Peer Review File

Manuscript Title: Environmental structure in the evolutionary differentiation of bird skeletons.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This manuscript represents a substantial contribution to the literature on avian macroevolution. The methods will, I suspect, become widely adopted (or at least they should). It seems to me that the findings are sufficiently novel and interesting for publication in Nature, as well. However, I have five general recommendations that might help to strengthen the manuscript. These fall broadly under the categories of discussing prior work (highlighting what is novel/surprising in the current study), discussing ways in which the analytical results could produce false signal, connecting the patterns discovered to avian organismal biology, more detailed discussion of particular clades within the avian radiation (to avoid being overly general), and depicting the results.

I realize that several times I ask for further discussion and that there is a word limit in play. But for review purposes I'd suggest adding that discussion to the MS for the time being. It can always be shifted to a supplement later.

First, with regard to prior work: I would like to see more specificity with regard to what exactly has been said about early niche partitioning following the K/Pg extinction, about modularity, etc., especially in the context of all birds. Several of the papers cited here deal with smaller clades within Aves. Moreover, how novel are the findings in this work that have to do with modularity? Clarke and Middleton, 2008, and various papers thereafter discuss postcranial modularity in pan-Aves, do they not?

Second, I would like to see some discussions of what conditions might cause the MDI comparison method used here to indicate a lack of early partitioning (as it does) when in reality such partitioning did occur. Are there such conditions? And what more specifically are the cutoffs, qualitative or quantitative, for conclusions of early vs. late partitioning?

Third, this manuscript is written from such an elevated perspective that sometimes it seems like the biology of the birds themselves gets a bit lost. And there's a lot that's potentially of interest in the results. I would like for instance to see some discussion of the aquatic realm as a more challenging or extreme environment, biomechanically and physiologically, compared to the ancestral environment that birds inhabited, and to what extant those divergent ecological demands influence the observed evolution of disparity. Are there similar patterns among other bird clades that exist in extreme environments or have extreme lifestyles?

Fourth, and perhaps most importantly, I found myself wishing for deeper dives into bird clades of interest, and specifically into the aspects of their anatomy and function that underly distinctively high or low disparity, regional skeletal diversification, etc. These clades could be those which one might have reason to believe are interesting a priori (hummingbirds, toucans, parrots, corvids, etc.) or those which the results show to be interesting.

Fifth, and finally, in keeping with the last couple of requests, I would like to see as a summary figure a tree large enough that you can label terminals and major clades, and some indication on that tree of points of interest in your results, maybe in a kind of infographic format. Such a summary figure would serve to reach a much broader audience than currently your graphics might.

I would happily review a revised version of this manuscript.

Referee #2 (Remarks to the Author):

Navalón et al. present a study of morphological evolution of the crown birds using 3D landmark-based dataset across the major regions of avian skeleton system. Their comparative phylogenetic analyses found little evidence of early partitioning of morphospace during the early radiation of the crown birds. Overall, I think this manuscript could make good contribution to our knowledge of avian evolution, and the data presented here are quite impressive (although further explanation or clarification should be provided, see below). However, there are a number of caveats relating to the major conclusion that are not sufficiently addressed or controlled, and the findings as well as the methods have not very significantly advanced our understanding of crown bird diversification and the avian diversification in general.

My major concern is about the raw landmark data. I have noticed that the recent study (Orkney et al., 2021, Nature Ecology & Evolution) collected nearly the same landmark and semi-landmark from the same anatomical regions, skull, pectoral girdle, forelimb/hindlimb, etc. Indeed, there were 149 species sampled in that study, fewer than in the current work. Given some of the co-authors participated in both studies, I wonder if some species or landmarks used here were originally from Orkney et al., 2021. If so, this should and must be addressed in the main text. And the patterns of modularity and integration across avian skeleton found in Orkney et al. (2021) should be comprehensively compared and considered here, although the former study addressed different questions.

Considering the large samples of crown birds that exhibit huge degree of morphological variation, it is a challenge to deal with the placements of landmark on homologous points if there are no such points between certain taxa, e.g., the numbers of caudal trabeculae of sternum vary greatly (one pair, two pairs, and even absent), the presence/absence of the keel, procoracoid process of the coracoid. In addition, the authors shall consider to explain the inclusion of some bony elements, but not others. For instance, it is generally accepted that the pedal digits, both shape and size, are ecologically informative, which are pertaining to the main issue of the present manuscript (e.g., ecological diversification or partitioning between terrestrial, aquatic and arboreal taxa).

Is it possible that the lack of signal of early partitioning in skeletal form and size stem from the mosaic nature of avian skeleton? Specifically, it is possible that early partitioning can only be inferred or demonstrated by certain elements, or regions such as rostrum, and that signals got washed out by other ecological conservative regions. Related to this issue, different clades presumably have lineage-specific pattern and mode of disparity. The manuscript here analyzed the four subclades of the crown birds, which overall show different ecological adaptations, but a handful of lineages across these subclades occupy the same ecological nich by using different combinations of morphologies. The authors do not specify the ecology of sampled taxa, but rather accept the crude classification of the four subclades, e.g., terrestrial, arboreal, diving. Therefore, is it possible the lack of signal of early partitioning in fact gets diluted when pooling into large samplings?

The title is also strange. What is evolutionary radiation of bird skeletons? Differentiation might be a more appropriate term here for changes of bird skeletons.

Finally, I should stress that the volume of 3D data present here certainly benefit the whole community, but major revision of the manuscript is in need. Hope the author find these comments helpful, and sorry for not being more positive.

Referee #3 (Remarks to the Author):

Review of "Environmental structure in the evolutionary radiation of bird skeletons".

This manuscript describes an impressively detailed dataset of bird skeletal measurements that span the breadth of the extant avian radiation. The data set is exceptional in terms of breadth of coverage and range of quantified anatomical features. The data are then used to explore the evolution of morphological disparity through time, in major lineages, and in different subsets of the skeletal data set. The exploration of the data is thorough and the manuscript is also elegantly illustrated - the figures look beautiful. There are two areas of the manuscript that I found unconvincing. The first relates to rejection of early niche partitioning in relation to the temporal pattern of skeletal evolution. The second concerns the framing of the results in the context of environmental structure.

1) Early niche partitioning describes a pattern of morphological differentiation in which the majority of disparity arises early in clade history such that disparity among clades exceeds disparity within clades. The data set, while hugely impressive in many respects, includes 231 species, which is <2.5% of the total extant diversity of birds. The sampling includes species from all major lineages of birds but very sparse sampling within those lineages. This seems problematic because a detectable pattern of early niche partitioning requires the possibility for disparity to be limited within clades. It is not clear that the degree of within clade sampling allows for anything other than the observed patterns. One way to think about this is to ask how would disparity change if we add more species with respect to the metrics used. From the methods I understand this to be the mean pairwise Euclidean distance between species. At the extremes, adding more species could reduce disparity if species are packed into existing limits (within the data set) or they could increase disparity if species are added that expand morphospace. While some degree of expansion is possible, given the extent of non-sampled species it seems more likely that adding species will pack more densely into morphospace and reduce disparity. Whether this would be sufficient to flip the pattern towards early niche partitioning is not clear but in the absence of data the rejection of niche partitioning is problematic. With the current data and analyses I am not at all convinced that the conclusions are robust.

It would be unreasonable to expect the sampling to be increased in scale by an order of magnitude. One approach to reassure the reader that there are not inherent biases away from niche partitioning would be via simulation. This could be done by simulating a range of early partitioning scenarios (ideally over multiple dimensions to mimic the data) on a complete phylogeny of birds. Then prune the tree to the 231 species present in the current analysis and run the same disparity analyses as those used in the manuscript. If early partitioning is not detectable then that would suggest that the sampling fraction is not sufficient. There may well be better ways to address this problem and this is just a suggestion of an approach that may help to better understand the limitations of the data.

Also related to measuring disparity, the MDI test has been challenged an alternative method, the rank envelope test, has been advocated as a more powerful approach (see Murrell 2018. Methods in Ecology and Evolution https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13006).

2) The second area of concern is on the framing of the manuscript around environmental structure. This is emphasised in the title, abstract and numerous places throughout the manuscript but is not well justified. The environmental component comes from the comparison between the landbirds (Telluraves) and waterbirds (Aequorlitornithes). There are of course broad ecological differences between these clades but also substantial variation within them and overlap between

them. There are no explicit analyses that include environment and so the environmental differences largely boil down to differences between two clades. Of course environment might be an important factor (perhaps the most important factor) but that cannot be concluded from these analyses. Although I think this is significant issue, it can be addressed by toning down some aspects of the writing (e.g. line 17-18 "We also demonstrate differences in the structure of the macroevolutionary landscape across Earth's major environments" over-reaches the analyses).

Minor comments:

Lines 46-48: explain what perspective is incomplete. The sentence is currently too vague to understand what knowledge gap is being referred to.

Lines 62-63: I suggest inserting "likely" before lived since there remains a reasonable possibility that the common ancestor of extant birds is older than 83 million years.

Lines 284-286: The mean pairwise Euclidean distance is used as the metric of disparity . This is one of a wide array of possible metrics that could be used. Different disparity metrics measure different features – some are ideal for describing volume, others are better for describing density of morphospace. Presumably it was chosen as the metric that can most directly test the hypotheses of interest. It would be helpful to include an explanation for why this metric is expected to measure the features that you aim to measure (although precisely what those features are is also not clear).

Figure 1: Panel b shows disparity through time with a null expectation and observed trends for four clades. It is not clear which clades can be compared to the null – I think it is only the Neornithes that relate to the null. This requires clarification in order to interpret the figures appropriately.

Author Rebuttals to Initial Comments:

We now carefully read the suggestions from the three reviewers, and we made substantial changes in the text and figures following the majority of their suggestions. We have also implemented some completely new analyses, hopefully also satisfying all their concerns regarding statistical robustness. Specifically, we did:

1. **We thought long and hard about the main comments of Referee 3. To address them, we included a novel batch of analyses exploring the timing of the establishment of the morphospace of modern avian diversity for each element, region, and type of morphological data in our dataset. The aim of this is two-fold:**

   i) To differentiate between early establishment of the morphospace versus early partitioning of the morphospace, two different aspects of the mode of evolution which are often conflated together. Previous studies tended to focus on early establishment of morphospace, whereas we initially had focussed on early partitioning of morphospace. Now we include both, which established a better dialogue with previous works.

   ii) To demonstrate that our sample (of 228 species) can detect patterns of evolution that were reported previously, based on thousands of species. For this we show that our data can replicate the early establishment of morphospace results from Cooney et al., (2017), who analysed the beak shape of more than 2,000 species of living birds. Our analysis produces very similar results (Fig. 1B), suggesting that our data has the power to replicate macroevolutionary patterns using more taxon-dense data. We believe this occurs because our sample is phylogenetically over-dispersed and captures effectively most of the lineages of modern birds including all the deepest nodes which are central to the hypotheses being tested. See more details on this below.

2. **To address Referee 1's request for more in-depth information about bid subgroups, we included plates for our analyses of patterns of partition of morphospace for all elements, regions, and types of morphological data (Figs. S17-S34) in the same format as Figure 1 in the main text. This is intended to add more granularity to our visual portrayal of patterns of morphological evolution so patterns among smaller groups are readily available for inspection to the reader that is interested in this. We also discussed these in greater detail in the main text too as suggested by Referee 1.**

3. **Referee 3 also questioned if differences in skeletal evolution between Aequorlithornithes (waterbirds) and Inopinaves (landbirds) could be attributed to their environmental differences, or not. To address this, we included analyses of relative disparities for other additional clades of water- and land-based groups. This shows that the dichotomy in patterns of disparity across the skeleton between waterbirds and landbirds holds between other lineages of largely terrestrial versus largely aquatic modern birds. Specifically, we compared relative disparities across all the skeletal partitions and types of data in galliforms versus anseriforms, and in gruiforms versus all non-landbird neoavian largely terrestrial clades (i.e., otidimorphs plus strisoreans). Analyses of these extra groups also helps to address Referee 1's request for further detail.**

4. **We spotted two taxonomic mistakes in our sample which led us to remove two passerine taxa from our sample (*Gerygone* and *Smithornis*). We reviewed our entire dataset again to makes sure that there were no further errors. We then repeated all the previous and novel analyses with the amended sample.**

More specific comments:

Referees' comments:

Referee #1 (Remarks to the Author):
This manuscript represents a substantial contribution to the literature on avian macroevolution. The methods will, I suspect, become widely adopted (or at least they should). It seems to me that the findings are

sufficiently novel and interesting for publication in Nature, as well. However, I have five general recommendations that might help to strengthen the manuscript. These fall broadly under the categories of discussing prior work (highlighting what is novel/surprising in the current study), discussing ways in which the analytical results could produce false signal, connecting the patterns discovered to avian organismal biology, more detailed discussion of particular clades within the avian radiation (to avoid being overly general), and depicting the results.
I realize that several times I ask for further discussion and that there is a word limit in play. But for review purposes I'd suggest adding that discussion to the MS for the time being. It can always be shifted to a supplement later.
First, with regard to prior work: I would like to see more specificity with regard to what exactly has been said about early niche partitioning following the K/Pg extinction, about modularity, etc., especially in the context of all birds. Several of the papers cited here deal with smaller clades within Aves. Moreover, how novel are the findings in this work that have to do with modularity? Clarke and Middleton, 2008, and various papers thereafter discuss postcranial modularity in pan-Aves, do they not?

>> We added some more information following the reviewer's suggestions (lines 56 to 74). We already had included the citation (Clarke & Middleton 2008) suggested by the reviewer in our previous version of the article. However, we cited again in the appropriate context the reviewer suggests (line 71). The other 3 articles cited in the same context dealt with patterns of phenotypic evolution of skeletal proportions in Neornithes (Nudds et al., 2004), waterbirds (Wang & Clarke, 2014) and landbirds (Crouch & Rickelfs, 2019). These are the three main clades examined in this paper, but only the latter of this articles (landbirds; Crouch & Rickelfs, 2019) quantitatively approached (in an indirect way and not tested) the question of the early partitioning of morphospace for these clades. All of them assumed this pattern of early partitioning was present in birds. However, they did not explicitly test this hypothesis, which we tested for the first time in our article. It is important to stress that our work is not about modularity, although it rests in the concept that modular evolution occurs, and that different parts can therefore show different evolutionary patterns (modularity is a specific, special case of this, and is not the focus of our work). This latter result we demonstrated comprehensively for the first time using a complex multidimensional quantification of skeletal morphology.

References:

Clarke, J. A. & Middleton, K. M. Mosaicism, modules, and the evolution of birds: results from a Bayesian approach to the study of morphological evolution using discrete character data. Systematic biology 57, 185-201 (2008).

Crouch, N. M. & Ricklefs, R. E. Speciation rate is independent of the rate of evolution of morphological size, shape, and absolute morphological specialization in a large clade of birds. The American Naturalist 193, E78-E91 (2019).

Nudds, R., Dyke, G. & Rayner, J. Forelimb proportions and the evolutionary radiation of Neornithes. Proceedings of the Royal Society of London. Series B: Biological Sciences 271, S324-S327 (2004).

Wang, X. & Clarke, J. A. Phylogeny and forelimb disparity in waterbirds. Evolution 68, 2847-2860 (2014).

Second, I would like to see some discussions of what conditions might cause the MDI comparison method used here to indicate a lack of early partitioning (as it does) when in reality such partitioning did occur. Are there such conditions? And what more specifically are the cutoffs, qualitative or quantitative, for conclusions of early vs. late partitioning?

>>The cutoff is 0. We explain this in the text when describing the results ("Negative MDI values indicate…" and "…positive MDI values, indicating…"), and also in the Methods: "Negative MDI values indicates that the empirical curve exhibits lower subclade disparities than the simulated curve

and indicates that subclades explore smaller morphospaces than expected given the diversity of the full clade, suggesting early partitioning of trait space, or constrained subclade evolution. Positive MDI indicates that subclades explore larger morphospaces than expected given the diversity of the full clade, suggesting little partitioning of avian trait space, and/or that constraints on evolution at the level of all birds are similar to those of its subgroups. Confidence intervals use the 95% and 5% percentiles of simulated disparities in place of the mean". The cutoff is not a hard boundary because we evaluated differences from zero using iterative simulations and confidence intervals. Essentially if the confidence interval spans only negative values that would represents statistically robust evidence for early partitioning. Our simulations are generated under a fully multivariate model of Brownian Motion (Clavel et al., 2015). We consider this approach the most robust available at the moment of writing this text (see below in responses to the Reviewer 3).

We do not think there are common conditions under which this test of early partitioning would fail. However, while conducting our revisions, we also realised that early partitioning of morphospace (i.e., total morphospace is partitioned early into smaller subspaces) can be confused with a different type of observation related to early establishment of the total morphospace (i.e., early achievement of a large portion of present-day disparity during early times). Therefore, to further clarify our results, we included a new batch of analyses exploring when the morphospaces for living avian disparity were established, for each element and region, and type of morphological data in our sample (Fig 1. Figs. S15 and S16 and see lines 104-115 in the main text for discussion of these novel results). Our approach mirrors recent studies (e.g., Cooney et al., 2017) and allows us to determine if extant disparity was achieved early, as per classic evolutionary hypotheses, or later over avian evolution as compared with a neutral fully multivariate model of evolution (Clavel et al., 2015). These outcomes can then be compared with our more detailed analyses on early vs late partitioning of morphospace and offer complimentary information on phenotypic evolution. For instance, the total morphospace for modern beak shape disparity was explored comparatively early in avian evolution (in agreement with the findings of Cooney et al., 2017), however this morphospace was not partitioned early into smaller subspaces and later significant expansions in beak morphology occurred within sublineages (Figure S17B).

These analyses are similar to previous studies and permitted us validating our approach by very closely replicating with our smaller sample the patterns of early establishment of beak morphologies shown by a previous study with a much larger sample (~2000 species; Cooney et al., 2017). This suggest our sample is a good representation of modern avian diversity. See our responses below for more details on this.

References:

Cooney, C. R. et al. Mega-evolutionary dynamics of the adaptive radiation of birds. Nature 542, 344-347 (2017).

Clavel, J., Escarguel, G. & Merceron, G. mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. Methods in Ecology and Evolution 6, 1311-1319 (2015).

Third, this manuscript is written from such an elevated perspective that sometimes it seems like the biology of the birds themselves gets a bit lost. And there's a lot that's potentially of interest in the results. I would like for instance to see some discussion of the aquatic realm as a more challenging or extreme environment, biomechanically and physiologically, compared to the ancestral environment that birds inhabited, and to what extant those divergent ecological demands influence the observed evolution of disparity. Are there similar patterns among other bird clades that exist in extreme environments or have extreme lifestyles?

>>We completely agree with the reviewer. Therefore, we included significantly more discussion of interesting groups and mechanical challenges of aquatic birds also in response to the reviewer's other comments (see below). Our new comparisons among other pairs of water-linked versus terrestrial lineages are intended to show generalities of evolving mainly linked to aquatic environments as

opposed to doing so in terrestrial environments (Figure 4, Extended Data figure 4). We delved more into these effects in the current version of the manuscript and our results suggest that, at least, variation in skeletal proportions in broadly aquatic lineages seem to be larger than in related terrestrial lineages, although our sample size for these additional clades is admittedly limited (lines 141 and 144-150). Our interpretation for this is not that the aquatic realm might be a more challenging or extreme environment (a desert can be as physiologically challenging for instance) but that the topological complexity of terrestrial environments can offer substantially more opportunities for finer niche partitioning than aquatic environments (this is hinted, for instance by classic studies on how different passerine species partition the use of the vegetation cover in temperate forests). We consider this is also reflected in our results in morphological patterns in landbirds, as this varies more in local aspects of shape variation and not much in skeletal proportions. Finally, while we consider interesting the comparison of aquatic ecologies with the ancestral crown bird ecology, our knowledge for the ancestral crown bird ecology is very limited and different studies yielded different results. For instance, interpreting at face value the fossil record and the phylogenetic placement of near crown stem or early crown birds reveals that most of these animals seem to be linked to aquatic environments, suggesting an aquatic-linked ancestral ecology (Benito et al., 2022). Conversely, studies based on ancestral state reconstruction of ecological traits using living species reveal that the most likely ecology for the ancestral crown bird is terrestrial (Field et al., 2018). To date, no study has included these two kinds of data in a more informative analysis. While we are working at the moment in a similar project, an analysis like this is completely out of the scope for this article.

References:

Benito J, Chen A, Wilson LE, Bhullar BA, Burnham D, Field DJ. 40 new specimens of Ichthyornis provide unprecedented insight into the postcranial morphology of crownward stem group birds. bioRxiv. 2022 Jan 1.

Field, D. J. et al. Early evolution of modern birds structured by global forest collapse at the end-Cretaceous mass extinction. Current Biology 28, 1825-1831. e1822 (2018).

Fourth, and perhaps most importantly, I found myself wishing for deeper dives into bird clades of interest, and specifically into the aspects of their anatomy and function that underly distinctively high or low disparity, regional skeletal diversification, etc. These clades could be those which one might have reason to believe are interesting a priori (hummingbirds, toucans, parrots, corvids, etc.) or those which the results show to be interesting.

We consider this request very reasonable, and we understand it (despite the space limitations). As a result, we included plates summarising the detailed outcomes of all our analyses of morphospace partitioning encompassing all anatomical partitions and morphological data (Figs. S17–S34) including labelling of all the clades and species and underlining some examples of clades showing extremely high disparity and low disparity for every kind of data and anatomical regions. Therefore, visual inspection of the patterns in the phylogeny should reveal a myriad of detailed results on specific clades and should stimulate more detailed research in specific lineages in the coming years (including some ongoing work by our own team). We included some more detailed comments on these patterns in the text (see lines 158-161 and lines 163-169). Furthermore, we explored in detail the relative disparities across all the skeleton in four additional clades (Extended Data Figure 4 and Fig. S34) and we briefly commented on them in the main text (144-150).

Fifth, and finally, in keeping with the last couple of requests, I would like to see as a summary figure a tree large enough that you can label terminals and major clades, and some indication on that tree of points of interest in your results, maybe in a kind of infographic format. Such a summary figure would serve to reach a much broader audience than currently your graphics might.

>> We agree such a figure would be interesting. Therefore, we now incorporated a figure like this in our supplementary information (Fig. S35) that we will hope will satisfy the reviewer's suggestion. The

figure includes our phylogenetic tree with the taxa names and clades labelled, it also includes a categorisation of broad ecological categories.

I would happily review a revised version of this manuscript.

Referee #2 (Remarks to the Author):

Navalón et al. present a study of morphological evolution of the crown birds using 3D landmark-based dataset across the major regions of avian skeleton system. Their comparative phylogenetic analyses found little evidence of early partitioning of morphospace during the early radiation of the crown birds. Overall, I think this manuscript could make good contribution to our knowledge of avian evolution, and the data presented here are quite impressive (although further explanation or clarification should be provided, see below). However, there are a number of caveats relating to the major conclusion that are not sufficiently addressed or controlled, and the findings as well as the methods have not very significantly advanced our understanding of crown bird diversification and the avian diversification in general.

>> We disagree with the assessment of this reviewer. We believe that that our research does advance understanding of crown bird evolution and avian diversification. Many ideas that were previously framed on qualitative feelings or 'expert knowledge' are explicitly tested for the first time using quantitative data at a transformative scale (e.g., early partitioning of morphological evolution in birds, the mosaic nature of skeletal variation in vertebrates or patterns of interzonal versus subzonal evolution). Our study places important constraints on what might be true or untrue about avian macroevolution and sets the stage for substantial future research. It is comparable to other research in avian macroevolution published in Nature (e.g., Cooney et al., 2017) and will have, at least, a similar impact on the field. The referee does not say what they think is not novel, so it is difficult for us to say more. But we are extremely surprised by this comment regarding a paper that has been more than five years from multiple people in progress and sets out to test some of the broadest ideas in vertebrate macroevolution by studying for the first time macroevolutionary patterns across much of the skeleton.

My major concern is about the raw landmark data. I have noticed that the recent study (Orkney et al., 2021, Nature Ecology & Evolution) collected nearly the same landmark and semi-landmark from the same anatomical regions, skull, pectoral girdle, forelimb/hindlimb, etc. Indeed, there were 149 species sampled in that study, fewer than in the current work. Given some of the co-authors participated in both studies, I wonder if some species or landmarks used here were originally from Orkney et al., 2021. If so, this should and must be addressed in the main text. And the patterns of modularity and integration across avian skeleton found in Orkney et al. (2021) should be comprehensively compared and considered here, although the former study addressed different questions.

>>Our dataset is an extended version of that used by Orkney et al. (2021). Details of morphological data are explained in detail in the dataset paper by Bjarnason & Benson (2021). Both are now clearly signposted in the main text Methods, right at the start. It should allow the referee and other readers to find information about the ontogeny of the dataset, and also how we made pan-avian landmarking decisions. We apologise for not making this clearer in the previous version of the manuscript. Our dataset includes 79 species more than the one used in Orkney et al., 2021), focusing on better representing diversity within passerine birds, a group that represents more than half the diversity of living birds and so was central to the questions of our current work.

>>Although we believe our dataset is impressive, as acknowledged by all the referees, and is substantially advanced from its previous form, it is not a 'dataset' paper (unlike Bjarnason & Benson, 2021) and this comment by Referee 2 is not a scientific concern *per se*. Our questions here, and therefore also our analyses, are of a very different form to those addressed by Orkney et al., 2021, who only investigated covariances among the shapes of different skeletal elements. We do not do that here, so it isn't clear how we would comprehensively compare and consider the results between the two papers. Having said that, we do make use of some of the findings of Orkney et al. Specifically, that terminal elements have high ecological signal – relevant to interpretation of the macroevolutionary dynamics shown in landbirds (Inopinaves) in the current paper. We think this is a fairly orthodox

approach to referencing past literature, similar to treatment of other previous studies here, and by other works.

References:

Orkney, A., Bjarnason, A., Tronrud, B. C. & Benson, R. B. Patterns of skeletal integration in birds reveal that adaptation of element shapes enables coordinated evolution between anatomical modules. Nature Ecology & Evolution 5, 1250-1258 (2021).

Bjarnason, A. & Benson, R. A 3D geometric morphometric dataset quantifying skeletal variation in birds. MorphoMuseuM 7 (2021).

Considering the large samples of crown birds that exhibit huge degree of morphological variation, it is a challenge to deal with the placements of landmark on homologous points if there are no such points between certain taxa, e.g., the numbers of caudal trabeculae of sternum vary greatly (one pair, two pairs, and even absent), the presence/absence of the keel, procoracoid process of the coracoid.

>>Thanks for this comment. We have presented the data and consideration regarding homologies in Bjarnason & Benson 2021. This includes addressing how to landmark the sternum, which is arguably the most difficult element in our dataset from the perspective of homology (and also see Lowi-Merri et al. 2021 for similar consideration). Furthermore, we use curves of semilandmarks to capture the shape of entire regions within skeletal elements which define a regional relationship of homology when more specific point homology is difficult to establish (e.g., caudolateral rim of the sternum as mentioned by the reviewer). This solution (Bookstein, 1997) is widely accepted in the field of shape analysis, and it is backed by a significant body of literature including seminal works which are routinely used to teach geometric morphometrics (e.g., Zelditch et al., 2012).

References:

Bookstein FL. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Medical image analysis. 1997 Apr 1;1(3):225-43.

Lowi-Merri TM, Benson RB, Claramunt S, Evans DC. The relationship between sternum variation and mode of locomotion in birds. BMC biology. 2021 Dec;19(1):1-23.

Zelditch ML, Swiderski DL, Sheets HD. Geometric morphometrics for biologists: a primer. academic press; 2012 Sep 24.

In addition, the authors shall consider to explain the inclusion of some bony elements, but not others. For instance, it is generally accepted that the pedal digits, both shape and size, are ecologically informative, which are pertaining to the main issue of the present manuscript (e.g., ecological diversification or partitioning between terrestrial, aquatic and arboreal taxa).

>>Our dataset is much larger and more inclusive than any previous studies, which mainly focussed on single elements, especially the skull. It is true that we did not succeed in landmarking every part of the skeleton. But we certainly have gone much further than anyone was able to previously. We regret that we could not go even further than we have, because not all of the specimens that we accessed were sufficiently complete. We have added a line to the Methods explaining that some parts were not landmarked for practical reasons, in spite of their potential ecological relevance (lines 251-253). This includes distal elements such as the pedal phalanges, which are often disarticulated or absent in museum specimens. We acknowledge we did not do everything possible here. But we did do a lot and there is scope for future analysis. For example, we are currently undertaking a separate study of pedal phalanges. This is more complicated than the 13 landmarked skeletal elements we digitised in our current manuscript because the foot is a complex, multi-element system, requiring additional considerations. However, for this study we digitised 207,245 individually placed coordinates (including

landmarks and semilandmarks) in total and that took us several years by a large team of people which underlines that studies such as ours are time consuming.

Is it possible that the lack of signal of early partitioning in skeletal form and size stem from the mosaic nature of avian skeleton? Specifically, it is possible that early partitioning can only be inferred or demonstrated by certain elements, or regions such as rostrum, and that signals got washed out by other ecological conservative regions.

>>Yes, it is possible that the lack of early partitioning of skeletal form (i.e., at the largest anatomical scale) results from mosaicism. However, we also analysed patterns of partitioning at smaller scales, down to the local shape deformations of individual elements, which seems to not be acknowledged by this reviewer. Following the reviewer's advice, we also included in this version of the paper analysis of the partitioning of morphospace in the beak region (=rostrum). Our results show that even the beak does not show a pattern of early partitioning of morphospace (although it shows an early establishment of morphospace revealed by our new analyses, see Fig. 1; lines 104-115; and the new section of Material and Methods. Patterns of establishment of morphospace). Furthermore, we found that evidence for early-partitioning was absent in general (except in landbirds for the combined form of the whole skeleton). Therefore, we are very confident that a lack of early partitioning in the avian skeleton or its parts cannot be explained solely by mosaicism. This is shown in Figures 2-4 in the main text, Extended Data Figure 3 and it is widely commented in the manuscript.

>>On a separate note, we believe that our paper represents the most comprehensive documentation of mosaicism of the avian skeletal morphology so far. We quantified both the timing of establishment of morphospace and, in much greater detail, the patterns of partitioning of morphospace and the relative disparities, three major aspects of the mode of evolution, across a wide range of anatomical scales in many skeletal elements.

Related to this issue, different clades presumably have lineage-specific pattern and mode of disparity. The manuscript here analyzed the four subclades of the crown birds, which overall show different ecological adaptations, but a handful of lineages across these subclades occupy the same ecological nich by using different combinations of morphologies.

>>We agree entirely with this comment. The fact that different clades have lineage-specific patterns is central to our conclusions. We show this for landbirds, waterbirds and passerines in great detail. The nature of macroevolutionary analysis is that to sample a clade well means sampling many members that clade. Therefore, the feasibilities of taxon sampling limit us to a focus on these major groups, which include most of bird diversity and seem, qualitatively, to be in stark contrast to each other. Certainly, there is scope for future analyses to probe more into other clades, or into subgroups of landbirds, waterbirds and passerines with more dense sampling within each clade. However, this will require many years of further work in data collection. Furthermore, our analyses are not strictly limited to landbirds, waterbirds and passerines. In fact, there is a lot of details and granularity our figures. We have done more to bring that granularity in response to the comments for Referee 1 (see above). Also please note that we included specific analyses of some other subclades such as gruiforms, anseriforms, galliforms, strisoreans and otidimorphs and sublineages within waterbirds (Extended Data Figure 4 and Fig. S34) which are nonetheless included in the patterns for neornithines (Figs 2-3 & Extended Data Figure 3). This leaves only palaeognathans as the single major clade for which we did not undertake specific analyses on. This is due to our limited sample of palaeognathans.

It is true that there is a handful (i.e., very small number, see below) of ecological convergences between landbirds and waterbirds, and maybe especially between landbirds and passerines. We then show that these groups have fundamentally different macroevolutionary dynamics. We propose ecological explanations for this. Those explanations rest on these groups being largely distinct, not entirely distinct with zero examples of convergence. We nuanced this in the current version (line 132). We added a qualitative category in one of the phylogenetic trees provided as Supplementary Information (Fig. S35) that encodes whether the taxon exhibits a water-linked autecology or not. This refers to whether the species in question uses water bodies as the main resource for food acquisition, breeding, or concealment. The information was extracted from the Handbook of the Birds of the World (Del

Hoyo et al., 2017), the standard reference for avian ecology. You can clearly see in that file that the majority of waterbirds exhibit water-linked ecologies with two exceptions (*Turnix* and *Pedionomus*) while landbirds exhibit mostly non-water-linked ecologies with four exceptions (*Pandion*, the related *Alcedo* and *Chlorocerylе*, and *Cinclus*).

References:

Del Hoyo, J. et al. Handbook of the Birds of the World Alive. Lynx Editions, Barcelona. (2017).

The authors do not specify the ecology of sampled taxa, but rather accept the crude classification of the four subclades, e.g., terrestrial, arboreal, diving. Therefore, is it possible the lack of signal of early partitioning in fact gets diluted when pooling into large samplings?

>>The major macroevolutionary hypothesis under discussion is the idea that birds show a signal of early partitioning following the K-Pg. By-definition, this hypothesis makes predictions of a signal at large phylogenetic and ecological scales (i.e., across all birds, or lineages within Neoaves). The hypothesis of early partitioning relates to evolution *among* ecotypes, not within them. For this reason, we specifically chose broad-brush ecological descriptors (see Fig. S35), within which there was divergence into narrower and more specific ecotypes. In that context, it doesn't make sense to say that a signal of morphological partitioning would be diluted when treated at large sampling scales – those are the scales at which the hypothesis predicts that the pattern should be present. We have tried to make this clearer in the current version of the manuscript (lines 63-74).

>>Note also a distinction between 'ecological partitioning' [used by the referee] and 'morphological partitioning' [used in our manuscript]. Our data are morphological, and we test the hypothesis of early partitioning of morphology. The distinction is important because there are various ways in which the two could be decoupled from each other (as discussed in our manuscript).

The title is also strange. What is evolutionary radiation of bird skeletons? Differentiation might be a more appropriate term here for changes of bird skeletons.

>>We agree with this comment, and we have changed the title accordingly.

Finally, I should stress that the volume of 3D data present here certainly benefit the whole community, but major revision of the manuscript is in need. Hope the author find these comments helpful, and sorry for not being more positive.

Referee #3 (Remarks to the Author):
Review of "Environmental structure in the evolutionary radiation of bird skeletons".
This manuscript describes an impressively detailed dataset of bird skeletal measurements that span the breadth of the extant avian radiation. The data set is exceptional in terms of breadth of coverage and range of quantified anatomical features. The data are then used to explore the evolution of morphological disparity through time, in major lineages, and in different subsets of the skeletal data set. The exploration of the data is thorough and the manuscript is also elegantly illustrated - the figures look beautiful. There are two areas of the manuscript that I found unconvincing. The first relates to rejection of early niche partitioning in relation to the temporal pattern of skeletal evolution. The second concerns the framing of the results in the context of environmental structure.

1) Early niche partitioning describes a pattern of morphological differentiation in which the majority of disparity arises early in clade history such that disparity among clades exceeds disparity within clades. The data set, while hugely impressive in many respects, includes 231 species, which is <2.5% of the total extant diversity of birds. The sampling includes species from all major lineages of birds but very sparse sampling within those lineages. This seems problematic because a detectable pattern of early niche partitioning requires the possibility for disparity to be limited within clades. It is not clear that the degree of within clade sampling allows for anything other than the observed patterns. One way to think about this is to ask how would disparity change if we add more species with respect to the metrics used. From the methods I understand this to be the mean pairwise Euclidean distance between species. At the extremes, adding more

species could reduce disparity if species are packed into existing limits (within the data set) or they could increase disparity if species are added that expand morphospace. While some degree of expansion is possible, given the extent of non-sampled species it seems more likely that adding species will pack more densely into morphospace and reduce disparity. Whether this would be sufficient to flip the pattern towards early niche partitioning is not clear but in the absence of data the rejection of niche partitioning is problematic. With the current data and analyses I am not at all convinced that the conclusions are robust.

>>Thank you for the detailed consideration of our approach. There is a lot potentially to discuss here and we have divided our response into separate comments to help with processing the information. We really hope to be able to convince the referee that what we did is acceptable — this project cost a lot of heart and soul, and we made our methodological decisions in good faith, considering exactly the sorts of points that the referee has put forward here. We provide various explanations below:

>>Firstly, the referee has written about our sample size in terms of absolute species counted. However, this is only indirectly relevant to this kind of analysis because the count of species sampled is not as important as the representation of nodes through time. Our current level of sampling does a great job of covering deep divergences in bird evolution up to around half-way through the Cenozoic. We cannot easily add the many (many) more species required to cover much more shallow divergences. However, the hypotheses under discussion predict early partitioning in the wake of the KPg mass extinction event — therefore with greatest relevance to the first half of the Cenozoic, and well within our sampling envelope. We also note that our results are based on relative disparities not absolute disparities and the relationships among disparities at least among the major clades recorded by deep divergences are likely to be retained considering our sampling. Indeed, our sampling follows that of the two most comprehensive phylogenomic studies which sampled most lineages of living birds (therefore, being an accurate representation of most deep nodes in the avian tree).

>>Nevertheless, we were interested to find ways of asking whether our sample was sufficient to replicate results of existing analyses of much larger species samples, to validate our approach. We therefore conducted an analysis of the establishment of the morphospace for only the beak region exactly analogous to the one presented in Cooney et al., 2017. The results, using the univariate simulate method used by Cooney et al, are shown in Fig. 1B. Our analysis produces largely the same pattern of early establishment of the morphospace to that of Cooney et al, in spite of the fact that Cooney et al analysed more than 2000 species (new Figure 1 in the main text). The main difference in findings is predicted above – we detect this pattern only during the first half of bird evolution whereas Cooney et al have a denser sample of younger divergences and find that it persisted for a bit longer in time (~three-quarters of bird evolution).

>>We can also provide empirical support that our sample is sufficient to recover evidence for early partitioning. For example, that we do find evidence for early-partitioning in the whole skeletal form in landbirds. This indicates that our data has the power to detect such patterns and also indicates substantial variation in the extent of support for different modes of evolution. For example, we find that landbirds show different very patterns to those seen in waterbirds. These observations are relevant to the suggestion that "It is not clear that the degree of within clade sampling allows for anything other than the observed patterns**".**

>> The similarities between results from our analyses and the ones provided by Cooney et al., 2017 suggest that choice of disparity estimator is not having a big effect on results. But we still want to include some discussion of that here too. We used mean pairwise Euclidean distances, not least because those were used in foundational work on the method we used (Harmon et al. 2003), but also that they are robust to variation in sample size (see below). In contrast, the referee describes a view that they believe these should be inversely correlated with sample size. That is not the case. In fact, compared to some other estimators of spread, they are relatively invariant to sample size (e.g., Cherry et al 1982, and various Foote papers, cited below; we choose these because they are foundational in the disparity literature), with similar performance to variance-based measures. This differs from the (poor) performance of range-based measures (and cell count measures), which by definition have a positive correlation with sample size. We agree in spirit with the referee's point that ideally, we would choose

an estimator that shows no such correlation with sample size. Indeed, that was the basis for our decision here to use the mean pairwise dissimilarity. This is different to what the referee has written: "adding more species could reduce disparity if species are packed into existing limits**". In fact, the reason that mean pairwise differences are invariant to sample size is that the set of Euclidean distances between a new datapoint and the pre-existing points within some sampling universe will be similar to those returned when each of the pre-existing points were added. So, disparity should be more-or-less the same as sample size increases, because new points are drawn from the same underlying distribution, with each new point providing a similar estimate of the mean pairwise dissimilarity of the distribution (citations that affirmed his general point are given below).**

>>A separate question concerns whether an effect such as that described by the referee would bias our analyses towards rejecting evidence for early partitioning. Analytically, early-partitioning is signified by low subclade disparity compared to whole-clade disparity. If the referee is correct and our disparity measure is negatively correlated with sample size, then that would be expected to bias analysis in favour of early partitioning, not against it, because subclades have smaller sample sizes than the whole clade. We also note that our analyses involve comparison to simulated distributions, and that estimators applied to those distributions should have the same properties in regard to sample size bias as for the empirical distributions. It is not clear to us, based on the referee's explanation, how adding taxa to the analysis, past some reasonable coverage of disparity specifically (rather than species richness) could flip the result to the opposite findings.

>>Finally, we also acknowledge the limitations of our methods and data. For instance, our taxon sample is relevant and has the power to detect patterns of early partitioning of disparity, but we doubt it will be of any use to detect patterns of late 'niche packing' within subclade morphospaces. Nevertheless, the statement of 'absence of data**' is difficult to interpret here as our study represents the most detailed quantification of skeletal vertebrate phenotype ever tackled.**

References:

Cherry, L. M., S. M. Case, J. G. Kunkel, J. S. Wyles, and A. C. Wilson. 1982. Body shape metrics and organismal evolution. Evolution 36:914-933.

Foote, M. 1991a. Morphologic patterns of diversification: examples from trilobites. Palaeontology 34:461-485.

Foote, M. 1991b. Morphological and taxonomic diversity in a clade's history: the blastoid record and stochastic simulations. Contributions from the Museum of Paleontology, University of Michigan 28:101-140.

Foote, M. 1992a. Rarefaction analysis of morphological and taxonomic diversity. Paleobiology 18:1-16.

Foote, M. 1992b. Paleozoic record of morphological diversity in blastozoan echinoderms. Proceedings, National Academy of Sciences, U.S.A. 89:7325-7329.

Foote, M. 1993a. Discordance and concordance between morphological and taxonomic diversity. Paleobiology 19:185-204.

Foote, M. 1993b. Contributions of individual taxa to overall morphological disparity. Paleobiology 19:403-419.

Additional references in the matter:

Ciampaglio, C. N., Kemp, M. & McShea, D. W. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27, 695-715 (2001).

Foote, M. The evolution of morphological diversity. Annual Review of Ecology and Systematics 28, 129-152 (1997).

Harmon, L. J., Schulte, J. A., Larson, A. & Losos, J. B. Tempo and mode of evolutionary radiation in iguanian lizards. Science 301, 961-964 (2003).

Wills, M. A., Briggs, D. E. & Fortey, R. A. Disparity as an evolutionary index: a comparison of Cambrian and Recent arthropods. Paleobiology 20, 93-130 (1994).

It would be unreasonable to expect the sampling to be increased in scale by an order of magnitude. One approach to reassure the reader that there are not inherent biases away from niche partitioning would be via simulation. This could be done by simulating a range of early partitioning scenarios (ideally over multiple dimensions to mimic the data) on a complete phylogeny of birds. Then prune the tree to the 231 species present in the current analysis and run the same disparity analyses as those used in the manuscript. If early partitioning is not detectable then that would suggest that the sampling fraction is not sufficient. There may well be better ways to address this problem and this is just a suggestion of an approach that may help to better understand the limitations of the data.

We broadly agree with the intention of this comment. Because of the nature of our analyses which use inferences about various nodes in the tree, the question is basically about when (measured in time towards the recent) does our sample run out of power? It would be easiest to answer this by comparison to an existing empirical study. Therefore, we conducted an analyses of early vs late establishment of the morphospace for the shape of the beak region in our sample of 228 taxa. The same analysis was conducted by Cooney et al., 2017 using a sample of more than 2,000 species of modern birds and a similar landmarking scheme to capture beak shape. Our results reveal a very closely similar pattern to the one presented in Cooney et al., 2017, but our analysis seems to run out of statistical power to detect this pattern around 40 Ma (i.e., about halfway through bird evolution), whereas theirs continued to detect it closer to the recet (up to about 15 Ma or so), consistent with their much better sample of younger nodes in the tree. This suggesting our sample and methods have the power to detect precise time patterns of disparity in the evolution of shape in birds (also see above for other arguments of why we consider our methods have the power to detect the patterns we comment on).

See this paragraph in the main text:

"We find that birds established a wide morphospace of overall skeletal form during their early evolution (Fig. 1), driven primarily by divergence of the head and forelimb morphology (Fig. S15). This pattern is also evident for some individual elements, including the skull and mandible (Fig. S16), as well as the local shape variation of the sternum and tarsometatarsus, and overall shape (but not local shape) of the humerus, radius and ulna (Fig. S16). Other elements, especially of the hindlimb and pectoral girdle, show no evidence of this (Fig. S15). Our finding of early establishment of bird beak morphospace directly replicates most of the pattern from Cooney et al.4, when we exactly replicate their approach (Fig. 1b). Cooney et al.4 analysed a much larger dataset (2,000 species), effectively sampling both old and young nodes in bird phylogeny, returning evidence for early establishment of bird morphospace up c.20 MYA (Fig 2A in Cooney et al.4). In contrast, our dataset is smaller and samples mostly older nodes and returns evidence for early establishment of bird morphospace for a slightly shorter duration, up to c. 30 MYA (Fig. 1b), or close to the margin of the 95% confidence interval up to c. 40 MYA when using multivariate simulations, different to Cooney et al.4. This shows that our dataset, in spite of its smaller size, has sufficient power to test hypotheses of early morphological evolution among birds, which is the focus of our analysis and interpretations. We would not expect out data to be so informative for more recent events in bird evolution."

Also related to measuring disparity, the MDI test has been challenged an alternative method, the rank envelope test, has been advocated as a more powerful approach (see Murrell 2018. Methods in Ecology and Evolution https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.13006).

>> We have read carefully the paper suggested by the reviewer. The article challenges the MDI-test (not the MDI as a variable though) as currently implemented as a statistical test to identify statistically

significant deviations from Brownian Motion in particular points in time. Our approach does not aim to test this. We are solely interested in the general shape of the empirical curve of disparity through time as compared with simulations. For this, the Morphological Disparity Index (MDI) calculated in the traditional way (which is also ours) and using the rank method suggested in the paper, would only produce marginally different values as evidenced in the Figure 4 of the paper the reviewer suggested. For this, we do not think using one method or the other for time averaging the simulated trends is a concern for our study. As such, we are particularly careful in our paper not to interpret local variations in the empirical curve at face value.

2) The second area of concern is on the framing of the manuscript around environmental structure. This is emphasised in the title, abstract and numerous places throughout the manuscript but is not well justified. The environmental component comes from the comparison between the landbirds (Telluraves) and waterbirds (Aequorlitornithes). There are of course broad ecological differences between these clades but also substantial variation within them and overlap between them. There are no explicit analyses that include environment and so the environmental differences largely boil down to differences between two clades. Of course environment might be an important factor (perhaps the most important factor) but that cannot be concluded from these analyses. Although I think this is significant issue, it can be addressed by toning down some aspects of the writing (e.g. line 17-18 "We also demonstrate differences in the structure of the macroevolutionary landscape across Earth's major environments" over-reaches the analyses).

We understand the concerns of the reviewer here and we significantly toned down some of our claims (see lines 17-18). We also included new analyses that test, where allowed by sample size, aspects of disparity between other water-and land-related groups (gruiforms vs non-landbird terrestrial neoavians; galliforms vs anseriforms, see Extended Data Figure 4). These analyses broadly confirm our initial findings, but we also flagged the potential exceptions to these broad generalities now, in our main text.

Regarding the distribution of ecologies among groups, we have added a new detailed tree figure to illustrate this (Fig. S35). It shows a very striking phylogenetic pattern. Contrary to the referee's comment, there is not really a substantial overlap between ecologies of landbirds and waterbirds: only a handful of lineages within waterbirds exhibit autecologies not linked to water environments, while only a handful of landbirds exhibit autecologies significantly linked to water environments (Fig. S35). Also, the striking observation that all these water-linked or terrestrial lineages are related is foundational to our current understanding of the relationships in birds (e.g., Jarvis et al., 2014; Prum et al., 2015). Although this is the most striking case as it represents the most lineages of water and land lineages, the pattern also characterises other parts of the tree (e.g., galliforms versus anseriforms, Extended Data Figure 4, Fig. S35). The strong phylogenetic signal of water vs terrestrial ecologies in birds is a well-recognised pattern in avian evolution and it has been discussed in every recent text in avian macroevolution (e.g., Avian Evolution, Gerald Mayr, 2016). Independent of that, hopefully you are convinced by the evidence-summary we provided now (Fig. S35).

References:

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh A. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014 Dec 12;346(6215):1320-31.

Mayr G. Avian evolution: the fossil record of birds and its paleobiological significance. John Wiley & Sons; 2016 Oct 31.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 2015 Oct;526(7574):569-73.

Minor comments:

Lines 46-48: explain what perspective is incomplete. The sentence is currently too vague to understand what knowledge gap is being referred to.

We agree the sentence was too vague, so we deleted it.

Lines 62-63: I suggest inserting “likely” before lived since there remains a reasonable possibility that the common ancestor of extant birds is older than 83 million years.

We agree with this, and we changed the manuscript accordingly.

Lines 284-286: The mean pairwise Euclidean distance is used as the metric of disparity . This is one of a wide array of possible metrics that could be used. Different disparity metrics measure different features – some are ideal for describing volume, others are better for describing density of morphospace. Presumably it was chosen as the metric that can most directly test the hypotheses of interest. It would be helpful to include an explanation for why this metric is expected to measure the features that you aim to measure (although precisely what those features are is also not clear).

See our previous comment on this topic and section Methods. Quantifying disparity in the main text. Broadly, we chose a disparity estimator that had the favourable property of being relatively insensitive to sample size. Our thinking about this is philosophically different to the referee because we aren’t trying to describe/measure features, but instead to estimate properties from an incomplete sample. Disparity here is an index of the spread of points in morphospace, that has a long history in the literature (see e.g., citations to Cherry, Foote and Harmon above). Other concepts may be difficult to quantify accurately or have negative statistical properties such as dependency on sample size. For example, range-based measures might be more intuitive as measures of the extent of morphospace but are highly sensitive to sample size and to outliers. And ‘density’ is a combined measure of species richness and phenotypic disparity, which has rarely been quantified in studies of disparity.

Figure 1: Panel b shows disparity through time with a null expectation and observed trends for four clades. It is not clear which clades can be compared to the null – I think it is only the Neornithes that relate to the null. This requires clarification in order to interpret the figures appropriately.

The reviewer is right. We have clarified this in the caption in Figure 1 (lines 481-483).

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

With apologies for the delay, I have now read over the authors' revised manuscript and supplement and the referee comments and responses.

I am largely satisfied with the authors' responses to my own comments. I would still like to see some discussion of possible ancestral bird ecology, probably in the introductory section.

With regard to referee 2's comments, I tend to agree that data overlap with Orkney et a. (2021) diminishes the novelty of this study somewhat. However, this work does involve a substantial expansion of that dataset and a considerably greater scope.

With regard to semilandmarks, they certainly have their own issues, but there is no widely accepted better solution. I would suggest that the authors conduct some analyses using only the fixed landmarks in their dataset and see whether these are consistent with results from their full landmark set.

With regard to non-inclusion of some skeletal elements, in general I agree with the authors that this is not an enormous problem. However, the referee is correct that pedal phalanges have been assigned particular significance in various studies of bird evolution. My hunch, though, is that inclusion of those structures would not alter the overall study results.

With regard to mosaicism, I believe that the authors have satisfactorily addressed this concern of the referee's.

With regard to referee 3's comments, the most significant seems to be about species sampling. I strongly support the authors' overall claim that adequate sampling of deep divergences requires a smaller species count than sampling of shallow divergences. However, the referees have no actual evidence beyond the authors' assertions that these divergences have indeed been captured. Can the authors use either recent large-scale phylogenetic analyses or some kind of reasonable estimation/supertree approach to estimate the total number of divergences among all birds they have captured and missed within various time bins, say 65-55, 55-45, 45-35, 35-25 ma, etc.? Moreover, is it possible to estimate or even speculate on the number of such divergences missed owing to extinction?

With regard to ecological categories, again especially considering Fig. S35 I tend to agree broadly with the authors. However, "water-linked" as in Fig. S35 does seem very broad, and I wonder whether the other referees do have a point with regard to using being part of a clade as a proxy for ecology. Could the authors at least partition the "water-linked" lifestyles into those that involve a good deal of time on land and those that do not?

Referee #2 (Remarks to the Author):

I appreciate the author's replies and revisions. I am convinced that the authors have done their best to respond to referee's comments. I have no further comment and would like to support the publication of this contribution.

Referee #3 (Remarks to the Author):

I appreciate the thoughtfulness and thoroughness of the responses to my comments. All of my concerns have been addressed, either by new analysis or with a detailed explanation in the response and appropriate changes to the paper.

The revision has raised one small issue that I can't quite figure out. Specifically, the null models in figure 1 include distributions derived based on multivariate estimation and univariate estimation. The confidence intervals are much broader for the multivariate null but it is not clear why this is the case. The input data are PC scores that should be orthogonal (although in phylogenetically transformed space there may be some correlation, though they would be weak). The variances for each axis should be similar for the multivariate (mvMORPH) simulations and the independent (phytools) simulations. I am probably just missing something but feel that this is something the authors should comment on.

Author Rebuttals to First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

With apologies for the delay, I have now read over the authors' revised manuscript and supplement and the referee comments and responses.

I am largely satisfied with the authors' responses to my own comments. I would still like to see some discussion of possible ancestral bird ecology, probably in the introductory section.

Thank you for your comments. We added a phrase to reflect the current state of our knowledge on this matter (see line 60). This comes from two main sources. First, analyses of ancestral ecology using living species within a phylogenetic framework reveal that the oldest nodes in crown birds are consistently predicted as non-arboreal (Field et al., 2018). Secondly, the majority of recent phylogenetic analyses reveal that most near-crown stem birds (and many early branching crown birds) seem to exhibit water-linked ecologies of some sort, suggesting a water-linked ecology for the ancestral crown bird. Combining these two, we can be only reasonably sure that the evidence indicates that the ancestral bird ecology might have not been arboreal, but we cannot say much more for sure. More work is needed on this matter.

References:

Field, D.J., Bercovici, A., Berv, J.S., Dunn, R., Fastovsky, D.E., Lyson, T.R., Vajda, V. and Gauthier, J.A., 2018. Early evolution of modern birds structured by global forest collapse at the end-Cretaceous mass extinction. Current Biology, 28(11), pp.1825-1831.

Benito, J., Chen, A., Wilson, L.E., Bhullar, B.A.S., Burnham, D. and Field, D.J., 2022. 40 new specimens of Ichthyornis provide unprecedented insight into the postcranial morphology of crownward stem group birds. bioRxiv.

With regard to referee 2's comments, I tend to agree that data overlap with Orkney et a. (2021) diminishes the novelty of this study somewhat. However, this work does involve a substantial expansion of that dataset and a considerably greater scope.

We are glad the reviewer sees the potential of our work!

With regard to semilandmarks, they certainly have their own issues, but there is no widely accepted better solution. I would suggest that the authors conduct some analyses using only the fixed landmarks in their dataset and see whether these are consistent with results from their full landmark set.

Thanks for this suggestion of ways to address the initial round of comments from referee 2. As far as we understood, reviewer 2, was concerned in the initial reviews for the homology of highly variable regions (e.g., caudolateral processes of the sternum) and how this is accounted for in geometric morphometrics. The use of sliding semilandmarks is a common practice used to circumvent these issues in the field of shape analysis since its formulation in the late 90s (Bookstein, 1997) and it is a widely recognised solution to capture the shape of structures that do not exhibit easily identifiable type I or II landmarks (e.g., Zelditch et al., 2012). In our case, removing semilandmarks would involve getting rid of more than half the information about three-dimensional geometry in most skeletal elements. We would prefer not to do this, because the resulting landmarks would be a very incomplete quantification of the structure of the bird skeleton. It is worth noting that our current approach is consistent with various other recent studies. For example, other similar recent studies have similar and (much) lower proportions of landmark to sliding semilandmark data (Cooney et al., 2017; Felice & Goswami, 2018). And referee 2 seems now satisfied with our approach. Therefore, we hope that our use of sliding semilandmarks will not be contentious.

References:

Bookstein F.L., 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med. Image. Anal. 1: 225–243.

Zelditch, M.L., Swiderski, D.L. and Sheets, H.D., 2012. Geometric morphometrics for biologists: a primer. academic press.

Felice, R.N. and Goswami, A., 2018. Developmental origins of mosaic evolution in the avian cranium. Proceedings of the National Academy of Sciences, 115(3), pp.555-560.

Cooney, C.R., Bright, J.A., Capp, E.J., Chira, A.M., Hughes, E.C., Moody, C.J., Nouri, L.O., Varley, Z.K. and Thomas, G.H., 2017. Mega-evolutionary dynamics of the adaptive radiation of birds. Nature, 542(7641), pp.344-347.

With regard to non-inclusion of some skeletal elements, in general I agree with the authors that this is not an enormous problem. However, the referee is correct that pedal phalanges have been assigned particular significance in various studies of bird evolution. My hunch, though, is that inclusion of those structures would not alter the overall study results.

We agree with the reviewer. As we said before, we are working on a similar project dealing with pedal ecomorphology in amniotes including birds. Therefore, this is definitely something we have thought deeply about and that we will be willing to explore in future work.

With regard to mosaicism, I believe that the authors have satisfactorily addressed this concern of the referee's.

Thank you!

With regard to referee 3's comments, the most significant seems to be about species sampling. I strongly support the authors' overall claim that adequate sampling of deep divergences requires a smaller species count than sampling of shallow divergences. However, the referees have no actual evidence beyond the authors' assertions that these divergences have indeed been captured. Can the authors use either recent large-scale phylogenetic analyses or some kind of reasonable estimation/supertree approach to estimate the total number of divergences among all birds they have captured and missed within various time bins, say 65-55, 55-45, 45-35, 35-25 ma, etc.? Moreover, is it possible to estimate or even speculate on the number of such divergences missed owing to extinction?

Thanks for the suggestion, this is a good idea that would illustrate the effect of taxon sampling on node age sampling through time. We now did the following (new Methods text): "To illustrate that our sample of extant bird species (N = 229) provides an effective sample of ancient bird divergences, we compared the count of phylogenetic lineages through time for a supertree of 9,993 extant bird species sourced from Cooney et al. 2017 (complete supertree from Jetz et al. 2012, using backbone from Prum et al., 2015) to the count of lineages in a pruned tree that includes only those species that we sampled. Results show that our species sample achieves a near-complete coverage of deep bird divergences up to around 30 million years ago (Figure S37) and only missing half the lineages at about 20 million years ago". Therefore, we consider our sample is apt to establish patterns or establishment and partitioning of morphospace during the first part of their history (up to about 40 million years ago) but cannot inform about patterns during the young history of birds, such as packing of morphospace.

References:

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M. and Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature, 526(7574), pp.569-573.

Oliveros, C.H., Field, D.J., Ksepka, D.T., Barker, F.K., Aleixo, A., Andersen, M.J., Alström, P., Benz, B.W., Braun, E.L., Braun, M.J. and Bravo, G.A., 2019. Earth history and the passerine superradiation. Proceedings of the National Academy of Sciences, 116(16), pp.7916-7925.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K. and Mooers, A.O., 2012. The global diversity of birds in space and time. Nature, 491(7424), pp.444-448.

With regard to ecological categories, again especially considering Fig. S35 I tend to agree broadly with the authors. However, "water-linked" as in Fig. S35 does seem very broad, and I wonder whether the other referees do have a point with regard to using being part of a clade as a proxy for ecology. Could the authors at least partition the "water-linked" lifestyles into those that involve a good deal of time on land and those that do not?

Thanks for this suggestion. We agree that the difference between water-based and terrestrial birds is gradational and that it would be nice if our categories reflected that. We made another supplementary figure to reflect a finer partition of water-linked ecologies (Figure S36). This follows a subdivision of semiaquatic ecologies that was suggested in a recent paper some of the authors of this study also co-authored (Fabbri et al. 2022 Nature 603 (7903), 852-857). It reflects whether a particular taxon frequently engages in subaqueous immersion for procuring food or concealment, or not. Our scoring reveals that approximately half of the water-linked birds do not engage in frequent subaqueous immersion stressing the discontinuities of water-linked ecologies in birds and supporting our findings that evolutionary variation in the skeleton of water-linked birds follows a different pattern than arboreal/terrestrial lineages.

Referee #2 (Remarks to the Author):

I appreciate the author's replies and revisions. I am convinced that the authors have done their best to respond to referee's comments. I have no further comment and would like to support the publication of this contribution.

Thank you very much for these words!

Referee #3 (Remarks to the Author):

I appreciate the thoughtfulness and thoroughness of the responses to my comments. All of my concerns have been addressed, either by new analysis or with a detailed explanation in the response and appropriate changes to the paper.

Thank you very much for recognising all the work that we put into this to explore the issues raised by your previous review. They made the paper significantly better than it was before.

The revision has raised one small issue that I can't quite figure out. Specifically, the null models in figure 1 include distributions derived based on multivariate estimation and univariate estimation. The confidence intervals are much broader for the multivariate null but it is not clear why this is the case. The input data are PC scores that should be orthogonal (although in phylogenetically transformed

space there may be some correlation, though they would be weak). The variances for each axis should be similar for the multivariate (mvMORPH) simulations and the independent (phytools) simulations. I am probably just missing something but feel that this is something the authors should comment on.

This is a good question and is one that we thought about too. The difference in width of the simulation envelope for the multivariate version versus the univariate version is very pronounced for beak shape. But it is less pronounced for many of the other shape, form, and proportions data that we analysed. This difference can only feasibly result from covariances among PC axes, which exist so far as comparative analyses are concerned because we used non-phylogenetic PCA. We think it is likely that the referee is right – that these covariances are small (i.e., axes are only slightly non-orthogonal) for most of our analyses, but are larger in some instances (e.g., for beak shape). We know that such covariances (=integration) can in theory cause more eccentric shape variation (Felice et al., 2018) and so could cause the widening of the simulation envelope we are seeing here. It is very hard to see what else could cause such an effect.

References:

Felice, R.N., Randau, M. and Goswami, A., 2018. A fly in a tube: macroevolutionary expectations for integrated phenotypes. Evolution, 72(12), pp.2580-2594.

Reviewer Reports on the Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I have no further comments on the manuscript. The authors have done an admirable job of addressing all concerns and not shying away from the prospect of additional analyses.

Referee #3 (Remarks to the Author):

I am satisfied with the revisions and response to my remaining question.