay7650.  
1059
- 1060 Podani, J. & Schmera, D. (2006). On dendrogram-based measures of functional diversity.  
1061 *Oikos*, 115, 179-185.  
1062
- 1063 Pollock, L.J., Thuiller, W. & Jetz, W. (2017). Large conservation gains possible for global  
1064 biodiversity facets. *Nature*, 546, 141.  
1065
- 1066 Reeves, J.C., Moon, B.C., Benton, M.J. & Stubbs, T.L. (2020). Evolution of ecospace  
1067 occupancy by Mesozoic marine tetrapods. *Palaeontology*, n/a.  
1068
- 1069 Reich, P.B., Ellsworth, D.S., Walters, M.B., Vose, J.M., Gresham, C., Volin, J.C. *et al.*  
1070 (1999). Generality of leaf trait relationships: A test across six biomes. *Ecology*, 80,  
1071 1955-1969.

















1072  
1073 Reich, P.B., Wright, I.J., Cavender-Bares, J., Craine, J.M., Oleksyn, J., Westoby, M. *et al.*  
1074 (2003). The Evolution of Plant Functional Variation: Traits, Spectra, and Strategies.  
1075 *International Journal of Plant Sciences*, 164, S143-S164.  
1076  
1077 Ricklefs, R.E. (2005). Small clades at the periphery of passerine morphological space.  
1078 *American Naturalist*, 165, 43-659.  
1079  
1080 Ricklefs, R.E. (2010). Evolutionary diversification, coevolution between populations and their  
1081 antagonists, and the filling of niche space. *Proceedings of the National Academy of*  
1082 *Sciences of the United States of America*, 107, 1265-1272.  
1083  
1084 Rodriguez, A. & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*,  
1085 344, 1492.  
1086  
1087 R ger, N., Condit, R., Dent, D.H., DeWalt, S.J., Hubbell, S.P., Lichstein, J.W. *et al.* (2020).  
1088 Demographic trade-offs predict tropical forest dynamics. *Science*, 368, 165.  
1089  
1090 Rundell, R.J. & Price, T.D. (2009). Adaptive radiation, nonadaptive radiation, ecological  
1091 speciation and nonecological speciation. *Trends in Ecology & Evolution*, 24, 394-399.  
1092  
1093 Sala, E., Mayorga, J., Bradley, D., Cabral, R.B., Atwood, T.B., Auber, A. *et al.* (2021).  
1094 Protecting the global ocean for biodiversity, food and climate. *Nature*, 592, 397–402.  
1095  
1096 Sanders, D., Th bault, E., Kehoe, R. & Frank van Veen, F.J. (2018). Trophic redundancy  
1097 reduces vulnerability to extinction cascades. *Proceedings of the National Academy of*  
1098 *Sciences*.  
1099  
1100 Savile, O.B.O. (1957). ADAPTIVE EVOLUTION IN THE AVIAN WING. *Evolution*, 11,  
1101 212-224.  
1102  
1103 Schluter, D. (1993). Adaptive Radiation in Sticklebacks - Size, Shape, and Habitat Use  
1104 Efficiency. *Ecology*, 74, 699-709.  
1105  
1106 Schneider, F.D., Morsdorf, F., Schmid, B., Petchey, O.L., Hueni, A., Schimel, D.S. *et al.*  
1107 (2017). Mapping functional diversity from remotely sensed morphological and  
1108 physiological forest traits. *Nature Communications*, 8, 1441.  
1109  
1110 Schrodte, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A. *et al.* (2015). BHPMF –  
1111 a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology  
1112 and functional biogeography. *Global Ecology and Biogeography*, 24, 1510-1521.  
1113  
1114 Somers, R.H. (1962). A New Asymmetric Measure of Association for Ordinal Variables.  
1115 *American Sociological Review*, 27, 799-811.  
1116  
1117 Sosiak, C.E. & Barden, P. (2021). Multidimensional trait morphology predicts ecology across  
1118 ant lineages. *Functional Ecology*, 35, 139-152.  
1119

















- 1120 Spehn, E.M., Hector, A., Joshi, J., Scherer-Lorenzen, M., Schmid, B., Bazeley-White, E. *et al.*  
1121 (2005). Ecosystem effects of biodiversity manipulations in European grasslands.  
1122 *Ecological Monographs*, 75, 37-63.  
1123
- 1124 Stubbs, T.L. & Benton, M.J. (2016). Ecomorphological diversifications of Mesozoic marine  
1125 reptiles: the roles of ecological opportunity and extinction. *Paleobiology*, 42, 547-573.  
1126
- 1127 Su, G., Logez, M., Xu, J., Tao, S., Villéger, S. & Brosse, S. (2021). Human impacts on global  
1128 freshwater fish biodiversity. *Science*, 371, 835.  
1129
- 1130 Thorndike, R.L. (1953). Who belongs in the family? *Psychometrika*, 18, 267-276.  
1131
- 1132 Thuiller, W., Guéguen, M., Georges, D., Bonet, R., Chalmandrier, L., Garraud, L. *et al.*  
1133 (2014). Are different facets of plant diversity well protected against climate and land  
1134 cover changes? A test study in the French Alps. *Ecography*, 37, 1254-1266.  
1135
- 1136 Trindade-Santos, I., Moyes, F. & Magurran, A.E. (2020). Global change in the functional  
1137 diversity of marine fisheries exploitation over the past 65 years. *Proceedings of the*  
1138 *Royal Society B: Biological Sciences*, 287, 20200889.  
1139
- 1140 Valentine, J.W., Collins, A.G. & Meyer, C.P. (1994). Morphological Complexity Increase in  
1141 Metazoans. *Paleobiology*, 20, 131-142.  
1142
- 1143 Villegger, S., Mason, N.W.H. & Mouillot, D. (2008). New multidimensional functional  
1144 diversity indices for a multifaceted framework in functional ecology. *Ecology*, 89,  
1145 2290-2301.  
1146
- 1147 Villegger, S., Novack-Gottshall, P.M. & Mouillot, D. (2011). The multidimensionality of the  
1148 niche reveals functional diversity changes in benthic marine biotas across geological  
1149 time. *Ecology Letters*, 14, 561-568.  
1150
- 1151 Violle, C., Thuiller, W., Mouquet, N., Munoz, F., Kraft, N.J.B., Cadotte, M.W. *et al.* (2017).  
1152 Functional Rarity: The Ecology of Outliers. *Trends in Ecology & Evolution*, 32, 356-  
1153 367.  
1154
- 1155 Walker, B.H. (1992). Biodiversity and Ecological Redundancy. *Conservation Biology*, 6, 18-  
1156 23.  
1157
- 1158 Westoby, M. (1998). A leaf-height-seed (LHS) plant ecology strategy scheme. *Plant and Soil*,  
1159 199, 213-227.  
1160
- 1161 Winemiller, K.O. (1991). Ecomorphological Diversification in Lowland Freshwater Fish  
1162 Assemblages from Five Biotic Regions. *Ecological Monographs*, 61, 343-365.  
1163
- 1164 Winemiller, K.O., Fitzgerald, D.B., Bower, L.M. & Pianka, E.R. (2015). Functional traits,  
1165 convergent evolution, and periodic tables of niches. *Ecology Letters*, 18, 737-751.  
1166
- 1167 Winemiller, K.O. & Rose, K.A. (1992). Patterns of Life-History Diversification in North  
1168 American Fishes: implications for Population Regulation. *Canadian Journal of*  
1169 *Fisheries and Aquatic Sciences*, 49, 2196-2218.

1170  
1171 Wright, I.J., Reich, P.B., Westoby, M., Ackerly, D.D., Baruch, Z., Bongers, F. *et al.* (2004).  
1172 The worldwide leaf economics spectrum. *Nature*, 428, 821-827.  
1173  
1174 Xu, D. & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of*  
1175 *Data Science*, 2, 165-193.  
1176  
1177 Zhu, L., Fu, B., Zhu, H., Wang, C., Jiao, L. & Zhou, J. (2017). Trait choice profoundly  
1178 affected the ecological conclusions drawn from functional diversity measures.  
1179 *Scientific Reports*, 7, 3643.  
1180  
1181  
1182

1183 **Table 1:** We compiled 30 published trait datasets including 17 assembled in the CESTES  
 1184 database (Jeliazkov *et al.* 2020). We selected datasets with at least 40 species and 10 traits for  
 1185 high-dimensional space and perform sub-selection analyses (Supplementary Figure 1). Further,  
 1186 to allow the calculation of trait dissimilarity between all species pairs and test of robustness to  
 1187 trait omission, we removed traits with missing information for more than 60% of species and  
 1188 species with more than 50% of missing information for their traits. This procedure altered 4 out  
 1189 of 30 datasets and on average removed 15% (7-25%) of species and 57% (22-79%) of traits per  
 1190 dataset. We provide below a description and a reference of each dataset with the geographic  
 1191 extent and location (Area), species richness (S), the number of traits (T), the percentage of  
 1192 missing values (%NA), the percentage of quantitative traits (%Q) (the others, 1-%Q, being  
 1193 categorical), the mean pairwise Kendall correlation between traits (Cor) and a unique icon for  
 1194 each taxon used in some other Figures.

Datasets	Taxon	Area	S	T	%NA	%Q	$\mu$ Cor	Icons
Biolog	Bacteria	Global	865	97	0	99	0.17	
Bartonova <i>et al.</i> 2016	Butterfly	Czech Republic	128	13	0	100	0.22	
BirdLife	Bird	Global	9297	20	0	100	0.12	
Carvalho <i>et al.</i> 2015	Stream fishes	Amazonia. Brazil	65	26	0	4	0.17	
Charbonnier <i>et al.</i> 2016	Bird	Europe	73	10	0	40	0.15	
Chmura <i>et al.</i> 2016	Plant	Poland	46	17	0	94	0.18	
Fish Base	Chondrichthyes	Global	969	14	23	79	0.21	
Clearly <i>et al.</i> 2016	Vertebrate	Jakarta Bay. Indonesia	165	15	0	87	0.43	
Coral Trait Database	Invertebrate	Global	802	12	25	42	0.12	
Diaz <i>et al.</i> 2008	Invertebrate	Segura River. Spain	208	62	0	0	0.12	
Eallonardo <i>et al.</i> 2013	Plant	New York State. US	41	11	0	55	0.24	
Toussaint <i>et al.</i> 2016	Freshwater-fish	Global	8134	10	3	100	0.10	
Fried <i>et al.</i> 2012	Plant	France	75	10	0	30	0.17	
Gibb <i>et al.</i> 2015	Spider	South-Eastern Australia	86	10	0	100	0.41	
Goncalves <i>et al.</i> 2014	Spider	Brazilian coast	112	21	0	95	0.32	
Jeliazkov <i>et al.</i> 2013	Macro-invertebrate	France	112	89	0	0	0.14	

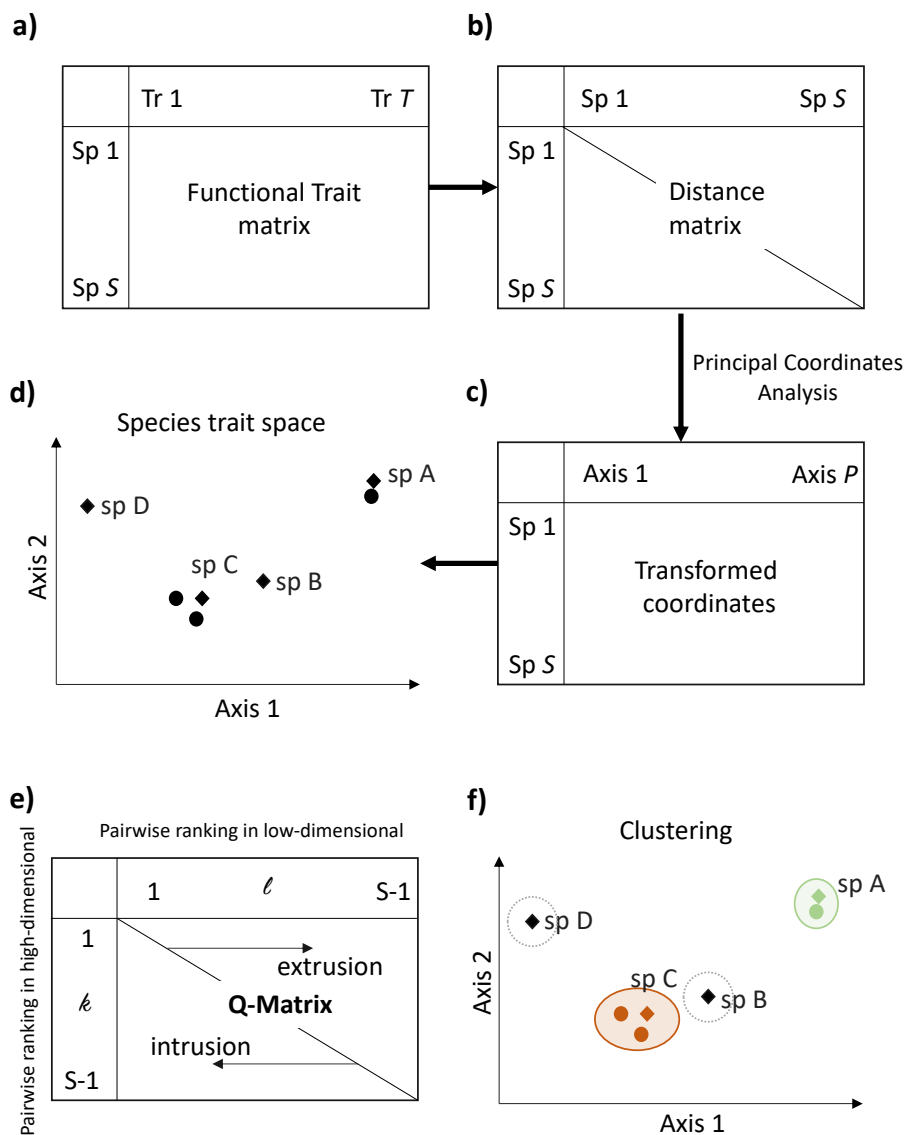
Krasnov <i>et al.</i> 2015	Ectoparasite	Palaearctic	177	12	0	100	0.17	
Loiseau <i>et al.</i> 2020	Terrestrial mammals	Global	4675	15	0	73	0.14	
McLean <i>et al.</i> 2018	Fish	North Sea. Atlantic	138	14	3	64	0.13	
Doledec <i>et al.</i> 2011	Stream macro-invertebrate	New-Zealand	495	59	0	0	0.14	
Pakeman <i>et al.</i> 2011	Plant	Scotland	148	28	0	36	0.14	
Kissling <i>et al.</i> 2019	Plant	Global	2557	22	28	82	0.17	
Pavoine <i>et al.</i> 2011	Plant	Algeria	56	14	0	29	0.15	
Rimet & Druart 2018	Phytoplankton	Temperate lakes	1222	15	0	40	0.30	
Thuiller <i>et al.</i> 2014	Plant	French Alps	3718	33	16	12	0.12	
Riberta <i>et al.</i> 2001	Beetle	Scotland	68	20	0	50	0.17	
Chapman <i>et al.</i> 2019	Thermal vent	Global	646	16	15	31	0.18	
USDA 2020	Plant	US	1876	20	6	90	0.09	
Villéger <i>et al.</i> 2012	Fish	Mexico	46	16	0	100	0.19	
Yates <i>et al.</i> 2014	Ant	New South Wales Australia	123	11	0	91	0.18	

1195

1196

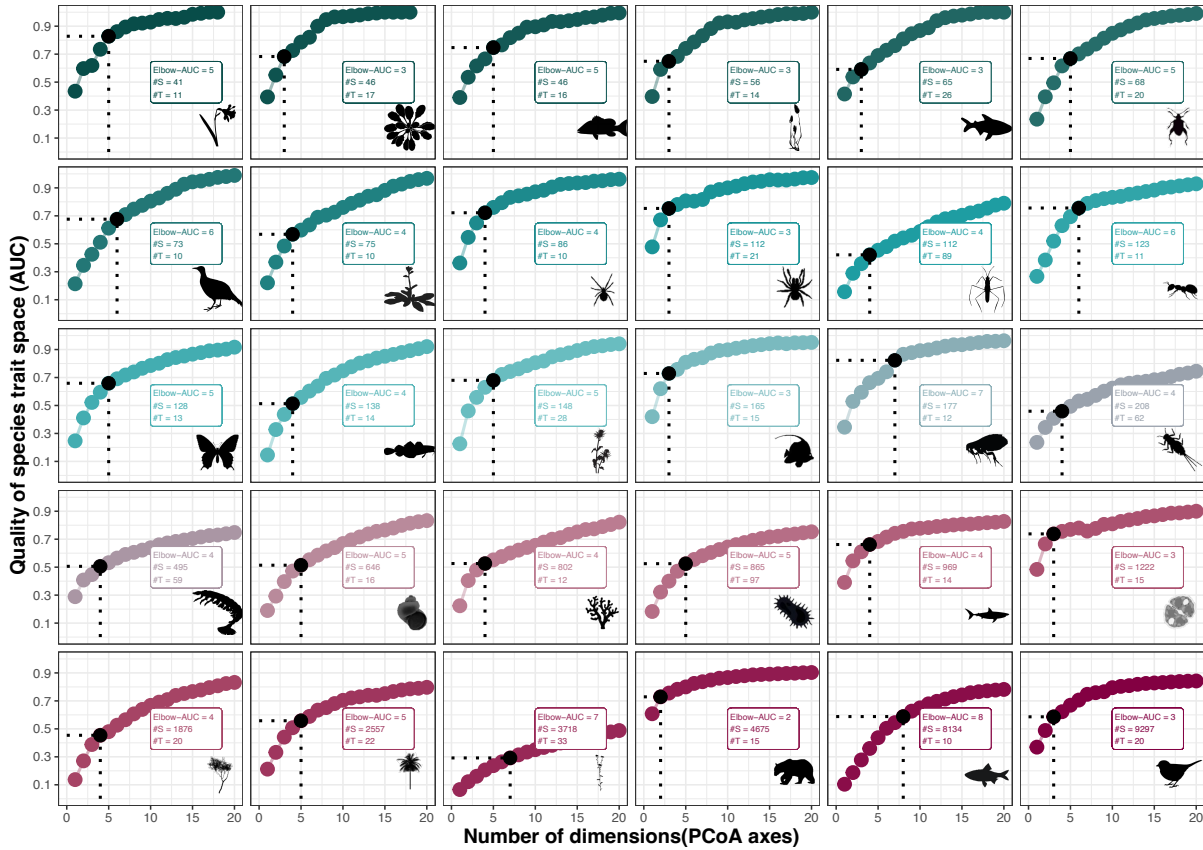
1197

1198 **Figure 1:** Theoretical example showing the different steps of our framework from species  
 1199 trait matrix (a) to species trait space (d) after calculating species pairwise distances (b) and  
 1200 extracting synthetic axes providing new species coordinates in a low-dimensional space (c).  
 1201 Then the ranking of species pairs in both high-dimensional (i.e. considering all traits so  
 1202 distance matrix b) and low-dimensional space (i.e. considering coordinates on few axes in c)  
 1203 can provide a  $Q$  matrix where the diagonal corresponds to all species pairs with a perfect  
 1204 match in their ranking in both spaces while off diagonal values correspond to mismatching  
 1205 species pairs in the co-ranking, i.e. species get closer in low-dimensional space (intrusion) or  
 1206 farther (extrusion) compared to their relative position in the high-dimensional space. A  
 1207 clustering algorithm isolates two unique species (species B and D) in the trait space (no  
 1208 neighbors within a given radius  $d_0$ ) and creates two clusters with 2 (green) and 3 (red) species  
 1209 (f). See Methods for details.



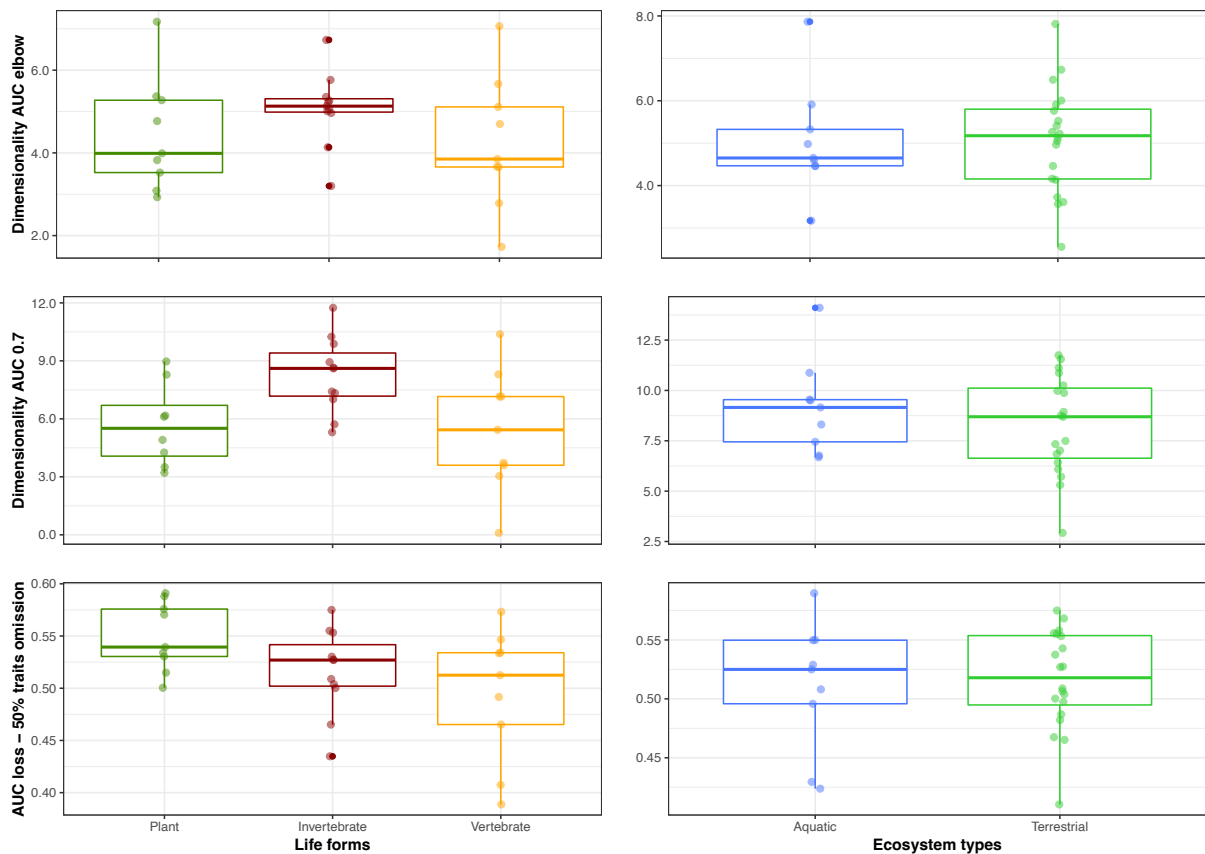
1210

1211 **Figure 2:** Influence of the number of dimensions (number of retained PCoA axes) used to  
 1212 build the 30 species trait spaces on the space quality assessed by the Area Under the Curve  
 1213 (AUC) criteria. The black dots and dotted lines correspond to the elbow-based optimal  
 1214 dimensionality for each dataset. The values indicate the elbow-based dimensionality, the total  
 1215 species richness (#S) and the total number of traits (#T) in each dataset. Datasets are ranked  
 1216 (top-left to bottom-right and from dark green to dark red) following the number of species.  
 1217



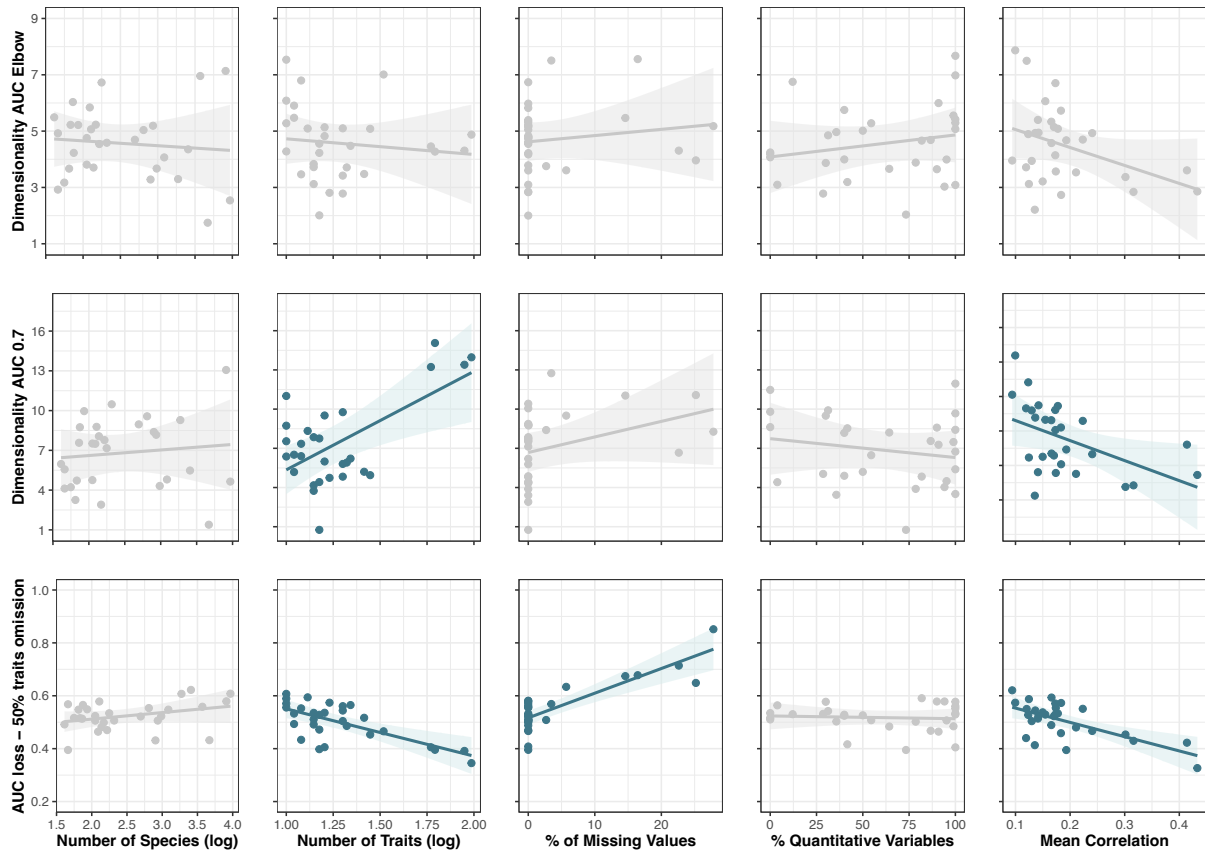
1218

1219 **Figure 3:** Partial plots showing the influence of the species life form (Plant, Invertebrate,  
 1220 Vertebrate) and ecosystem type (Aquatic, Terrestrial), while controlling for the five dataset  
 1221 quantitative characteristics, on species trait space dimensionality measured with the elbow-  
 1222 based (first row) or threshold-based (second row) AUC criteria. The third row shows trait  
 1223 space robustness, in terms of AUC loss, to trait removal or omission (50%) according to the  
 1224 two factors being tested. Related statistics are reported in the Supplementary Table 1, the  
 1225 effects are all non-significant.  
 1226



1227

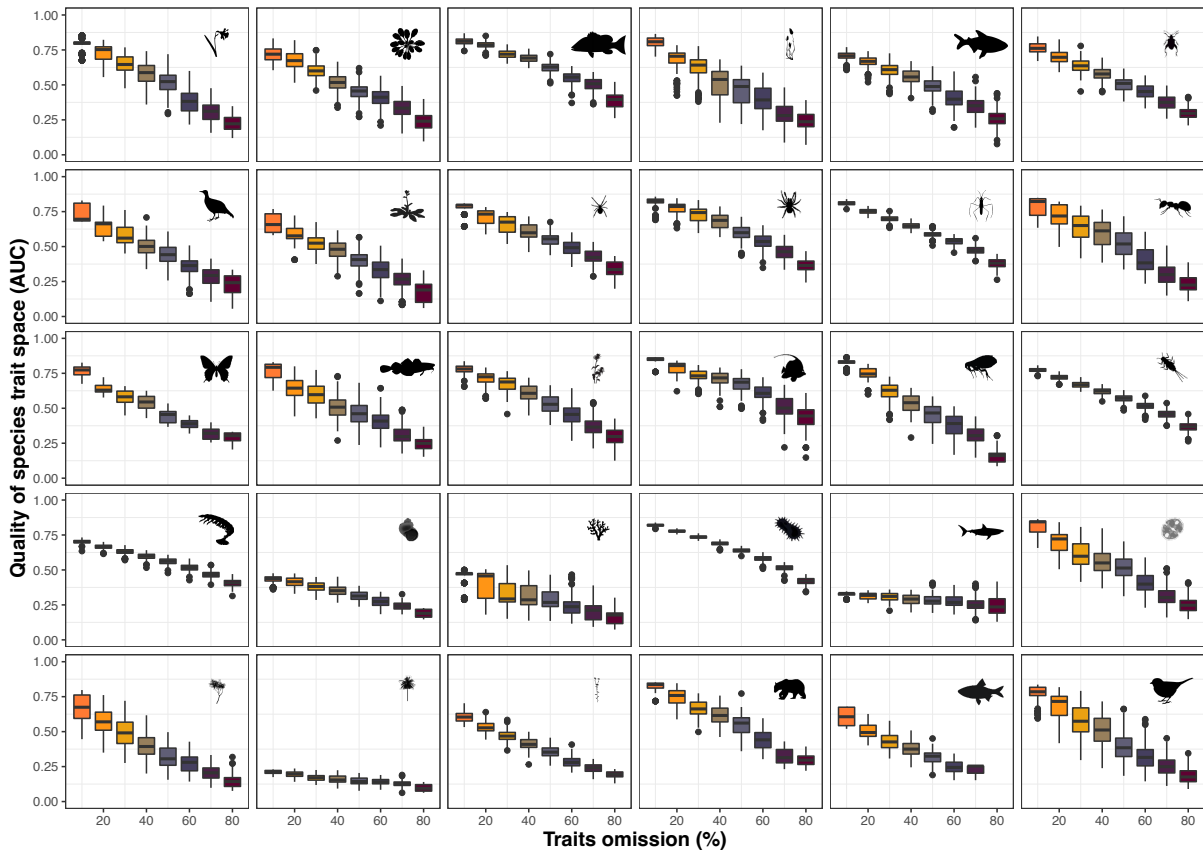
1228 **Figure 4:** Partial plots showing the influence of the five trait dataset characteristics on species  
 1229 trait space dimensionality measured with the elbow-based (first row) or threshold-based  
 1230 (second row) AUC criteria. The third row shows trait space robustness, in terms of AUC loss,  
 1231 to trait removal or omission (50%) according to the five characteristics. Only significant  
 1232 ( $p < 0.05$ ) relationships are colored the others are grey. Related statistics are reported in  
 1233 Supplementary Table 2.  
 1234



1235

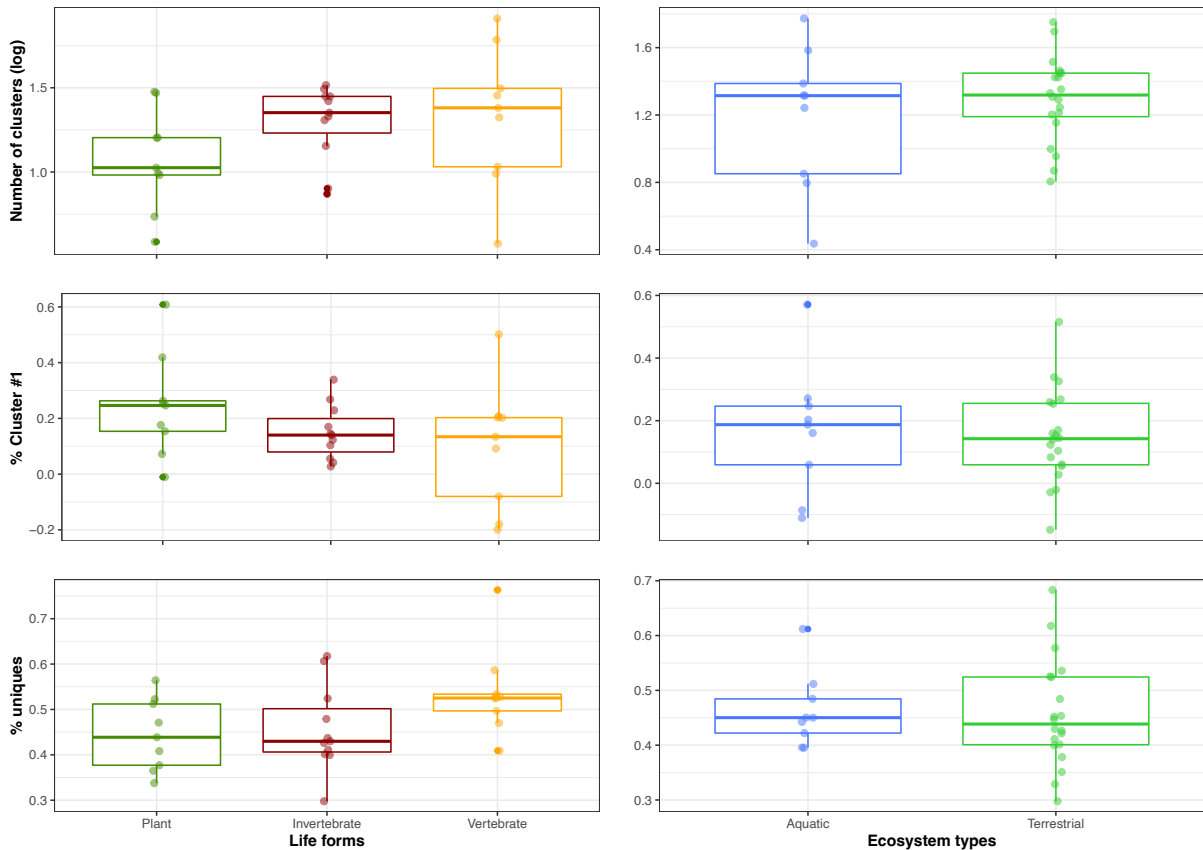


1236 **Figure 5:** Influence of the percentage of traits omission (between 10% and 80%) on the  
1237 quality of the trait space in terms of AUC when representing species in a trait space of lower-  
1238 dimensionality. For this, we randomly removed traits 100 times for each level of omission to  
1239 obtain the boxplots across the 30 datasets ranked by the total number of species (top-left to  
1240 bottom right). For 0% of trait omission AUC is 1.  
1241



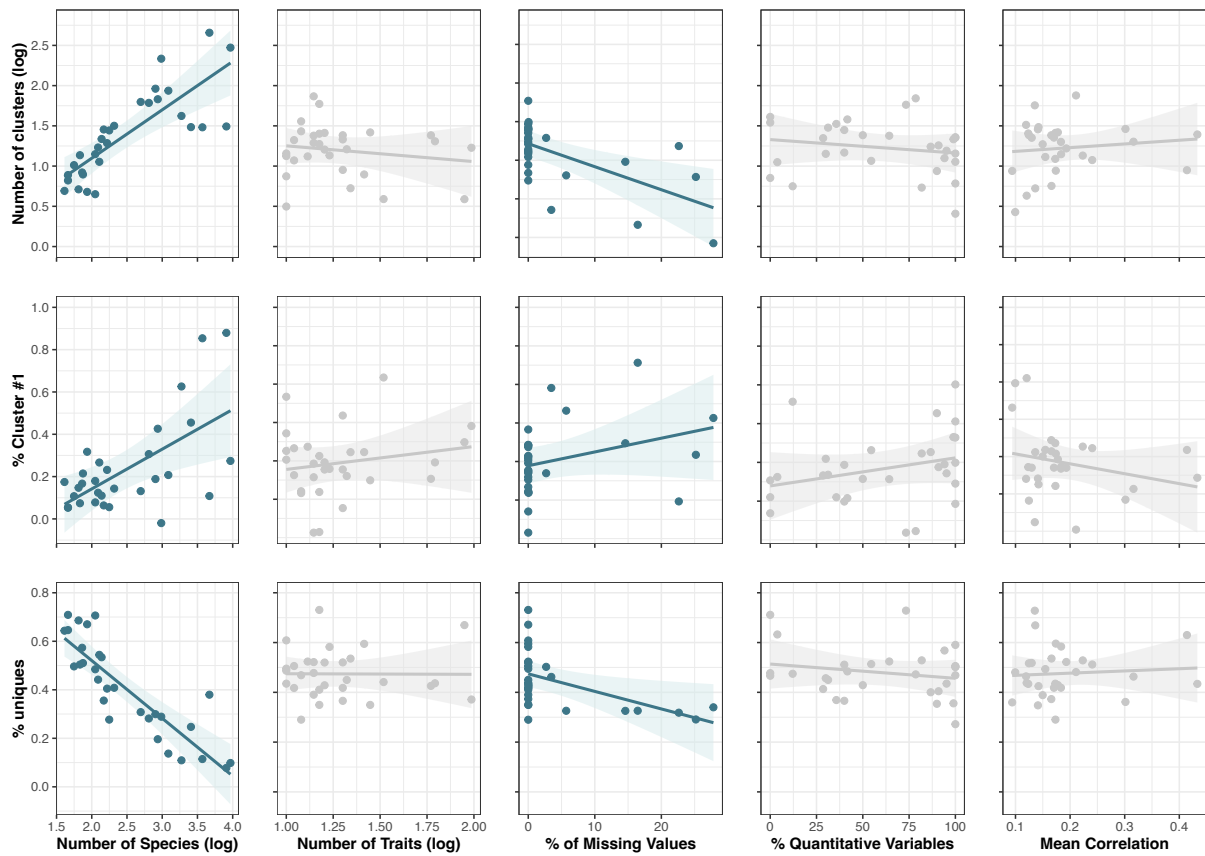
1242

1243 **Figure 6:** Partial plots showing the influence of the species life form (Plant, Invertebrate,  
 1244 Vertebrate) and ecosystem type (Aquatic, Terrestrial), while controlling for the five dataset  
 1245 quantitative characteristics, on the log-number of species clusters (first row), the proportion of  
 1246 species packed in the first or dominant cluster (second row) and the proportion of unique  
 1247 species so those isolated in the trait space (third row). Related statistics are reported in the  
 1248 Supplementary Table 3, the effects are all non-significant.  
 1249



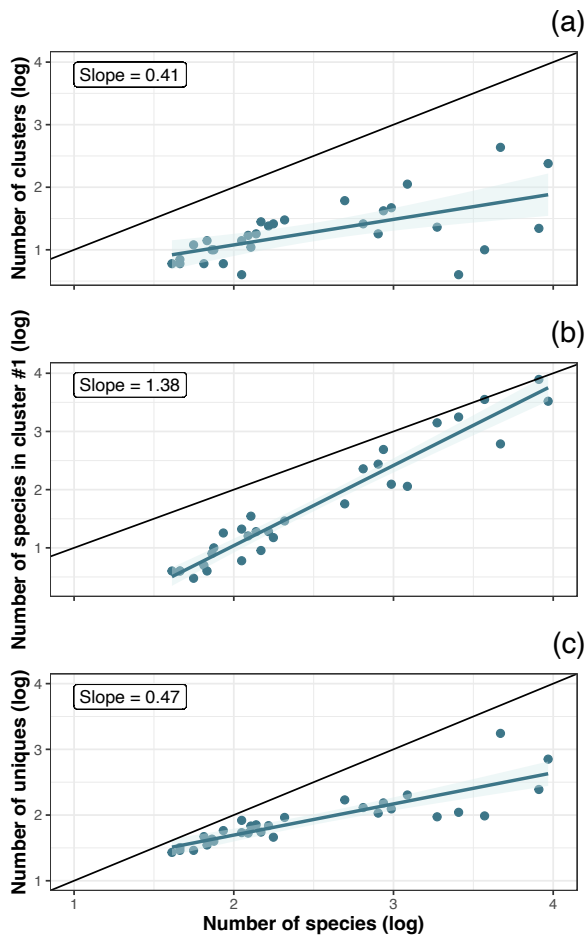
1250

1251 **Figure 7:** Partial plots showing the influence of the five trait dataset characteristics on the  
 1252 log-number of species clusters (first row), the proportion of species packed in the first or  
 1253 dominant cluster (second row) and the proportion of unique species so those isolated in the  
 1254 trait space (third row). Only significant ( $p < 0.05$ ) partial relationships are blue plain dots and  
 1255 lines, others are in grey. Related statistics are reported in Supplementary Table 4.  
 1256



1257

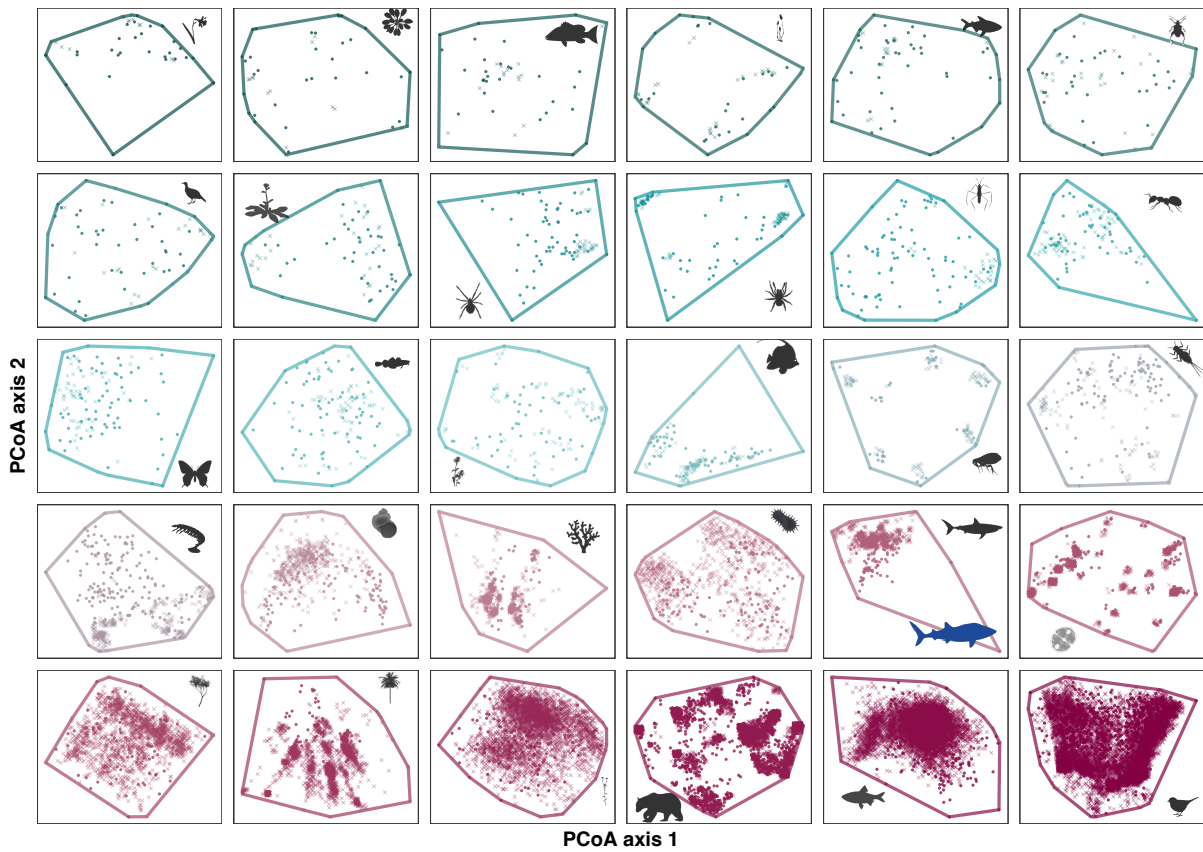
1258 **Figure 8:** Partial log-log relationships between the number of groups clustered by the fast  
1259 search and find of density peaks algorithm (a), the number of species in the most dominant  
1260 cluster (b) and the total number of unique species so those not being part of any group (c), and  
1261 the number of species in the 30 datasets. Slopes of the log-log relationships, so exponents of  
1262 power laws, are reported.  
1263



1264

1265 **Figure 9.** Trait spaces for the 30 datasets where the two axes come from Principal Coordinates  
1266 Analyses (PCoA) representing the distribution of species according to their trait values. Species  
1267 colored in dark are detected as statistically and ecologically unique species by the fast search  
1268 and find of density peaks algorithm. The whale shark (*Rhincodon typus*) is highlighted in blue  
1269 being highly distinct and unique in its clade. Datasets are ranked (top-left to bottom-right and  
1270 from dark green to dark red) following the number of species.

1271  
1272



1273