The Shape of Penguins in Four Dimensions

Assessing macroevolutionary shifts in a constructional morphology framework

A thesis presented in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

in

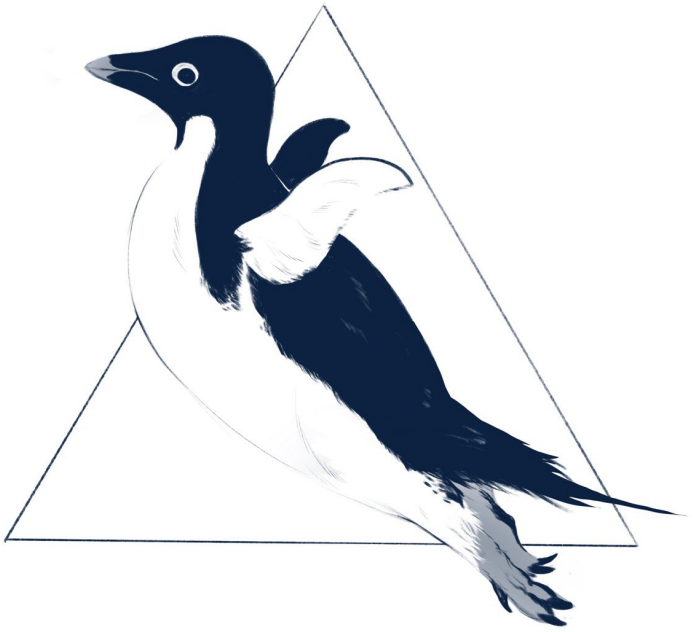
Zoology

at Massey University, Albany Campus,

New Zealand

Simone Giovanardi

2021



There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

Charles Darwin

Sono diventati un po piccoli, sono diventati.

Anonymous

From horses we may learn not only about the horse itself but also about animals in general, indeed about ourselves and about life as a whole.

George Gaylord Simpson

Abstract

Describing the morphology of a clade through deep time provides rich insight into the drivers that shaped modern diversity. It is in this context that Adolph Seilacher developed the constructional morphology concept, in order to describe which are the major pressures acting on an organism morphological appearance. The three components or apices of constructional morphology are the evolutionary history of the organism, the constraints placed on the structure, and the opportunity provided by adaptation. The structure of this thesis is based on the constructional morphology concept with each chapter focusing on the impact of each of the three components of constructional morphology with the scope to provide a novel approach to quantify morphological macroevolution. After introducing concepts of constructional morphology, geometric morphometrics and Bayesian statistics in Chapter 1, Chapter 2 presents an analysis to estimate the historical apex: a phylogenetic analysis based on the synthesis of previous published matrices, as well as a description of the giant fossil penguin *Kairuku waewaeroa*. The resulting phylogenetic tree indicates that the penguin evolutionary history was characterised by many monophyletic large groups that challenges previous results and indicate that the body plan of extinct penguins could be more diverse than previously thought. Chapter 3 focused on the structural apex, aiming to provide a generalisable Bayesian approach to estimate the size of extinct giant penguins in the context. By measuring the total volume of the femur and the humeral articular facet of the coracoid it was possible to generate two sets of models that together provided novel evidence in favour of reduced body mass estimates for giant penguins when compared with prior published estimates. Moreover,

although the two sets of estimates are derived from two distinct features, the body mass estimates from the two models tend to converge, providing confidence in the accuracy of the Bayesian-informed method. Chapter 4 presents and investigation into the impact of adaptation on two separate locomotory modules, the humerus and the tarsometatarsus, using 3D geometric morphometric techniques. Comparing morphological rates of change reveals a steady rate decrease in the humerus and more heterogeneous rates for the tarsometatarsus. Similar results are obtained by estimating the morphospaces for humeri and tarsometatarsi from hypothetical ancestors using a penalized likelihood approach. The synthesis that this constructional morphology framework approach provides highlights the important relationship between shape and size, showing how size can be a driver of morphological innovation. More importantly, the results of this thesis highlight the relevance that constructional morphology still has today, and how it can be integrated into palaeontology and evolutionary biology studies through the use of advanced statistical techniques. A constructional morphology approach is not solely applicable to penguins and may be extended to a broad range of groups of organisms, contributing thus to better understand the underlying forces that shaped the origins of modern biota.

Acknowledgments

It goes without saying that I would never been able to finish this thesis if not for the support of many people, I will do my best to not leave any name behind. These last three years have been challenging not only to PhD students, it is thus crucial to acknowledge the following people and institutions. A huge acknowledgment goes to Massey University that funded my research and supported me with the the doctoral scholarship. Many thanks goes to Linh Mills that was always available when I was in doubt for any sort of administrative paperwork. I am grateful to Anil Malhotra and Tony Shi and to all the IT team at Massey for their support and for granting me to use the Massey University cluster server, without it this thesis would have featured way less phylogenetic trees.

Many thanks goes also to the natural sciences and ecology team here at Albany, Odette Howarth for being available to prepare biological specimens, James Dale for his supervision, Aaron Harmer for giving his thoughts on this thesis and Dave Aguirre and Heather Hendrickson for their precious comments at the earlier stages of this project. I shared this journey with many other researchers and students that I have to mention, Emma Holvast, Jacques DeSatge, Emma Feenstra, Heshani Edirisinghe, Akshya Ilangovan, Abigail Kuranchie, Jessica Patiño Pérez, Mehrnaz Tavasoli, Vanessa Arranz, Hayden Pye, Michelle Roper, Sam Caurruthers and Wesley Webb, I can ensure that I learned something from all of you. A special thank goes also to Enzo Reyes, a talented PhD student, a flatmate but most importantly a good friend.

Where this project would be without fossils? Huge thanks goes to the team

at the Waikato museum Te Whare Taonga o Waikato: Salina Ghazally, Pauline Farquhar, Steve Chappell, Jon Primmer, Rodney Cook, Stephen Pennruscoe, Anita Robertson and Cherie Meecham, for being always available to open the doors of the museum to study *Kairuku waewaeroa*. Huge thanks goes also to Mike Safey and the Junats for donating the fossil to science and to Chris Templer for his work in preparing the specimen. Many thanks goes to Paul Scofield and Vanessa De Pietri at the Canterbury Museum for allowing us to scan many of their extinct and extant seabirds. I am grateful to Ewan Fordyce also for the specimens that provided us and to Marcus Richards for his kindness and his knowledge of fossil penguins. Matt Rayner and Ruby Moore at The Auckland War Memorial Museum Tāmaki Paenga Hira were essential in providing a lot of small femurs. A special thanks goes to Alan Tennyson, Felix Marx and Thomas Schultz for their availability and for sharing the specimens at the Te Papa museum. I'm also thankful to Te Papa for having involved me in their excavations in the last years, searching fossils is definitely one of my favourite activities.

A huge thank goes to Daniel Ksepka for having co-authored the description of *Kairuku waewaeroa*, for his insights on penguin osteology and for the comments that provided me on this thesis.

I'm grateful for the supervision provided by Emanuel Tschopp that always found the time for a little chat and for his comment on this project. The greatest of all acknowledgements goes to Daniel Thomas, without him as main supervisor I would not be writing these words today. His experience, his ideas and his support have been fundamental for my development and the development of this text. I could not have hoped for a better mentor and friend.

Thanks to the skillful Italian team of peer reviewers: Fabrizio, Elio, Paride, Cisco, Ciosso, Bonne, Roger, Joel and Crunch. Lastly I want to thank some people on the other side of the world that were always with me during this three years: Alice, Mamma, Papá, Nonna e Nonno.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The constructional morphology framework

Constructional morphology is a model introduced by German Paleontologist Adolph Seilacher (Seilacher, 1970) with the aim of explaining the morphological disparity among organisms (Seilacher & Gishlick, 2019). The concept assumes that similarities occurring between different biological entities are mainly the consequence of three factors: 1) shared ancestry (e.g. homology), when two distinct organisms share a character inherited from their most recent common ancestor (Darwin, 1872; Panchen, 1999; Wagner, 1989); 2) adaptation, when two distantly related organisms have independently evolved a similar feature for analogous functions; 3) structure, when features in different organisms exist because of analogous generating mechanisms (i.e. the same physical constraint).

Constructional morphology provides the framework for the investigation presented in this thesis. The three components of constructional morphology (shared ancestry, adaptation, structure) are modelled as the corners of a ternary diagram that explains the origins of a phenotype. A phenotype that is the result of equal parts ancestry, adaptation and structure would be placed in the centre of that triangle. A phenotype that is purely a consequence of constraint would plot at the tip of the "structure" corner, and likewise for ancestry and adaptation (Seilacher, 1970). Seilacher and Gishlick (2019) later included the effect of environment in determin-

ing the appearance of a phenotype and expanded the framework into a tetrahedral structure termed morphodynamics (Fig. 1.1). Here we focus on the constructional morphology version of the diagram to assess paleobiological questions.

Figure 1.1: Graphical representation of the morphodynamics framework. Constructional morphology represents one face of a three dimensional tetrahedron. From Seilacher and Gishlick (2019)

The constructional morphology concept has itself been adapted through time, keeping almost the same underlying structure but receiving "aptive triangle" as a synonym (McGhee, 1999). Stephen Jay Gould in his book "The Structure of Evolutionary Theory" (Gould, 2002) popularised the aptive triangle concept but this idea was presented to the scientific community well before by Gould and Lewontin (1979). The main purpose of Gould and Lewontin (1979) was to criticise the view of the so-called "adaptationist programme", also known as the Neo-Darwinian current. This school of thought greatly emphasised the role of adaptation in the origin and function of morphological features in organisms. According to researchers who subscribed to the "adaptationist programme", all phenotypic characters were "molded" entirely under selective pressure, and thus were adaptive in nature (Costa & Bisol,

1978; Rudwick, 1964; Shea, 1977). In an attempt to contrast this view, Gould and Lewontin (1979) argued that phenotypes can be interpreted as the result of at least three distinct shaping factors, using the model from Seilacher (1970) as a counter argument.

Indeed, adaptation still plays a role in this framework but represents only a single vertex, the one linked to the functional aspect of the organism. Another vertex can be conceptualised as the structural constraints that act on the phenotype. The third vertex accounts for the historical dimension, acknowledging that specific biological features are are simply inherited (Gould, 2002) (Fig. 1.2).

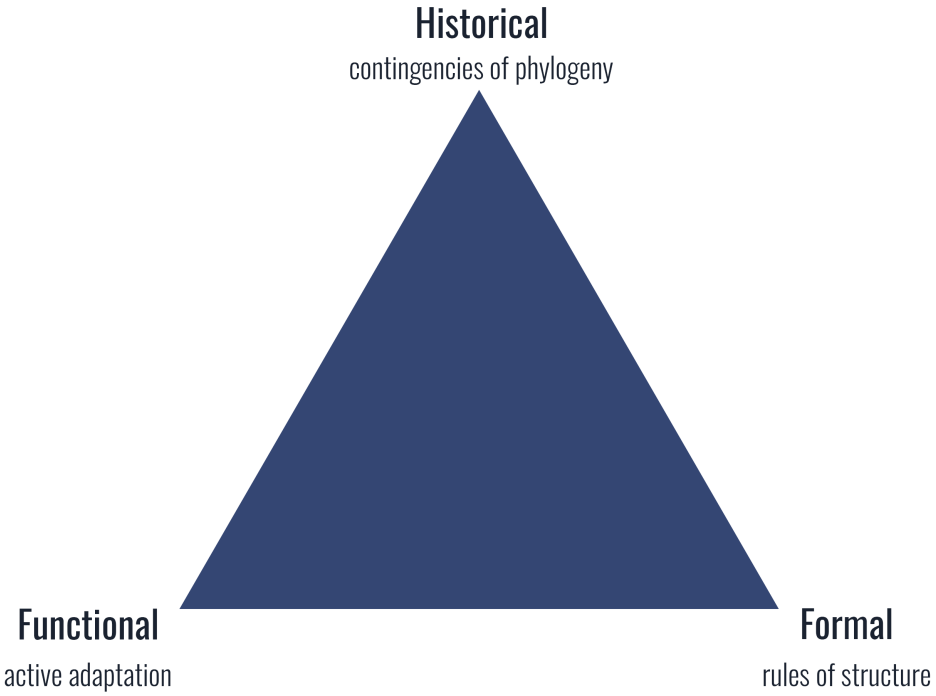


Figure 1.2: The aptive triangle concept. Modified from Gould (2002).

While Gould (2002) acknowledged the strong influence that the constructional morphology framework (Seilacher, 1970) had on his idea, another major source of inspiration for Gould was the Galton Polyhedron metaphor (Galton, 1894). Galton was trying to describe the complex relationship between genotype and phenotype and developed a model which described organisms as polyhedra located on a plane. The plane represented a morphospace, that is an hypothetical surface where all potential phenotypes can be projected, with each morphology bearing its own unique

set of coordinates. Ideally, similar phenotypes should be closer, and an increase in morphological dissimliarity corresponds to an increase in morphospace distances. The face on which the polyhedron lies is the actual phenotype of a given organism. Under the Galton Polyhedron metaphor, natural selection acted to push the polyhedron from one face to another, thus forcing phenotype to change through time. The concept at the core of this representation (Fig. 1.3) was that species cannot occupy an infinite continuum of forms, but rather a limited number of discrete phenotypes dictated by internal constraints.

Figure 1.3: Representation of the Galton polyhedron. From Held (2009).

Constructional morphology sensu Seilacher (1970) is a valuable concept that continues to guide evolutionary research due to it is intuitive combination of phylogeny, physiology and environmental constraint (Cubo, 2004; Gould, 2002; McGhee, 1999). Seilacher intended his idea to be a framework for inspiring research rather than a theory on its own, but nevertheless he also advocated the potential for constructional morphology to reveal evolutionary patterns (Briggs, 2017; Seilacher, 1984). Seilacher's concept is now used as a context for testing evolutionary hypotheses (Cubo et al., 2008).

A possible issue here is that phenotypes are the outcomes of complex biological interactions between several networks placed at different hierarchical levels. The nature of these networks can be genetic, developmental/morphogenetic, metabolic as well as external factor like trophic and environmental networks. For this reason

being able to identify one aspect of an organism as an incontrovertible product of a single vertex is not plausible. However, as an assumption it should be acknowledged that when a researcher is studying a specific phenotype they are not looking directly at one of the vertices. Instead, they are "sampling" a point within the area of the constructional morphology triangle (i.e. ternary diagram) (Cubo, 2004). This single point will be the result of the vertices operating together, but with three different magnitudes as in a three coordinate system. An important consideration discussed by Cubo (2004) is that the three vertices are "fuzzier" than earlier thought. Understanding the contribution of the "targeted" vertices is instructional, while keeping in mind that a pure attribution is not possible. This view doesn't undermine the validity of the general framework since in most scientific disciplines it is extremely difficult, or even impossible, to completely isolate an observation from all potential confounding factors (Worrall, 2002).

This thesis embraces the constructional morphology concept as a guideline to investigate how the three fundamental aspects (adaptation, structure, phylogeny) can be jointly studied. Each chapter will have one vertex of the triangle as the primary focus. This thesis will present mostly morphological data and thus explore the phenotypic aspects of organisms.

1.2 Geometric morphometrics

1.2.1 Historical perspective

Phenotypes were the main sources of information for understanding evolutionary relationships and biodiversity before the rise of molecular methods (reviewed in Adams et al., 2004). Studies of morphological phenotypes underwent a "quantification revolution" (Bookstein, 1998) in the early early 20^{th} century when measurements and statistical methods gained popularity for assessing inter- and intra-specific differences. This revolution was accompanied by the rise of new research disciplines including morphometry (J. S. Huxley, 1932; Pearson, 1901; Thompson, 1945). The

main focus of morphometry was to investigate how biological shape changes in relation to environmental variables or other factors (reviewed in Dryden and Mardia, 1998). Early morphometrics research was often based on linear measurements, ratios of measures, angles between points, and counts of discrete features (Rohlf & Marcus, 1993). Measurements were often analysed with bivariate or multivariate statistical methods to assess covariation between the morphometric measurements to understand the interactions with external factors (Blackith & Reyment, 1971).

Shape would become important for phenotype studies in the mid 20^{th} century, with foundational studies like Raup (1967). Shape is defined as "...all the geometric information that remains when location, scale and rotational effects are filtered out from an object..." (Kendall, 1977). However, this definition of shape exposes an underlying weakness in the general morphometric framework. Between traditional morphometricians there was little or no general agreement on which method to adopt to fully separate size and shape (as reviewed in Adams et al., 2004; Jungers et al., 1995; Sundberg, 1989). Moreover, the measurements collected from a specimen were not always taken between homologous points, and consequently the values could imply different biological meanings in different specimens. Lastly, the ability to visualise the results of these methods were heavily limited, often resulting in massive tables of numbers that could be difficult to interpret by researchers (as reviewed by Zelditch et al., 2012).

A second revolution then happened, the "morphometric revolution" (Rohlf & Marcus, 1993), representing the culmination of a series of advances that aimed to overcome the weaknesses of traditional morphometry. Landmark-based geometric morphometrics represented a new synthesis that used both robust mathematical processes along with new visualisation techniques that made the interpretation of results easier. The goal of geometric morphometrics was to determine a priori a series of homologous features across a dataset termed landmarks. These landmarks were consequently collected on a set of 2D figures or 3D objects. Every specimen thus resulted in a matrix $k \times d$, where k is the number of landmarks and d is the

number of dimensions for the coordinate system (Bookstein, 1991).

Homology is a core principle for geometric morphometrics where it was less important to traditional morphometics. Indeed, one of the key assumptions for geometric morphometrics is that corresponding landmarks across a dataset of specimens should meet the identity principle (Zelditch et al., 2012). This principle is crucial because the rationale behind geometric morphometrics is not anymore to analyse distances and ratios between specific points but rather the differential distribution of these points as a whole. Changes between specimens can be represented and quantified as functions that deform space between landmarks (Rüber & Adams, 2001), in a similar fashion as the transformation grids proposed by Thompson (1945) at the beginning of the 20^{th} century. Homologous points instrumentally inform the researcher about the implied alterations occurring between them and the decision on which landmarks to include should be made on the least possible number of points that sufficiently describes the target shape (A. Watanabe, 2018; Zelditch et al., 2012).

1.2.2 Procrustes Superimposition

It may help to visualise a given shape as a single point in a hyperdimensional space. The number of dimensions in this hyperdimensional space is equal to $k \times d$, and each shape within the same dataset occupies a precise position in this space (Klingenberg, 2020; Zelditch et al., 2012). Hence, to quantify the differences between two shapes one would eventually compute the distance between the two "shape vectors". However, biological structures described in vector form first require mathematical transformations to be applied to those vectors before distances between them would represent differences in shape. This is because the original descriptions of a biological structure using a vector of landmarks also includes information about the size, location and rotation of those structures. In order to filter out non-shape information one of the first steps is to compute the centroid for every observation (i.e. a centroid for every set of landmarks from a single specimen) (Bookstein, 1991). This

centroid will represent a point at which coordinates are equal to the average set of points calculated over all landmark coordinates. Next, the size of the centroid is calculated by taking the square root of the summed squared distances between every landmark and its corresponding centroid. This will result in a value proportional to inter-landmark distances (Bookstein, 1991). Then all coordinates in the vector are updated as the centroid sizes are rescaled to unit size, to control for effects of size on shape (Zelditch et al., 2012). Centroid coordinates are then subtracted from all landmarks to remove the effect of location (Bookstein et al., 1985; Zelditch et al., 2012). The vectors are consequently centered around the same Cartesian origin, removing the effect of location. With size and location removed, the vectors are aligned as new shape vectors in what is called "pre-shape" space. Coordinates at this point can be interpreted as positioned over an hypersphere of radius one (Klingenberg, 2020; Zelditch et al., 2012).

The new standardised set of coordinates are then rotated. Pairs of "pre-shape" specimen configurations are chosen with one considered the target configuration. The second configuration is then rotated on the centroid such that the sum of squared distances between homologous landmarks is minimised. The square root of the remaining inter-landmark offset left after the rotation is known also as partial Procrustes distance (Webster & Sheets, 2010) and the procedure just explained is known as partial Procrustes superimposition. The process can be extended to more than two landmark sets, by iteratively repeating the rotational stage to match the target shape, and then averaging all coordinates. These steps are repeated through an algorithm until the consensus between stages is reached. This method is called generalised Procrustes analysis (GPA) (as reviewed in Mitteroecker and Gunz, 2009).

Several superimposition methods have been developed (Dryden & Mardia, 1998; Goloboff & Catalano, 2016; Rohlf & Slice, 1990; Theobald & Wuttke, 2006) and each is based on distinct assumptions and diverse procedures. Nevertheless, the rationale behind these methods is that differences found after a superimposition should be attributed to shape dissimilarities (Zelditch et al., 2012).

Note that superimpositions put all shape configurations in a space of $kd-d-1-\frac{d(d-1)}{2}$ dimensions. The new space is known as shape space, and as for the pre-shape space, all shape vectors can be represented as single points lying on the surface of a hypersphere. The shape space is not Euclidean, thus distances computed across specimens need to be corrected via a subsequent projection on an Euclidean space tangent to the hypersphere in correspondence to the target configuration (Rohlf, 1999). The reprojection however inexorably deforms the true distances occurring between shapes, in the same way that flat maps deform continents displaced on a geodetic surface. One way to reduce this distortional effect would be to use the averaged mean shape from all configurations in a given analysis as a target shape (Bookstein, 1996). It is worth noting that an analysis of mammalian skull shape variation (Marcus et al., 2000) found that in comparison to true Procrustes distances, the reprojecton held only a limited amount of induced deformation and that most of the amount of distortion was found between greatly diverging forms that one should probably not be included in the same analysis (i.e the shape space is sensitive to outliers).

1.2.3 Shape Analysis

After shape information has been extracted with the Procrustes superimposition process (along with scale, location and rotation corrections), the entire set of shape vectors can be analysed with multivariate methods. Popular current methods include principal component analysis (PCA) (Sundberg, 1989) or canonical variates analysis (CVA) to visualise variation among specimens or groups, or ing's T^2 tests or multivariate analysis of variance (MANCOVA) to compare the effects of independent categorical variables on shape (Zelditch et al., 2012). Furthermore, multivariate regression models based on geometric morphometric variables (i.e. "shape regression") can be used to measure the magnitude of a response variable as it relates to shape (Klingenberg, 2016; Monteiro, 1999). Other approaches may measure the amount of integration within a given shape through the assignation of sets of

landmarks to separate modules (Goswami & Polly, 2010; Klingenberg & Marugán-Lobón, 2013). A module is a discrete anatomical unit within a body that may vary or evolve semi-independently in respect to other such units (Goswami & Polly, 2010). By assessing modularity (i.e. independence of anatomical features) and integration (i.e. correlation between anatomical features) it is possible to measure the degree of semi-independence between modules in a current population (Goswami & Finarelli, 2016), whereas in a macroevolutionary context shape modularity can assess whether separate modules evolved at different rates (Bardua et al., 2020; Bardua, Wilkinson, et al., 2019; Felice et al., 2020; A. Watanabe et al., 2019). Geometric morphometrics connects back to the idea of morphospace first introduced with the Galton Polyhedron model (Galton, 1894). As mentioned above, morphospace can be interpreted as a space in which different phenotypes can have unique sets of coordinates, in a similar manner as different shapes occupy different locations in the hypersphere of the shape space (Klingenberg, 2020). Even if shape space cannot be fully interpreted as a whole due to its high dimensionality, differences between phenotypes can be visualised with different tools like PCA (Sundberg, 1989), phylogenetic principal component analysis (PhyPCA,Revell, 2010), phylogenetically aligned component analysis (PACA, Collyer and Adams, 2021), and non-metric multi-dimensional scaling (NMDS, Kruskal, 1964). With these methods it is possible to identify the areas of a morphospace that some phenotypes occupy and quantify the degree of similarity (or dissimilarity) between different groups of organisms with respect to the structures of those organisms represented in the morphospace. The shape analysis that will be performed in this thesis will rely on the Procrustes method highlighted above and will take advantage of traditional PCA to visualise and estimate the occupation of morphospace through time.

Geometric morphometrics is the gold standard when it comes to shape analysis (Adams et al., 2013). With a robust mathematical framework and intuitive data visualisation it is possible to correlate shape data with factors that are external (or internal) to a target organism, allowing us to quantify the effects that the vertices

of constructional morphology have on shape determination.

1.3 Bayesian Statistics

1.3.1 Limits of frequentist statistics

Bayesian statistics is a broad field that encompasses a multitude of statistical approaches that are often presented in contraposition to the frequentist "school" (McElreath, 2020). There are many differences among these schools of thought but probably the most important is the interpretation of probability. Under a frequentist approach, probability can be interpreted as the absolute frequency (hence the name) or tendency for an event to occur; in a Bayesian framework, probability is the expectation that any given observer may have for that same event to occur (Cox, 1946). In a Bayesian sense, probability represents the quantification of a personal belief (de Finetti, 1970) that can vary through time and subsequently be updated on the basis of available evidence.

We often attribute two lines of thought to frequentist statistics: the concepts introduced by Fisher including the well known ideas of null-hypothesis, the significance level along with the p-value (Fisher, 1973) and the influences of Neyman and Pearsons with the ideas of alternative hypothesis and the notion of power (Neyman & Pearson, 1933; Neyman & Pearson, 1928). When a researcher is interested in reporting results in the frequentist framework it is usually required to present a null hypothesis (H_0) and at least one alternative hypothesis (H_1). Once the data has been collected a test can be conducted as to whether the data fits a given distribution. The results of the test should assess whether or not to reject H_0, and usually this is done with p-values. If a p-value falls outside a given confidence interval (traditionally set to 95%) then H_0 can be rejected "safely". Although this approach contributed immensely to the development of verifiability standards in a scientific context it has also raised some serious issues that may have undermined the accountability of the majority of published research (Ioannidis, 2005). First,

p-values are often interpreted as the probability that H_0 is false but this is not the case; in fact, *p*-values better describe the extremeness of the observed results in an H_0 scenario (Ioannidis, 2005). The probability that H_0 fails to be rejected should be seen as $P(H_0|\text{ data})$, i.e. the probability of H_0 given the observed data (the "posterior probability"). The posterior probability accounts for the "prevalence" of H_0 or how much we weigh this particular hypothesis in relation to any other possible hypothesis (Lesaffre & Lawson, 2012). However, probabilities are not usually assigned to different hypotheses in frequentist statistical analyses (Lesaffre & Lawson, 2012).

Another issue that needs to be addressed is the link between *p*-values and the publication bias. This bias influences researchers, reviewers and editors in the whole scientific community by reporting or ignoring part of the statistical tests that do not find significant differences between groups or end up deliberately contradicting the expected results (Dickersin et al., 1987; Ioannidis, 2005). This bias often leads to a magnification of the recovered results and seriously undermines the replicability of the majority of experiments (Ioannidis, 2005).

In addition, hypotheses are often rejected with respect to an entirely arbitrary 95% confidence span that has no intrinsic connection to the distribution of the observed data indicating this confidence interval is sufficiently "safe" (Morey et al., 2016). Note that in this thesis I will use the equally arbitrary value of 89% for the credible intervals reported here. Credible intervals are often used in Bayesian analyses and have a slightly different interpretation than confidence intervals. Whereas credible intervals make a specific statement that describes observer uncertainty around a parameter value, confidence intervals capture whether the value for a model parameter falls within a given range with a degree of confidence (Gelman et al., 2013; Gelman et al., 2008; Kruschke, 2015). Although frequentist statistical methods enabled invaluable progress in science it is still important to acknowledge the drawbacks that come along with these methods. Poor understanding of *p*-values, publication biases, and the lack of a solid foundation for hypothesis rejection represent some of the limits that may be bypassed with the help of Bayesian statistics.

1.3.2 Bayes theorem

In Bayesian statistics, hypotheses are neither rejected nor failed to be rejected, but rather each hypothesis or scenario (hereafter "model") is attributed a degree of uncertainty, and this uncertainty can be transferred to the parameters of the model (Kruschke, 2015). A parameter in Bayesian terms is an unmeasured variable to which a probability distribution can be assigned to help quantify our degree of belief (McElreath, 2020). Although Pierre-Simon Laplace developed most of the statistical theory behind modern "Bayesian statistics" (McGrayne, 2011), the school is named for Thomas Bayes, a theologian who derived the theorem of conditional probability (Eq.1.1) (Bayes, 1763).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

Consider two events (event A and event B) that may occur with probability P(A) and P(B), respectively. The probability that one event may occur given that the other has already occurred is the posterior probability, P(A|B) (often referred to as simply the posterior). Such probability is proportional to the likelihood of that same event, i.e. P(B|A) multiplied by the probability of that event alone P(A) (Eq. 1.1 numerator). If this quantity is divided by the marginal probability P(B) (often referred to as simply the marginal) then the posterior probability can be obtained. Bayes theorem can usefully be extended to observed data and proposed models. The posterior represents the probability of the model given the data (P(model|data), and as stated above, this likelihood is the probability of the data given that specific model(P(data|model)). P(A) is the probability of the model (i.e. P(model)) (Eq. 1.2.)

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{P(\text{data})} \tag{1.2}$$

In this case the marginal probability P(data) is the probability of observing the collected data under all available models (Eqs. 1.3 and 1.4)

$$P(\text{data}) = \sum^{n} P(\text{data}|\text{model}_n) \tag{1.3}$$

or

$$P(\text{data}) = P(\text{data}|\text{model}_1) + P(\text{data}|\text{model}_2) + ... + P(\text{data}|\text{model}_n) \tag{1.4}$$

With this premise, Bayesian statistical methodology enables multiple hypotheses to be evaluated simultaneously. By assigning different weights to each of these separate hypotheses (i.e. separate models), a study can forego hypothesis rejection and instead produce more nuanced outcomes that take into account the overall degree of uncertainty. The way in which the weight of the model or the prior probability (often referred to as simply the prior) can be determined is discussed in the next section.

1.3.3 Priors

Although there are no controversies around the validity of Bayes theorem, one of the major criticisms of the Bayesian workflow is the subjectiveness involved in selecting priors. One of the requirements of calculating a posterior probability under Bayes theorem (Eq. 1.2) is to first assign a probability to the model P(model) before collecting any data. As mentioned above in section 1.3.2, this type of probability is known as a prior, and there needs to be a prior associated with all parameters of a given model (McElreath, 2020). Priors can be interpreted as the belief of an observer and for this reason they are ultimately subjective (Gelman & Nolan, 2017). Priors embed this subjectiveness into equation 1.2, meaning that the posterior probability is influenced by the observer's subjectiveness. An obvious caveat is that the final results of a Bayesian analysis are strongly influenced by decisions made by the researcher. One way of limiting issues associated with subjectivity is to carefully select weakly-informative priors in a manner that the validity of the overall analysis is not undermined, and indeed, there is substantial literature centered on prior selection (Gelman et al., 2013; Gelman et al., 2008; Kruschke, 2015). In fact, by

placing more attention on prior "tuning" the researcher is forced to reason more on the structure of the model itself, to ensure that a prior can be adequately justified to reflect the observer's belief. I personally argue that the robustness of the results depends more on the goodness of the data collected rather than prior specifications. Moreover, frequentist approaches are not entirely devoid from subjectiveness given that many aspects of every statistical analysis require some amount of subjective decisions (how to collect data and which data to collect, outliers management, test choice, p-value cut-off, etc.).

1.3.4 Melanistic penguins and Bayesian updating

Underpinning Bayesian analysis is the concept of updating probabilities after collecting new evidence (Gelman & Nolan, 2017). To better explain the concept of updating probabilities, a small aside will be taken to present an example modified from Gelman and Nolan (2017) to suit the subject of this thesis. Melanism is a condition in which an animal may tend to have an overproduction of melanin pigments in the integument, and therefore often appear darker if not completely black (Grouw, 2017). Let us assume that we are interested in determining the incidence of this condition in a population of wild penguins. We embark on a boat and start to capture penguins in a given stretch of sea. Given that melanistic penguins are dark also on the ventral section of the body they need to be captured and examined to be assessed on their condition, meaning that the assessment of melanism in the population is updated at each capture. We will assume that penguin assessment can be modeled as a Bernoulli distribution with an outcome of 1 that has probability p_1 and an outcome of 0 that has probability 1-p_1 The better way to figure this distribution is through the example of a flipping coin for which we do not know the exact probability of head and tail coming out from a draw. We also are going to assume that each draw is independent and that the same individual will not be counted more than once. In the following example p_1 will represent the proportion of melanistic penguins in the whole population and will effectively represent the parameter of

this model. Our complete ignorance about the proportion of melanistic penguins in the population will be converted in an uninformative flat prior, in this case a uniform distribution between 0 and 1 (Fig.1.4.1, dashed line). At the first catch the penguin is not melanistic so the numbers can be inserted in the Bayes formula for this scenario (Eq. 1.2) as a Bernoulli distribution with size 1 and an outcome of 0. However, given that p is a continuous value between 0 and 1, the outcome of the Bayes formula is a probability density function (Fig.1.4.1 solid line). The function is skewed toward low values of p but allows for uncertainty at the higher end (Fig.1.4.1 solid line). At the second catch the penguin is again not melanistic. The Bayes formula can be updated but this time rather than using the uniform prior from our initial draw, we instead use the probability density function from the previous step. The iterative calculation of the Bayes formula using an updated input term is an essential element of Bayesian statistics in terms of updating our belief after collecting new evidence, meaning in practice the posterior of an initial analysis may become the prior for a subsequent one (Kruschke, 2015).

The experiment draws ten more penguins from the population which includes three melanistic penguins from the 12 sampling events (Fig.1.4.3 to 1.4.12). The resulting posterior probability density function after these 12 sampling events for the parameter p_1 is centered on 0.25 and has 89% credible intervals spanning from 0.11 and 0.48. One of the aspects to note is how the probability density function changes through samples, with initial samples impacting the overall shape of the function the most whereas the last few draws have a relatively minor impact on function shape. Note how the high probability for extremely low values of p_1 drops as soon as one melanistic penguin is caught with the curve assuming more of a bell shape (Fig. 1.4.4). This highlights that even with a highly skewed prior (Fig. 1.4.4 dashed line) the posterior can change dramatically when evidence derived from new data is presented. Note also how each sample has an impact over the position of the bell with each melanistic penguin "pushing" the peak toward higher values of p_1. Moreover, with increasing sample size, the diameter of the bell tends to shrink (Fig.

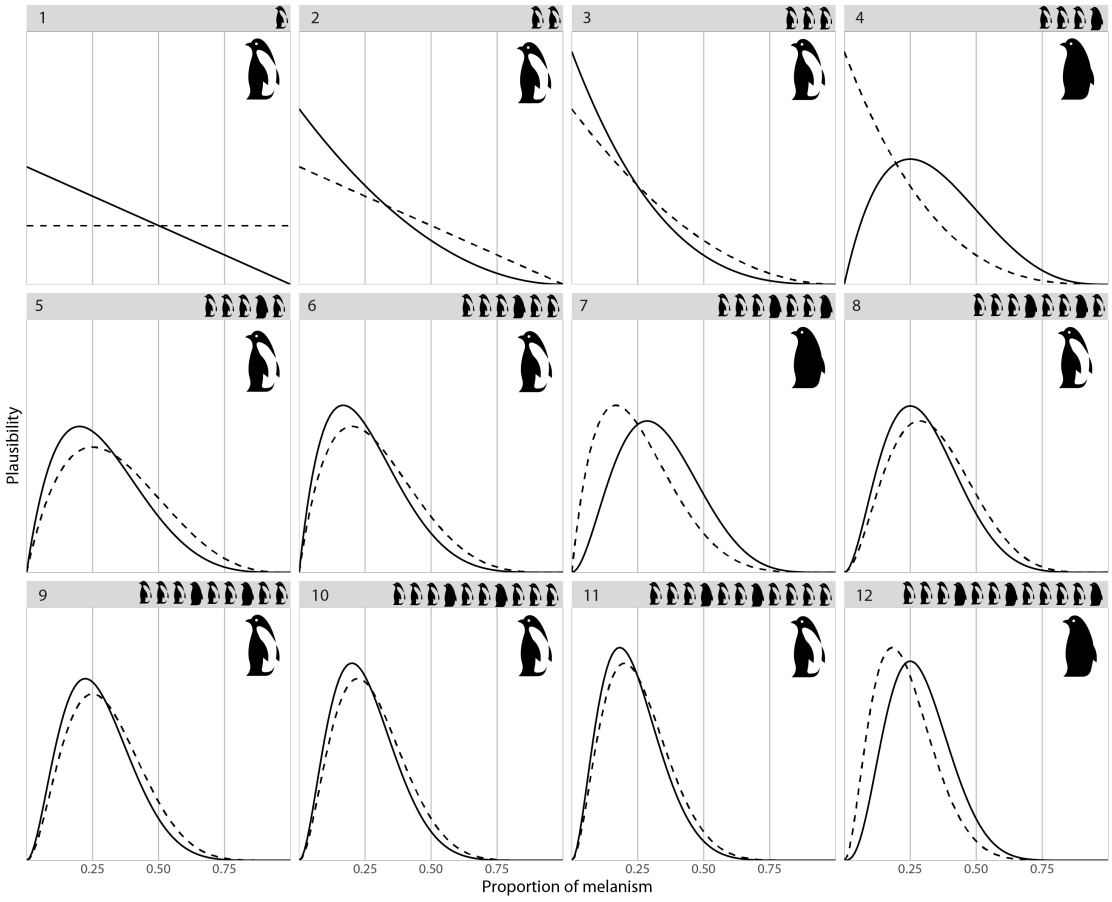


Figure 1.4: Representation of a Bayesian workflow. Each plot represents the posterior density function (solid line) and the prior density function (dashed line) for the proportion of melanistic penguins in the sampled population. The experiment is run for 12 samples, silhouettes in each plot denote the type of penguin sampled at that draw and the smaller silhouettes describe the structure of the whole sample. Note that the prior in each cell is deriving from the posterior of the previous cell except for the first cell where the prior is an uniform distribution between 0 and 1.

1.4.4 to 1.4.12), implying a decrease in overall parameter uncertainty along with the increasing sample size (McElreath, 2020). At this point the researcher may decide that the amount of collected data is enough and then conclude the experiment or continue sampling from the population until the estimated distribution of p_1 reaches a sufficiently-reduced uncertainty range. Note also that the outcome of this analysis can be performed in a single run once that the totality of the data has been collected and the results would not change so long that the initial prior distribution remain the same (McElreath, 2020). Now that the structure of Bayesian data analysis has been clarified we can consider to point out how the posterior probability and the marginal probability can be estimated correctly.

1.3.5 Posterior estimation

Methods of data analysis based on Bayes theorem present some challenges that are not encountered with analyses using "frequentist" methods. As outlined in previous sections, parameters may often assume values along a probability density function (Eq. 1.5; compare with Eq. 1.2)).

$$f(\theta|\text{data}) = \frac{f(\text{data}|\theta)f(\theta)}{\int f(\text{data}|\theta)f(\theta)d\theta} \tag{1.5}$$

If we are trying to compute the posterior for a model with a single parameter θ (e.g. the example of the melanistic penguin frequency stated above) then the denominator takes the form of an integral (note that the numerator of the equation remains mostly unchanged from equation 1.2 with point probabilities simply substituted with their respectives probability density functions). However, if the complexity of the problem increases by introducing a second parameter ϕ (e.g. modelling more than a variable, like trying to also estimate the average body mass of the penguin population in addition to melanism frequency) then equation 1.5 becomes equation 1.6.

$$f(\theta, \phi|\text{data}) = \frac{f(\text{data}|\theta, \phi)f(\theta, \phi)}{\int_\theta \int_\phi f(\text{data}|\theta)f(\theta)d\theta d\phi} \tag{1.6}$$

Now the integral is folded two times, one for each of the model parameters, meaning that it marginalises over all possible value combinations of θ and ρ. These types of integrals become computationally intractable as the number of parameters grow. Multilevel models like phylogenetic inference that often deal with hundreds or more parameters are not possible (Nalborczyk et al., 2019; Wiley & Lieberman, 2011). One of the most common ways that researchers deal with this type of challenge is with the use of a Markov Chain Monte Carlo (MCMC) algorithm (Metropolis et al., 1953). Markov Chain Monte Carlo algorithms sample directly from the posterior distribution by comparing between two possible states (Eq. 1.7) rather than focusing on computing the denominator as in equation 1.5 and equation 1.6.

$$\frac{f(\theta^*, \phi^*|\text{data})}{f(\theta, \phi|\text{data})} = \frac{\frac{f(\text{data}|\theta^*,\phi^*)f(\theta^*,\phi^*)}{\int_\theta \int_\phi f(\text{data}|\theta)f(\theta)d\theta d\phi}}{\frac{f(\text{data}|\theta,\phi)f(\theta,\phi)}{\int_\theta \int_\phi f(\text{data}|\theta)f(\theta)d\theta d\phi}} = \frac{f(\text{data}|\theta^*, \phi^*)f(\theta^*, \phi^*)}{f(\text{data}|\theta, \phi)f(\theta, \phi)} \tag{1.7}$$

Once the posteriors between two different sets of parameter values are calculated (Eq. 1.7 θ, ϕ versus θ^*, ϕ^*) the denominators with integrals are cancelled from the equation (Eq. 1.7 second and third blocks), resulting in a ratio R equivalent to the likelihood odds times the prior odds (Metropolis et al., 1953). The MCMC algorithm needs values for all parameters of the model for the initial iterations (here in this example θ and ϕ) and these values can either be supplied or randomly generated. A new set of parameter values is subsequently proposed (in this example θ^* and ϕ^*) and the ratio R is computed. If the proposed values produce an increase in the posterior (i.e. $f(\text{data}|\theta^*, \phi^*)f(\theta^*, \phi^*) > f(\text{data}|\theta, \phi)f(\theta, \phi)$ they will constitute the new state of the Markov Chain which is then stored into memory ("sampled") (Metropolis et al., 1953). If $f(\text{data}|\theta^*, \phi^*)f(\theta^*, \phi^*) < f(\text{data}|\theta, \phi)f(\theta, \phi)$ then the new values are accepted with probability R, meaning that parameter values that bring a slight decrease in the posterior probability are accepted more often than values that greatly decrease the posterior (Metropolis et al., 1953). If these steps are repeated for enough samples the MCMC should reach the global optimum for all model parameters. Markov Chain Monte Carlo methods have demonstrated their robusticity in statistical inference (Brooks et al., 2011).

A metaphor commonly used to describe the behavior of a MCMC is that of a robot climbing over a landscape with many hills in search of the tallest peak (e.g. Lewis, 2001b). The robot is completely blind and can only measure the height of the land surface where it is currently standing, and the height of the land surface one step away. If the neighboring spot is higher then the robot immediately jumps over to that spot. If the neighbouring point is lower then the robot draws a random number between 0 and 1; if this number is greater than the ratio between the current height and the proposed height then the robot jumps; otherwise the proposal is refused. This type of behavior ensures that the robot doesn't limit itself to exploring just the "peaks" in this landscape to avoid the risk of becoming stuck in a local optimum.

By occasionally walking downhill the robot can do a better job of exploring the landscape and searching for the tallest possible peak.

The Metropolis MCMC algorithm (Metropolis et al., 1953) described above is amongst the first to be developed, and has been followed by numerous improvements to sampling and overall. For example, the Metropolis-Hastings algorithm allows proposals (i.e. the set of parameters values proposed at each iteration of the algorithm) that are asymmetric (Hastings, 1970), the Gibbs sampler introduced adaptive proposals (Geman & Geman, 1984; Plummer, 2003), and the Hamiltonian or Hybrid MCMC simulates Hamiltonian physics to sample from the posterior (Carpenter et al., 2017; Neal, 1994). This thesis will rely on the Metropolis-Hastings algorithm for phylogenetic inference and evolutionary rates estimates (Chapter 2 and Chapter 4; Höhna et al., 2016; Rannala and Yang, 2007) and Hamiltonian MCMC for the size estimations models (Chapter 3; Carpenter et al., 2017). Although MCMC allows the researcher to overcome what would be classified as an intractable mathematical problem (the marginal estimation; (Brooks et al., 2011)) these types of algorithms are still more computationally expensive compared to traditional frequentists approaches (Quiroz et al., 2019). The recent diffusion of Bayesian statistics in the scientific community is thus likely driven by increasing availability of computational resources given that the foundational math was developed before the rise of widespread computing (Cox, 1946; de Finetti, 1970).

MCMC length is also an important factor to understand if the chain reached the global optimum for all given parameters. There are three important properties that suggest that an MCMC has identified a global optimum for a given parameter: Stationarity, convergence and sufficient mixing (Gelman et al., 2013; Kruschke, 2015; McElreath, 2020). Stationarity is the relative stability of parameters of a MCMC throughout most of its length. Stationarity can be checked using traceplots (Fig. 1.5). The traceplot for a particular parameter records the parameter value for all samples from a MCMC. When interpreting a traceplot it is often desirable to have a "noisy" chain stable around a value (Fig.1.5A) rather than a fluctuating one (Fig.

1.5B) because in the former it likely means that the MCMC is sampling from an optimum point whereas in the latter the MCMC is still exploring the parameter space (Brooks et al., 2011; McElreath, 2020).

If two independent chains converge to the same or nearly identical set of parameter estimates (i.e. "convergence"; McElreath, 2020) it may indicate that the global optimum was reached. Given that often a MCMC starts far from the global optimum value then the initial phases are not stationary, because the sampling is still coming from the exploratory phase. To solve the issue these samples are usually removed after the analysis and the portion of excluded samples become known as the "burn-in phase" (e.g. Brooks et al., 2011).

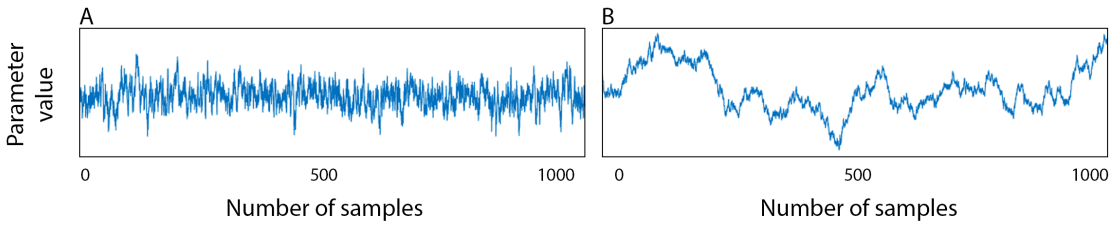


Figure 1.5: Traceplots from an hypothetical run of two different MCMCs. The parameter value is presented on the vertical axis whereas the number of samples (i.e. "chain length") is presented on the horizontal axis. A) Fluctuation of a functional MCMC with stationary and random oscillation around a parameter value. B) A more incoherent oscillation through time implying that the chain did not reach an optimum value.

Mixing describes the scenario where variation is consistently observed in parameter values across the chain length (i.e. that the MCMC continues to explore the parameter space). In practice, mixing is demonstrated by a MCMC that doesn't return the same parameter value sequentially for long periods of time (i.e. doesn't produce a traceplot with one or few horizontal lines). Visual inspection of traceplots is usually the first step in determining if mixing occurred but there are many statistical tools that assess if a MCMC sufficiently explored the posterior. For example, one of the tools used to explore mixing is the measure of the effective number of samples for an MCMC (also called effective sample size or ESS) (Carpenter et al., 2017; Drummond & Rambaut, 2007). This effective sample size metric helps to quantify the amount of autocorrelation between samples (McElreath, 2020) and can

be calculated by dividing the total number of samples over the design effect (i.e. the ratio between the variances of two estimators for any parameter that needs to be calculated; Kish, 1965). Recall that samples in an MCMC are not independent from one another given that the probability to accept any proposed value at any given time is influenced by the state at which the MCMC resides at that stage. ESS is thus an estimation of the samples that are independent in the chain, usually the higher the value the better it is for the chain itself (Drummond & Rambaut, 2007; McElreath, 2020). Another tool used to assess chain performance is the $\hat{R}$ measure, which is also known as the Gelman Rubin convergence test (Gelman & Rubin, 1992). The $\hat{R}$ measure compares if separate chains reached a sufficiently similar value. Stationarity, convergence and mixing will all be evaluated in this thesis as Bayesian methods of analysis are used to describe multiple aspects of the evolutionary history of penguins.

1.4 Penguins

Penguins (Aves: Sphenisciformes, Spheniscidae) have many remarkable aspects that together make them an ideal candidate for this thesis. Penguins are a clade of flightless birds that are well-adapted for life in water (Davis & Darby, 2012) as one may easily interpret by looking at their osteological traits.

Underwater flight for penguins is driven by flipper-like wings that generate thrust during both the downstroke and the upstroke (Clark & Bemis, 1979; Lovvorn et al., 2001). Such peculiarity is achieved through stiffened wing joints (Ksepka & Ando, 2011) and a hypertrophied musculus (hereafter m.) supracoracoideus that connects the humerus to the keel (Schreiweis, 1982). Along with alcids (Charadriiformes, Alcidae), petrels (Procellariiformes, Procellariidae) and dippers (Passeriformes, Cinclidae), penguins are considered wing-propelled divers, contrary to the majority of aquatic birds that rely mainly on their legs and feet for underwater motion (foot-propelled divers; e.g. Podicipediformes and Gaviiformes) (Clifton et al., 2018; Hinić-Frlog & Motani, 2010). Underwater flight represents an exaptation for

the bird wing (i.e. an adaptation towards a new function) and this markedly different functional role is an excellent candidate to investigate the adaptational vertex of the constructional morphology triangle.

The penguin hindlimb is conversely less-specialised for swimming and doesn't function in propelling the bird through water but instead acts like a rudder (Ksepka & Ando, 2011). By having a limited role underwater and an ambulatory function on land it is difficult to assume a strong selective pressure on the hindlimb as a locomotory module when compared with the wing (Chapter 4). Although the walking gait of a penguin is more energetically expensive when compared to other birds (Pinshow et al., 1977), it is important to not be trapped in the pitfall of associating it with evolutive inefficiency. It has been proven that the waddling pace contributes to recover mechanical energy during strides (Griffin & Kram, 2000), and that the structural constraint is given by the shortness of the leg. One of the principal functions of the hindlimb is to support the mass of the penguin on land when the animal no longer benefits from the buoyancy of water. Hence, the hindlimb is an interesting region of the penguin body to study the effect of physical constraint in controlling skeleton shape (Chapter 3).

Lastly, penguins confirm their status of outliers among birds thanks to the richness of their paleontological record (Chapter 2). Fossil species outnumber living species by at least three times (Ksepka & Ando, 2011), an amount that keeps growing even at present day (Blokland et al., 2019; T. L. Cole et al., 2019; Giovanardi et al., 2021; Mayr et al., 2019). Even though one should acknowledge that the meaning of species in a paleontological context has a different nuance to the biological species concept (Allmon & Yacobucci, 2016; Tschopp et al., 2021), it is still important to understand the reasons behind this unusual pattern. Simply stated, marine ecosystems show higher fossilisation rates when compared to terrestrial environments (Levin & King, 2016). Larger and more dense bones are also positively biased in the fossil record and penguins show osteosclerosis, a reduction of the medullary cavity inside their bones (Ksepka et al., 2015). The penguin fossil record spans from almost the

beginning of the Paleogene period (65.5 Ma - 23.03 Ma) with the oldest fossil penguins around 62 million years old (Blokland et al., 2019; Mayr, De Pietri, Love, Mannering, & Scofield, 2017; Slack et al., 2006). Penguins are known from most of the Cenozoic (65.5 Ma to present) and even Holocene subfossil remains younger than 100 thousand years reveal extinct species of penguin (T. L. Cole et al., 2019). This remarkable sampling allows paleontologists to identify major macroevolutionary shifts that the clade experienced over the last 60 million years. A pattern that emerges is that the crown clade originated relatively recently (roughly 15 Ma), with most fossil penguin species instead identified as stem taxa (Gavryushkina et al., 2017). The crown clade is identified as all of the living members of a group, the most recent common ancestor that those living members share, and then all of the descendants of that most recent common ancestor (both living and extinct). A stem member of a clade is any fossil that is more closely related to one particular crown clade than to any other but which itself does not have living descendants. By walking through time across the history of penguins we can see a drastic transformation in the bauplan (e.g. the body plan). This transformation can be described as series of modular steps that allow us to infer the rate at which the phenotypic change occurred (Chapter 4).

1.5 Aims

Of all wing-propelled divers only penguins exhibit a detailed record through time, enabling detailed analysis of the secondary adaptations to water that have arisen in birds, and encouraging comparisons with other major clades of marine tetrapods. The use of geometric morphometrics in studies of penguins has mostly been limited to inferring the ecology or taxonomic identification of extinct penguins (Acosta Hospitaleche & Tambussi, 2006; Chávez-Hoffmeister, 2020; Jadwiszczak, 2020; Jadwiszczak & Mörs, 2019), whereas Bayesian analyses in penguin research have only been used for phylogenetic inferences (Blokland et al., 2019; Gavryushkina et al., 2017; Thomas et al., 2020). This thesis will use geometric morphometrics and

Bayesian methods to study the evolution of selected phenotypes in fossil and modern species of penguin.

The broad aim of this thesis is to study shape changes in penguins as they became more efficient wing-propelled divers, beginning shortly after the macroevolutionary transition from their aerially-volant ancestors. Addressing the topic from the perspective of the constructional morphology concept described by Seilacher (1970), a core goal is to provide the first quantification of evolutionary rates for key morphological changes (Simpson, 1944). Using extinct taxa as interpolation points it is possible to effectively determine how these rates changed during time and how these changes were distributed across the penguin body. More importantly, establishing these evolutionary rates of morphological change for penguins represents a new way of studying major morphological shifts, which is a methodology that can be extended to other groups of extinct or extant taxa.

The three specific aims of this thesis are linked to each apex in the constructional morphology framework:

- **History**: Use parsimony- and Bayesian-based methods to infer the evolutionary history of penguins. Describe a new species of penguin to contribute to this understanding of evolutionary history (Chapter 2).

- **Structure**: Demonstrate that femur volume is mostly the consequence of structural constraint but is also influenced by evolutionary history and ecology (Chapter 3).

- **Adaptation**: Reveal how different regions of the penguin body have been subject to different rates of evolution through time (Chapter 4).

Figure 1.6: Reconstruction of a group of the early Paleocene penguin *Kaiika maxwelli* preying on a school of fish.

Chapter 2

Historical apex

2.1 Introduction

Within the aptive triangle metaphor of Seilacher (1970), the historical apex would ideally represent the impact that phylogeny has over any given aspect of a biological entity (Gould, 2002; Seilacher & Gishlick, 2019). One way that this apex can be interpreted is the tendency for two closely related species to have more in common than two species chosen at random (Felsenstein, 1985). Thus in order to grasp the impact that shared evolutionary history may have over any given trait it is crucial to be able to estimate accurate phylogenies prior to any comparative assessment. Since the seminal work of Hennig et al. (1966) that aimed to bring together evolutionary biology and taxonomy, huge efforts have been made to establish a solid foundational method for estimating phylogenies (Felsenstein, 2004; Heath et al., 2014; Rannala & Yang, 1996; Swofford et al., 1996; Wiley & Lieberman, 2011). Today, the software packages that allow us to use these techniques are common, easy to access (Bouckaert et al., 2014; Felsenstein, 1993; Goloboff & Catalano, 2016; Höhna et al., 2016), and widely used across the life sciences and even further (Maurits et al., 2017). The starting point to infer a phylogeny is often a morphological matrix or molecular sequence alignment, where each row represents a taxon or specimen and each column is a homologous character or site within the sequence. Data in each column usually has a limited range of possible variables (i.e. states within each character), and a

frequent goal of phylogenetic inference is to estimate an evolutionary bifurcating tree connecting all taxa that is the most likely given the observed pattern of shared variables. Taking advantage of the extensive methodological background available (e.g. Goloboff and Catalano, 2016; Heath et al., 2014), this chapter aims to define the context in which penguins evolved, in order to use the resulting framework as a scaffold for analyses in the next two chapters.

2.1.1 Publication of *Kairuku waewaeroa*

The main body of work for the following chapter is presented in 'A giant Oligocene fossil penguin from the North Island of New Zealand' published on 17/09/2021 in Journal of Vertebrate Paleontology. I collected the data from the fossil penguin and from other comparative specimens, and I conducted the majority of the data analysis and wrote the initial draft manuscript. Daniel Thomas and Daniel Ksepka contributed to manuscript writing. The published paper is a formal part of this chapter, but is attached here in Appendix A because of the convention to not name a new species in a thesis chapter (ICZN, 1999), and because the paper includes writing contributions from Thomas and Ksepka. Here I provide a brief summary of the paper and encourage the reader to refer to the manuscript in the appendix for the full work. This paper describes a novel penguin species, *Kairuku waewaeroa*, as a member of a clade that was widely distributed across Zealandia within the Oligocene (33.9 to 23.03 Ma, Cohen et al., 2013). The taxon described in the paper was several million years older than previously described *Kairuku* species, and a parsimony-based phylogenetic analysis returned *Kairuku waewaeroa* at the base of the *Kairuku* clade. Given the completeness of the specimen the description of the taxon is also important in explaining how the bauplan (i.e. the overall body plan) of fossil penguins changed through time. *Kairuku waewaeroa* shows overall body proportions closer to living penguins suggesting that the distinctive bauplan exhibited by *Kairuku grebneffi* and *Kairuku waitaki* (Ksepka et al., 2012) may have been exclusive to these two species rather than being shared across all giant fossil

penguins, and thus that diversity among fossil penguins was greater than previously expected.

2.1.2 Penguin evolutionary history explored with Bayesian methods

After the submission of Giovanardi et al. (2021) I continued to develop the phylogenetic framework of this thesis, to keep the results of Chapter 3 and Chapter 4 up to date with the current state of knowledge. Recent advances in this field include additional descriptions of extinct penguins (Acosta Hospitaleche et al., 2019; Acosta Hospitaleche et al., 2021; Blokland et al., 2019; Chávez-Hoffmeister, 2020; Jadwiszczak et al., 2021; Mayr et al., 2021; Mayr et al., 2019; Richards, 2019; Thomas et al., 2020) that were often accompanied by updated morphological matrices. For example, Blokland et al. (2019) combined data from multiple sources (Bertelli & Giannini, 2005; Chávez Hoffmeister et al., 2014; Chávez-Hoffmeister, 2014; Degrange et al., 2018; Ksepka et al., 2012) to generate a matrix with 284 morphological characters scored for 89 taxa. Matrices published by Thomas et al. (2020) and Giovanardi et al. (2021) added 32 new characters to the matrix from Degrange et al. (2018). As an effort to fulfill to the principle of "total-evidence" driven inference (Ronquist et al., 2012; Wiley & Lieberman, 2011), the current chapter sought to merge the matrix of Giovanardi et al. (2021) with that of Blokland et al. (2019). The parsimony phylogenetic analysis in Giovanardi et al. (2021) was consequently redeveloped, and a second phylogenetic analysis using Bayesian fossilised birth death tree (FBDT) models was also developed (Heath et al., 2014). This family of models incorporate information about specimen geological ages on branch tips along with speciation and extinction rates (Gavryushkina et al., 2014), and relax the assumption of not sampling from the tree's branches, allowing fossil taxa to be recovered as potential ancestors of any given lineage (Cau, 2017).

Parametric phylogenetic inferences (maximum likelihood and bayesian) were once restricted to molecular data but have now became more common in paleontol-

ogy after discrete morphological dataset were able to be used (Geisler et al., 2011; Lee & Worthy, 2012; Lewis, 2001a; Mongiardino Koch & Thompson, 2021; Šmíd & Tolley, 2019). Although there is still debate over the influence of missing data and the effect of parsimony optimised datasets in likelihood model-based inferences (Goloboff et al., 2018; O'Reilly et al., 2016; Wiens & Morrill, 2011), there is increasing evidence showing the strengths of Bayesian phylogenetic analyses applied to paleontology (Matzke & Irmis, 2018; Vernygora et al., 2020; A. M. Wright & Hillis, 2014). Thus rather than focusing the debate on which phylogenetic inference method is more accurate it would be worthwhile to perform both non-parametric (parsimony) and parametric (likelihood or Bayesian) analysis and then report both results. Reflecting on similarities and differences in the phylogenetic topologies resulting from the parsimony and Bayesian analyses provides more insight into the evolutionary implications of each method, as well as the importance of assumptions unique to each method. For these reasons thus this chapter includes both a parsimony-based and a FBDT model phylogenetic analyses, performed on an updated matrix of penguin diversity. Note that penguin datasets have previously been studied using both parsimony and FBDT models (Degrange et al., 2018; Gavryushkina et al., 2017; Thomas et al., 2020) showing that thanks to an extensive taxon sampling lasting for a window of more than 60 million years (Slack et al., 2006), penguins represent a good candidate for dated phylogenetic inference.

2.2 Materials and methods

2.2.1 Updated matrix

The matrix used in the description of *Kairuku waewaeroa* (Giovanardi et al., 2021) includes 282 characters and 76 taxa and derived from the matrix in Degrange et al. (2018) to which were added 31 characters from Thomas et al. (2020). The Giovanardi et al. (2021) matrix was then updated to include 30 additional characters for the humerus and tarsometatarsus from Chávez-Hoffmeister (2014) and

Chávez Hoffmeister et al. (2014), which were scored according to phylogenetic analyses published in Blokland et al. (2019). 15 taxa were added to the Giovanardi et al. (2021) matrix: *Arthrodytes andrewsi*, *Crossvallia unienwilla*, *Crossvallia waiparensis*, *Eudyptes atatu*, *Eudyptes calauina*, *Eudyptes warhamii*, *Kaiika maxwelli*, *Kumimanu biceae*, *Kupoupou stilwelli*, *Megadyptes antipodes waitaha*, *Megadyptes antipodes richdalei*, *Paraptenodytes brodkorbi*, *Paraptenodytes robustus*, *Pseudaptenodytes macraei*, *Platydyptes amiesi*. The 15 taxa added to the updated matrix were scored for all characters used in the description of *Kairuku waewaeroa*. Characters were scored as missing entries for all instances where it was not possible to correctly infer the states of characters (e.g. hindlimb characters for *Kaiika maxwelli*). Following a recent description of a skull referred to *Anthropornis grandis* (Acosta Hospitaleche et al., 2019), this taxon was recoded with cranial characters in the updated matrix. Character scoring was performed in `Mesquite` (Maddison and Maddison, 2019 v. 3.7). The update resulted in a morphological discrete matrix of 312 characters scored for 92 taxa (See supplementary file `Chapter_2_Morphological_Matrix.nex`). The discrete character matrix was then combined with the mitogenomic alignments currently available for all extant penguins published by T. L. Cole et al. (2019). The phylogenetic assessment included a parsimony-based analysis and a Bayesian FBDT model inference. Prior to the Bayesian analysis the matrix was further modified with the removal of all procellariiform taxa due to the assumption that only fossils that are part of the ingroup should represent the least derived taxa in the phylogeny (Gavryushkina et al., 2017). The deletion of the Procellariiform outgroup resulted in an inflation of invariant characters that were informative to define clades outside of Sphenisciformes. Invariant characters can be quickly removed from a matrix in `Mesquite` (Maddison & Maddison, 2019) and resulted with the loss of 47 characters (See supplementary file `Chapter_2_Reduced_Matrix.nex`). Fossil dates included in the model were taken from the most updated information currently available (Blokland et al., 2019; Gavryushkina et al., 2017; Thomas et al., 2020).

2.2.2 Phylogenetic software

The parsimony-based analysis was performed in `TNT` (Goloboff and Catalano, 2016, v 1.5) on the matrix (see supplementary file `Chapter_2_Parsimony.tnt`) with equal weighting and non-molecular multistate characters treated as ordered if at least one of the states could be interpreted as transitional between the remainder. After the preliminary most parsimonious trees (MPT) were found with the fast search algorithm, a subsequent search was run until memory overflow (9999 trees) with the TNT command `bbreak`. This subsequent search collapsed nodes with no support (Tschopp et al., 2015). To gain further insights from the polytomies of the consensus topology, an "iterative positional congruence reduced" (iterPCR) was conducted to identify the taxa whose phylogenetic position varied the most causing a loss in node resolution (Pol & Escapa, 2009). Positional congruence is a way to assess taxa instability and can be computed by first splitting two phylogenetic trees into all possible triplets of taxa (quadruplets in case of unrooted trees). Then to assess the positional congruence of a target taxon all triplets that include that taxon and share the same topology between the two sets are counted and then divided by the total number of triplets. This results in a proportion spanning from 0 to 1 with taxa that exhibit more unstable behavior having lower positional congruence values. Given that for a large phylogeny this task can be extremely demanding in terms of computational time, the key principle of iterPCR is calculating positional congruence at the polytomies and the deriving branches of a consensus tree. The iterPCR was performed within the `TNT` software package (Goloboff & Catalano, 2016; Pol & Escapa, 2009). The FBDT model inference was performed in `RevBayes` (Höhna et al., 2016, v.1.1.1), including three partitions of information into the analysis: a molecular partition (i.e. the mitogenomic alignment), a morphological partition (i.e the morphological matrix), and the stratigraphic occurrence ranges of the taxa. During the Markov Chain Monte Carlo (MCMC) sampling, the likelihood of the combined partitions is evaluated for a given phylogenetic tree topology and a given set of parameters.

2.2.3 MCMC - General and FBDT framework

The MCMC model is described below (Fig. 2.1) and follows the visualisation from Höhna et al. (2014) where each parameter is represented as a node. Each node has inputs given by its parent nodes and produces outputs directed to a child node. Nodes are subdivided into four types (constant, stochastic, deterministic and clamped) and more nodes can be clustered within a loop plate to iterate more parameters during each round of the MCMC (Fig. 2.1). During the MCMC sampling run, the value of a node is updated at each iteration following the distribution given for the priors, and these values are used to compute the likelihood (Höhna et al., 2016). Constant nodes are equivalent to a constant value throughout all the analysis and are usually parent nodes of prior distributions of the analysis. Stochastic nodes instead are drawn at each MCMC iteration from a distribution given by its parent. Deterministic nodes are nodes for which the value is entirely defined by parent nodes (e.g. a sum or a ratio between two nodes). Clamped nodes are the nodes that represent the observed data (i.e taxa dates, the morphological matrix and the mitogenomic alignment). The likelihood of the parameters is computed at these nodes in the graph. The loop plate represents a `for` loop in `RevBayes` and is usually necessary when a given value needs to be computed for several entries during a single MCMC round (i.e. assessing when the proposed tree topology is consistent with the dates availables from the stratigraphic dataset).

Once specified, the model structure of the MCMC was set up to run for 1×10^8 samples. The analysis was performed on a multi-cluster machine at Massey University and took roughly 40 days of computation.

The FBDT node is a stochastic node that needs five parameters as inputs (Höhna et al., 2016): a deterministic diversification rate λ, a deterministic turnover rate μ, a stochastic fossilisation rate Ψ (Eq. 2.5), a constant sampling probability ρ, and the stochastic origin time $\boldsymbol{O}$. λ is given by the difference between the stochastic speciation rate d and the stochastic extinction rate r (Eq. 2.1) and μ is the ratio between extinction rate and speciation rate (Eq. 2.2). Following Thomas et al.

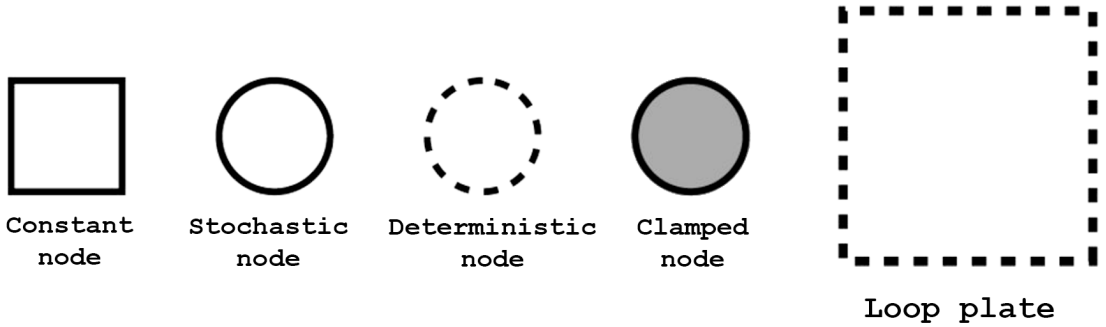


Figure 2.1: Graphical legend for the Markov Chain Monte Carlo model, following the scheme presented in Höhna et al. (2014). Constant nodes are represented by squares with solid lines, interpreted as prior information. Stochastic nodes are represented by solid line circles, whereas deterministic nodes have a dashed outline. Circles filled with grey background represent clamped nodes where likelihood is computed. Loop plates are represented as larger squares that contain multiple nodes.

(2020) priors for d, r and Ψ were given as an exponential distribution with mean 1.0 (Eqs. 2.3 to 2.5). Note that the origin was specified as a uniform distribution between 130 and 65 Million years ago (Eq. 2.6), and that ρ (Eq. 2.7) was set fixed as 1 given that all living taxa have been sampled (Thomas et al., 2020).

$$\lambda = d - r \tag{2.1}$$

$$\mu = \frac{r}{d} \tag{2.2}$$

$$d = \text{Speciation rate} \sim \text{Exponential}(1.0) \tag{2.3}$$

$$r = \text{Extinction rate} \sim \text{Exponential}(1.0) \tag{2.4}$$

$$\Psi \sim \text{Exponential}(1.0) \tag{2.5}$$

$$\rho = 1.0 \tag{2.6}$$

$$\boldsymbol{O} \sim \text{Uniform}(130\ \text{Ma} - 65\ \text{Ma}) \tag{2.7}$$

The MCMC process is represented as a directed acyclic graph (DAG) representation of the FBDT node as presented in Figure 2.2, following the scheme presented in Höhna et al. (2014). Consistency of the phylogenetic tree T is tested in relation to fossil occurrences. If dates on the tree do not match the observed time pattern

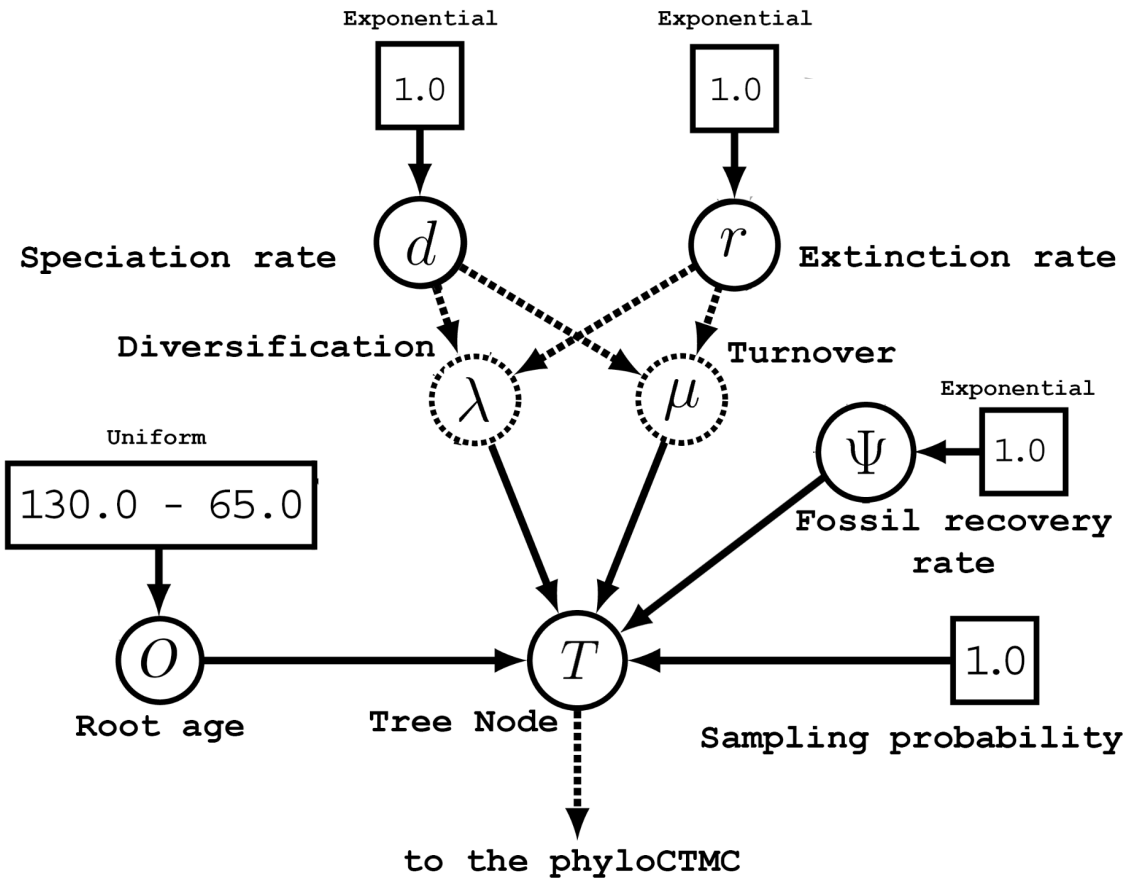


Figure 2.2: Tree node structure, following the scheme presented in Höhna et al. (2014). Both speciation rate d, extinction rate r and fossil recovery rate ψ were assumed to have an exponential distribution with mean of 1.0. Sampling probability was set to 1.0 since all currently known species of penguin are present in the matrix. The origin for the clade was assumed to be between 130 and 65 million years ago.

then the tree is rejected (i.e. if a given taxon is found outside its given temporal range) (Fig. 2.3). For this operation, each proposed tree from the FBDT node comes with series of fossil age offsets. These offsets are compared to the lower observed stratigraphic range (a), and to the upper boundary of the stratigraphic range (b), for each fossil in the dataset (Fig. 2.3).

2.2.4 MCMC - molecular partition

The likelihood for the molecular partition is calculated at a specific clamped node known as phylogenetic continuous time Markov chain (`PhyloCTMC`, Fig. 2.4). Four input parameters are needed to compute the likelihood for the molecular partition: a phylogenetic tree T (here deriving from the tree node above specified), a deter-

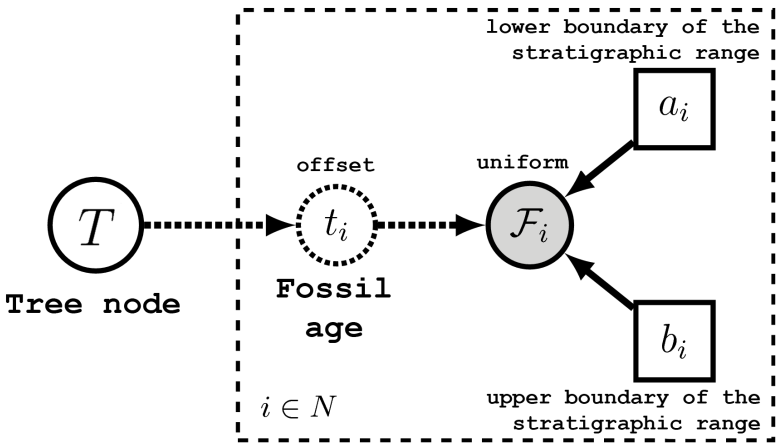


Figure 2.3: Schematic showing how fossil age is incorporated into the Markov Chain Monte Carlo model (Höhna et al., 2014; Stadler et al., 2018). The tree node provides a series of fossil age offsets and the tree is then rejected if fossil age occurrences are inconsistent with the lower and upper boundaries of the stratigraphic range.

ministic transition probability matrix Q (Eq. 2.8), the deterministic among sites variation sr (Eq. 2.11), and a stochastic molecular clock rate r (Eq. 2.13). The Q matrix is composed of the base frequencies π estimated from a Dirichlet distribution of four values (Eq. 2.9) that are each multiplied by the specific base transition rate $\boldsymbol{r}$, modeled from a Dirichlet distribution of six values (Tavaré, 1986) (Eq. 2.10). The resulting Q matrix is defined as general time reversible (GTR). The among-site variation sr is defined by a discretised gamma distribution with $\alpha{=}\beta$ and four discrete categories. The α prior is defined by an exponential distribution with a mean of 1.0 (Eq. 2.12). Note that here a specific molecular clock rate r_i is given to each branch of the phylogenetic tree (Eq. 2.13), corresponding thus to a relaxed clock model (Drummond et al., 2006). The molecular clock rates are all drawn from an exponential distribution with mean $1 \div v$ with v defined as an hyperprior drawn from an exponential with mean 1.0 (Eq. 2.12).

$$Q_{\text{GTR}} = \begin{bmatrix} . & \boldsymbol{r}_{AC}\pi_C & \boldsymbol{r}_{AG}\pi_G & \boldsymbol{r}_{AT}\pi_T \\ \boldsymbol{r}_{AC}\pi_A & . & \boldsymbol{r}_{CG}\pi_G & \boldsymbol{r}_{CT}\pi_T \\ \boldsymbol{r}_{AG}\pi_A & \boldsymbol{r}_{AG}\pi_C & . & \boldsymbol{r}_{GT}\pi_T \\ \boldsymbol{r}_{AT}\pi_A & \boldsymbol{r}_{AT}\pi_C & \boldsymbol{r}_{GT}\pi_G & . \end{bmatrix} \tag{2.8}$$

$$\pi = \text{Base frequencies} \sim \text{Dirichlet}(1, 1, 1, 1) \tag{2.9}$$

$$r = \text{Base substitution rate} \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1) \tag{2.10}$$

$$sr \sim \text{Discrete gamma}(\alpha, \alpha, 4) \tag{2.11}$$

$$\alpha \sim \text{Exponential}(1.0) \tag{2.12}$$

$$r_i = \text{Clock rate for branch } i \sim \text{Exponential}(1/v) \tag{2.13}$$

$$v = \text{Average clock rate} \sim \text{Exponential}(1.0) \tag{2.14}$$

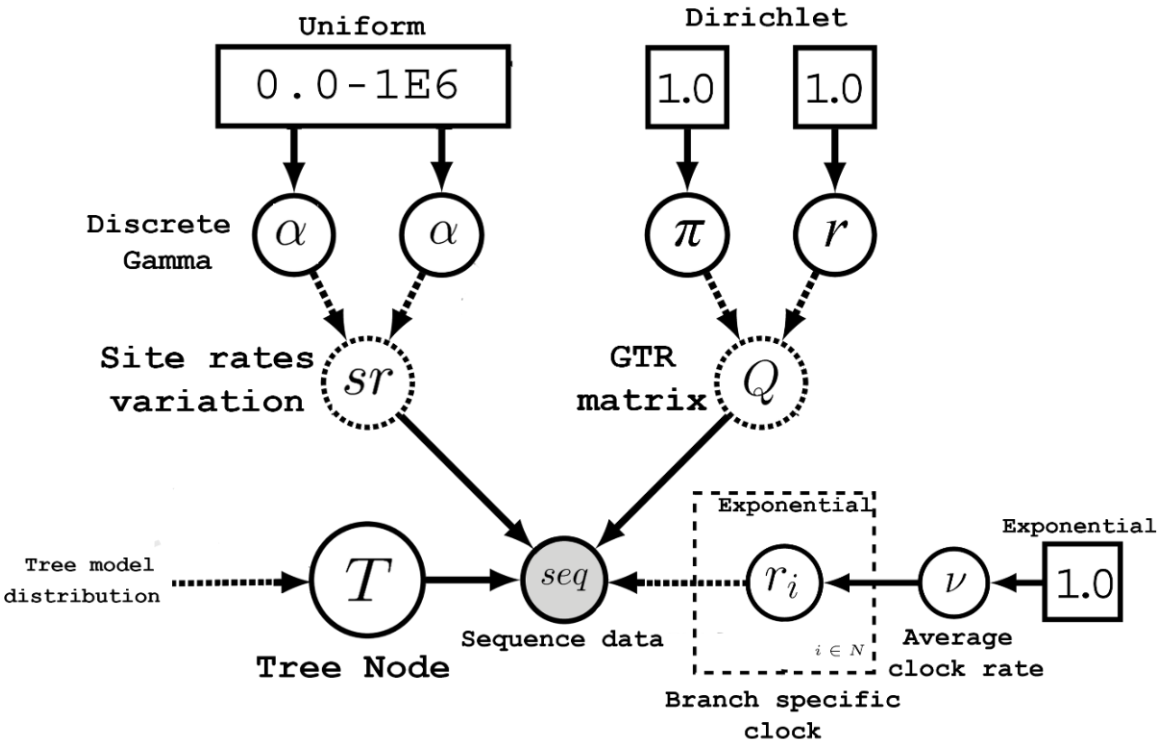


Figure 2.4: Molecular partition structure, following the scheme presented in Höhna et al. (2014). The substitution model was defined by the general time reversible (GTR) with both stationary densities π, and rates of transition, $\boldsymbol{r}$, defined by Dirichlet distributions set to 1.0 (Tavaré, 1986). The among-site variation sr was assumed to have a discretised gamma distribution with identical α and β parameters, assumed to be distributed on an exponential of mean 1.0 (Yang, 1994). The model allowed each branch to have it's own rate drawn from an exponential distribution (after Drummond et al., 2006).

2.2.5 MCMC - morphological partition

Like the molecular partition described above, the morphological partition is computed at a `PhyloCTMC` clamped node (Fig. 2.5). The morphological partition follows the same structure as the molecular partition, needing the same four types of input parameters (T, Q, sr and r) though here the parameters are modeled differently. Here the transition matrix Q for the morphological partition is modeled following the Mk. model for discrete characters (Jukes & Cantor, 1969; Lewis, 2001a). The model assumes equal transition probabilities among all character states regardless of the number of the states (Eq. 2.15). The variation among characters sr has been modeled in the same way as for the molecular partition (Eqs. 2.16 and 2.17), whereas the morphological clock r here is assumed to be uniform across the tree with a single parameter drawn from an exponential distribution with a mean of 1.0 (Eq. 2.18).

$$Q_{\mathrm{JC}} = \begin{bmatrix} -\mu_0 & \mu_{01} & \dots & \mu_{0n} \\ \mu_{10} & -\mu_1 & \dots & \mu_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n0} & \mu_{n1} & \dots & -\mu_n \end{bmatrix} \tag{2.15}$$

$$sr \sim \text{Discrete gamma}(\alpha_{\text{Morpho}}, \alpha_{\text{Morpho}}, 4) \tag{2.16}$$

$$\alpha_{\text{Morpho}} \sim \text{Exponential}(1.0) \tag{2.17}$$

$$r = \text{Morphological clock rate} \sim \text{Exponential}(1.0) \tag{2.18}$$

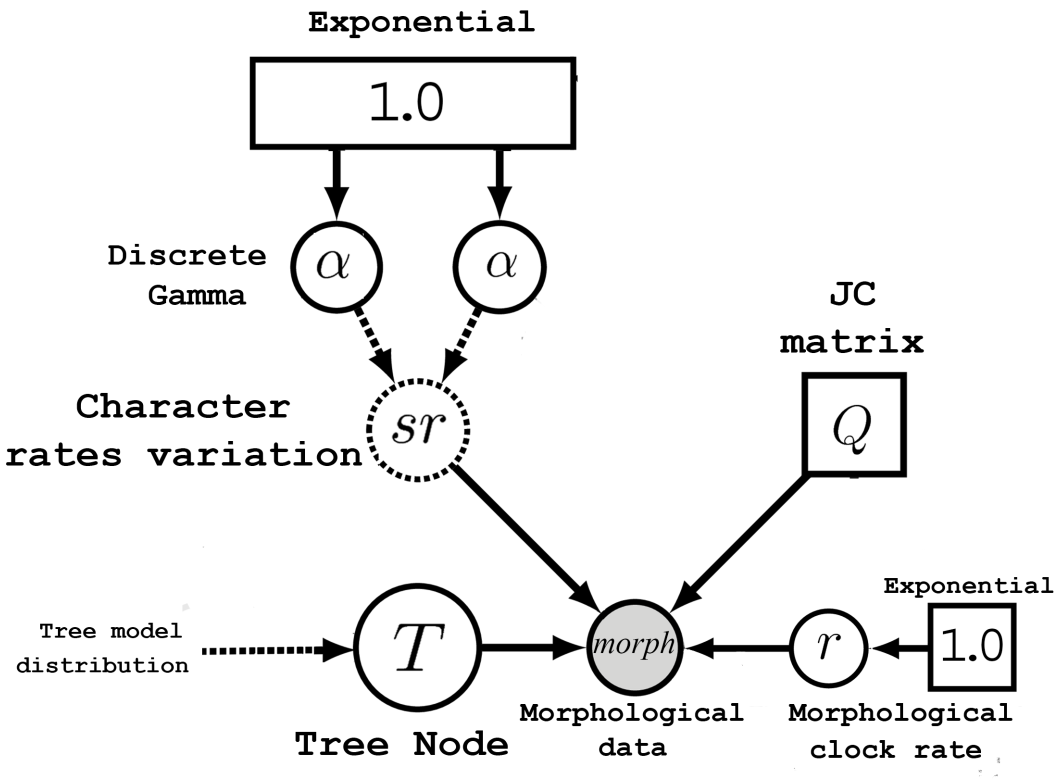


Figure 2.5: Morphological partition structure, following the scheme presented in (Höhna et al., 2014). The Q node was defined by the Jukes-Cantor model as formalized in the Mk. model (Jukes & Cantor, 1969; Lewis, 2001a). Priors for the rate variation among characters sr was identical to the priors used for the molecular partition, whereas the morphological clock r was modeled under the assumption of uniform rates across the phylogeny.

2.3 Results

2.3.1 Parsimony analysis

The initial parsimony search found 259 most parsimonious trees (MPTs) with 10,646 steps each, then a consensus tree was obtained from an extended search until memory overflow (Fig. 2.6). The iterPCR step revealed that the most unstable taxa were *Crossvalia unienwilla*, *Delphinornis larseni*, *Eudyptes atatu*, *Eudyptes calauina*, *Icadyptes salasi*, *Kaiika maxwelli*, *Kumimanu biceae*, *Notodyptes wimani* and *Pygoscelis grandis* and these taxa were then pruned to re-estimate the consensus tree (Fig. 2.7).

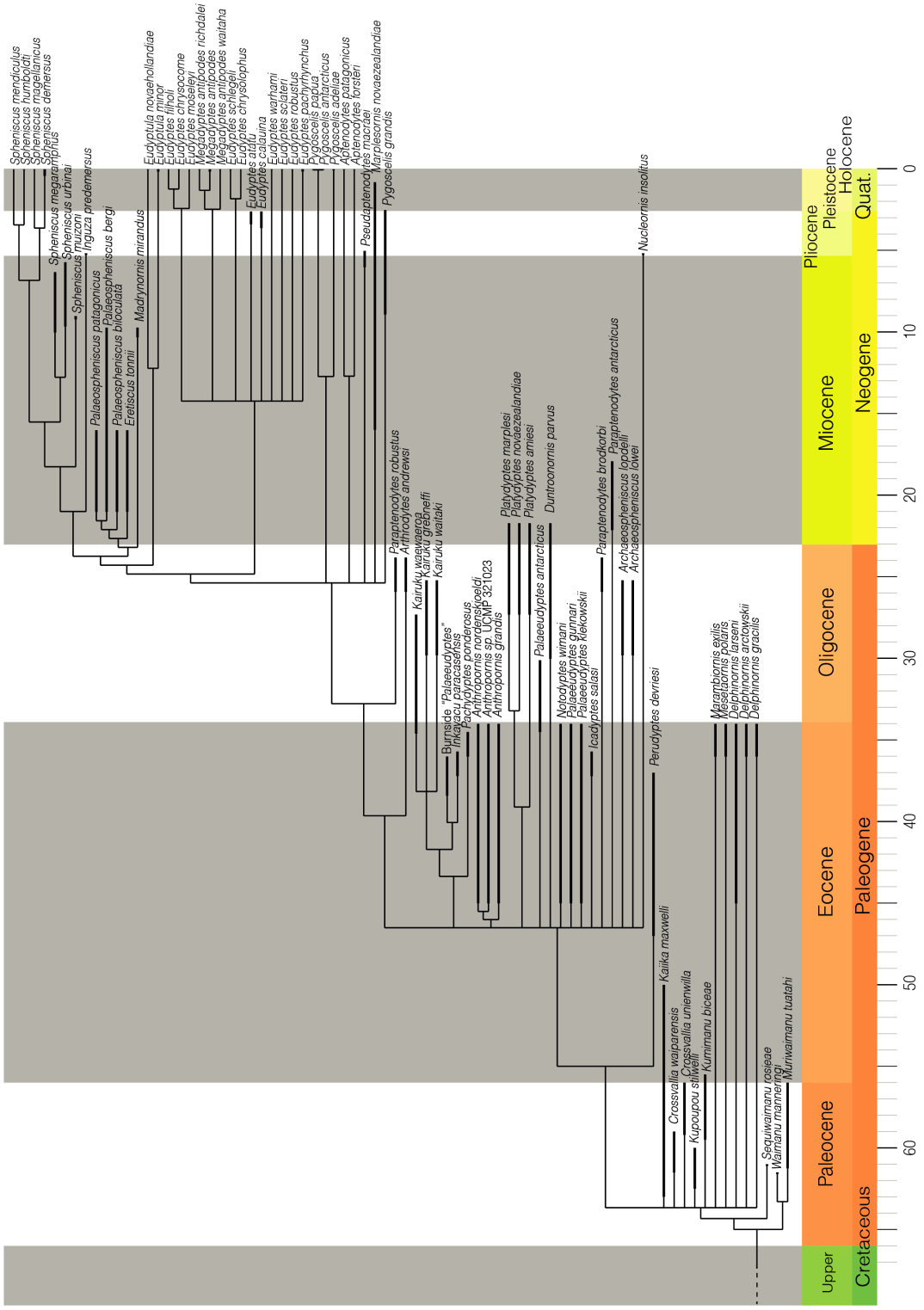


Figure 2.6: Tip time calibrated consensus tree for the parsimony analysis based on the updated matrix.

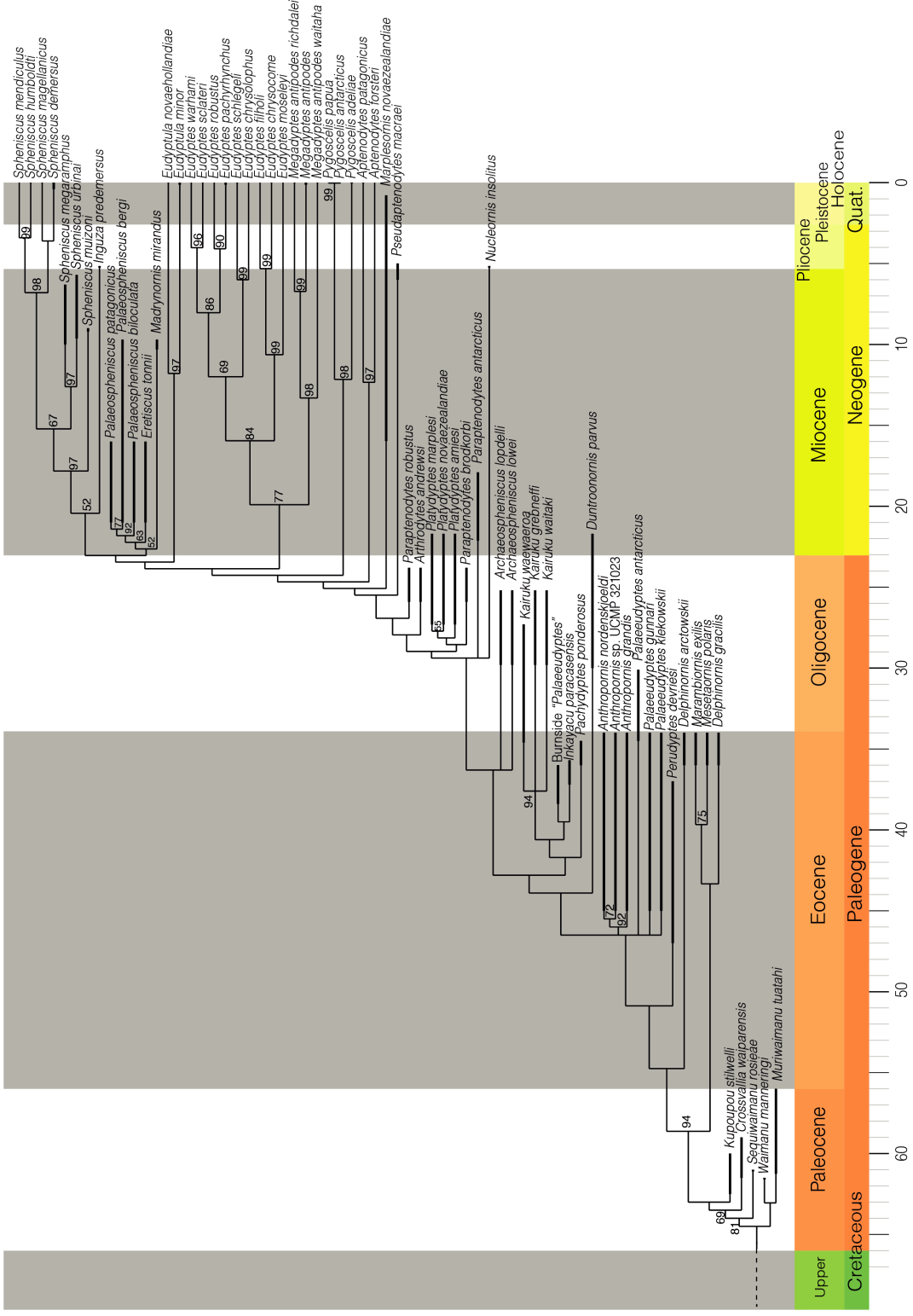


Figure 2.7: Tip time calibrated pruned consensus tree for the parsimony analysis based on the updated matrix. Numbers at the nodes represent the bootstrap values greater than 50 based on 200 replicates.

2.3.2 Fossilized Birth Death Tree analysis

Given the length of the total chain log file, the samples were subsetted to 63,771 (0.1 of the total length) to reduce file size and ease parameter summarisation. A burn-in of 25% of the resulting chain length was then set. Some of the sampled parameters of the analysis exhibited relatively low effective sample sizes (ESS) (i.e. origin time and age of the crown radiation, Table 2.1 and some branch rates from the molecular partition, Supplementary file `Chapter_2_Log_summary.csv`). The low ESS means that perhaps the space parameter was not sampled long enough during the MCMC to draw confident conclusions (Drummond & Rambaut, 2007). Importantly though, the majority of parameters reached an ESS greater than 200 indicating that the MCMC reached the global optimum for these parameters (Bouckaert et al., 2014; Drummond et al., 2006). Both d and r exhibited values around 0.5 and had similar credible intervals. In a similar manner also mitogenomics average rates and the morphological rates shared similar mean values, with mitogenomic rates having broader credible intervals (Table 2.1). The topology of the phylogenetic tree from the Bayesian analysis differs in several key ways from the parsimony analysis described above. Most importantly, *Palaeospheniscus* is outside the crown radiation in the maximum credibility clade (MCC) tree (Fig. 2.8) and in the maximum a posteriori (MAP) tree (Fig. 2.9). *Madrynornis* remains within the crown in both the MCC and MAP trees. As a consequence of placing *Palaeospheniscus* deeper in the phylogeny, the origin of the crown clade moves from earliest Miocene in the parsimony tree to middle Miocene in the Bayesian results (around 13 to 17 Ma; Table 2.1; Figs. 2.8 and 2.9). Importantly for this study, a single clade is recovered that includes many giant penguins ranging from the Eocene up to the Oligocene (similar to the result from the parsimony inference). Other genera like *Platydyptes*, *Archaeospheniscus*, and *Paraptenodytes* are not part of any major radiation but instead form a single monophyletic clade. *Delphinornis* is recovered closer to the crown than in previous reports (cf. Giovanardi et al., 2021; Ksepka et al., 2012; Mayr, Scofield, et al., 2017; Thomas et al., 2020) and *Nucleornis* is placed in a clade with *Paraptenodytes*

Table 2.1: Parameter estimates from the Fossilized birth death analysis. Estimates have been computed from a subset of 63,771 samples from the total chain with a burn-in of 25% of the total length. For each parameters are presented mean, standard deviation (SD), 89% credible intervals (lower and upper) and effective sample size (ESS)

	Mean	**SD**	**Lower**	**Upper**	**ESS**
Posterior	-52253.73	25.22	-55398.91	-52217.03	166.60
Likelihood	-52160.87	9.94	-55315.16	-52146.89	355.60
Prior	-92.86	24.04	-129.72	-43.64	169.80
Diversification λ	0.01	0.02	-0.03	0.05	20153.50
Turnover μ	0.98	0.05	0.91	1.05	13834.50
Speciation rate d	0.52	0.10	0.38	0.69	372.90
Extinction rate r	0.51	0.10	0.37	0.68	376.70
Fossilization rate Ψ	0.03	0.01	0.02	0.04	435.90
Age crown radiation (Ma)	15.64	1.24	13.77	17.64	45.30
Origin time O	96.75	17.19	71.72	124.91	28.50
Number of total sampled ancestors	4.14	1.56	2.00	7.00	308.50
α	0.17	0.01	0.16	792494.11	10214.70
α_{morpho}	1.51	0.33	1.07	2.07	9775.10
Mitogenomic average rate v	0.01	0.00	0.01	0.01	115.40
Morphological rate r	0.01	0.00	0.01	0.02	126.60

(this result was not observed in Ksepka and Thomas, 2012). Here Sphenisciformes is hypothesised to have originated in the Late Cretaceous (Table 2.1, Fig. 2.9 and 2.8), representing so far one of the oldest estimates for the origin of the clade (cf. Blokland et al., 2019; Gavryushkina et al., 2017; Thomas et al., 2020).

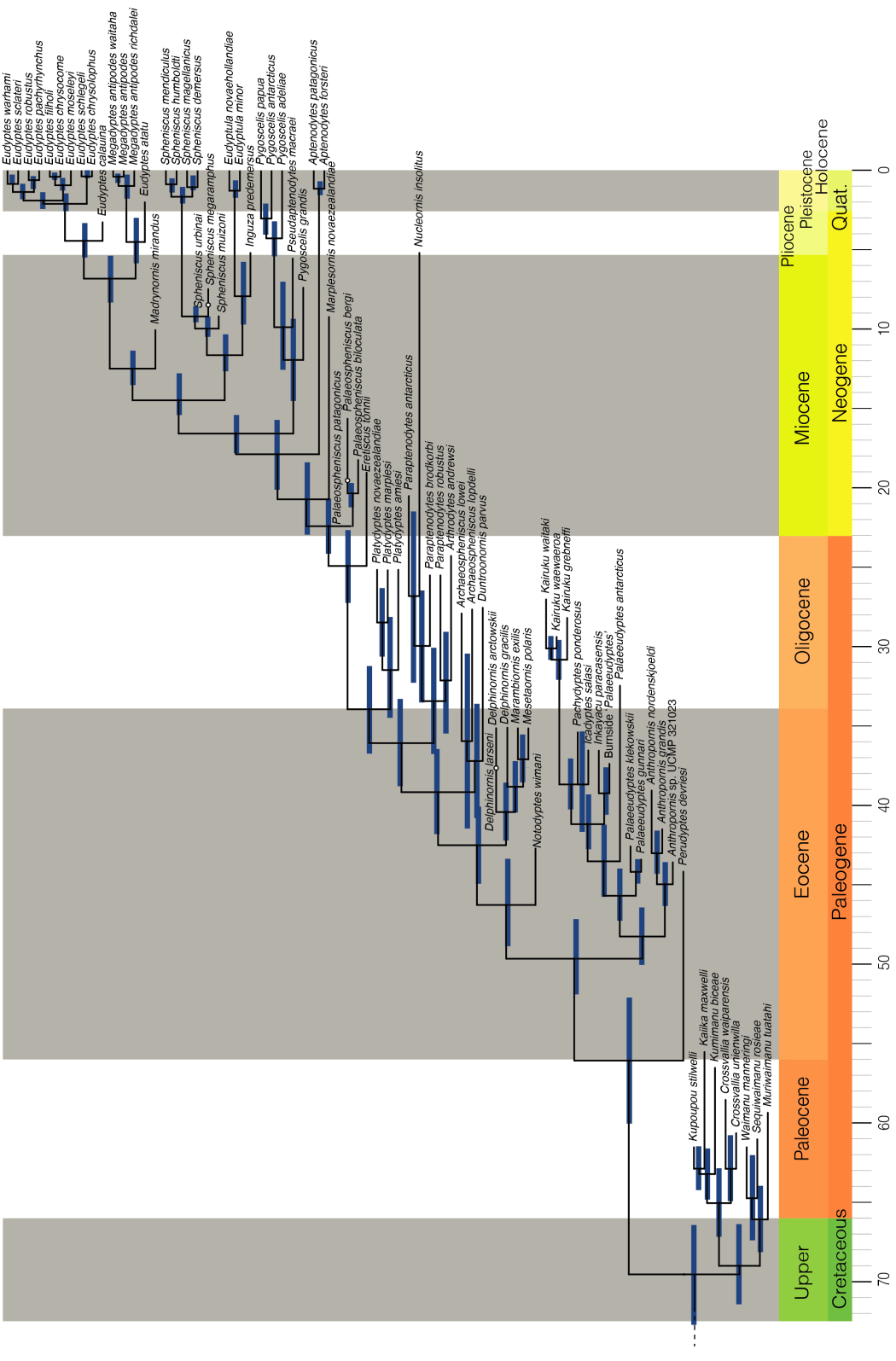


Figure 2.8: Maximum clade credibility (MCC) phylogenetic tree resulting from the fossilised birth death tree model. The bars at the nodes represent the 95% highest posterior density interval (HPDI) for the node date.

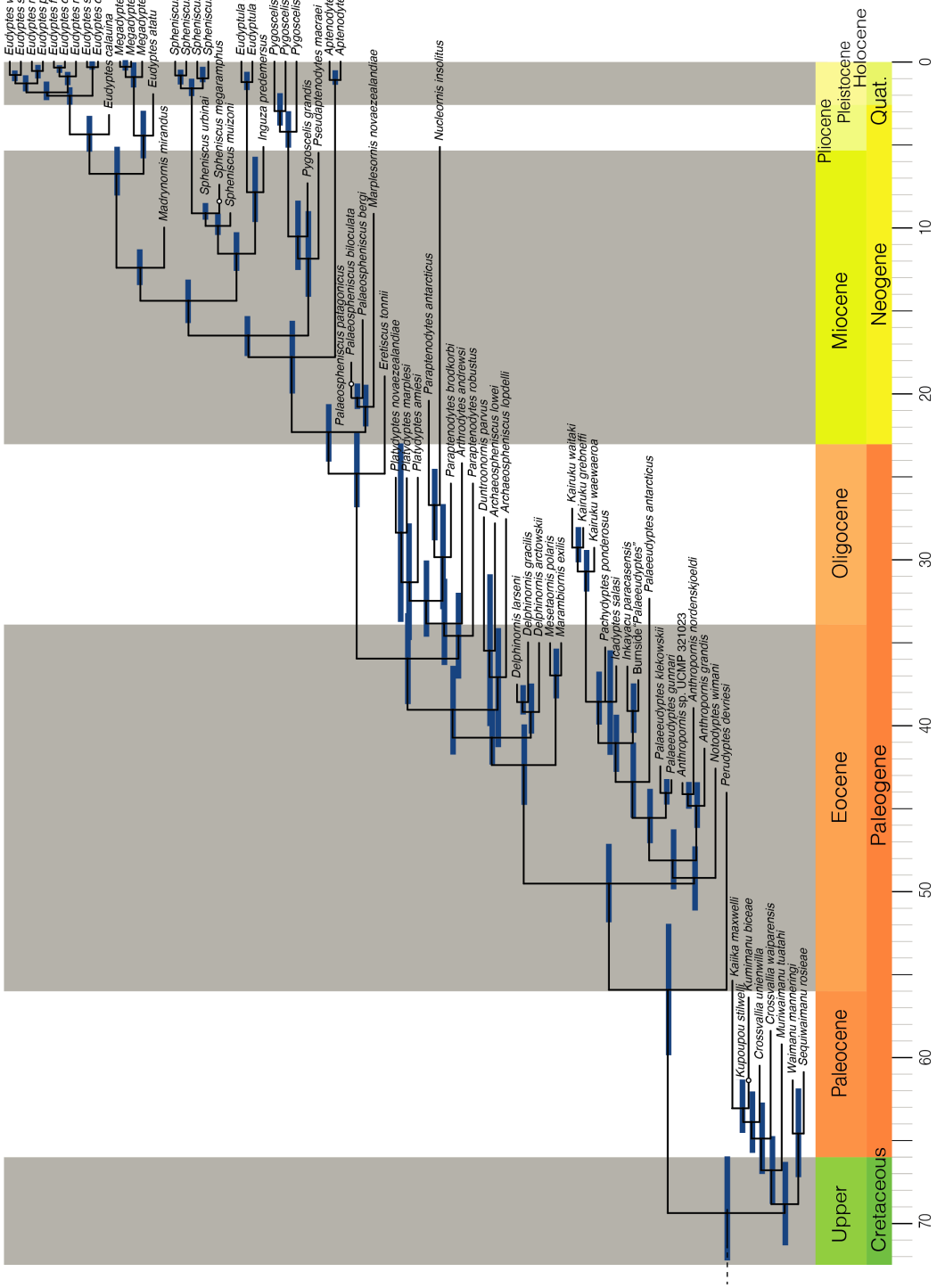


Figure 2.9: Time calibrated maximum a posteriori (MAP) phylogenetic tree resulting from the FBDT model. The bars at the nodes represent the 89% highest posterior density interval (HPDI) for the node date.

2.4 Discussion

The evolutionary hypotheses emerging from both the parsimony and Bayesian phylogenetic analyses share several key topological traits and differ in some aspects from previous analyses (Degrange et al., 2018; Gavryushkina et al., 2017; Ksepka et al., 2012). The most significant divergence between the parsimony-based result presented here and previous studies is the inclusion of *Palaeospheniscus* as a member of the crown, which pushes back the origin of the crown clade at least before the 19 million years ago mark and substantially inflates implied ghost lineages as a result (a ghost lineage is the predicted extension of a lineage beyond the range of fossil evidence for that lineage). *Paleospheniscus* was incorporated into the crown clade primarily because of similarities with *Madrynornis*, in addition to synapomorphies shared with crown taxa. Bootstrap values indicate stronger support for sister relationships between *Paleospheniscus* and *Madrynornis* compared with the crown clade itself.

The fossilised birth death tree (FBDT) returns a phylogenetic hypothesis with a topology and node ages more consistent with previous studies (Degrange et al., 2018; Gavryushkina et al., 2017; Ksepka et al., 2012). However, an important feature of the FBDT phylogeny described here is a "giant penguin" clade. If the monophyly of giant penguins is confirmed in future studies it may suggest that gigantism evolved among penguins fewer times than previously expected (Clarke et al., 2007; Ksepka et al., 2012; Mayr et al., 2019; Mayr, Scofield, et al., 2017), meaning that many resemblances between these taxa may be due to shared inherited traits rather than ecological convergence. It is worth noting that the phylogenetic position of *Palaeeudyptes* in the 'giant penguin' clade is relatively basal. Since its first description based on a single fragmentary element (T. H. Huxley, 1859), the referral of additional specimens to this taxon has been challenging and not universally supported (see Ksepka et al., 2012). The specimens that are lumped into *Palaeeudyptes* often lack distinctive features (Marples & Finlay, 1952). If the phylogenetic position of *Palaeeudyptes*, which is presented here as a stemward and relatively undifferen-

tiated giant penguin, is confirmed in future studies, then this more-basal position may help explain this taxonomic history. Specifically, *Palaeeudyptes* may have been defined using plesiomorphic features and thus not sufficiently reflect a pattern of derived traits to define a real clade (Hennig et al., 1966).

The 'giant penguin' clade recovered here is interesting for the evolution of *Kairuku* (Thomas & Ksepka, 2016). As with the phylogeny presented in Giovanardi et al. (2021), the FBDT still finds a monophyletic grouping for all *Kairuku* although with possibly different relationships among the three named species (Fig. 2.9 and 2.8). *Kairuku* would here be considered the final lineage of 'giants' before their extinction in the late Oligocene. With the description of *K. waewaeroa* it is evident that *K. grebneffi* and *K. waitaki* evolved a peculiar bauplan compared to their closer relatives. The shorter stature of *K. grebneffi* and *K. waitaki* could indicate the presence of different environmental pressures for the giant penguins of Zealandia during the Oligocene (Ando & Fordyce, 2014). The FBDT analysis also suggests that all Paleocene taxa belong to a single clade within Sphenisciformes. Moreover, these numerous taxa near the root of the phylogenetic tree may explain why the Bayesian analysis in this chapter returns such an early origin for the whole Sphenisciformes clade compared to previous estimates (96 Ma here, in contrast to 70 Ma in Thomas et al., 2020 and around 60 Ma in Gavryushkina et al., 2017). The current analysis includes eight early taxa into the morphological matrix compared to previous time-calibrated analyses where only *Waimanu*, *Muriwaimanu* (Gavryushkina et al., 2017) and *Sequiwaimanu* (Thomas et al., 2020) were included. Given that the estimated rates of speciation remain unchanged, the Bayesian inference would expect an earlier origin to account for a greater number of species.

It's worth noting that the same updated matrix produced two slightly different hypotheses under two different methods of phylogenetic inference. Explanations for discrepancies between these hypotheses, other than differences between parsimony- and likelihood-based computations, may be brought back to the inclusion of the time variable in the analyses. The Bayesian analysis is much more sensitive to

inflations from ghost lineages, penalising trees that do not fit the stratigraphy as well if they are not supported by a consistent amount of morphological and molecular evidence (Lee & Palci, 2015). Such behaviour may also explain the reason behind the shifting position of the *Delphinornis* clade to a younger age, given their extremely fragmentary nature their phylogenetic position will be impacted much more by the available temporal information.

2.5 Conclusion

The current analysis provides a different perspective on the phylogeny of living and extinct penguins. Although evolutionary relationships among crown species were mostly unchanged compared to previous analyses (T. L. Cole et al., 2019; Gavryushkina et al., 2017; Thomas et al., 2020) relationships among many stem penguins showed a different pattern compared with previous hypotheses (Giovanardi et al., 2021; Ksepka et al., 2012). Rather than a fully pectinated phylogenetic tree, penguin lineages may have faced at least two major radiations: one in the Paleocene and one during the Eocene. These results demonstrate the impact that character selection can have on a phylogenetic analysis. With the phylogenies generated in this chapter it is now possible to assess the impact of phylogeny over any given set of traits shared among penguins. From here we proceed in investigating the impact that the two other vertices of the aptive triangle have had on the evolutionary history of Sphenisciformes.

Figure 2.10: Reconstruction of a group of *Delphinornis gracilis* sorrounding two *Anthropornis grandis*. Both these taxa were found on the Antarctic Seymour island from Eocenic rocks.

Chapter 3

Structural apex

3.1 Introduction

The size range of all known taxa encompasses at least 21 orders of magnitude (Robinson et al., 1983). Species at the extreme ends of this size spectrum are subject to completely different physio-ecological pressures dictated by the physical laws of the surrounding environment (Willmer et al., 2009). Size consequently determines how an organism is able to interact with the environment and is thus one of the most studied traits both generally in biology, and specifically in evolutionary research (Blanckenhorn, 2000; Chown & Gaston, 2010; Larramendi et al., 2020; Maurer et al., 1992; Pimiento et al., 2019; Slater et al., 2017). Knowledge about the body size of organisms within a clade can provide a better understanding of macroevolutionary events including for example the secondary adaptation to new environments observed in several vertebrate lineages (Benson et al., 2018; Finarelli, 2008). Within living vertebrates alone, body mass ranges from only a few grams (Rittmeyer et al., 2012) up to several tons (Sears & Perrin, 2009), with implications for respiration, nutrition, heat management, structural support, and much more. Consequently, when analysing evolutionary events it can be important to measure the size of both extant taxa as well as provide accurate estimates for extinct taxa (McClain & Boyer, 2009). However, the incomplete preservation of fossils makes the estimation of size for ancient life challenging. As a result, great efforts have been made to accurately

estimate body mass of extinct vertebrates by describing scaling relationships for measurable morphological features (Anderson et al., 1985; N. E. Campione & Campione, 2020; N. E. Campione & Evans, 2020; N. E. Campione et al., 2014; Field et al., 2013).

Developing accurate and well-constrained morphological scaling rules for estimates of body mass is especially important when the range of body masses of an extinct study group exceeds the range of body masses observed in close living relatives. As an example, non-avian theropods could reach dramatic sizes up to 6.4-7.9 tons (Ibrahim et al., 2020) compared to living theropods with the common ostrich *Struthio camelus* that can reach up to 156 kg (Davies et al., 2002). Penguins and their stem-taxa show a similar pattern, with size variation in living penguins relatively limited compared to the extinct forms. The body mass of extant penguins ranges 1 kg for little penguins (*Eudyptula* spp.) up to 34 kg for emperor penguins (*Aptenodytes forsteri*) (Stonehouse, 1975). Extinct taxa exhibited a much broader range of body sizes (and therefore body masses) compared with the crown-clade ranging from small, *Eudyptula*-sized species (e.g. *Eretiscus tonni* and "*Pakudyptes hakataramea*") up to the "giant" forms already introduced in the previous chapter of this thesis (Acosta Hospitaleche, 2016; Acosta Hospitaleche & Alicia, 2004; Ando, 2007; Clarke et al., 2010; Ksepka et al., 2012; Simpson, 1981). Previous studies have estimated the body masses of extinct penguins (Jadwiszczak, 2001; Livezey, 1989) using regressions based solely on extant penguins and thus have often extrapolated well beyond the range of body sizes represented by living penguins. Taking a broader perspective, however, we can recognise that penguins are still subjected to similar constraints (Bright et al., 2016; Demery et al., 2021; Heers & Dial, 2015) that act over crown group birds (Aves). Thus, by first exploring general scaling rules across the whole avian clade, it would be possible to gain information that can be used to predict the body masses of particular avian clades (Gelman et al., 2012). By incorporating information from a wide diversity of birds when developing the morphological scaling rules for penguins we can provide for the first time constraints

around the estimates of body masses for extinct penguins that were larger than any living penguin species.

Including a set of morphological measurements from a wide diversity of birds to define a set of scaling rules would also contribute to solve another limitation: extant penguins have an uneven size distribution as they are represented by a few smaller species (i.e. $\sim$ 2 kg or less), many medium-sized species (between $\sim$2 and $\sim$13 kg), and one large species (*Aptenodytes forsteri*; $\sim$34 kg). The consequence of such an uneven size distribution is that the emperor penguin bears a considerable amount of leverage (Young, 2019), meaning that it has a greater influence on body mass scaling relationships than any other living species of penguin and therefore biases the inferences for any large- or giant-sized penguin. Given the lack of opportunity for validating body mass estimates for "giant" penguins in particular, and the often fragmentary preservation of penguin fossils, there is a high risk of overestimating the body mass of large extinct penguins when making predictions based on scaling rules developed entirely from living penguins. This risk is especially pertinent when we consider that some "giant" penguins have been found to be more slender than we might have otherwise expected (Giovanardi et al., 2021; Ksepka et al., 2012). By incorporating data for more birds into the scaling rules for penguins we can reduce the leverage that any one living penguin species has on the body mass predictions for extinct species.

A challenge of building a model for estimating body mass that includes a broad amount of avian diversity arises in the selection of a biological measurement that would make penguins comparable to all other birds. We must be mindful that while birds generally have a conserved body plan compared with other tetrapod clades (e.g. lissamphibians, lepidosaurs, synapsids), penguins do have a body plan that is dissimilar from most other birds (Ksepka & Ando, 2011; Simpson, 1946). Moreover, penguin bones are more dense than those of aerially-volant birds (Larramendi et al., 2020), and no other living avian group is as well adapted to life in water, meaning that any model informed by body masses outside of Sphenisciformes has

the potential to underestimate penguin body mass. The challenge that this chapter aims to overcome is to find the right balance among overestimation of body mass (i.e. estimates of giant penguin body mass based entirely on living penguins) and underestimation of body mass (i.e. allowing non-penguin birds to have too strong of an influence on estimates of penguin body mass) and thus generate and validate an informed model that may produce reliable estimates.

The body mass of many groups of living birds has previously been shown to be strongly correlated with the width of the humeral articular facet (HAF) on the coracoid (Field et al., 2013) (Fig. 3.1). The great majority of birds in the dataset for the Field et al. (2013) study were volant taxa where the coracoid HAF is an extremely important articular surface for locomotion. The strong correlation between coracoid HAF and body mass is best explained by the structural or constraint apex of the constructional morphology triangle (Seilacher & Gishlick, 2019). Specifically, greater body masses require proportionately greater amounts of thrust to achieve flight (Tobalske, 2007), and thus it is reasonable to think that HAF size undergoes little or null ecological overprinting as stated by the original study authors (Field et al., 2013). In this sense, the constructional morphology framework may represent a guideline to quantify trait evolvability (sensu Klingenberg, 2005) providing thus reliable correlates for size.

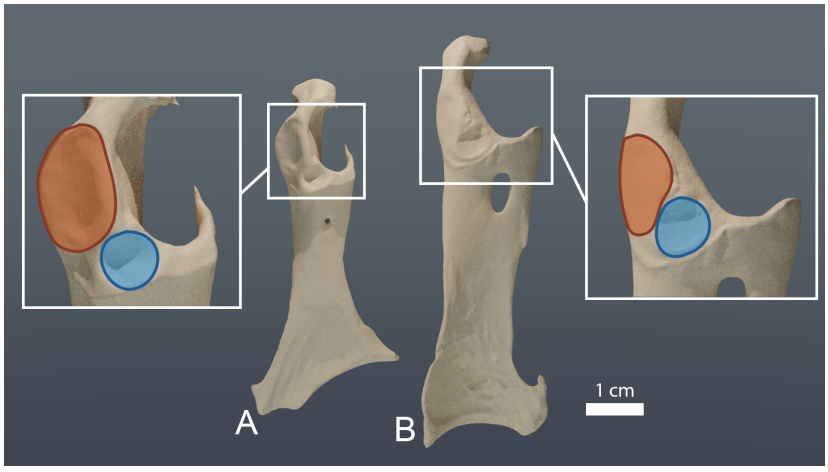


Figure 3.1: Coracoids of A) *Platalea regia* (left, OR29403) and B) *Eudyptes pachyrhynchus* (right OR29154) seen in dorsal views. The zoomed section of both bones shows inferred humeral articular facet highlighted in red and the scapular cotyle highlighted in blue. Figure uses 3D meshes and not photographs.

Importantly for the current chapter, Field et al. (2013) did not include extant penguins, allowing the possibility that the scaling relationship between body mass and coracoid HAF may only hold for birds that fly through air and not in water (i.e. a physio-ecological relationship for birds that only fly in a low-viscosity fluid). As described in Chapter 1: Introduction, penguins are wing-propelled divers (Simpson, 1946) meaning that they use their forelimbs to generate thrust like most birds. However, given the increased density of water compared to air we anticipate that the coracoid HAF-body mass relationship for penguins may lie on a different regression line when compared with the regression line for birds that are volant in air. Stated another way, a penguin may have a larger or smaller coracoid HAF compared to an aerially-volant bird with the same body mass because of the additional thrust required to overcome the drag of water. Hence, when developing a model for estimating body mass from coracoid HAF, we should aim to 1) include a broad diversity of birds in addition to penguins so that we can estimate the body masses of the full range of penguins including the extinct "giants", while also 2) allowing the model to select the best regression parameters for the focal group of birds. The current chapter aims to expand the dataset of Field et al. (2013) to include penguins and therefore investigate the relationship between coracoid HAF and body mass for penguins.

This chapter will also aim to develop a second set of models that focus on the femur in addition to the coracoid. Measurements from femora have previously been demonstrated to correlate well with body mass estimates of quadrupedal and bipedal tetrapods (N. E. Campione & Campione, 2020; N. E. Campione & Evans, 2020; N. E. Campione et al., 2014). As the autopodial bone in the hindlimb of birds and other tetrapods, the femur needs to bear the weight of the animal in order to resist the gravitational forces from the body to the ground (Bessho et al., 2007; N. E. Campione et al., 2014). Penguins and many other birds may principally swim but they still rely on hindlimbs to support their upper body mass when they walk. Similar to the relationship between coracoid HAF and body mass described above,

the proportions of femora are likely tightly constrained by the body mass of the bird (Gould & Lewontin, 1979; Seilacher & Gishlick, 2019). A femur that is too small for the body mass of the bird would not adequately support the weight of the animal, and a femur that is too large for the body mass of the bird would impose an unnecessary mass burden. Moreover, femur shape is highly conserved among clades. Selecting the femur among the elements of the hindlimb is a sensible starting point in terms of comparability across birds compared with the tibiotarsus and tarsometatarsus, where the latter two tend to show more diverse morphologies between avian groups (Baumel & Witmer, 1993).

Previous studies that estimate the body masses of extinct birds tended to use linear measurements from bones, like circumferences and widths (e.g. Anderson et al., 1985; N. E. Campione and Evans, 2020; Field et al., 2013). However, with the ongoing digital revolution of the paleontological field (Sutton et al., 2016; Tocheri, 2009), size and shape-accurate digital replicas of the bones are easily maneuverable in a virtual environment, often allowing easier access to intricate measurements that can be more difficult to achieve in the real world (Brassey, 2017). Specifically, measuring the volume of a digital replica is a relatively easy task implemented in several software packages compared to previous methods for volume estimation (Colbert, 1962). The increase in bone volume and body mass are both cubic compared to linear measurements (Willmer et al., 2009). Hence, we anticipate that body mass will correlate more closely with femur volume than with any one linear measurement of the femur. The statistical models developed in this chapter will therefore focus on the total volume of the femur as well as the linear measurement from coracoid HAF (the latter for consistency with the Field et al., 2013 dataset). Note that total volume includes the bone, marrow and air spaces within the bone despite the different densities of these materials, and their different proportions between avian groups.

The body mass estimation methods developed for femora and coracoid HAF will use Bayesian methods (for background see Chapter 1: Introduction, and for a phy-

logenetic application see Chapter 2). Recall that the use of Bayesian Markov chain Monte Carlo (MCMC) approaches in paleontology can be extended to many statistical applications. Specifically for this chapter, the traditional maximum likelihood approaches for fitting a linear model to a dataset (Chambers, 1992; Wilkinson & Rogers, 1973) can be reframed in a Bayesian context by including prior information on any parameter in the fit (e.g. intercept and slope), and thereby allowing a posterior probability of the model to be calculated (Kruschke, 2015; McElreath, 2020). Given the iterative nature of MCMC the output of any Bayesian fitted model will be a distribution of values rather than pointwise estimates, allowing the quantification of our degrees of uncertainty for any given prediction. The calculation of this distribution is one of the key reasons for using a Bayesian method in this chapter. Previous studies that have estimated the body masses of extinct penguins have presented only a single body mass estimate without an indication of error around the estimate. This chapter will produce a distribution of body mass estimates for each extinct species to better convey the uncertainty around the estimate (van de Schoot et al., 2021).

Another advantage of using Bayesian methodology is the accessibility of a wide variety of programming languages (Stan, Jags, Bugs, Carpenter et al., 2017; Lunn et al., 2009; Plummer, 2003) that support many customisation options allowing the user to control for a range of predictors that may affect the response variable. One of the important predictors of body mass that will be explored in this chapter will be phylogenetic covariance given that the dataset of collected body mass and bone measurements (e.g. femur volume) includes taxa with different degrees of shared evolutionary history (Felsenstein, 1985). Previous studies have demonstrated that not accounting for phylogenetic covariance may bias parameter estimates in comparative studies (Garland & Adolph, 1994; Losos, 1994; Martins, 1996; Stone, 2011). Hence, we should expect that including phylogenetic information may also lead to less biased predictions (Garland & Ives, 2000). Such hypotheses will be tested in the current chapter with a set of Bayesian models that will take into account

different degrees of phylogenetic information of growing complexity. At the basal-most level there will be a simple model allowing for different intercepts for each major avian order. A more complex model will account for covariation among observations in the form of a Brownian motion model of evolution (Felsenstein, 1985), whereas the most complex implementation of trait evolution will be the equivalent of a Ornstein-Uhlenbeck gaussian process (Lande, 1976). The grand aim of this chapter is to develop and search through a series of methods, and ultimately find the best tool for making realistic body mass estimates for extinct penguins.

3.2 Material and Methods

3.2.1 Dataset

Humeral articular facet of the coracoid

The dataset of coracoid HAF measurements from Field et al. (2013) was supplemented with data from an additional 14 extant penguin species. Nine coracoid HAF measurements from extinct species of penguin were collected for use in body mass estimations. Specimens were provided by the Field Museum of Natural History, Chicago, USA; Massey University, Auckland, New Zealand; Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand; Otago University Geology Museum, Dunedin, New Zealand; Waikato Museum Te Whare Taonga o Waikato, Hamilton, New Zealand. All accession numbers for the dataset generated for this study are provided in the supplementary file `Chapter_3_Specimen_ID_and_Weights.csv` along with measurements, associated data including source institution, and method of 3D mesh generation. Coracoid HAF was measured following the methods described in Field et al. (2013), which involves measuring the maximum width along the facet. Measurements from extant penguins ranged from 7.2 mm (*Eudyptula minor*) up to 22.8 mm (*Aptenodytes forsteri*). Measurements were collected from digital replicas of specimens using `HP DAVID 3` software (HP Inc., Palo Alto, CA, USA) and were checked against caliper measurements of physical specimens.

Femora

Total internal volume of the femur was collected from a dataset of 319 digital replicas of femora, with one femur each from 319 species in 63 families and 27 orders. Femora ranged in size from 10 mm long for 8.37 mm^3 (piwakawaka New Zealand fantail *Rhipidura fuliginosa*) to 32 cm long for 926.6 cm^3 (ostrich *Struthio camelus*). The dataset of extant species only included bones from adult birds with no damage, and no pathology reported in the museum label. The sex of the individuals was not recorded for this study because the majority of specimens did not report sex on the specimen label. An additional dataset of digital replicas was generated for femora from 12 extinct species for use in body mass predictions. Specimens for both the extant and extinct datasets were provided again by the institutions cited above and The Auckland War Memorial Museum Tāmaki Paenga Hira, Auckland, New Zealand, Canterbury Museum, Christchurch, New Zealand. The dataset also included meshes from Morphosource.org from the oMega project funded by the Idaho museum of Natural History. Meshes were generated using two methods, surface scanning and computed tomography (CT). Surface scanning was performed with a HP 3D camera (DAVID-CAM-3.1-M; HP Inc., Palo Alto, CA, USA) and a HP 3D HD camera along with a K132 + DLP projector (Acer Incorporated) and meshes finalised in the `HP DAVID 3` software (v. 3.10.4.4657; HP Inc., Palo Alto, CA, USA), or with the Creaform HandySCAN 3D laser surface scanner (Creaform, Levis, Canada) with resolution varying from 0.1 to 0.3 mm and meshes finalised in the `VXelements` software suite (`VXelements` and `VXmodels` v. 8.1, Creaform, Levis, Canada). Meshes generated from CT data were obtained using a protocol adapted from an existing method (Buser et al., 2020), in the software package `SlicerMorph` (Rolfe et al., 2021, release 29738) developed for `Slicer` (Fedorov et al., 2012; v. 4.11.2) to convert image stacks of scan data into a finalised surface mesh. All meshes were manually inspected in a digital environment, and non-manifold meshes were corrected. Subsequently models were decimated depending on the number of polygons, ranging from a decimation of 2% of the original vertex count (smallest meshes were

4.7 x 10^5 vertices) up to to 0.1% of the original vertex count (biggest meshes were 3.4 x 10^6 vertices). Manifolding and decimation steps were both performed in `Blender` (Blender Online Community, 2020, v. 2.9.1). Femur volume was calculated in the `R` environment (R core Team, 2021; v. 4.0.5) using the function `rgl::vcgVolume` (v 0.103.5) (Fig. 3.2).

Body mass

Body mass data used in this study were mostly from the dataset assembled by Dunning (2007). The Dunning (2007) dataset is one of the most extensive reviews of bird body size that has been completed and contains information for over 8700 species gathered across more than 1000 publications. The body mass for each species was provided as a mean across observations from variable numbers of individuals, and in some cases separate means for males and females. However, given that sex was not identifiable from the majority of specimens included in this chapter, as well as not always reported in the original source (Dunning, 2007), the species mean was used as the main response variable. The Dunning (2007) dataset was here supplemented with body masses from Vasilakis et al. (2016) for cinereous vulture *Aegypus monachus*, McNab and Ellis (2006) for south island takahē *Porphyrio hochstetteri*, and Jenkins and Veitch (1991) for tīeke saddleback *Philesturnus rufusater*.

Taxonomic order parameter used in Bayesian models

Taxonomic information for each observation was from the International Ornithological Congress World Bird Master List version 11.1 (www.worldbirdnames.org). Taxonomic order was used as a predictor variable in several analyses. However, given that Palaeognathae would have resulted in multiple families with only few if not single entries, it was decided to treat the clade as a single group. The resulting groups were: "Accipitriformes", "Anseriformes", "Apodiformes", "Charadriiformes", "Columbiformes", "Coraciiformes", "Cuculiformes","Eurypygiformes", "Falconiformes", "Galliformes", "Gaviiformes", "Gruiformes", "Otidiformes", "Palaeognathae", "Passeri-

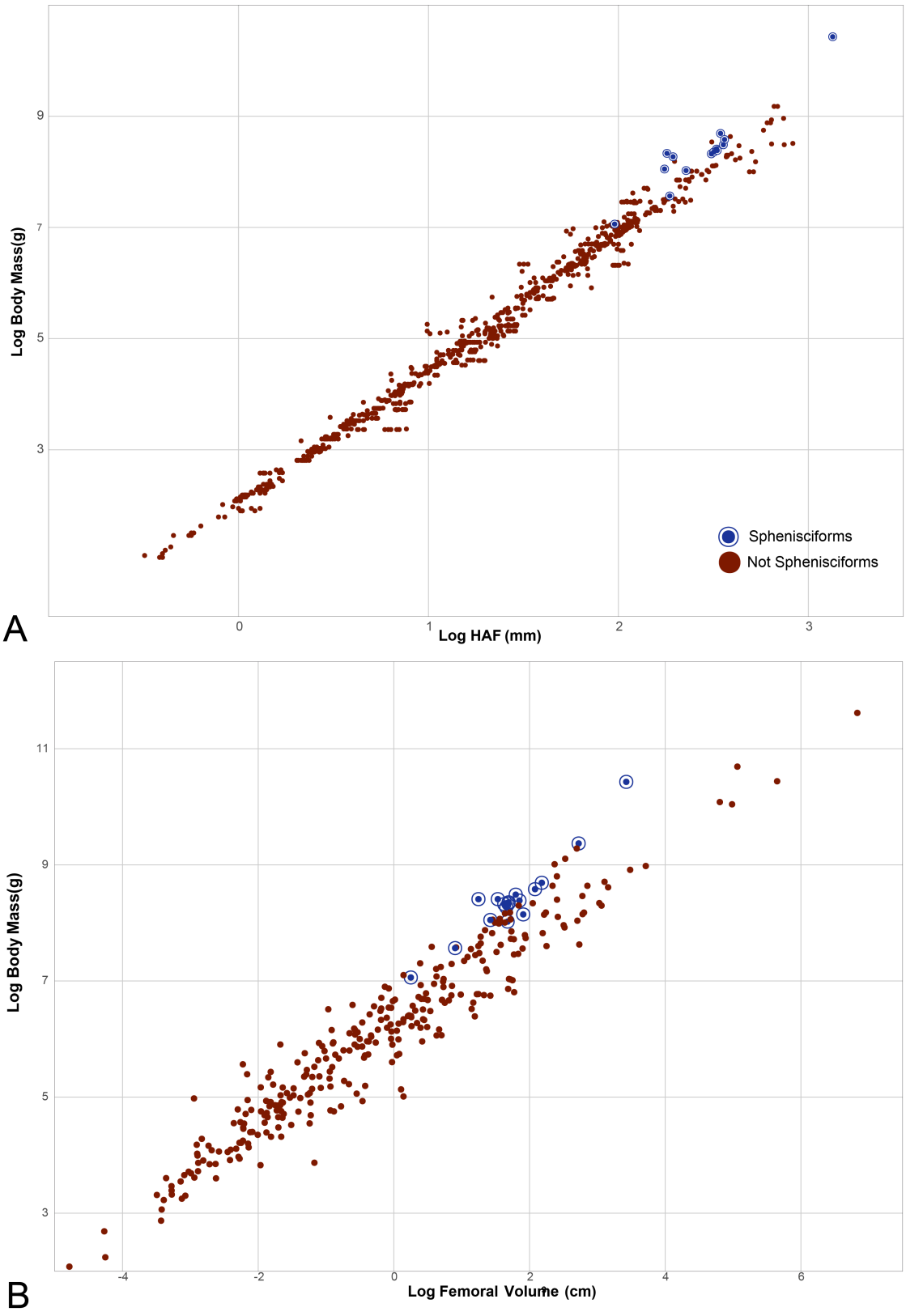


Figure 3.2: Relationship between log-transformed body mass (g) and A) log-transformed coracoid humeral articular facet (HAF) (mm) and B) log-transformed femoral volume (cm^3). Blue circles represent observations from penguins.

formes", "Pelecaniformes", "Phaethontiformes", "Phenicopteriformes", "Podicipediformes", "Procellariformes", "Psittaciformes", "Sphenisciformes", Strigiformes" and "Suliformes".

Ecological parameter used in Bayesian models

Ecological descriptors for the species used in this study derived from Pigot et al. (2020). Here the ecological parameters are referred to as the variable "Trophic Niche". Ten trophic niches were identified for the birds used in this study: "Aquatic predator", "Frugivore", "Granivore", "Aquatic herbivore", "Terrestrial herbivore", "Invertivore (i.e. carnivore that eats invertebrates)", "Nectarivore", "Omnivore", "Scavenger and "Vertivore (i.e. carnivore that eats vertebrates)".

Phylogeny

A phylogenetic supertree was constructed for use in several of the body mass prediction models (see Femur models VI, VII, VIII and IX below). The supertree was a maximum clade credibility (MCC) tree calculated from the pseudo-posterior sample of trees available from Jetz et al. (2012) (hereafter MCC-Jetz supertree). The MCC-Jetz supertree is the result of an extensive phylogenetic analysis based on sequence data and relaxed phylogenetic clock models. A subsample of 100 random trees was downloaded from birdtree.org, the online repository of the Jetz et al. (2012) results, pruned down to the 319 species in the femur dataset developed for this chapter (using the pruning option in the "taxonomy' section of birdtree.org). The root-to-tip distances of the MCC-Jetz supertrees were then calculated in `R` with the `adephylo::distRoot` (Jombart et al., 2010,v.1.1), as well as the root-to-tip distance of the Sphenisciformes clade. Sphenisciformes was subsequently pruned from the supertrees and substituted with the fossilised birth-death (FBD) tree calculated in Chapter 2. Branch lengths of the FBD tree were normalised to the root-to-tip distance of the Sphenisciformes clade pruned from the MCC-Jetz supertree computed in the previous step. The phylogenetic position of *Aptenodytes ridgeni* and 5 addi-

tional extinct taxa for which body masses will be estimated needed to be grafted onto the FBD tree prior to joining it to the supertree as these taxa were not present in the FBD tree analysis performed in Chapter 2. *Aptenodytes ridgeni* was assigned to the same phylogenetic and temporal position as in the analysis of Thomas et al. (2020), as a 0 length branch found at 6.3 million years ago topologically prior to the *Aptenodytes forsteri* and *Aptenodytes patagonicus* split. The other five fossils were added on a polytomy found at the base of the branching between the *Kairuku* clade and more crownward penguins. The original pruned MCC-Jetz supertree is presented in figure 3.3 along with the modified version obtained from the steps mentioned above.

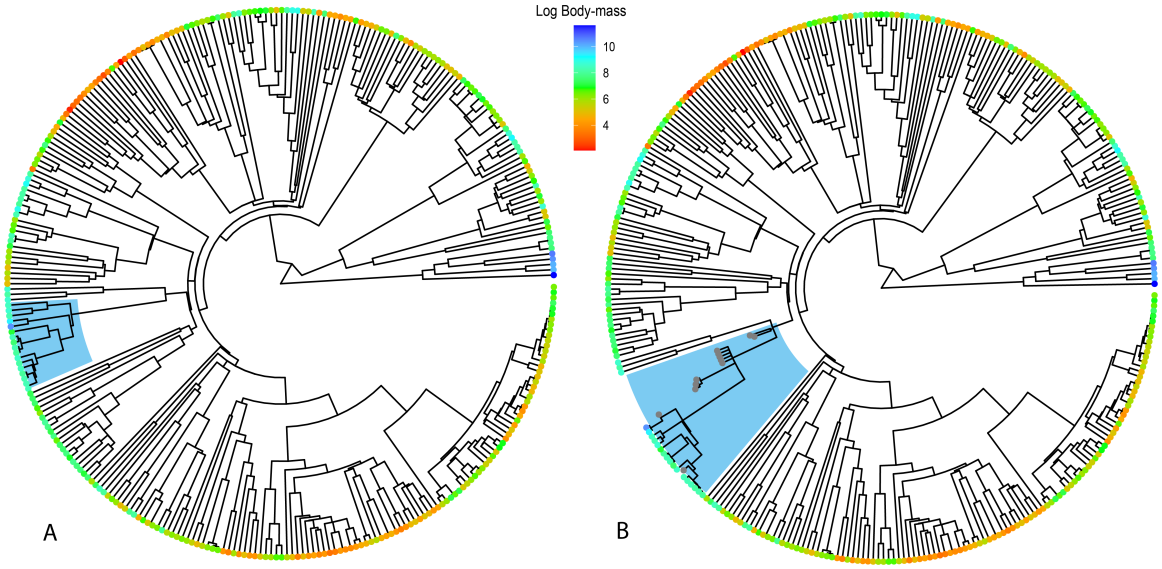


Figure 3.3: Bird phylogenies modified from Jetz et al. (2012). A) The maximum clade credibility (MCC) tree estimated from the 100 trees sampled from Jetz et al. (2012). B) The supertree used for analyses in this chapter including fossil tips. Tip color denotes the log of body mass. Sphenisciformes highlighted in blue.

3.2.2 Bayesian modelling

Humeral articular facet of the coracoid

As stated above, the aim of this chapter is to build statistical models that can predict the body mass of a bird using information from a bone. Here will follows the

description of the structures of two Bayesian models developed for predicting body mass from the humeral articular facet of the coracoid (HAF model I and HAF model II). Both models used the coracoid HAF and body mass datasets described above. These datasets were log transformed in `R` to distribute size variation more evenly across the range of observations. These models both used Markov chain Monte Carlo (MCMC) methods and were implemented in `Stan` (Carpenter et al., 2017) through the `R` package `brms` (Bürkner, 2017). Unless stated otherwise, the MCMC for HAF models were set to a total length of 9000 draws and samples were set to half of the chain total length (default options in `Stan`; Bürkner, 2017; Carpenter et al., 2017). Both models assumed that each body mass value represented a mean sampled from a normal distribution (Smith, 1973) (Eq. 3.1). The key difference between HAF model I and HAF model II was that only HAF model II included information about the taxonomic order of the birds in the analysis. After the parameters were estimated for each of model mean, standard deviation, and 89% credible intervals were calculated in `R`.

HAF model I (Coracoid HAF only) Designed to be the simplest model with HAF measurements as the only predictor. Estimates of mean body mass (μ_i) are performed using an equation in the form of a traditional linear model, with an intercept ($\alpha_{\text{Intercept}}$) added to a slope ($\beta_{\text{HAF}}$) that is multiplied by the area of the humeral articular facet of the coracoid (HAF_i) (N. E. Campione & Evans, 2020; Chambers, 1992)(Eq. 3.2). Priors for the intercept ($\alpha_{\text{Intercept}}$) and slope ($\beta_{\text{HAF}}$) were sampled from normal distributions with a mean of 0 and standard deviations of 2 and 1, respectively (Eqs. 3.3 and 3.4). Priors for the standard deviation (σ) associated with the body mass were sampled from an exponential distribution (Eq. 3.5). Sampling from an exponential distribution is common with Bayesian linear modelling approach because it ensures the positivity of the σ parameter and reduces the risk of biasing the analysis (Kruschke, 2015; McElreath, 2020).

$$\text{Body mass}_i \sim \text{Normal}(\mu_i, \sigma) \tag{3.1}$$

$$\mu_i = \alpha_{\text{Intercept}} + \beta_{\text{HAF}}\text{HAF}_i \tag{3.2}$$

$$\alpha_{\text{Intercept}} \sim \text{Normal}(0, 2) \tag{3.3}$$

$$\beta_{\text{HAF}} \sim \text{Normal}(0, 1) \tag{3.4}$$

$$\sigma \sim \text{Exponential}(1) \tag{3.5}$$

HAF model II (Order, random) Differs from HAF model I by including information about order-level groupings (Eq. 3.6), to observe if the inclusion of taxonomic information would substantially improve model results. As with HAF model I, estimates of mean body mass (μ_i) are performed using an equation in the form of a traditional linear model, with an intercept (α_{Order}) added to a slope (β_{HAF}) that is multiplied by coracoid HAF (HAF_i) (Eq. 3.6). However, HAF model II replaces the intercept with a numerical value used to represent the taxonomic order of the bird (α_{Order}) in the form of a multilevel approach of nested priors ($\overline{\alpha}$, σ_α) (Nalborczyk et al., 2019)(Eqs. 3.7–3.9). The multilevel approach of nested priors is the Bayesian equivalent to including a random effect into a generalised linear model, resulting in a model with varying intercepts (McElreath, 2020). What this translates into is that each avian order will have its own intercept α_{Order}, and all of these intercepts will be extracted from a normal distribution with mean $\overline{\alpha}$ and standard deviation σ_α (Eq. 3.7). The main difference with the distribution of priors in HAF model I is that instead of having fixed values, $\overline{\alpha}$ and σ_α are subjected to two further priors. $\overline{\alpha}$ is expected to be drawn from a normal distribution with mean 0 and standard deviation 2 (Eq. 3.8) and σ_α from an exponential distribution with mean 1 (Eq. 3.9).

$$\mu_i = \alpha_{\text{Order}} + \beta_{\text{HAF}}\text{HAF}_i \tag{3.6}$$

$$\alpha_{\text{Order}} \sim \text{Normal}(\overline{\alpha}, \sigma_\alpha) \tag{3.7}$$

$$\overline{\alpha} \sim \text{Normal}(0, 2) \tag{3.8}$$

$$\sigma_\alpha \sim \text{Exponential}(1) \tag{3.9}$$

Femur

As introduced above, femora support the weight of the bird and are therefore potentially informative elements for modelling body mass. Described here are the structures of nine Bayesian models developed for predicting the body mass of a bird from the total volume of the femur (Femur models I–IX). Models used the femur volume and bird body mass datasets described above, where both femur volume and body mass were log transformed in `R`. As above for the HAF models, the nine femur models used MCMC methods and were implemented in `Stan` (Carpenter et al., 2017) through the `R` package `brms` (Bürkner, 2017). The MCMC for the femur models were set to a total length of 9000 draws and samples were set to half of the chain total length as the default option for `Stan` (Bürkner, 2017; Carpenter et al., 2017). Unless stated otherwise, all femur models assumed that the body mass of each species was distributed normally around a mean (Eq. 3.10), where the mean is linked to the structure of a traditional linear model (Lindley & Smith, 1972; Smith, 1973). The relationship between femur volume and bird body mass appears to be less constrained than the relationship between HAF area and body mass (compare Fig. 3.2 A and B), allowing for the possibility that other potential variables could improve estimation accuracy. Hence, nine femur models were explored instead of just two models for the relationship between HAF and body mass. After the parameters were estimated for each of model mean, standard deviation, and 89% credible intervals were calculated in `R`.

Femur model I (Femur volume only) Designed to be the simplest model with femoral volume as the only predictor. Echoing HAF model I and equation 3.2, here estimates of mean body mass (μ_i) are performed using an equation in the form of a traditional linear model with an intercept ($\alpha_{\text{Intercept}}$) added to a slope ($\beta_{\text{HAF}}$) that is multiplied by femoral volume (Vol_i) (N. E. Campione & Evans, 2020; Chambers, 1992) (Eq. 3.11). Priors for the intercept (α) and slope (β_{Vol}) were sampled from normal distributions with a mean of 0 and standard deviations of 2 and 0.5, respectively (Eqs. 3.12 and 3.13). Priors for the standard deviation (σ) associated with the body mass were sampled from an exponential distribution (Eq. 3.14). All femur models used these default algorithms for generating priors except where stated otherwise.

$$\text{Body mass}_i \sim \text{Normal}(\mu_i, \sigma) \tag{3.10}$$

$$\mu_i = \alpha_{\text{Intercept}} + \beta_{\text{Vol}}\text{Vol}_i \tag{3.11}$$

$$\alpha_{\text{Intercept}} \sim \text{Normal}(0, 2) \tag{3.12}$$

$$\beta_{\text{Vol}} \sim \text{Normal}(0, 0.5) \tag{3.13}$$

$$\sigma \sim \text{Exponential}(1) \tag{3.14}$$

Femur model II (Order, fixed) Femur model II includes information about order-level groupings (Eq. 3.15) as a fixed intercept to observe if the inclusion of taxonomic information substantially improves model results. Estimates of mean body mass (μ_i) are made using a linear model. Taxonomic order is the intercept for the model and is represented by a numerical value (fixed intercept; α_{Order}) drawn from a normal distribution with a mean of 0 and a standard deviation of 2 (Eq. 3.16). The α_{Order} intercept is added to a slope (β_{Vol}) that is multiplied by femoral volume (Vol_i) (Eq. 3.15). Each order has a different intercept value estimated from the α_{Order} prior.

$$\mu_i = \alpha_{\text{Order}} + \beta_{\text{Vol}}\text{Vol}_i \tag{3.15}$$

$$\alpha_{\text{Order}} \sim \text{Normal}(0, 2) \tag{3.16}$$

Femur model III (Order, random) Femur model III is similar to femur model II but instead uses the multilevel method for calculating the intercept that was described for HAF model II to account for the variation among taxonomic orders (Nalborczyk et al., 2019)(Eqs. 3.17–3.19). Bird order was used as a predictor variable as for femur model II, but instead of a fixed intercept, femur model III uses varying intercepts, the Bayesian equivalent to random effects (Nalborczyk et al., 2019)(see description of HAF model II for more detail). Other than a greater flexibility in terms of modelling (Gelman et al., 2012), the hierarchical structure of femur model III should help also to avoid divergent transitions (i. e. chains that fail to sample correctly from the posterior) when sampling is performed during the MCMC process (Carpenter et al., 2017).

$$\alpha_{\text{Order}} \sim \text{Normal}(\overline{\alpha}, \sigma_\alpha) \tag{3.17}$$

$$\overline{\alpha} \sim \text{Normal}(0, 2) \tag{3.18}$$

$$\sigma_\alpha \sim \text{Exponential}(1) \tag{3.19}$$

Femur model IV (Trophic niche, fixed) Femur model IV took an approach similar to femur model II but instead of focusing on taxonomic information, this model used trophic niche to test whether the adaptation to a given niche could improve the overall model predictive power. Estimates of mean body mass (μ_i) are made using a linear model where a numerical value representing trophic niche ($\alpha_{\text{Trophic Niche}}$) is the intercept drawn from a normal distribution with a mean of 0 and a standard deviation of 2 (Eq. 3.20). The $\alpha_{\text{Trophic Niche}}$ intercept is added to a slope (β_{Vol}) that is multiplied by femoral volume (Vol_i) (Eq. 3.21).

$$\mu_i = \alpha_{\text{Trophic Niche}} + \beta_{\text{Vol}}\text{Vol}_i \tag{3.20}$$

$$\alpha_{\text{Trophic Niche}} \sim \text{Normal}(0, 2) \tag{3.21}$$

Femur model V (Trophic niche, random) Femur model V again explored trophic niche information as a predictor variable. As in femur model III here ecological information was used in a multilevel model to approximate the inclusion as varying intercepts (Eqs. 3.22–3.24), with each trophic niche having a different value drawn from nested priors (see femur model III and HAF model II for more detail).

$$\alpha_{\text{Trophic Niche}} \sim \text{Normal}(\overline{\alpha}, \sigma_\alpha) \tag{3.22}$$

$$\overline{\alpha} \sim \text{Normal}(0, 2) \tag{3.23}$$

$$\sigma_\alpha \sim \text{Exponential}(1) \tag{3.24}$$

Femur model VI (Brownian motion) Femur model VI tested whether phylogenetic information given as a tree topology would provide a more robust explanation of inter-specific differences in relationships between body mass and femur volume when compared with using order as a numerical value (i.e. femur model II and III). Femur model VI took a more complex approach than the fixed discrete character in femur model II and III and instead describes correlation induced by the phylogeny as an **R** matrix (Eq. 3.27). The **R** matrix is derived from the tree topology above specified with the function `ape::corBrownian` (Paradis and Schliep, 2019, v. 5.0) in the `R` programming language.

The **R** matrix is a symmetric square matrix of size $n \times n$ where n is the number of total taxa in the dataset with each cell representing the grade of phylogenetic relatedness among each pair of taxa. This **R** matrix represents the amount of correlation under the assumption of a Brownian motion model of evolution (Martins & Hansen, 1997). Body mass was assumed to be distributed on a multivariate

normal, an extension in vector form of a traditional normal distribution (Symonds & Blomberg, 2014) (Eq. 3.25). A covariance matrix Σ is derived from the $\mathbf{R}$ matrix multiplied by a diagonalised σ matrix (Garland & Ives, 2000; Martins & Hansen, 1997) (Eq. 3.26). Mean body mass (μ_i) was estimated using a linear model as specified in femur model I (Eq. 3.28). Note that body mass estimates of extinct taxa were first included as missing data in femur model VI (also in femur models VII–IX) before running the models using the `brms` package in R to estimate the body masses of extinct taxa. Consequently, body mass estimations and model parameters are estimated jointly during the MCMC.

$$\text{Body mass}_i \sim \text{MVNormal}(\mu_i, \Sigma) \tag{3.25}$$

$$\Sigma = \begin{bmatrix} \sigma & 0 & 0 & \dots & 0 \\ 0 & \sigma & 0 & \dots & 0 \\ 0 & 0 & \sigma & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma \end{bmatrix} \mathbf{R} \begin{bmatrix} \sigma & 0 & 0 & \dots & 0 \\ 0 & \sigma & 0 & \dots & 0 \\ 0 & 0 & \sigma & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma \end{bmatrix} \tag{3.26}$$

$$\mathbf{R} = \begin{bmatrix} 1 & \text{cor}_{1,2} & \text{cor}_{1,3} & \dots & \text{cor}_{1,n} \\ \text{cor}_{2,1} & 1 & \text{cor}_{2,3} & \dots & \text{cor}_{2,n} \\ \text{cor}_{3,1} & \text{cor}_{3,2} & 1 & \dots & \text{cor}_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cor}_{n,1} & \text{cor}_{n,2} & \text{cor}_{n,3} & \dots & 1 \end{bmatrix} \tag{3.27}$$

$$\mu_i = \alpha_{\text{Intercept}} + \beta_{\text{Vol}} \text{Vol}_i \tag{3.28}$$

Femur model VII (Trophic niche, Brownian motion) Femur model VII is an increase in complexity compared with previous models. Here, femur model VII tests whether both phylogenetic information and trophic niche information together provide a more robust explanation of inter-specific variation in the relationships between body mass and femur volume. The covariance structure of femur model

VII remains identical to femur model VI, but the linear part of the model (Eq. 3.29) includes the trophic niche for each bird as a varying intercept as seen in femur model V.

$$\mu_i = \alpha_{\text{Trophic Niche}} + \beta_{\text{Vol}}\text{Vol}_i \tag{3.29}$$

$$\alpha_{\text{Trophic Niche}} \sim \text{Normal}(\overline{\alpha}, \sigma_\alpha) \tag{3.30}$$

$$\overline{\alpha} \sim \text{Normal}(0, 2) \tag{3.31}$$

$$\sigma_\alpha \sim \text{Exponential}(1) \tag{3.32}$$

Femur model VIII (Ornstein-Uhlenbeck) The aim of femur model VIII was to observe whether a more complex model of trait evolution could provide an improved explanation for the variation between body mass and femur volume. Body mass is estimated using a multivariate normal distribution using a **K** matrix instead of using the fixed covariation **R** matrix as in femur model VII (McElreath, 2020)(Eq. 3.34 versus 3.27). Each cell of this square $n \times n$ symmetric **K** matrix will be proportional to the phylogenetic distance occurring among each pair of taxa, computed from the topology of the phylogenetic tree specified with the `R` command `ape::cophenetic` (Paradis & Schliep, 2019) (Eq. 3.34). The rate of covariation decay requires the estimation of two additional parameters η and ρ, in that manner this formalisation reflect an Ornstein-Uhlenbeck model of trait evolution as a Gaussian process (Lande, 1976)(Eqs. 3.36 and 3.37). As for femur model VI, the mean of the multivariate normal μ is linked to a traditional linear model (Eq. 3.35).

$$\text{Body mass}_i \sim \text{MVNormal}(\mu_i, \mathbf{K}) \tag{3.33}$$

$$\mathbf{K} = \begin{bmatrix} \eta^2 & \eta^2 e^{\rho^2 \text{D}_{1,2}} & \eta^2 e^{\rho^2 \text{D}_{1,3}} & \dots & \eta^2 e^{\rho^2 \text{D}_{1,n}} \\ \eta^2 e^{\rho^2 \text{D}_{2,1}} & \eta^2 & \eta^2 e^{\rho^2 \text{D}_{2,3}} & \dots & \eta^2 e^{\rho^2 \text{D}_{2,n}} \\ \eta^2 e^{\rho^2 \text{D}_{3,1}} & \eta^2 e^{\rho^2 \text{D}_{3,2}} & \eta^2 & \dots & \eta^2 e^{\rho^2 \text{D}_{3,n}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta^2 e^{\rho^2 \text{D}_{n,1}} & \eta^2 e^{\rho^2 \text{D}_{n,2}} & \eta^2 e^{\rho^2 \text{D}_{n,3}} & \dots & \eta^2 \end{bmatrix} \tag{3.34}$$

$$\mu_i = \alpha_{\text{Intercept}} + \beta_{\text{Vol}} \text{Vol}_i \tag{3.35}$$

$$\eta \sim \text{Half Normal}(1, 0.25) \tag{3.36}$$

$$\rho \sim \text{Half Normal}(3, 0.25) \tag{3.37}$$

Femur model IX (Trophic niche, Ornstein-Uhlenbeck) Femur model IX is a more complex version of femur model VIII that includes trophic niche information alongside trait evolution modelling. As in the previous model here an Ornstein-Uhlenbeck model of covariation was implemented with the addition of trophic niche as a predictor variable formalised as varying intercept, see femur models V and VII for more information. Note that although the set of equations used here (Eq. 3.35 to Equation 3.38) is identical to the set in femur model VII, here the model differs in how the covariance among observations is treated. In femur model VII covariance is represented by the Σ matrix (Eq. 3.26) whereas here the covariance is represented by the $\mathbf{K}$ matrix (Eq. 3.34).

$$\mu_i = \alpha_{\text{Trophic Niche}} + \beta_{\text{Vol}} \text{Vol}_i \tag{3.38}$$

$$\alpha_{\text{Trophic Niche}} \sim \text{Normal}(\overline{\alpha}, \sigma_\alpha) \tag{3.39}$$

$$\overline{\alpha} \sim \text{Normal}(0, 2) \tag{3.40}$$

$$\sigma_\alpha \sim \text{Exponential}(1) \tag{3.41}$$

Mass estimation from HAF and femur models

Body mass for *Eudyptes atatu*, *Kairuku grebneffi*, *Kairuku waewaeroa*, *Kairuku waitaki*, *Kumimanu biceae*, *Kupoupou stilwelli*, *Pachydyptes ponderosus*, *Platydyptes marplesi*, *Sequiwaimanu rosieae* was estimated using both the HAF models, and body mass for *Aptenodytes ridgeni*, *Eudyptes atatu*, *Kairuku grebneffi*, *Kairuku waewaeroa*, *Kairuku waitaki*, *Muriwaimanu tuatahi*, *Palaeeudyptes* sp. (GL429 Burnside), *Sequiwaimanu rosieae*, and four Sphenisciformes indet. (Seymour OU22195, Seymour REF11, Seymour REF31 and DM1449 the Seal rock "*Palaeeudyptes*") were estimated using each of the femur models. Note that body mass estimates were made for five species (*Eudyptes atatu*, *Kairuku grebneffi*, *Kairuku waewaeroa*, *Kairuku waitaki* and *Sequiwaimanu rosieae*) using both femur and HAF models. Body mass predictions were made with the command `brms::predict.brmsfit` (Bürkner, 2017).

3.2.3 Model evaluation

PSIS-LOO

Model reliability was determined by calculating both log pointwise predicted densities values (ELPPD, Vehtari et al., 2017), as well as Pareto smoothed importance-sampling leave-one-out cross-validation values (PSIS-LOO, Vehtari et al., 2017). Whereas ELPPD are estimated from the log likelihood for each observation given the model parameters, the PSIS-LOO is an information criterion that is analogous to the popular Akaike information criterion (AIC), the Generalized Information criterion (GIC) and the Watanabe-Akaike information criterion (WAIC)(Merkle et al., 2019). For traditional leave-one-out cross validation (LOO) the often computationally-expensive models that are being evaluated need to be re-fitted multiple times in order to estimate overall accuracy (Zhang & Yang, 2015). In contrast, the main advantage of PSIS-LOO is that a close estimate of LOO can be produced without refitting the whole model. With PSIS-LOO the performance of each model is com-

pared and the model that performs better in predicting the result is assigned a score of 0. Models that performed less well because they are over- or under-parameterised are assigned lower values (McElreath, 2020).

Body mass re-estimation

In addition to PSIS-LOO, model performance was determined by estimating body masses for a test set of 24 femora from extant species for which the mean body mass is known. These 24 species were excluded from the study and each of the nine femur models were used to predict body mass using the available information. Given that the main focus for this chapter is to test accuracy for Sphenisciformes, the test set included a large penguin (king penguin *Aptenodytes patagonicus*), a medium-sized penguin (erect-crested penguin *Eudyptes sclateri*), and a small penguin (little penguin *Eudyptula minor*) to cover most of the size range for extant penguins. A set of 22 taxa were then randomly assigned to the test set, including: parakeet auklet *Aethia psittacula*, razorbill *Alca torda*, roroa great spotted kiwi *Apteryx haastii*, cattle egret *Bubulcus ibis*, Chatham Island snipe *Coenocorypha pusilla*, metallic pigeon *Columba vitiensis*, wandering albatross *Diomedea exulans*, white-face heron *Egretta novaehollandiae*, nankeen kestrel *Falco cenchroides*, kelp gull *Larus dominicanus*, wonga pigeon *Leucosarcia melanoleuca*, Melanesian megapode *Megapodius eremita*, house sparrow *Passer domesticus*, Otago shag *Phalacrocorax chalconotus*, Bounty shag *Phalacrocorax ranfurlyi*, takahē *Porphyrio hochstetteri*, Cassin's auklet *Ptychoramphus aleuticus*, piwakawaka New Zealand fantail *Rhipidura fuliginosa*, common tern *Sterna hirundo*, buff-breasted sandpiper *Tryngites subruficollis*, painted buttonquail *Turnix varius*.

3.3 Results

3.3.1 HAF modelling

Overall results remain virtually unchanged from Field et al. (2013) with a β for HAF model I estimated at 2.45 ± 0.01 and an $\alpha_{\text{Intercept}}$ of 1.99 ± 0.01 (2.44 and 2.0 respectively in Field et al., 2013). Hence, the addition of coracoid HAF measurements from penguins to the Field et al. (2013) dataset had negligible impact on the model fit (Fig. 3.4B) and both the Bayesian and maximum likelihood methods converge to the same results. β parameters remain almost identical across both HAF model I and HAF model II (Table B.1, Fig. 3.4B) with major differences referring mainly to the estimated intercepts of Sphenisciformes. Whereas the $\alpha_{\text{Intercept}}$ for HAF model I is 1.99 ± 0.01 and the $\overline{\alpha}$ of HAF model II is 2.04 ± 0.04, implying that the grand mean of the intercept is almost unchanged among the two models, the $\alpha_{\text{Sphenisciformes}}$= 2.44 ± 0.01 is significantly greater, with negligible overlap among the distribution of these parameters (Tables B.1 and 3.4). Consequently, the intercept for penguins (HAF model II) lies clearly on a different value as predicted from the different medium in which they fly (i.e. the different structure constraint applied to coracoid HAF). HAF model II provides a better explanation for the variation between coracoid HAF and body mass compared with HAF model I (Table 3.1), given the substantial separation among ELPDs with no overlap among distributions. Recall that HAF model II includes both HAF measures and bird order as predictor variables, meaning that taxonomic information substantially increases the goodness of fit of the model.

Table 3.1: Estimated log pointwise densities (ELPD) differences from the leave one one (LOO) estimates between the better model and the target model (ELPD differences) along with standard error (ELPD Standard error for models based on HAF). Higher values in ELPD differences should reflect overall better performing models in relation to their number of parameters (Vehtari et al., 2017).

	ELPD differences	**ELPD Standard error**
HAF model II	0	0
HAF model I	-151.6	19.8

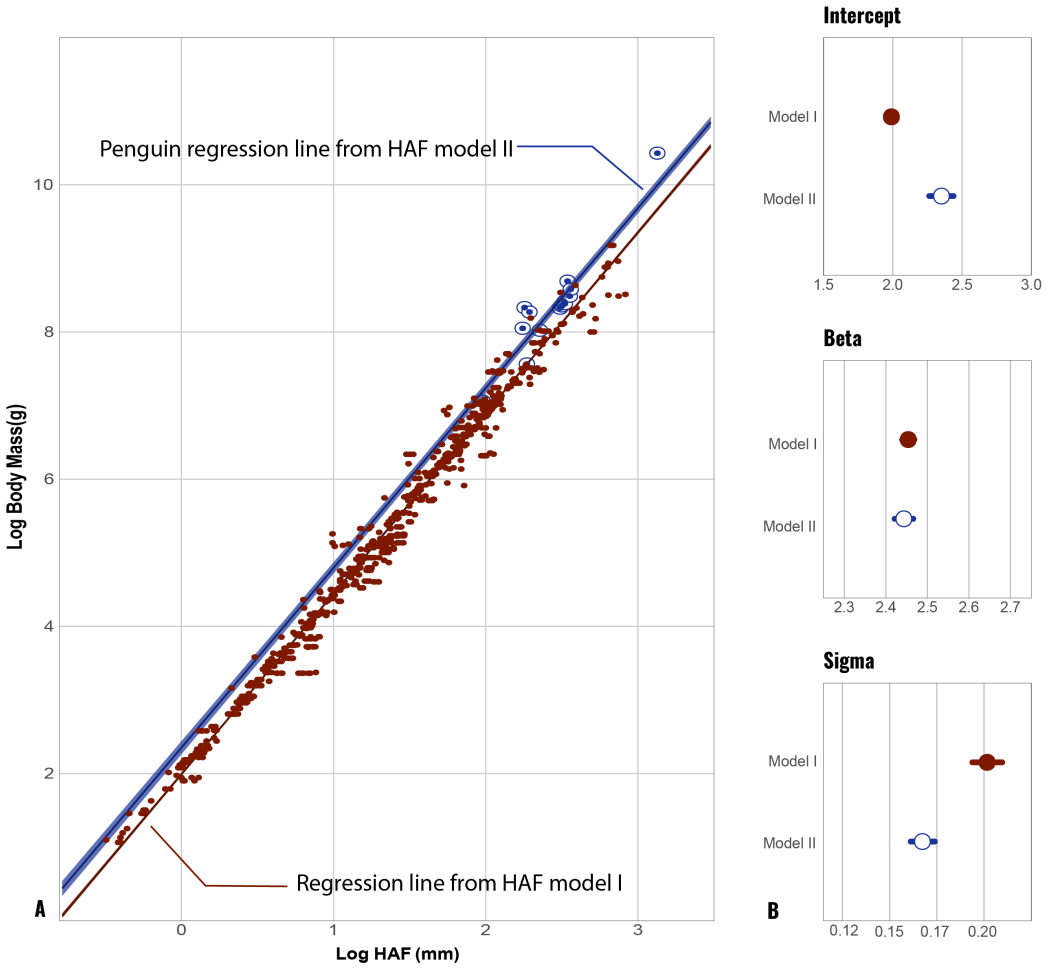


Figure 3.4: A) Relationship between the length of the log-transformed coracoid humeral articular facet (HAF; mm) and log-transformed body mass (grams). Blue circled points represent measurements from penguins whereas red dots represent all other bird groups. Red line is the fit estimated from HAF model I whereas the blue line represents the Sphenisciformes fit for HAF model II. B) Parameter estimates for HAF models I and II. Horizontal lines represent 89% credible interval red dots represent parameters from HAF model I empty blue dots represent parameters from HAF model II.

3.3.2 Femur modelling

β parameters from femur models I–V are around 0.8 with a substantial drop to roughly 0.3 for femur models VI and VII (Brownian motion) and a subsequent increase to 0.7 for femur models VIII and IX (Ornstein Uhlenbeck) (Fig. 3.5, Table B.1). A similar pattern emerges also for intercepts α with all models exhibiting values around 6.0-6.9 except for femur models VI and VII that have lower α values

between 3.1-5.4. σ parameters are found between 0.3 and 0.4 for all models except again for Brownian motion models VI and VI that have values around 2.4-2.6. Femur models VI and VII (i.e. traits modelled across a phylogeny according to Brownian motion) show greater uncertainty in parameter estimates, with increased credibility intervals, and with a substantial drop of β and α and an increase of σ (light blue dots in Fig. 3.5, Table B.1). Femur models VIII and IX (i.e. traits modelled across a phylogeny with Ornstein-Uhlenbeck) show a substantial reduction in uncertainty ranges compared to Brownian Motion models but still the distribution of estimated parameters is greater than what is observed in all other models (Blue dots in Fig. 3.5, Table B.1).

Table 3.2: Estimated log pointwise densities (ELPD) differences from the leave one one (LOO) estimates between the better model and the target model (ELPD differences) along with the standard error (ELPD Standard error for models based on Femur Volume). Higher values in ELPD differences should reflect overall better performing models in relation to their number of parameters (Vehtari et al., 2017).

	ELPD differences	**ELPD standard error**
Femur model III (Order, random)	0	0
Femur model II (Order, fixed)	-0.5	2.1
Femur model V (Ecology, random)	-51.9	12.6
Femur model IV (Ecology, fixed)	-54.3	12.6
Femur model I (Femur volume only)	-100.8	14.9
Femur model IX (Ornstein-Uhlenbeck + Ecology)	-226.3	19.9
Femur model VII (Ornstein Uhlenbeck)	-230.6	19.9
Femur model VII (Brownian motion + Ecology)	-341.7	21.9
Femur model VI (Brownian motion)	-341.9	21.9

Femur model III includes both volume and taxonomic order as random intercepts and provides the best explanation for the variation between femur volume and body mass when compared with each other femur model in this study (estimated PSIS-

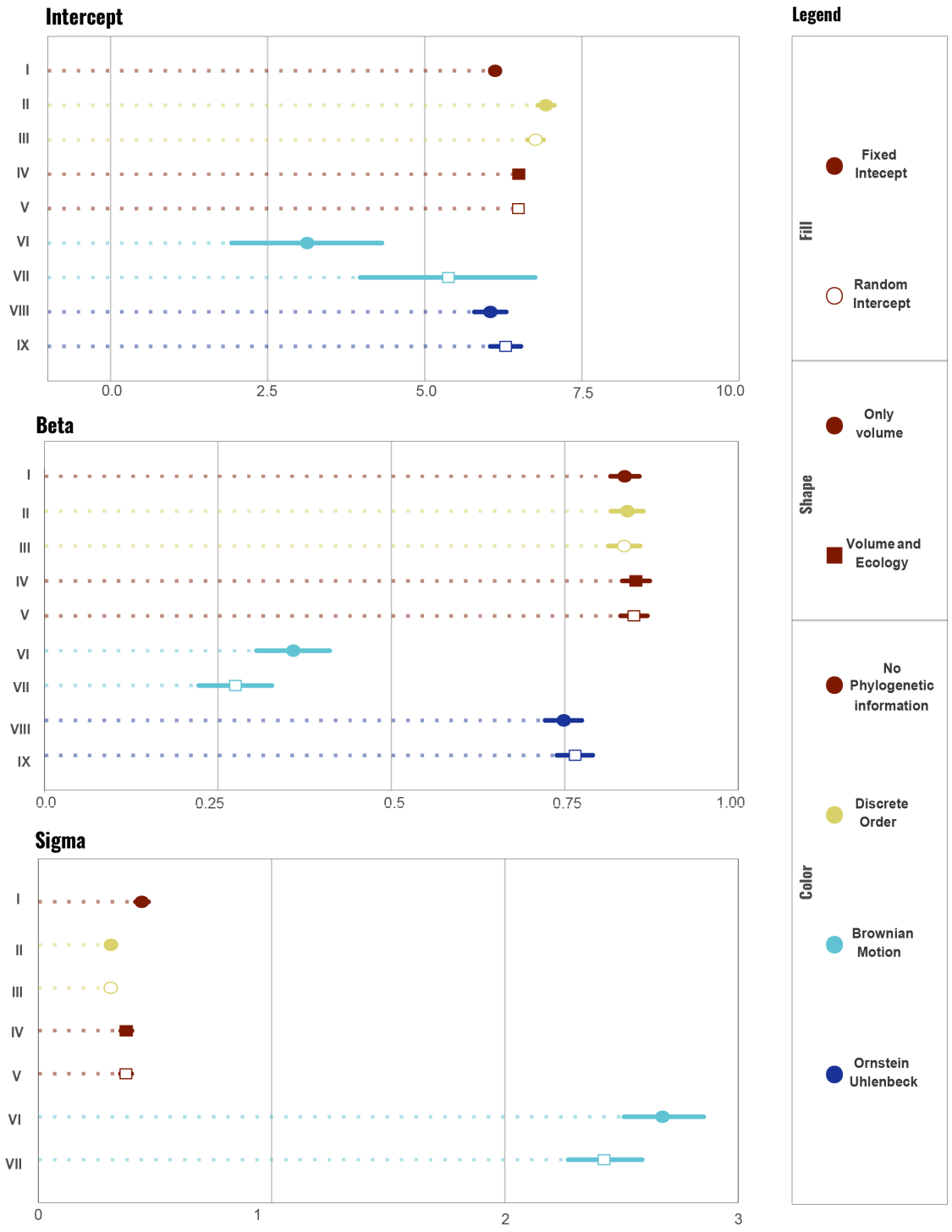


Figure 3.5: Parameter estimates for femur models. Each parameter is presented in its own box, each point represents the estimated average of the parameter value and lines describe the 89% credible interval of parameter distribution. Color and shapes coding are illustrated in the legend on the right.

LOO; Table 3.2). However, given that femur models II and III exhibit similar ELPD values distributions, we cannot assume that femur model III is substantially better than femur model II. Rather, we can instead be confident that femur model III outperforms all other models, given the substantial gap between ELPD differences and standard errors. (Table 3.2). Femur model VI and VII (i.e. traits modelled across a phylogeny according to Brownian motion) and femur model VIII and IX (i.e. Ornstein-Uhlenbeck) were not as successful at explaining the relationship between body mass and femur volume.

3.3.3 Test set model validation

Body mass estimates for the 24 extant species used to test each model recall the pattern identified by PSIS-LOO estimates with femur models II and III producing the most accurate estimates (Fig. 3.6-3.8, Table B.2). With respect to figures 3.6-3.8, known body mass for each taxon is depicted as a vertical line and each dot represents the mean from the posterior, colored according to the type of model. Body mass estimates from femur models VI and VII (i.e. Brownian motion models; Fig. 3.6-3.8 light blue lines, Table B.2) show the greatest uncertainty in body mass estimates. The uncertainty seems to be reduced in the Ornstein-Uhlenbeck framework (Fig. 3.6-3.8, dark blue lines Table B.2) but not substantially. Focusing on penguins, the best fitting femur models II and III slightly under-estimated body masses for *Aptenodytes patagonicus* (10.26 kg estimated versus 11.72 kg) and slightly over-estimated body masses for *Eudyptula minor* (1.26 kg estimated versus 1.16 kg observed) and *Eudyptes sclateri* (5.0 kg estimated versus 3.45 kg observed). Estimates for other birds exhibited overall a similar pattern with femur models II and III being the ones with the closest estimates to the observed values. It is also worth noting that uncertainty on estimates for model II and III tend to increase when there are limited observations for any given avian order, as can be seen for the *Falco cenchroides* (Falconiformes). Excluding *Falco cenchroides* during the model evaluation step mean that the model included only one species within Falconiformes

(*Falco novaeseelandiae*), and hence the greater uncertainty with $\alpha_{\text{Falconiformes}}$.

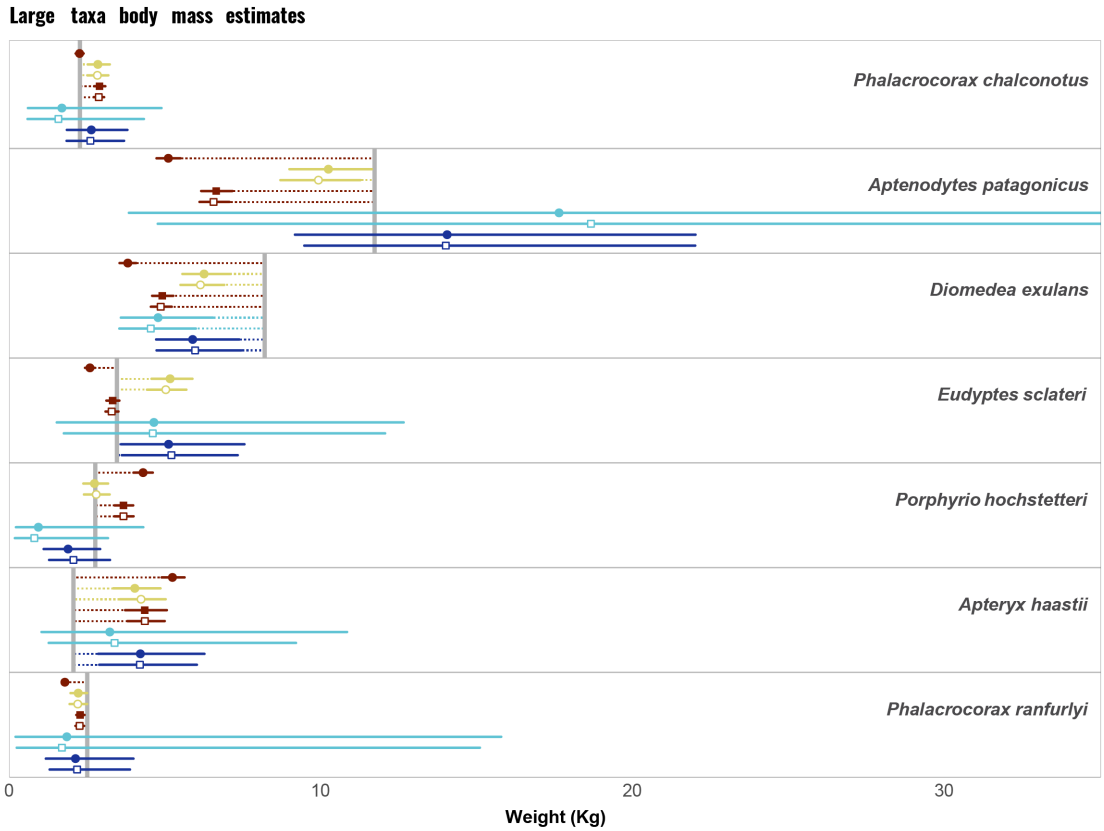


Figure 3.6: Body mass estimates for the largest taxa (greater than 2 kg) for the femur validation test set. Each point represents the mean of the estimated weight and lines describe the 89% credible interval of parameter distribution. All nine model estimates are represented and grouped for the species. Colour and shape codings follow the legend in figure 3.5. Briefly, red dots represent femur model I, IV and V (i.e. no taxonomic information), yellow represents femur model II and III (i.e. taxonomic order defined as a discrete variable), light blue represents femur model VI and VII (i.e. traits modelled using Brownian motion), and dark blue represents femur model VII and IX (i.e. traits modelled using Ornstein-Uhlenbeck). Vertical gray lines represent the observed weight for each taxon and dotted line connects the median to the observed weights.

3.3.4 Fossil estimates

Comparing body mass estimates for fossil penguins from HAF models I and II shows unsurprisingly that including taxonomic information (i.e. HAF model II) consistently increases body mass estimates for fossil penguins (Table 3.3 and Fig. 3.10). Such a pattern is a direct consequence of the differences between $\alpha_{\text{Intercept}}$

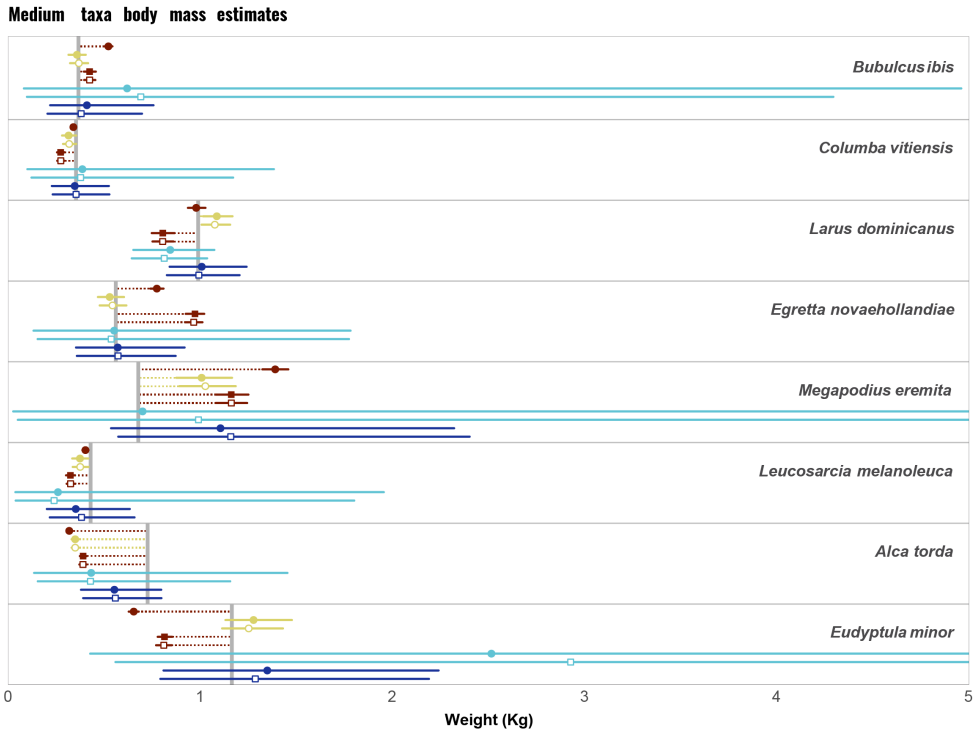


Figure 3.7: Body mass estimates for medium-sized taxa (ranging from 300 g up to 2 kg) for the femur validation test set. See figure 3.6 caption for interpretation.

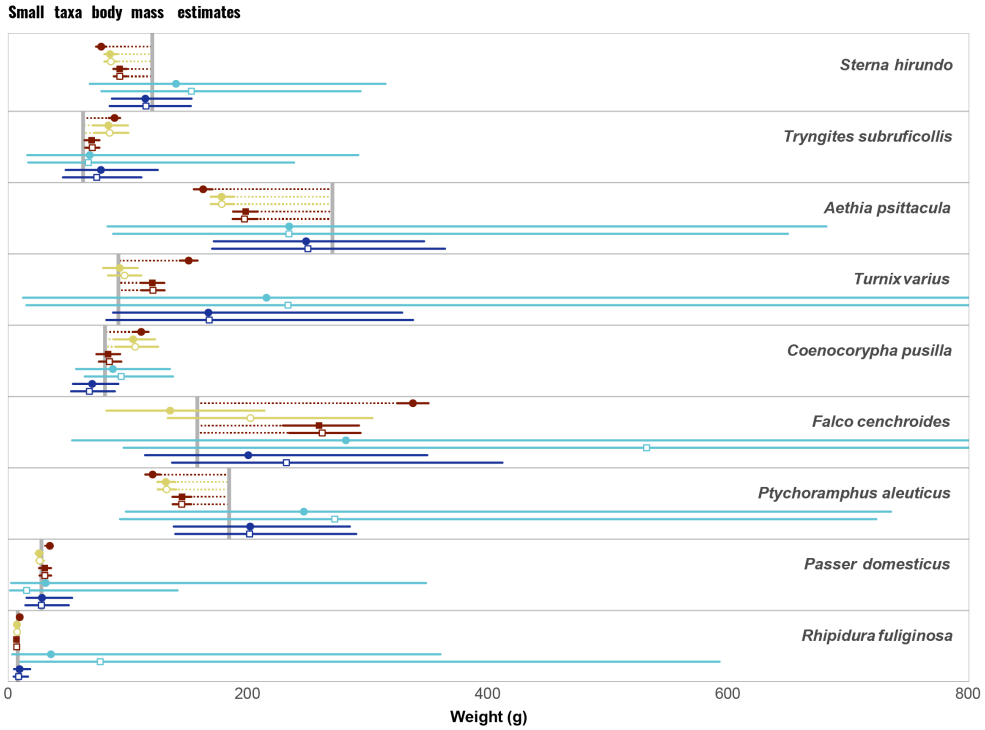


Figure 3.8: Body mass estimates for small-sized taxa (lower than 300 g) for the femur validation test set. See figure 3.6 caption for interpretation.

(HAF I) and $\alpha_{\text{Sphenisciformes}}$ (HAF II) explained above. The lowest body mass estimates from HAF model II are attributable to *Eudyptes atatu* $\sim$ 3.2 kg whereas *Kumimanu biceae* seems to be the largest among all sampled fossils $\sim$ 79.9 kg. One critical aspect to note is the relatively reduced size attributed to *Kairuku* penguins, with *Kairuku grebneffi* and *Kairuku waitaki* expected to have a mean body mass of 34.1 and 32.2 kg respectively, which is close to an emperor penguin even though *Kairuku* are morphologically clearly greater in size (humeral length of *Kairuku grebneffi* 177.6 mm compared with 121.9 mm for *Aptenodytes forsteri*; Giovanardi et al., 2021; Ksepka et al., 2012; Thomas and Ksepka, 2016).

Table 3.3: Weight estimates for the HAF models for a set of extinct fossil penguins. Measurements are expressed in kilograms (kg), mean represents the average from the posterior weights estimates, whereas upper and lower denote the 89% credible interval.

Species	HAF model	Mean	Lower	Upper
Eudyptes atatu	I	2.3	2.2	2.3
	II	3.2	2.9	3.4
Kairuku grebneffi	I	24.6	23.8	25.3
	II	34.1	31.6	36.7
Kairuku waewaeroa	I	33.5	32.5	34.6
	II	46.4	43.1	50.0
Kairuku waitaki	I	23.2	22.5	23.9
	II	32.2	29.9	34.6
Kumimanu biceae	I	57.8	55.8	59.9
	II	79.9	73.9	86.1
Kupoupou stilwelli	I	4.0	4.0	4.1
	II	5.7	5.3	6.1
Pachydyptes ponderosus	I	44.4	42.9	45.9
	II	61.4	56.9	66.1
Platydyptes marplesi	I	3.3	3.3	3.4
	II	4.7	4.3	5.0
Sequiwaimanu rosieae	I	14.5	14.1	14.9
	II	20.2	18.8	21.7

Body mass estimates for the nine femur models for fossil taxa (Fig. 3.9, Table B.3) show a similar pattern of distribution to what is observed in the validation section in term of uncertainty for phylogenetic models (Fig. 3.6-3.8, Table B.2). The uncertainty of the Brownian Motion methods is even more pronounced here with only relatively recent taxa (*Eudyptes atatu* and *Aptenodytes ridgeni*) showing

credible intervals in a reasonable range (Fig. 3.9, light blue dots). Moreover these models exhibit a drop in body mass estimates for older taxa. Estimates of body mass for fossil taxa from femur models compared with estimates from HAF models tend to give consistent results (Fig. 3.10) except for *Sequiwaimanu rosieae*.

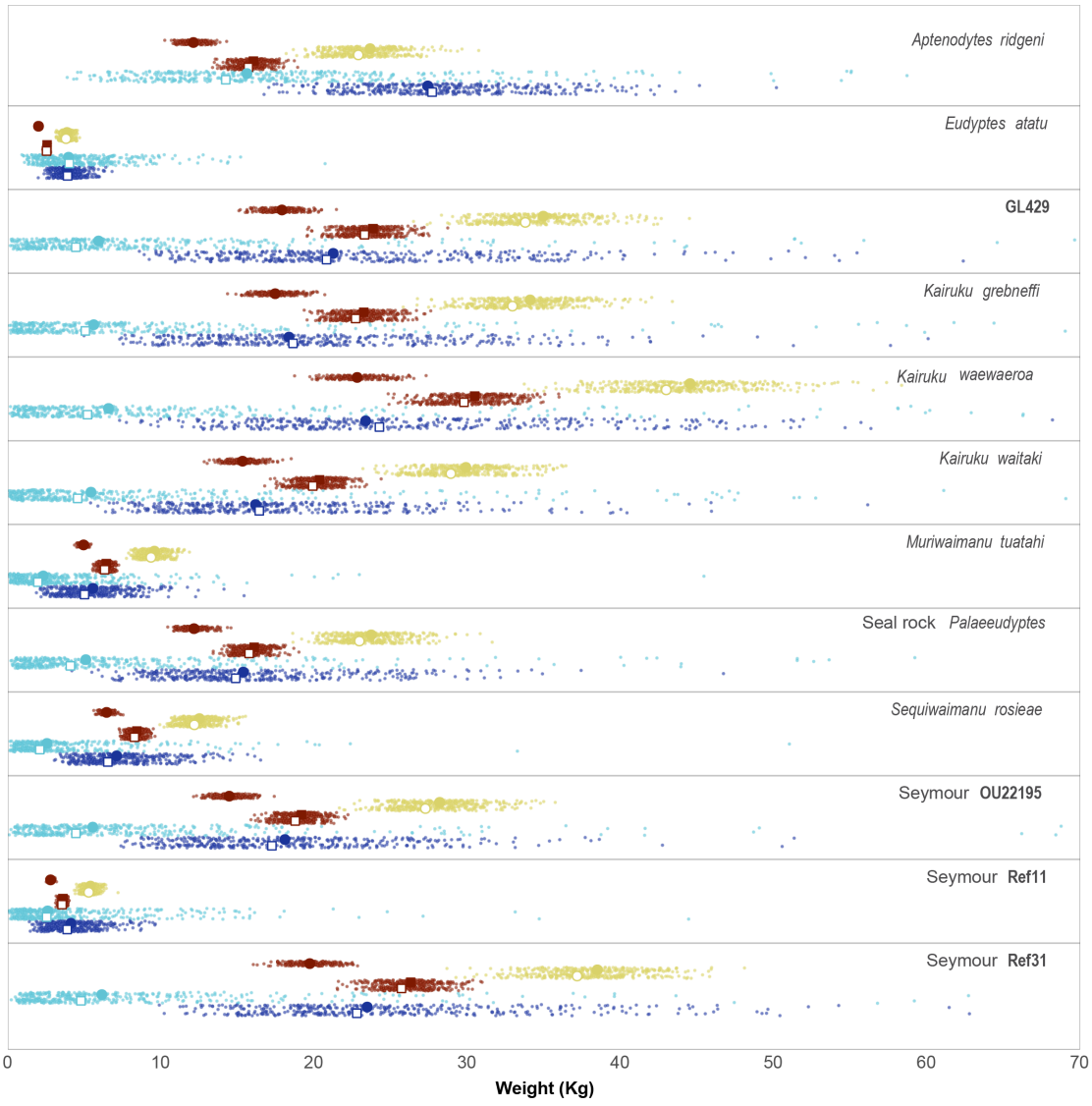


Figure 3.9: Body mass estimates of the 12 fossil specimens. Each semi-transparent small point represents a weight drawn from the posterior of each model. Bigger points represent the median of each distribution. Colour and shape codings follow the legend in figure 3.5. Briefly, red dots represent models with no taxonomic information, yellow with taxonomic order defined as discrete variables, light blue show Brownian motion models, and dark blue show Ornstein-Uhlenbeck models.

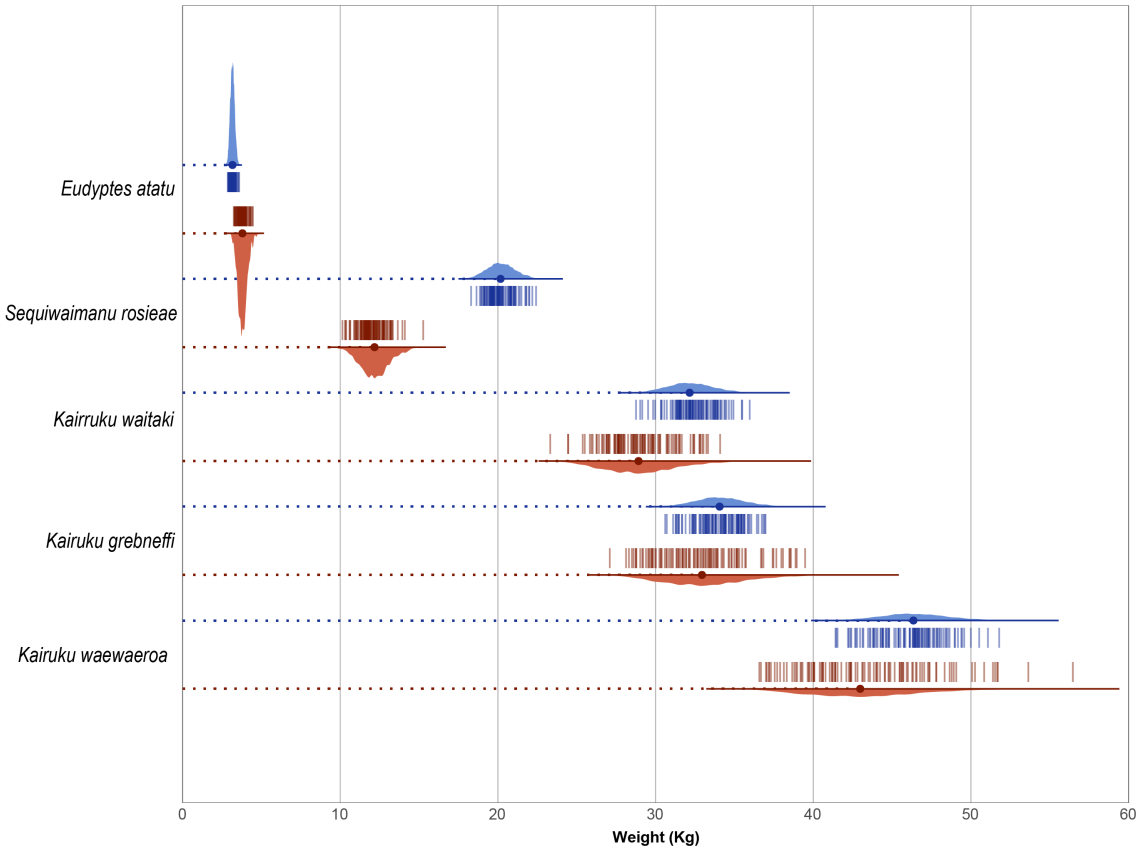


Figure 3.10: Raincloud plot of body mass estimates compared between humeral articular facet (HAF) and femur. Body mass distribution estimated from best performing models (i.e. blue: HAF model II; red: femur model III). Smaller coloured lines show a sample of 100 draws from each model.

3.4 Discussion

The use of Bayesian methods to estimate the body mass of birds showed that both the length of the coracoid humeral articular facet (HAF) and femur volume can be a valid proxy for mass estimation. However, depending on the type of information added to the model as a predictor variable, the results and performance of the model do vary (where performance is measured as the ability for the model to explain the variance in the dataset). Models that used taxonomic order as a categorical value tended to perform the best (e.g. femur model II and III), and these models were used to generate body mass estimates for extinct penguins.

HAF measurements from penguins were added to the Field et al. (2013) dataset which had only previously included data from any aerially-volant birds. The ad-

dition of data from penguins did not substantially change the general relationship between HAF and body mass (HAF model I; Fig. 3.4). Penguins by themselves have a higher intercept for the relationship between HAF and body mass (i.e. high initial body mass) when compared with the combined dataset of all other birds. Importantly though, the gradient for the HAF and body mass relationship in penguins is approximately similar to the gradient for birds in general, implying that HAF measurements do represent a structural constraint for penguins. As suggested by Simpson (1953), one may interpret the phenomenon as an example of different "adaptive ridges", where birds in different environments (i.e. aerial vs. marine) both increase HAF size with increasing body mass but from a different baseline to the relationship.

Model parameters show that femur volume and body mass grow in a similar scale as predicted (Fig. 3.5)(β close to 1.0) whereas instead body mass grows at as a cubic function of HAF measurements (β close to 2.0). Such a scaling pattern emerges also from previous studies, with circumferences of femur, and humerus for quadrupeds (N. E. Campione & Evans, 2020; Field et al., 2013). There is substantial uncertainty for the inferred parameters (and consequently body mass estimates) of Brownian Motion models (femur model VI and VII). The results from the Brownian Motion models may imply that residuals for these models are more dispersed and do not reflect the pattern inferred by the phylogeny. Brownian motion models are mostly informed by the phylogenetic covariance, as seen by the decrease with the parameter β estimates, meaning that at greater phylogenetic distance the model is not able to correctly estimate the body mass. Such an interpretation may explain the drop in body mass estimates observed in older fossil penguins with femur model VI and VII (Fig. 3.9). This pattern of increased uncertainty is partially restrained with the femur models that take advantage of the Ornstein-Uhlenbeck framework to model trait evolution (Fig. 3.9, dark blue dots), meaning that complex models of evolution may represent a more accurate device to deal with trait modelling.

One of the main aspects that seems to emerge is that including phylogenetic

information in the form of a tree topology in the models may not always significantly improve the results (N. E. Campione & Evans, 2020; Revell, 2010). However, including evolutionary information in the form of a categorical descriptor of taxonomic order appears to improve model accuracy with little overfitting risk as seen in femur models II and III, and even for datasets that show relatively reduced point dispersion like the HAF-body mass dataset (Fig. 3.2A and Fig. 3.4A). Phylogenetic information in the form of a phylogeny does not provide the same type of improvement (see above) and instead substantially increases parameter uncertainty, and therefore increases uncertainty around body mass estimation. Models that include taxonomic order as a discrete category (femur models II and III) are optimal for modelling the body mass of penguins but may be overly "rigid" for orders of birds that encompass high variability in HAF length and body mass. Here we see the advantage of evaluating multiple models with different predictor variables and selecting the most study-appropriate model from among them with PSIS-LOO estimates. These estimates help to select the best overall model, but particular clades may show more accurate predictions with models that do not necessarily perform better. As an example, femur models II and III consistently under-estimate the body mass of several species within Alcidae: razorbill *Alca torda*, parakeet auklet *Aethia psittacula* and Cassin's auklet *Ptychoramphus aleuticus* (Fig. 3.6-3.8). This model behaviour can be explained by the large variation of HAF length and body mass within Charadriiformes, which is the discrete taxonomic information that informs the intercept of femur model III. Numerous small taxa (especially from Charadrii and Scolopacii) "pull" the mass estimates for large charadriiforms like the alcids towards lower values because they all share the same intercept informed by taxonomic order. However, models which include phylogenetic information through an Ornstein-Uhlenbeck approach (e.g. femur models VIII and IX) provide more accurate estimates of alcid body mass despite having lower overall performance. It is thus extremely important to account for the intra-group variability if the predictions are meant for a specific clade. One option for the researcher interested in the mass

estimation of an heterogeneous group could be to use lower taxonomic levels (e.g. family instead of order). The main advantage of the Bayesian modelling approach is that it potentially allows for nesting subgroups into bigger groups (i.e. taxonomic families inside orders), thus generating a relatively simple hierarchical structure that does not require a complex phylogenetic framework. Given that penguins are the focal group for the current study (i.e. one family Spheniscidae within one order Sphenisciformes), the generation of this type of model goes beyond the scope of the chapter but may represent a valid tool for future research.

Estimated body masses of fossil penguins from HAF model II and femur model III (i.e. the models with the best PSIS-LOO values within their respective groups) tend to give similar results (Fig. 3.10), even with taxa that are larger than any living species (e.g *Kairuku* compared with emperor penguin; Ksepka et al., 2012). For the body mass estimates of fossil penguins made in this study only *Sequiwaimanu rosieae* showed a substantially different prediction between HAF model II and femur model III (Fig. 3.10, Tables 3.3 and B.3). *Sequiwaimanu rosieae* is around 61 million years old and is one of the earliest known penguins (Mayr, De Pietri, Love, Mannering, & Scofield, 2017). HAF model II predicts a body mass for this taxon as 20.2 kg whereas femur model III predicts a body mass of 12.6 kg. The two different body mass values estimated for *Sequiwaimanu rosieae* are based on two different locomotory modules (i.e. forelimb and hindlimb) and may be indicating different rates of evolution in different regions of the body (modular evolution; Lü et al., 2010). The earliest stem-lineage penguins from the Paleocene and Eocene epochs were morphologically distinctive from Oligocene to recent penguins (Mayr et al., 2020; Slack et al., 2006). Recalling the adaptive valley metaphor cited above (Simpson, 1953), *Sequiwaimanu* and closely related taxa may not be identified properly by the models due to their position between the "adaptive peaks", a section of the adaptive landscape that is not explored by extant taxa.

Body mass estimates for fossil penguins using HAF models I and II yield mixed results. Livezey (1989) estimated a body mass of 54 kg for *Pachydyptes ponderosus*,

close to the 60 kg average estimated in the current study, but differences between estimates from this study and previous ones tend to increase with increasing bone size. For example, *Kumimanu biceae* was estimated to be around 101 kg by Mayr, Scofield, et al. (2017) using a femur length simple linear model, whereas here the body mass is estimated around 80 kg using coracoid HAF.

Here the current study uses a dataset that encompasses several bird orders, whereas previously published studies tended to focus only on penguins. Not accounting for variation outside of Sphenisciformes is equivalent to assuming that penguins are inherently different from other birds, and thus escape from all avian scaling rules. The hierarchical multilevel model approach established here with Bayesian methods allows us to discriminate between different bird orders but still accounts for the greater extra-group variability of all birds thanks to the varying intercepts. As a consequence, each bird order can exhibit a given ratio of femur volume or coracoid HAF length to body mass without the need of subsetting the dataset.

3.5 Conclusion

Though size prediction of extinct taxa will probably never have the privilege of any sort of confirmation from direct measurements (N. E. Campione & Evans, 2020), this chapter aimed to at least offer an informed estimate based on the mix of structural constraints and phylogenetic bracketing. Being able to provide body mass measurements by developing models for different bones provides an opportunity to compare a key life history trait for fossil taxa that are missing different elements. Given the convergence of body mass predictions from both coracoid HAF length and femoral volume it is shown here to be possible to rely on separate body parts to generate similar body mass predictions. As these models improve and evolve, hopefully the body mass distributions estimated here may contribute to the priors of future body size-related analyses that deepen our knowledge of extinct life. The results emerging here enrich our current understanding of extinct penguins by showing that femur

size conforms to the structural apex of Seilacher's aptive triangle. Even though the body mass predictions made in this study are more conservative that previous estimates (e.g. Livezey, 1989; Mayr, Scofield, et al., 2017), they nevertheless underline the greater variability in body form that penguins had in deep time.

Figure 3.11: Reconstruction of *Kairuku waewaeroa*, used as the official reconstruction for the media release of Giovanardi et al. (2021) and accepted as Journal of Vertebrate paleontology volume 41 no. 3 cover art.

Chapter 4

Functional apex

4.1 Introduction

4.1.1 Adaptation and evolutionary rates

The adaptation vertex of the aptive triangle metaphor represents the last aspect that this thesis aims to investigate (Seilacher, 1970). Adaptation is a core concept in evolutionary biology, and although providing a precise and incontrovertible definition of adaptation may go beyond the scope of this thesis, adaptation can be broadly interpreted as an inherited feature of an organism that contributes to actively increasing the reproductive fitness of that organism (Darwin, 1872). Specific adaptations tend to be favoured or selected in response to precise environmental pressures (Futuyma & Kirkpatrick, 2017). Hence, describing adaptations can be extremely helpful in paleontology because they can indirectly provide information about the nature of the pressures that shaped the evolutionary trajectory of a particular group of organisms (Simpson, 1944).

Lineages of organisms generation after generation can be shaped by selection (although see Gould and Lewontin, 1979), and thus undergo an amount of change per given time unit, resulting in a "rate of evolution" (Felice et al., 2020; Puttick et al., 2014; Simpson, 1944, 1953; Wang & Lloyd, 2016). The concept of evolutionary rate was originally introduced by George Gaylord Simpson in an attempt to describe how

quickly any given trait of an organism may change over the generations (Simpson, 1944). The concept is now commonplace, with numerous studies revealing the rates at which a lineage may evolve in terms of their molecular sequences (Ho & Duchêne, 2014; Ho et al., 2011) or morphological traits (Lloyd et al., 2012; Pender et al., 2021). One of the areas in which rates have proven their solidity is within the field of systematic biology, with the now well-established concept of the molecular clock (Bromham & Penny, 2003; Thorne et al., 1998; Zuckerkandl & Pauling, 1965). A molecular clock can be used to estimate divergence times between a set of taxa (Beavan et al., 2021; dos Reis et al., 2012; Weir & Schluter, 2008). By assuming that mutation rates are consistent through time it can be expected that differences among molecular sequences are indicative of branching events on a phylogeny (Bromham & Penny, 2003). Using fossil calibrations to constrain the dates of branching events (Parham et al., 2011), it is possible to infer a dated phylogenetic tree for a clade of interest. Since its formalisation the molecular clock model has undergone great improvement, allowing the assumption of a constant rate of mutation to be relaxed (Drummond et al., 2006) and extending the clock model to phenotypic datasets (Álvarez-Carretero et al., 2019; Lee et al., 2014). This last aspect is crucial when the researcher aims to include into an analysis fossils that do not have molecular data.

Establishing evolutionary rates for morphological data is a first step in seeking evidence for adaptation, but presents interesting challenges depending on the type of morphological data that are being used. Although divergence dates can be calculated using discrete characters (Lee et al., 2014), the major drawback is that character definition is often subjective, with the consequence that if the same set of taxa is used to independently build two separate character matrices there is the risk of discrepancies in divergence estimates (Brazeau, 2011). In this sense geometric morphometric datasets provide a more rigorous and quantitative approach, albeit possessing some challenges that need to be addressed. A geometric morphometric dataset is usually an array composed of n observations distributed over p traits

(Adams & Collyer, 2019; Felsenstein, 1988). The data within this matrix is influenced by both the non-independence of observations due to shared ancestry (i.e. the phylogeny; see Chapter 2) and the non-independence among traits (i.e. within- and between-element character covariance) (Álvarez-Carretero et al., 2019; Felsenstein, 1988). In this sense the lack of complete independence among characters may be interpreted in the constructional morphology framework as a quantification of the structural constraints acting over a given set of traits. In fact it has been demonstrated that trait covariance matrices help to estabilish trait evolvability (Cheverud, 1982; Love et al., 2021). Geometric morphometric data could be used to infer evolutionary relationships, but the inherent covariance among characters could lead to incorrect inference of models of trait evolution if the data were used without any further pre-treatment (Adams & Collyer, 2019; Felsenstein, 1988; Lande, 1979; Uyeda et al., 2015). Hence, the correlation structure would need to be estimated simultaneously with the phylogeny while performing any sort of comparative or phylogenetic analysis (Felsenstein, 1988). However, such a task cannot be easily solved by modern computing machines (Felsenstein, 2002).

4.1.2 R rotation method

Even if practical solutions to the problem of having high correlation within geometric morphometrics datasets has yet to be found, a recent development by Álvarez-Carretero et al. (2019) has offered an alternative method for at least partly adjusting for character correlation. Álvarez-Carretero et al. (2019) proposed a method for computing the matrix of character correlation **R** on a population of organisms from the same species, producing an estimate of character correlation that is unbiased by phylogenetic confounds[1]. Once **R** is obtained, it can be multiplied to the dataset to "rotate" it in a similar fashion as a principal component analysis (PCA) does. Adopting this technique allows the character data from the dataset to be reoriented orthogonally to the major axes of variation and offers a way to mitigate the effects

[1]Note that in this context the matrix **R** assumes a different meaning than the phylogenetic correlation matrix used in the Brownian motion models in previous chapter.

in character covariance (Álvarez-Carretero et al., 2019). There are caveats with this **R** rotation method, including assuming that correlation among traits is constant throughout a phylogeny, but it does at least offer a way to control for character covariance. For example, a study might simply apply PCA to a character matrix and attempt to use principal component one scores to represent morphological traits in an evolutionary rate analysis (e.g. Felice et al., 2021). However, an advantage over a traditional PCA approach is that the **R** rotation method is estimated independently from the comparative dataset studied (Uyeda et al., 2015).

The **R** rotation method from Álvarez-Carretero et al. (2019) enables a molecular clock-style analysis to be extended to multivariate morphological datasets like those from geometric morphometric analyses. In this sense we can develop a 'morphological clock', enabling us to determine whether any given shape evolved continuously along a phylogeny, or to discover if a rate of morphological evolution experienced any drastic changes within a phylogeny. Moreover, if this 'morphological clock' approach is applied to different traits then it may help to determine whether these traits exhibited similar rates of morphological evolution or if they differed. Returning to the constructional morphology framework, if a trait has experienced a faster rate of evolution than another then we may have evidence that the trait was subject to an evolutionary driver, and thus that it represents an adaptation. In addition, compared to other methods of morphological trait evolution, the morphological clock allows the use of the entire geometric morphometric dataset with relatively little use of computational power, hence not forcing the researcher to discard variation in the dataset that is often covered by the lower eigenvectors of a PCA (Felice et al., 2021).

4.1.3 Wing propelled divers

Penguins can be considered a valid candidate to use these methods for several reasons. As described by Simpson (1946), penguins have aerially-volant ancestors that experienced a strong selective pressure toward aquatic habits, causatively-shaping

their bone morphology. This morphology of wing propelled divers (WPDs) is a complex set of traits that are correlated across the skeleton although likely arising under different rates of morphological change (and hence subject to different selection pressures for adaptation). Whole-bone three dimensional geometric morphometrics (3DGM) is advantageous for this study system given that 3DGM enables analyses of complex shapes without the need to subdivide shape into discrete character subsets.

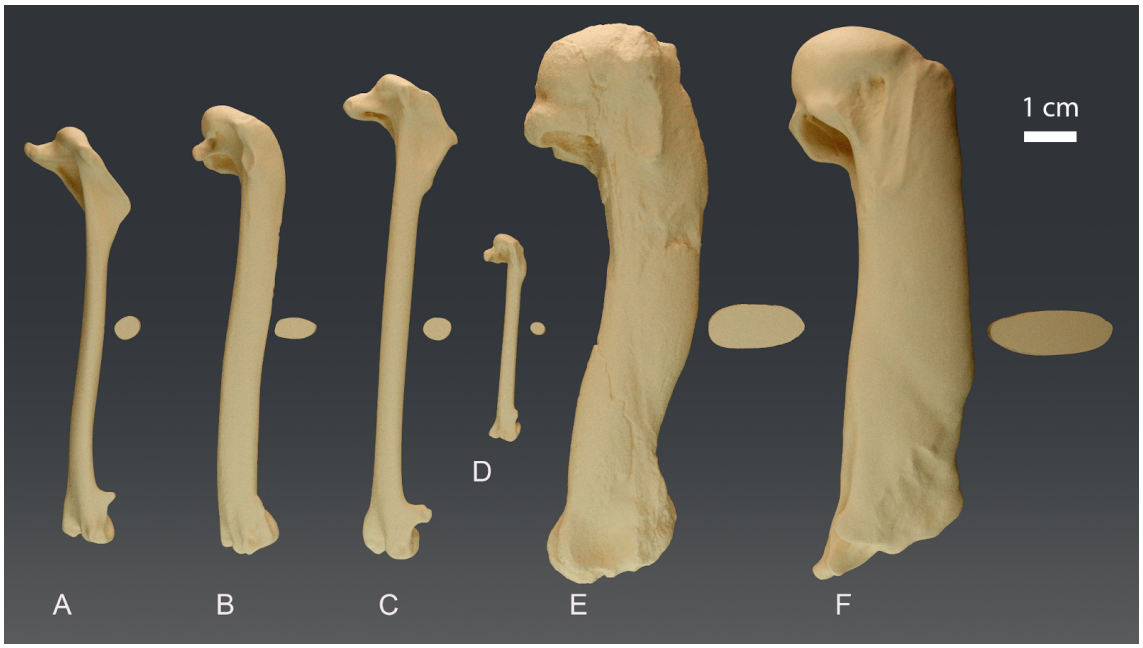


Figure 4.1: Humeri from wing propelled divers and close relatives. A) INMH 3554, kittiwake *Rissa* sp.; B) KU 66938 thick-billed murre *Uria lomvia*; C) OR 029759b, northern fulmar *Fulmarus glacialis*; D) ORN 120642, common diving petrel *Pelecanoides urinatrix*; E) NMNZ S.47304 Spheniscifformes sp.; F) CM 2013-1-257, king penguin *Aptenodytes patagonicus*. The shape on the right of each humerus represents the midshaft cross section. Inferred wing propelled divers are B, D, E and F, all bones are to scale

The set of characteristics that WPDs exhibit is extensive (see Mayr et al., 2021; J. Watanabe et al., 2020), but if we reframe the main focus on the humeurs (Fig. 4.1), then WPDs tend to converge toward similar structures in a clear example of convergent evolution (Losos, 2011). By comparing the humerus from a thick-billed murre (Charadriiformes, *Uria lomvia*, Fig. 4.1B) and from a kittiwake (Charadriiformes, *Rissa sp.*; Fig. 4.1A), then one of the most striking differences is the flatness of the humeral shaft in the former. In *Rissa* the humeral shaft is circular and in *Uria* the shape is more ellipsoidal, recalling the condition found in ancestral and

recent penguins (Fig. 4.1E-F). The humerus of a diving petrel (Procellariiformes, *Pelecanoides urinatrix* Fig. 4.1D) shows a midshaft profile that is similar to that of a northern fulmar (Procellariiformes, *Fulmarus glacialis* Fig. 4.1C) but the humeral head exhibits a raised and elongated scar for the musculus supracoracoideus that recalls the condition found in auks (Alcidae; Fig. 4.1B) and penguins (Fig. 4.1E-F)(J. Watanabe et al., 2020). The enlargement of the attachment of the m. supracoracoideus can be viewed as another of the WPDs key features given that this muscle needs to generate more force in the relocation of the wing after the downstroke in WPDs as they generate thrust during both wing strokes (Johansson & Aldrin, 2002; Ksepka & Ando, 2011; J. Watanabe et al., 2020). Although there is no sharp distinction between WPDs and other groups of seabirds, but rather a gradient-like distinction, it is still possible to recognise penguins as a group that have accumulated WPD traits through time.

Penguins of course also have a more extensive temporal sampling of species than any other WPD group (described in Chapter 1: Introduction). Although other avian orders may offer a broader selection of living taxa, very few can compete with Sphenisciformes in terms of fossil sampling (Jadwiszczak, 2009; Ksepka & Ando, 2011; Mayr, 2016). Applying a morphological clock to WPD traits may reveal whether rates of evolution in the locomotory modules of penguins have changed across time. The body plans and individual bone shapes of the earliest penguins were substantially different compared with more recent species (Mayr, De Pietri, Love, Mannering, & Scofield, 2017; Mayr et al., 2021; Mayr, Scofield, et al., 2017; Slack et al., 2006), to the extent where some authors have questioned their phylogenetic identity as Sphenisciformes (Mayr, 2005; Mayr et al., 2021). The estimation of evolutionary rates may determine whether there has been an increase in rates of morphological change over time as we move from the earliest species in the clade towards younger taxa, and if forelimb bones were subjected to stronger evolutionary pressure compared with hindlimb bones in these wing propelled divers.

4.1.4 Morphospace for humeri and tarsometatarsi through time

The morphological clock method of calculating rates of species divergence from the shapes of humeri and tarsometatarsi will be used in conjunction with a novel approach to estimate the morphospace for these bones through time. As described in Chapter I: Introduction, morphospace is a hyperdimensional space where each observed phenotype occupies a precise location with specific coordinates. Morphospaces have a long history of application in paleontology and comparative studies (Croft et al., 2018; Raup, 1967; Valkenburgh, 1985; Walton & Korn, 2018). Coupling a morphospace with fossil ages can be useful to describe the evolutionary trends in shape that a clade has experienced (Larson et al., 2016; Sibert et al., 2018; Simpson, 1944). However, describing evolutionary adaptation by studying shift in morphospace requires a large dataset that spans a sufficiently long time window, to ensure that the pattern of morphospace occupation is not a product of biased undersampling. The apparent absence of taxa from a particular region of a morphospace may be due to a poorly preserved paleontological record rather than the morphological diversity of a clade. Although Sphenisciformes has a rich fossil record compared to other avian clades, many fossil penguins are only described from incomplete skeletons, which means there are still limits to the number of specimens with comparable skeletal elements that can be included in a shape study.

In an attempt to overcome the issue using a "Bayesian"-inspired philosophy, uncertainty limits for the shape of humeri or tarsometatarsi inferred for hypothetical ancestors will be gained from the range of tree topologies generated in Chapter 2. Here I will estimate the "unseen" shapes of humeri and tarsometatarsi for hypothetical penguin ancestors using ancestral state reconstruction. To marginalise over phylogenetic uncertainty a set of comparative models will be fitted to a subsample of the posterior trees from the fossilised birth death models calculated in Chapter 2 rather being fitted to just a single consensus tree (Guillerme & Healy, 2014). Shape information for the internal nodes of each tree will be inferred and the morphospace

occupation will be based on the cloud of possible shapes available through time.

Morphological clock and morphospace methods will be used to analyse the evolution of humeri and tarsometatarsi of penguins. Humeri and tarsometatarsi are the most-frequently described skeletal elements from fossil penguins as a consequence of being among the most common bones to be recovered (Chávez Hoffmeister et al., 2014), and therefore provide the largest dataset available among fossil penguin bone. Furthermore, by selecting elements from two separate locomotory modules we aim to reduce the risk of redundant covariation among the elements. Although note that living organisms are complex systems that can be thought of as a series of interconnected networks (Goswami & Polly, 2010; Klingenberg & Marugán-Lobón, 2013), and thus complete independence between characters cannot occur (see above). However, modularity assumes that some traits tend to be more strongly interconnected than others generating a series of semi-autonomous units within an organism that may evolve at different rates (Bardua et al., 2020; Bardua, Wilkinson, et al., 2019; Felice et al., 2020; A. Watanabe et al., 2019). The ideal approach when using datasets that come from different shapes would be to merge them into a single dataset controlling for redundant covariance and hence being able to treat the entire set of traits as a whole (Collyer et al., 2020). However, achieving such a goal is not a trivial problem and there is still a debate whether these approaches may represent a valid morphometric approach (Collyer et al., 2020; Rhoda et al., 2021). Moreover, the use of a method that combines the shapes of both the humerus and tarsometatarsus would need to be extremely robust to missing information (Arbour & Brown, 2014). This is because comparatively few fossil specimens preserve both humerus and tarsometatarsus, whereas many more preserve just one of these elements. Shape datasets for the humerus and tarsometatarsus will therefore include different fossil species, meaning that evolutionary rates for the shape of the humerus and for the shape of the tarsometatarsus will be performed separately. Comparisons between these separate analyses must therefore be restricted to "homologous nodes" where the same shared common ancestor occurs in each analysis.

For this analysis we will assume that the shapes of the humerus and tarsometatarsus within the same individual have no covariation. Even if this may seem implausible, the functional differences of the wing and leg in a penguin allow for the possibility that the humerus and tarsometatarsus are only weakly connected through evolutionary covariation. Morphological clock and morphospace methods in concert may be used to detect differential rates of evolution among wing and foot modules in penguins, and most importantly, may help to understand how these modules adapted to life in water during the evolutionary history of penguins

4.2 Materials and Methods

4.2.1 Meshes

A dataset of 65 digital replicas of humeri and 57 digital replicas of tarsometarsi were used for this study (for accession details and scanning method see supplementary file `Chapter_4_Specimen_ID.csv`). Left bones were preferred, and when the left was not available the right was digitised and the mesh was mirrored using `Blender` (Blender Online Community, 2020, v. 2.9.1). Adults with no evident impairment were preferentially selected when digitising bones of modern species. The humeri spanned in proximodistal length from 40.9 mm for kororā little blue penguin *Eudyptula minor* (OR30224) up to 178.2 mm for *Kairuku waewaeroa* (WM 2006/1/1). The tarsometatarsi spanned in proximodistal length from 19.8 mm for *Eudyptula minor* (OR029118) up to 76.8 mm for *Waimanu manneringi* (CM zfa35). The institutions that provided the specimens were: American Museum of Natural History, New York, NY, USA; Canterbury Museum, Christchurch, New Zealand; Field Museum of Natural History, Chicago, IL, USA; Massey University, Auckland, New Zealand; Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand; Otago Museum, Dunedin, New Zealand; Otago University Geology Museum, Dunedin, New Zealand; Waikato Museum Te Whare Taonga o Waikato, Hamilton, New Zealand.

The complete dataset was initially divided into two groups by bone type: humerus

dataset and tarsometatarsus dataset. Each of these datasets was then divided into two main sub-blocks: the "comparative" block that included all digital replicas from fossils and from most of the living taxa, and the "covariation" block that included only digital replicas from *Eudyptula minor* specimens. The comparative humerus dataset included 30 specimens and the covariation humerus dataset included 30 specimens. The comparative tarsometatarsus dataset included 25 specimens and the covariation tarsometatarsus dataset included 31 specimens. Note that one *Eudyptula minor* specimen (NMNZ 18415) was present in both the comparative and the covariation dataset for both the humerus and the tarsometatarsus. Both the humerus and tarsometatarsus from the same individual were digitised when possible (See supplementary file `Chapter_4_Specimen_ID.csv`). The same individuals contributed to both the humerus and tarsometatarsus covariation datasets. Lastly the datasets included also the additional humeri of BMBH RMA43 *Anthropornis nordenskjoeldi*, NMNZ S.47304 Sphenisciformes indet., OU22168 *Kairuku sp.*, NMNZ DM1449 Sphenisciformes indet. "Seal rock specimen", OU21977 "*Pakudyptes hakataramea*" and the tarsometatarsii of OU22127 *Palaeeudyptes antarcticus*, and OU22181 *Palaeeudyptes gunnari*. Due to their uncertain taxonomic status and/or their fragmentary nature these specimen were used to build a "post-hoc" dataset to be used in a second stage to evaluate the accuracy of morphospace projection methods.

As for femora and coracoids from the previous chapter, the 3D digital replicas of bones (i.e. 3D meshes) used for the 3D geometric morphometrics analyses in this chapter were mostly generated using 3D surface scanning. Meshes were generated using either a 1) HP 3D camera (DAVID-CAM-3.1-M; HP Inc., Palo Alto, CA, USA) with a HP 3D HD camera along with a K132 + DLP projector (Acer Incorporated), where meshes were finalised in the `HP DAVID 3` software (`HP DAVID 3` software, version 3.10.4.4657; HP Inc., Palo Alto, CA, USA); or 2) with a Creaform HandySCAN 3D laser surface scanner (Creaform, Levis, Canada) with resolution varying from 0.1 to 0.3 and meshes finalised in the `VXelements` software suite (`VXelements` and `VXmodels` version 8.1, Creaform, Levis, Canada). Two bro-

ken humeri (NMNZ S.47308 *Kupoupou stilwelli* and DM13309 *Pygoscelis papua*) required further restoration steps into the VXelements software environment that involved the alignment and the welding of the proximal and the distal parts.

4.2.2 Landmarks and specimens

Landmarks were placed on the 3D digital replicas of humeri and tarsometatarsi in both the comparison and covariation datasets as well as on the specimens in the post-hoc datasets. The landmarking scheme for each bone was designed to optimise the shape variation preserved across the sometimes broken fossil specimens (Zelditch et al., 2012). The landmarking scheme used fixed landmarks and semilandmark curves to maximise anatomical recoverability across all specimens at the cost of reducing the total number of landmarks per bone. Surface semilandmarks were not used here, although they are commonly used in other studies (e.g. Bardua, Felice, et al., 2019; Bardua et al., 2020; Felice et al., 2021; Felice et al., 2020. Surface semilandmarks may capture a great amount of shape variation but were not appropriate for the current dataset due to the poor preservation of some specimens (e.g. NMNZ S.47308 *Kupoupou stilwelli* humerus, OU22127 *Palaeeudyptes antarcticus* tarsometatarsus, CM 2016-6-1 *Sequiwaimanu rosieae* humerus). Landmark registration was performed in `SlicerMorph` (Rolfe et al., 2021) which is an add-on package for 3D `Slicer` (Fedorov et al., 2012; v.4.11.2). Landmark sets for the humerus and tarsometatarsus are described in full in Appendix C (see also Figs. 4.2–4.3).

4.2.3 Morphological clock analysis

Landmarks from the comparative and covariation datasets for both the humerus and tarsometatarsus were exported from `Slicer` in .json format and converted to .pts using custom written code for the `R` environment (R Core Team, 2021; v. 4.0.5). Landmarks were read into `R` using `Morpho::readpts` (Schlager, 2017; v 2.8.1) which generates a matrix for each .pts file. Both the humerus and tarsometatar-

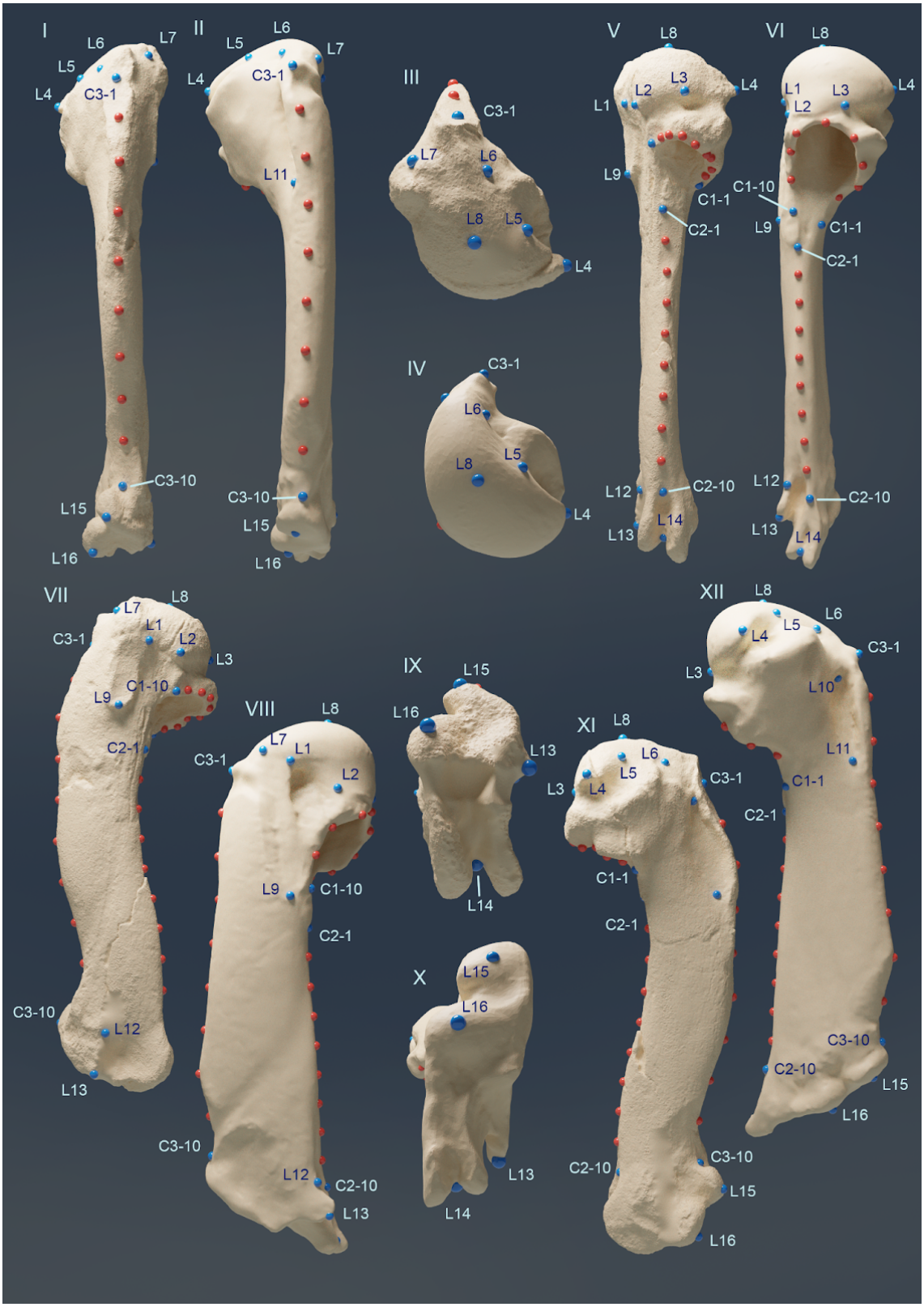


Figure 4.2: Landmark configurations for humeri. Humeri seen in cranial (I-II), proximal (III-IV), caudal (V-VI), dorsal (VII-VII), distal (IX-X) and ventral (XI-XII) views. Humeri from AV19569 *Aptenodytes forsteri* (II,IV,VI,VIII,X,XII) and from NMNZ S.47304 Sphenisciformes indet. (I,III,V,VII,IX,XI). Bones are not to scale.

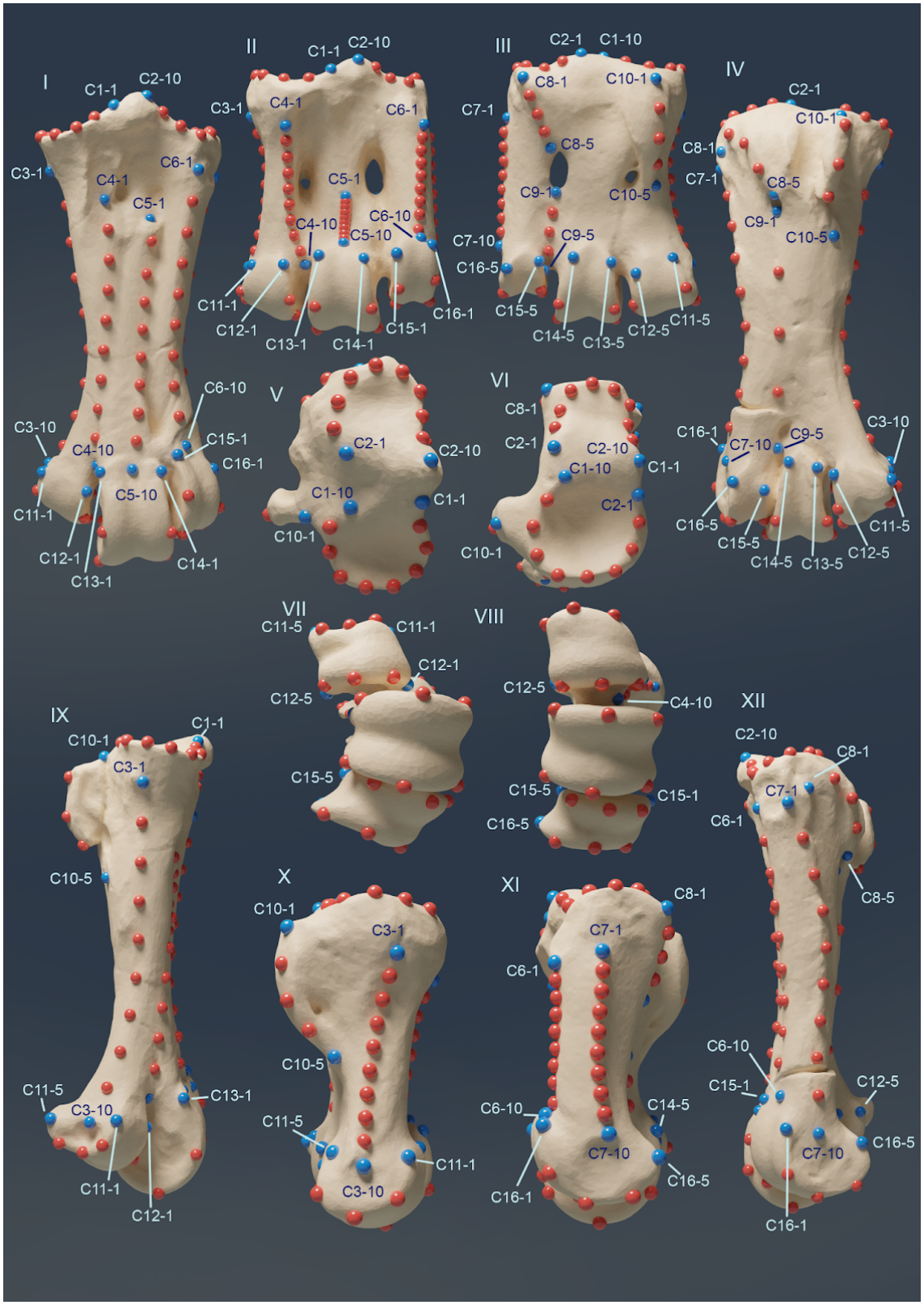


Figure 4.3: Landmark configurations for tarsometarsi. Tarsometatrsi seen in dorsal (I-II), plantar (III-IV), proximal (V-VI), distal (VII-VIII), medial (IX-X) and lateral (XI-XII) views. Tarsometarsi from NM23039 *Aptenodytes forsteri* (II,III,VI,VIII,XI) and from CM zfa35 *Waimanu manneringi* (I,IV,V,VII,IX,XII). Bones are not to scale

sus comparative datasets were Procrustes superimposed using `Morpho::procsym`, which ensures that only shape information is preserved in the dataset and removes all effects of size, rotation and location. The humerus and tarsometatarsus covariation datasets were then aligned to their corresponding Procrustes-transformed comparative datasets using `Morpho::align2procsym`. `Morpho::align2procsym` aligns the shapes of a new dataset (in this case the covariation dataset) to the shape of a target dataset (the comparative dataset) without the need to perform a full Procrustes superimposition a second time. Applying `Morpho::align2procsym` to a covariation dataset follows the recommendation of Álvarez-Carretero et al. (2019), and here ensures that inter-specific distances between shapes in the comparative datasets are not affected by the greater opportunity to resolve intra-specific distances between shapes in the *Eudyptula* covariation datasets.

Before the landmark data are used as a morphological clock (i.e. to estimate species divergence times) it is necessary to multiply the comparative dataset by the Cholesky decomposition of the inverse of the character correlation matrix $\mathbf{R}$ (for a full mathematical description on likelihood computation and the steps required see Álvarez-Carretero et al., 2019; Felsenstein, 1973; Freckleton, 2012). However, as the number of characters increases over the number of sampled individuals the estimated $\mathbf{R}$ matrix tends to have a determinant equal to 0, and in such cases the matrix is not invertible (Adams, 2014; Goolsby, 2016; Schäfer & Strimmer, 2005). The solution proposed by Álvarez-Carretero et al. (2019) to solve this issue is to calculate the linear shrinkage estimate $\mathbf{R}^*$ (Eq. 4.1).

$$\mathbf{R}^* = \delta\mathbf{I} + (1 - \delta)\widehat{\mathbf{R}} \tag{4.1}$$

Here $\mathbf{I}$ is an identity matrix in which all diagonal elements are 1 and off diagonal elements are 0. δ is a shrinkage parameter and assumes a value between 0 and 1. $\widehat{\mathbf{R}}$ is the estimated $\mathbf{R}$ matrix. The parameter δ controls the amount of shrinkage that a matrix is subjected to; when $\delta = 0$ $\mathbf{R}^*$ is equal to $\widehat{\mathbf{R}}$, and when $\delta = 1$ $\mathbf{R}^*$ is equal to the identity matrix. δ effectively adjusts the values of $\widehat{\mathbf{R}}$ to obtain a $\mathbf{R}^*$

that can be inverted (with $\delta \neq 0$) and thus allows the dataset to "rotate". The δ parameter was estimated with the `corpcor::cor.shrink` function (Schäfer and Strimmer, 2005; v1.6.9) to obtain $\mathbf{R}^*$ both for the humerus and the tarsometatarsus covariation datasets. Following the recommendation in Álvarez-Carretero et al. (2019) the variance of each character in both datasets was scaled to ensure that all variables had equal variance.

To perform the morphological clock analysis the procedure followed dos Reis et al. (2018) using the `MCMCtree` package in the `pamlX` software environment (Yang, 2007; v 1.3.1). The analysis required a MCMCtree alignment file and a control file. The `MCMCtree` alignment file specifies the $\mathbf{R}^*$ transformed dataset along with a tree topology (i.e. first calculated in Chapter 2 and elaborated below) and a series of dates associated with the tips (see below). The control file contains information about prior settings and specific file paths that `MCMCtree` requires. The `MCMCtree` alignment files were generated with the function `mcmc3r::write.morpho` and the control files were generated with `mcmc3r::ctlMCMCtree` (dos Reis et al., 2018; v. 0.4.3).

The tree topology used to estimate the morphological clock was based on the maximum clade credibility (MCC) tree estimated from the posterior distribution of trees from Chapter 2. *Aptenodytes ridgeni* was grafted onto the phylogenies used for the tarsometatarsus divergence analyses. The position of *Aptenodytes ridgeni* in these phylogenies was made to match the position found in Thomas et al. (2020) at 7.14 million years ago (Ma) (i.e. topologically prior to the *Aptenodytes forsteri* and *Aptenodytes patagonicus* split on all trees). *Aptenodytes ridgeni* was given a branch length of 0 million years. The phylogenies had all taxa pruned from them that were not part of the comparative dataset they were being used for in order to be correctly handled by `MCMCtree`. Thus 46 taxa were removed for the humerus analysis and 51 taxa for the tarsometatarsus and edge length was removed in order to retain a simple bifurcating topology. To calibrate the `MCMCtree` analysis the software requires the tips to have a fixed date, however, fossils can have uncertain

age ranges rather than a single definitive value. The value used for each of the dated taxa reflected the date that the tip exhibited on the MCC tree calculated in Chapter 2 (See table 4.1).

Table 4.1: Dates used on the tip of the morphological clock analysis derived from the median age retrieved on the dated phylogeny from Chapter 2

Species	**Million years ago**
Aptenodytes forsteri	0.00
Aptenodytes patagonicus	0.00
Aptenodytes ridgeni	7.14
Archaeospheniscus lowei	27.87
Eudyptes atatu	1.96
Eudyptes chrysolophus	0.00
Eudyptes filholi	0.00
Eudyptes pachyrhynchus	0.00
Eudyptes robustus	0.00
Eudyptes schlegeli	0.00
Eudyptes sclateri	0.00
Eudyptula minor	0.00
Eudyptula novaehollandiae	0.00
Icadyptes salasi	36.49
Kaiika maxwelli	55.49
Kairuku grebneffi	28.14
Kairuku waewaeroa	29.34
Kairuku waitaki	26.98
Kupoupou stilwelli	61.49
Megadyptes antipodes	0.00
Megadyptes antipodes spp. *waitaha*	0.00
Muriwaimanu tuatahi	59.32
Pachydyptes ponderosus	35.36
GL429 *Palaeeudyptes* sp.	37.33
Platydyptes amiesi	25.11
Platydyptes novaezealandiae	25.11
Pygoscelis adeliae	0.00
Pygoscelis antarcticus	0.00
Pygoscelis papua	0.00
Sequiwaimanu rosieae	60.99
Spheniscus demersus	0.00
Spheniscus humboldti	0.00
Spheniscus magellanicus	0.00
Spheniscus mendiculus	0.00
Waimanu manneringi	61.49

Three control files were generated for each of the humerus and the tarsometatarsus datasets. Each control file represented a different mode of character evolution following the three types of clock implemented in `pamlX`: the strict clock, the uncorrelated clock, and the auto-correlated clock. The strict clock assumes that the rate of evolution is the same along all branches (Thorne et al., 1998). The uncorrelated clock (or independent clock) assumes independent rates for each branch (Drummond et al., 2006; Rannala & Yang, 2007). The auto-correlated clock assumes that each parent branch rate affects the rates observed on descendant branches (Rannala & Yang, 2007).

Settings for the priors for the control files followed dos Reis et al. (2018) and Álvarez-Carretero et al. (2019). The parameters for the fossilised birth death process were set to $\lambda = \mu = 1$, $\rho = 0$ and $\psi = 0.001$. The variance σ^2 of the clock rate was set as a gamma distribution with parameters $\alpha = \beta = 2$. The root age was set as a uniform distribution with a soft bound between 61 and 180 million years ago. A preliminary maximum likelihood analysis was performed on the simple bifurcating topology mentioned above with the observed morphological alignment, to set the priors of the mean substitution rate. The analysis was run using `CONTML` from the `PHYLYP` package (Felsenstein, 1993, v.3.698), and produced the unrooted trees reported in Fig. 4.4. The mean substitution rate prior was set to be a gamma distribution with $\alpha = 2$ and a $\beta = \alpha$ /observed rates. The humerus dataset analysis consequently had a $\beta = 6.84$, and the tarsometatarsus dataset had a $\beta = 17.53$. The mean of the gamma distribution is α /β. With the priors described above we ensure that the gamma distribution is centered with a mean at the observed rates of morphological evolution.

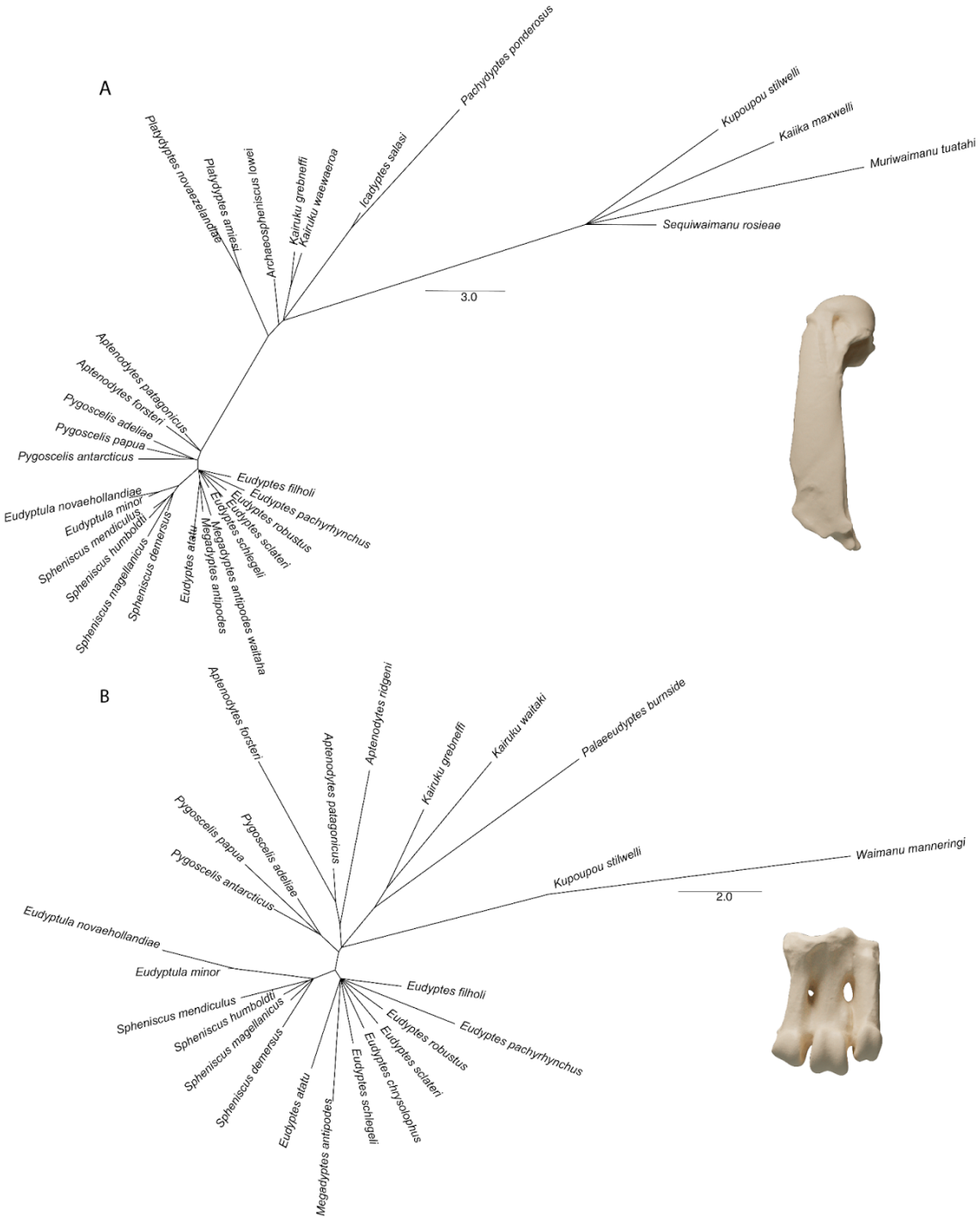


Figure 4.4: CONTML preliminary trees. Maximum likelihood trees estimated from the humerus (A) and tarsometatarsus (B) comparative datasets.

The Markov Chain Monte Carlo (MCMC) for all analyses were set to a total length of 4×10^5, a burn-in of 25%, and with the Finetune parameters set as default from `MCMCtree`.

Bayes factors (BF) were computed to perform model selection (Lavine & Schervish, 1999), with BF defined as the ratio in equations 4.2 and 4.5.

$$\mathrm{BF}(M_0, M_1) = \frac{\text{posterior odds}}{\text{prior odds}} \tag{4.2}$$

$$\text{posterior odds} = \frac{P(\mathrm{M}_0|\mathbf{X})}{P(\mathrm{M}_1|\mathbf{X})} \tag{4.3}$$

$$\text{prior odds} = \frac{P(\mathrm{M}_0)}{P(\mathrm{M}_1)} \tag{4.4}$$

With M_0 and M_1 as two models with a given set of estimated parameters (e.g. humerus data analysed with strict clock vs. humerus data analysed with uncorrelated clock), and with X as the observed data (i.e. the given alignment). $\mathrm{P}(\mathrm{M}_i|\mathrm{X})$ represents the probability of the model given the observed data (i.e. the posterior) and $\mathrm{P}(\mathrm{M}_i)$ is the prior probability. Equations 4.2, 4.3 and 4.4 are then used to calculate the Bayes factor with equation 4.5.

$$\mathrm{BF}(M_0, M_1) = \frac{P(\mathrm{M}_0|\mathbf{X})}{P(\mathrm{M}_1|\mathbf{X})} \div \frac{P(\mathrm{M}_0)}{P(\mathrm{M}_1)} \tag{4.5}$$

The resulting ratio is commonly used to decide which of the two models is preferred in terms of goodness of fit. Higher values ($\mathrm{BF} > 10$) tend to favor M_0 whereas lower values ($\mathrm{BF} < 1$) tend to favor M_1.

Estimating BF requires calculating the marginal probability of the data P(X) to produce $\mathrm{P}(\mathrm{M}_i|\mathrm{X})$, as per Bayes rules (Eq. 1.2). However, as mentioned in Chapter 1 Introduction, calculation of the marginal probability requires the resolution of extremely complex integrals, an intractable task that can be sidestepped by using algorithms like stepping-stone sampling (Baele et al., 2013; Xie et al., 2011). The goal here is to compute the probability of the observed data between the prior and the posterior to estimate the marginal distribution P(X). This approach involves it-

eratively re-running a series of MCMC-like chains that sample from a number of β_{BF} discrete distributions between the prior and the posterior distribution (Xie et al., 2011). For example, consider a scenario where β_{BF} is eight. In this scenario eight separate sampling events will be performed. The likelihood of these eight sampling events (Eq. 1.2) will be 8/8 the first time, 7/8 the second time, 6/8 the third time, and so on. In practice, during the first round we are sampling directly from the posterior distribution and during the last round we are sampling from the prior distribution, but more importantly, we will end up with six distributions that are a mixture of the prior and posterior to different degrees. With this set of distributions at hand it is in theory possible to estimate the marginal likelihood (Lavine & Schervish, 1999). Even if the stepping-stone sampling overcomes a challenging problem it is still computationally expensive and requires re-running the same analysis multiple times. Following (dos Reis et al., 2018), the number of β_{BF} points between distributions was set to eight for the analyses in this chapter, and the exponent for stepping-stone β generation was set to five. The procedure to generate the stepping-stone control files was performed in `R` with `mcmc3r::make.beta` and `mcmc3r::stepping.stones`. After the stepping-stone sampling was performed the BF for all three models (strict, uncorrelated and auto-correlated clocks) of both the humerus and tarsometatarsus comparative datasets was computed with `mcmc3r::bayes.factors`.

An important consideration for the morphological clock method applied here is the differential sampling of taxa among the humerus and the tarsometatarsus data partitions. To test whether any pattern emerging from the morphological clock analysis was due to an artificial undersampling of dated tips the analysis was repeated over the humerus dataset but with *Archaeospheniscus lowei*, *Platydyptes amiesi* and *Platydyptes novaezelandiae* pruned from the tree. These three taxa do not have tarsometatarsi so by removing them from the humerus comparative dataset for this trial I arrived at a more similar number of taxa to the tarsometatarsus comparative dataset. This pruning trial ensured that the phylogenetic trees used

for both the humerus and tarsometatarsus were similar by not including any dated tips on the tree branch between the origins of the Oligocene giants clade and the origin of the crown clade. The priors of the humerus comparative dataset were unchanged after these three taxa were pruned except for β, which was recalculated to be 7.25 in the new dataset. MCMC and burn-in lengths were left unchanged and analyses were performed for all three models.

4.2.4 Morphospace occupation through time and penalised likelihood analysis

As an additional way to quantify rates of evolution within Sphenisciformes the Procrustes-transformed landmarks from the humerus and tarsometatarsus comparative datasets were used to estimate the shape exhibited by internal nodes of the phylogenetic tree. This ancestral shape estimation used the penalised likelihood method described by Clavel et al. (2019). Briefly, the penalised likelihood (PL) method is a regularisation technique used in many statistical fields (S. R. Cole et al., 2014) to constrain the maximum likelihood estimation toward realistic values by penalising parameter estimates that are thought to be unrealistic. In this specific case the advantage of PL is the estimation of $\mathbf{R}$. Although $\mathbf{R}$ was calculated from a population of samples in the morphological clock analysis (i.e. covariation dataset), in a PL framework $\mathbf{R}$ is estimated directly from the comparative dataset. The PL methods in this case favour estimates of $\mathbf{R}$ that are symmetric positive definite and hence a matrix that can be invertible (Clavel et al., 2019). The penalised likelihood method has previously been applied to highly multivariate data in a macroevolutionary context Dellinger et al. (2019) and Eliason et al. (2020). Note that this analysis uses just the superimposed landmarks resulting from generalised Procrustes analysis and not the rotated dataset involving the $\mathbf{R}^*$ matrix. The function `RPANDA::fit_t_pl` (Morlon et al., 2016; v. 1.9) was used to fit the Procrustes-transformed landmark data on a given phylogenetic tree. Following recommendation in Clavel et al. (2019) the chosen penalty method for this dataset

was the Quadratic Ridge. The PL approach implemented by RPANDA supports the parameter estimation of four different models of trait evolution: Brownian motion, Ornstein-Uhlenbeck, early burst and Pagel's lambda. These four models of trait evolution were fitted over a sample of 100 trees from the posterior distribution of phylogenies estimated from Chapter 2 and then on each of these trees was measured the fit of all four models (see below). In order to perform the PL analysis each tree from the 100-tree sample was pruned to remove taxa that were not present in either the humerus or tarsometatarsus comparative dataset (depending on which analysis the tree was being used for). Each internal node (inferred ancestor) was attached to the phylogeny as a tip by duplicating the node and binding that inferred taxon in the same position as the original node with a branch length of 0. In addition, *Aptenodytes ridgeni* was grafted onto the tree following the description above. After performing the pruning and grafting modifications to all phylogenies in the 100-tree sample the four evolutionary models were fitted to each tree for both the humerus and the tarsometatarsus comparative datasets. The generalised information criteria (GIC) for each evolutionary model was calculated for each of the 100 trees to determine which of the four models best described the observed data. Ancestral state estimation was then performed for the best fitting model of trait evolution. This resulted in a set of shapes for each node in the sampled trees.

Recall that the comparative datasets for the humerus and the tarsometatarsus contain slightly different sets of taxa, and thus the phylogenetic trees used to analyse these datasets have different topologies. To compare the results from the humerus and tarsometatarsus analyses four nodes that are common to the topologies of all trees for both datasets were defined (Fig. 4.5). These nodes are a) Spheniscifformes MRCA, i.e. the most recent common ancestor (MRCA) of all penguins (the taxon assemblage that includes *Waimanu*, *Muriwaimanu*, *Kupoupou* and *Sequiwaimanu*) and all other penguins b) Oligocene MRCA, i.e. the most recent common ancestor to the Oligocene giants and the crown taxa (the taxon assemblage that includes *Kairuku* , *Icadyptes*, *Pachydyptes* and the Burnside "*Palaeeudyptes*" plus all living

penguins) c) crown-penguin MRCA, i.e. the most recent common ancestor of the crown taxa, d) the most recent common ancestor for *Eudyptes* (as a proxy for tip morphology). The morphological distance (i.e. Procrustes distance) between these nodes was calculated for each tree, and then this morphological distance was divided by the temporal distance in millions of years (Eq. 4.6) to estimate the rate of morphological evolution in these three segments of penguin evolutionary history.

$$\text{Rate of evolution} = \frac{\text{Morphological distance}_{\text{Procrustes distance}}}{\text{Temporal distance}_{\text{Million years}}} \tag{4.6}$$

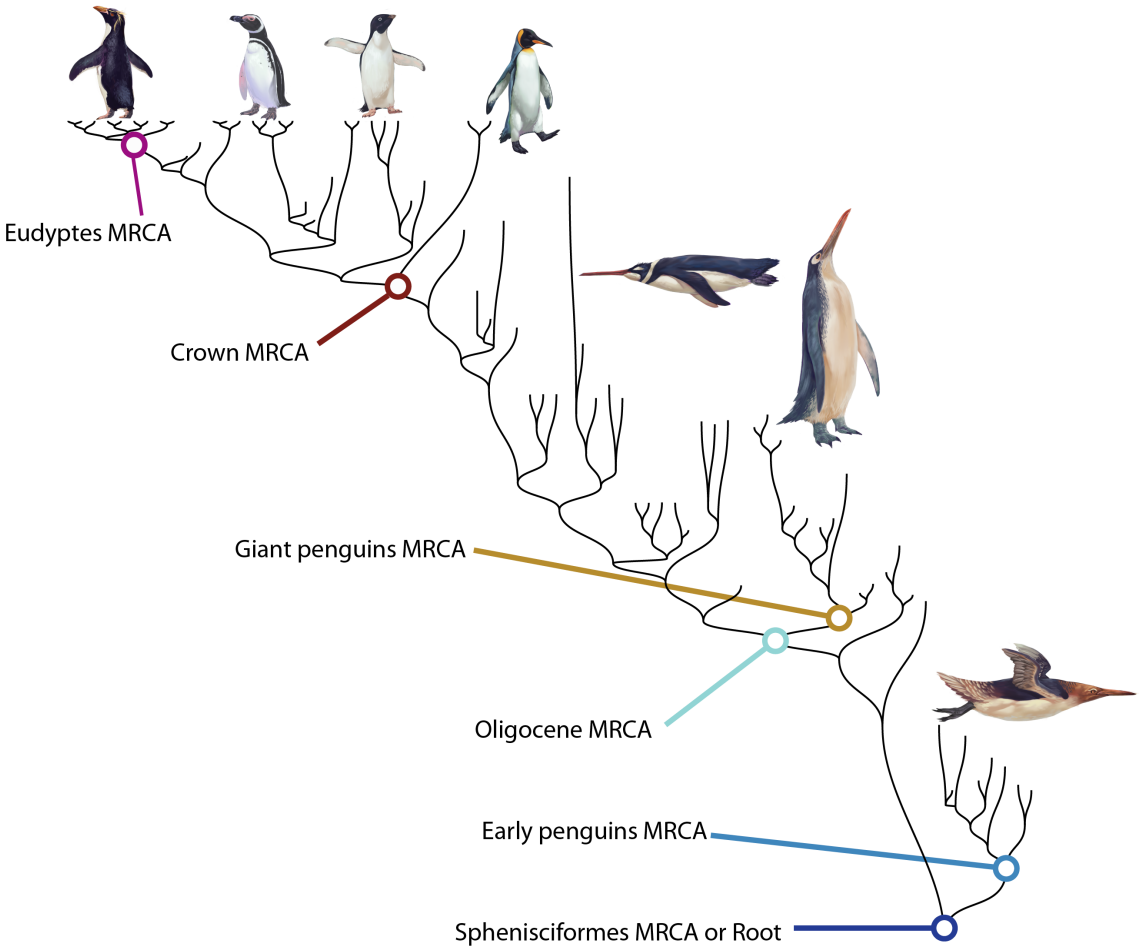


Figure 4.5: Schematised maximum clade credibility tree from the fossilised birth death tree analysis in Chapter 2. Each node used in this chapter has been marked with the corresponding name.

4.2.5 Morphospace prediction

Separate principal component analyses were performed on the Procrustes-transformed landmarks from the humerus and tarsometatarsus comparative datasets to visualise the locations of these bones in their respective morphospaces through time. The use of a traditional principal component analysis was preferred over the more recently developed phylogenetic principal component analysis (PhyPca; Revell, 2010) and phylogenetically-aligned component analysis (PACA; Collyer and Adams, 2021) because many of the inferred ancestral shapes derive from different phylogenetic topologies. Hence, accounting for phylogeny when constructing a morphospace could potentially bias the projection of the specimens into a common morphospace. The PCA was performed on both comparative datasets, then once the principal component scores were generated all ancestral shapes computed in previous steps were projected into the ordination space given by the principal components. In order to do so the ancestral shapes were combined in a single data frame that matched the structure of the original comparative dataset (i.e. column order). The resulting data frame was scaled with the original PCA scales and then multiplied by the matrix of variable loadings (i.e. the matrix whose columns contain the eigenvectors of the PCA). This approach of shape re-projection was preferred over a PCA on the whole combined datasets to ensure that the shape of ancestors did not bias the results. The age of each shape inferred at each node was estimated from the node age in the corresponding tree. Age information from the trees was then used to generate plots of morphospace location through time. Lastly, incomplete fossils and specimens for which there was no phylogenetic information available were projected into the ordination space generated above with the PCA to assess where these specimens were located in relation to their inferred age. Missing landmarks were estimated using `geomorph::estimate.missing` with the thin plate spline method (Adams and Otárola-Castillo, 2013; v. 3.3.1).

4.3 Results

4.3.1 Model performance

Evolution of the shape of the humerus and tarsometatarsus across the maximum clade credibility (MCC) tree was best explained by an evolutionary model where rates of change were uncorrelated between branches (as determined by Bayes factors; Table 4.2). The Bayes factor for the uncorrelated clock rates model tends to 1.0 for the humerus comparative dataset and is equal to 0.99 for the tarsometatarsus comparative dataset meaning that both the strict and the auto-correlated clock are substantially less-consistent with the data (Table 4.2).

Table 4.2: Bayes factors (BF) results from the three types of morphological clock analysis performed for both the humerus (HUM) and tarsometatarsus (TMT) comparative datasets. Along with BF are reported the log of the BF (Log BF), the probability of the model (Pr), the log likelihood (LogLik) and standard errors (se).

		Strict Clock	**Uncorrelated Clock**	**Auto-Correlated Clock**
HUM	BF	1.63×10^{-41}	1.00	3.55×10^{-3}
	Log BF	-93.92	0	-5.64
	Pr	1.63×10^{-41}	9.96×10^{-1}	3.54×10^{-3}
	LogLik	-13521.1	-13524.9	-13785.3
	se	0.025	0.043	0.047
TMT	BF	1.15×10^{-118}	0.99	1.40×10^{-2}
	Log BF	-271.56	0	-4.27
	Pr	1.14×10^{-118}	9.86×10^{-1}	1.38×10^{-2}
	LogLik	-14563.7	-14292.1	-14296.4
	se	0.0046	0.0069	0.010

4.3.2 Divergence time

Ages for the origins of the crown-penguin MRCA were around 33 Ma for the humerus comparative dataset and 56 Ma for the tarsometatarsus dataset when estimated using the morphological clock method. Origins for the Sphenisciformes MRCA and giant penguin MRCA were 146 Ma and 64 Ma, respectively for the humerus dataset. For the tarsometatarsus dataset the origins for the Sphenisciformes MRCA and giant penguin MRCA were 135 Ma and 58 Ma. These node age estimates are

older than the dates for the same nodes estimated from the phylogenetic analysis performed in Chapter 2 (Fig. 4.6, Fig. 4.7 and Table 4.3). For comparison, the origin of the crown clade was estimated to be around 15 Ma in the analysis performed in Chapter 2 (Table 2.1). Furthermore, the age estimates for the nodes in the divergence estimates calculated with the morphological clock method show that deeper nodes exhibit greater amounts of age discordance whereas instead nodes placed closer to a dated tip seems to agree more (Fig. 4.7 and Table 4.3). The humerus dataset morphological clock analysis consistently exhibits younger ages for nodes that are younger than the crown-penguin MRCA when compared with the phylogeny calculated in Chapter 2. Likewise, nodes that are older than the crown-penguin MRCA tend to have older dates in the humerus dataset morphological clock analysis compared with the phylogeny calculated in Chapter 2. The shape of the posterior densities for the age of the Sphenisciformes MRCA shows an abrupt truncation of the lower end for both humerus and tarsometatarsus morphological clock analyses (Fig. 4.7-Origin). This abrupt truncation suggests that the clock pushes the root to much older estimates. To test whether the prior on the root might also bias the age distributions around other nodes the analysis was repeated by allowing a time window for the Sphenisciformes MRCA of 61 to 350 Ma instead of 61 to 180 Ma. The revised analysis with the extended root age shows that all node positions except the root were unchanged, implying that the uncertainty is affecting only the root (Fig. C.1).

Table 4.3: Means of clade origins (millions of years ago) estimated from the uncorrelated morphological clock divergence analyses. Lower and upper refers to the 89% credible interval. Note that *Megadyptes* is not included here because there is a single living species and an insufficient number of specimens from *Megadyptes antipodes waitaha* were available for analysis.

Node	**Dataset**	**Mean**	**Lower**	**Upper**
Aptenodytes MRCA	Humerus	13.21	1.88	30.23
Aptenodytes MRCA	Tarsometatarsus	15.48	3.90	31.49
Pygoscelis MRCA	Humerus	19.39	8.83	31.22
Pygoscelis MRCA	Tarsometatarsus	19.16	6.48	35.58
Eudyptes MRCA	Humerus	13.85	6.38	22.63
Eudyptes MRCA	Tarsometatarsus	19.26	10.88	29.21
Eudyptula MRCA	Humerus	2.86	0.12	9.03
Eudyptula MRCA	Tarsometatarsus	5.18	0.67	13.24
Crown-penguin MRCA	Humerus	33.60	22.40	46.65
Crown-penguin MRCA	Tarsometatarsus	56.40	38.23	77.76
Oligocene MRCA	Humerus	74.71	56.30	98.63
Oligocene MRCA	Tarsometatarsus	76.98	55.92	104.92
Giant penguins MRCA	Humerus	64.53	49.21	84.00
Giant penguins MRCA	Tarsometatarsus	58.71	44.53	77.84
Early penguins MRCA	Humerus	94.93	76.55	119.42
Early penguins MRCA	Tarsometatarsus	79.51	65.21	103.60
Sphenisciformes MRCA or Root	Humerus	146.10	102.16	179.67
Sphenisciformes MRCA or Root	Tarsometatarsus	135.45	87.43	177.82

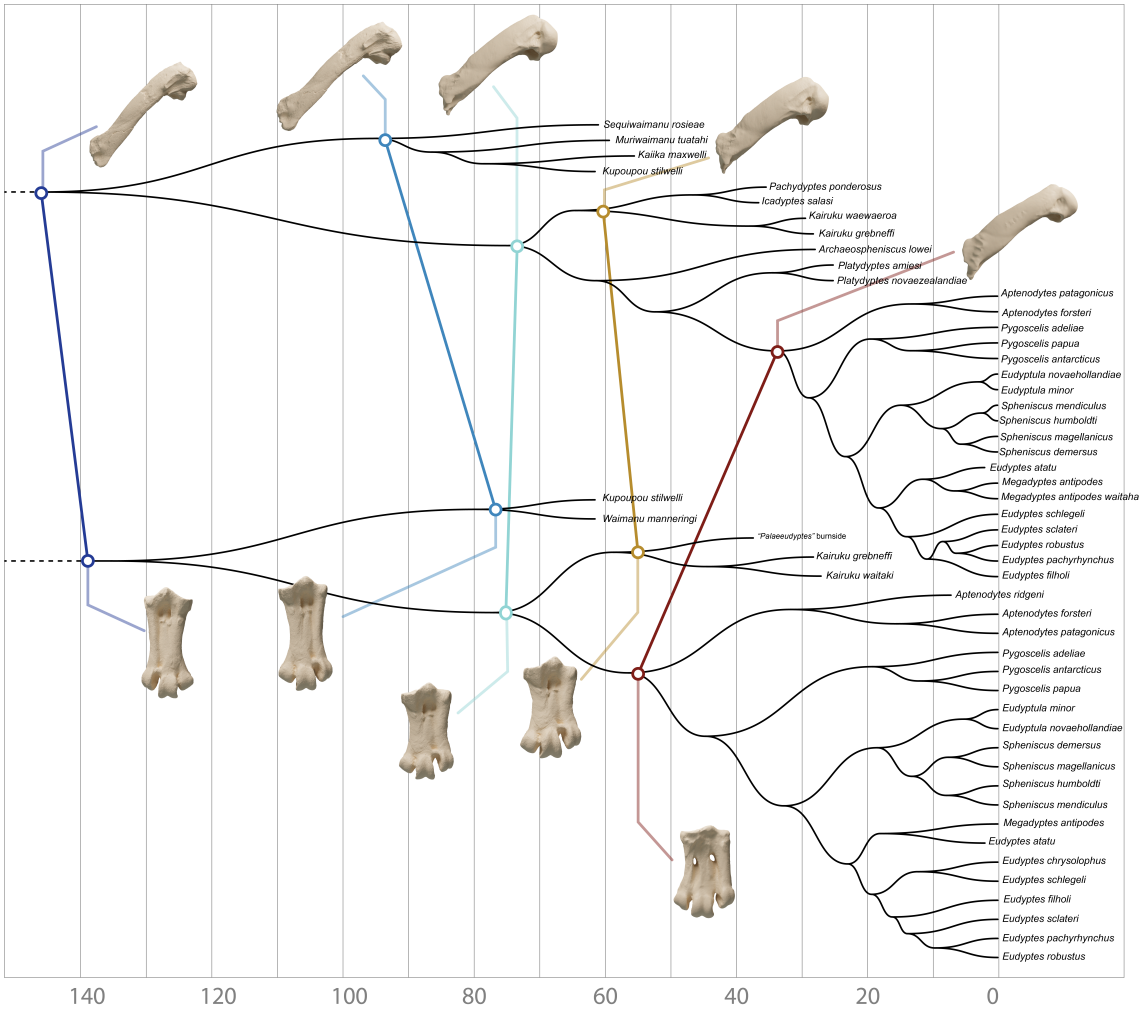


Figure 4.6: Node ages estimated using the morphological clock method applied to the humerus (upper) and tarsometatarsus (lower) comparative datasets. Numbers on the horizontal axis are in millions of years. Segments connect "homologous nodes". Ancestral shapes inferred from penalised likelihood models are shown for selected nodes.

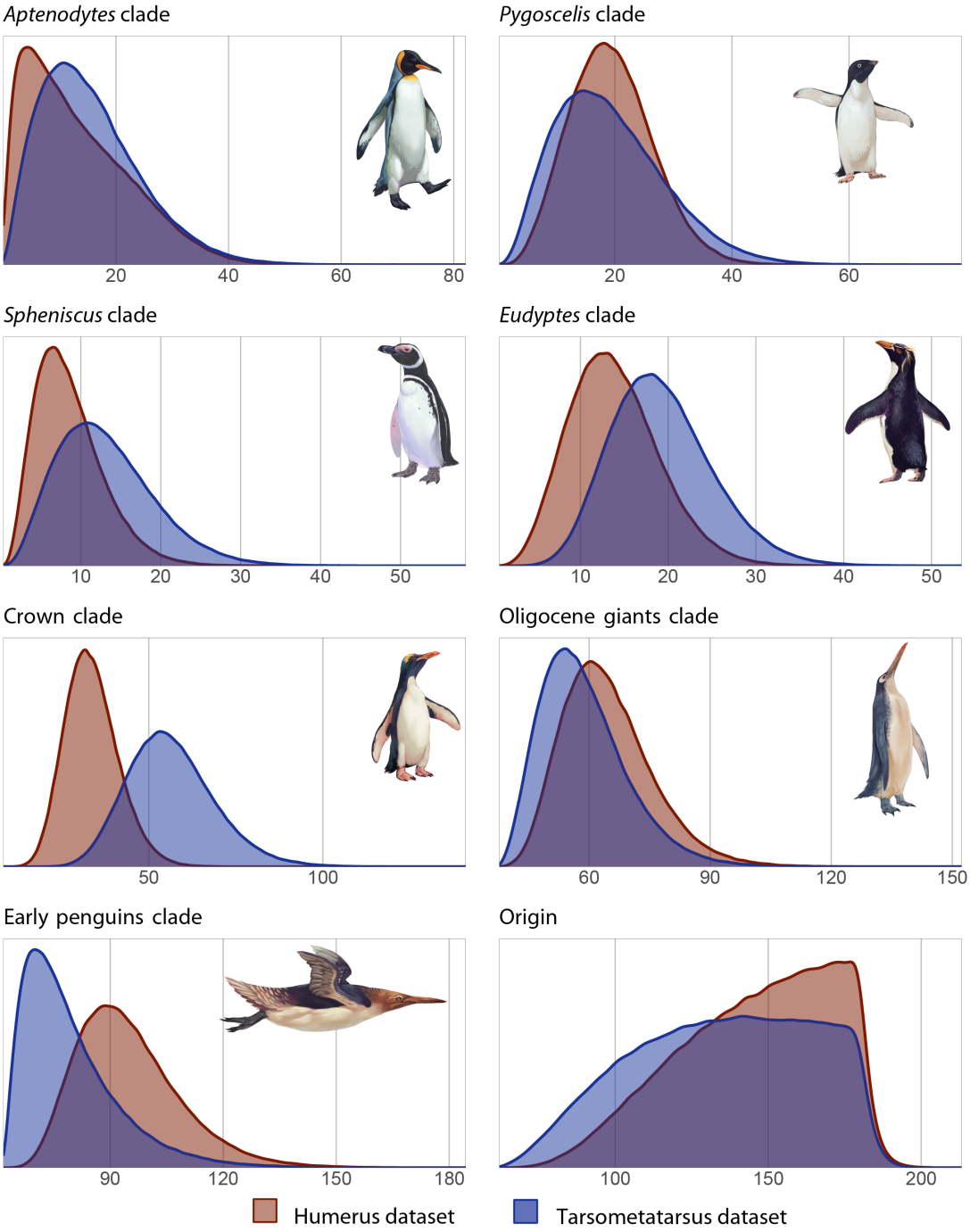


Figure 4.7: Distributions of ages of selected clades in the morphological clock analyses performed on the humerus (red) and tarsometatarsus (blue) comparative datasets. Plots represent densities resulting from the Markov Chain Monte Carlo independent rates analyses. Numbers on the horizontal axes represent time in millions of years.

Reducing the humerus dataset did not result in major differences between estimated times of clade origin, besides a slight thickening of the tail of each distribution (Fig. 4.8, reduced humerus datasets are shown in yellow). The thicker tail means that the reduced humerus dataset tends to push age estimates for the origins of clades slightly back in time compared with the complete humerus dataset (Arcila et al., 2015). The origin of the crown clade has an estimated mean around 33.6 Ma in the complete humerus dataset compared with 40.7 Ma in the reduced dataset (Fig. 4.7 and Fig. 4.8). The MRCA for the Oligocene giant penguins is estimated at 74.7 Ma in the complete humerus dataset and 81.03 Ma in the reduced humerus dataset (Tables 4.4). With the Oligocene giant penguin clade as the only exception, differences between node ages tended to be greater between the humerus and tarsometatarsus dataset when compared with node age differences between the complete humerus and reduced humerus datasets (Table 4.4). Hence, the major differences between estimated rates of morphological evolution are not exclusively due to differential sampling of datasets.

Table 4.4: Average clade origins differences comparing the morphological clock analyses between the complete humerus dataset to the reduced humerus dataset and to the tarsometatarsus dataset. Lower and upper refers to the 89% credible interval.

Node	Compared datasets	Mean difference	Lower	Upper
Aptenodytes MRCA	Complete humerus and reduced humerus	-1.59	-25.36	20.66
Aptenodytes MRCA	Complete humerus and tarsometatarsus	-2.27	-22.43	18.67
Pygoscelis MRCA	Complete humerus and reduced humerus	-2.68	-20.26	14.37
Pygoscelis MRCA	Complete humerus and tarsometatarsus	0.23	-19.17	17.84
Eudyptes MRCA	Complete humerus and reduced humerus	-1.42	-13.86	10.65
Eudyptes MRCA	Complete humerus and tarsometatarsus	-5.41	-17.81	6.73
Eudyptula MRCA	Complete humerus and reduced humerus	-0.47	-8.47	6.85
Eudyptula MRCA	Complete humerus and tarsometatarsus	-2.32	-11.18	5.49
Spheniscus MRCA	Complete humerus and reduced humerus	0.21	-8.29	9.02
Spheniscus MRCA	Complete humerus and tarsometatarsus	-4.55	-16.71	6.59
Crown-penguin MRCA	Complete humerus and reduced humerus	-7.11	-28.62	12.80
Crown-penguin MRCA	Complete humerus and tarsometatarsus	-22.80	-47.02	-0.44
Oligocene MRCA	Complete humerus and reduced humerus	-6.33	-42.51	27.30
Oligocene MRCA	Complete humerus and tarsometatarsus	-2.27	-35.95	29.75
Giant penguins MRCA	Complete humerus and reduced humerus	-0.70	-27.54	25.10
Giant penguins MRCA	Complete humerus and tarsometatarsus	5.82	-18.74	30.25
Early penguins MRCA	Complete humerus and reduced humerus	-0.59	-31.84	30.35
Early penguins MRCA	Complete humerus and tarsometatarsus	15.42	-14.68	44.76
Sphenisciformes MRCA or Root	Complete humerus and reduced humerus	-1.69	-58.12	54.42
Sphenisciformes MRCA or Root	Complete humerus and tarsometatarsus	10.66	-51.40	72.03

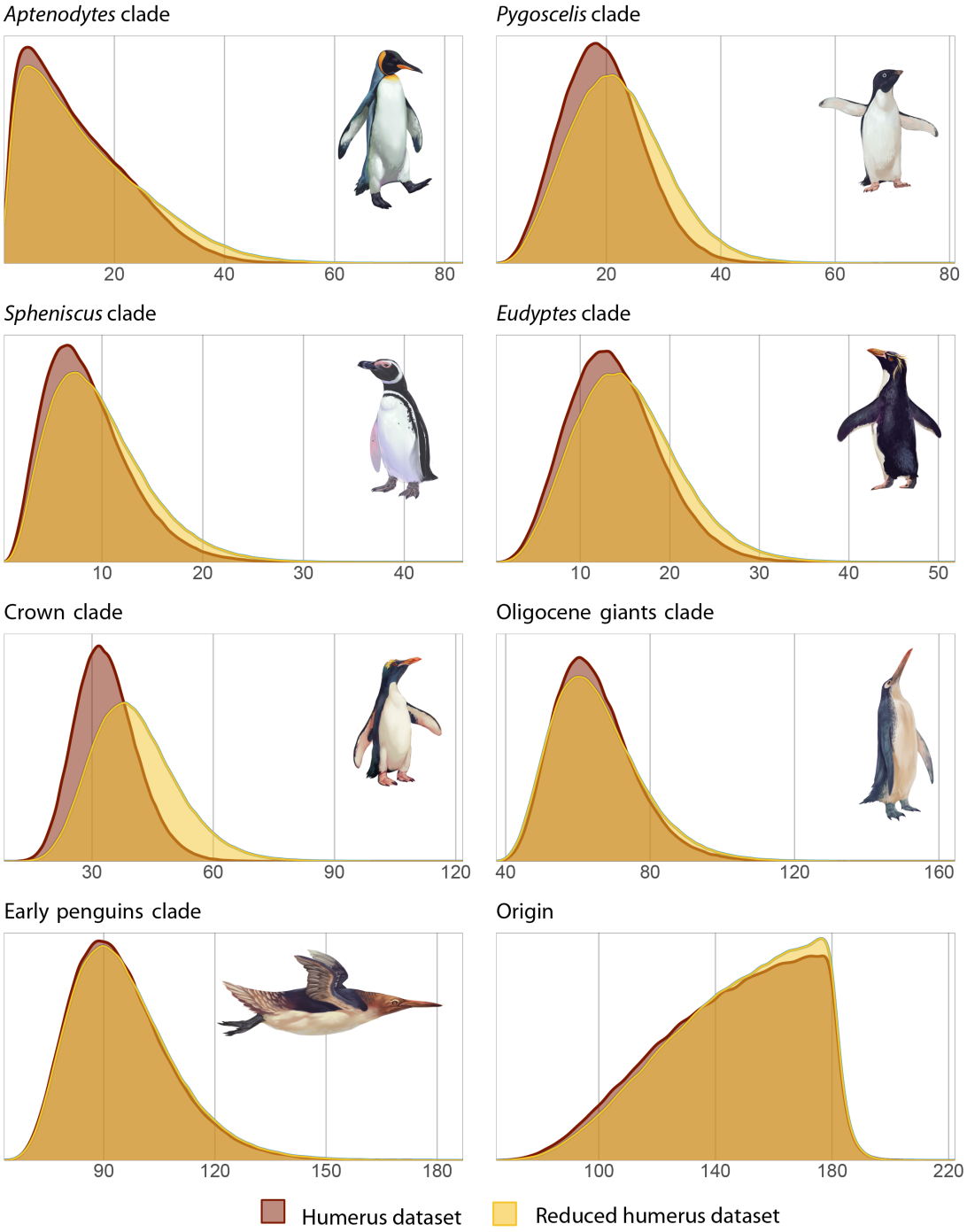


Figure 4.8: Distributions of ages of selected clades in the morphological clock analyses performed on the complete humerus (red) and reduced humerus (yellow) comparative datasets. Plots represent densities resulting from the Markov Chain Monte Carlo of independent rates analyses. Values on the horizontal axes represent time in millions of years.

4.3.3 Penalised likelihood framework

The generalised information criteria (GIC) from fitting the penalised likelihood models of evolution showed that the Pagel's lambda model tended to perform the best over the subsample of 100 phylogenetic trees (Fig. 4.9). The Ornstein-Uhlenbeck model was the second best performing model. The Brownian motion and early burst models exhibited identical distributions of GIC values (Fig. 4.9) and had the lowest goodness of fit of all four modes of evolution.

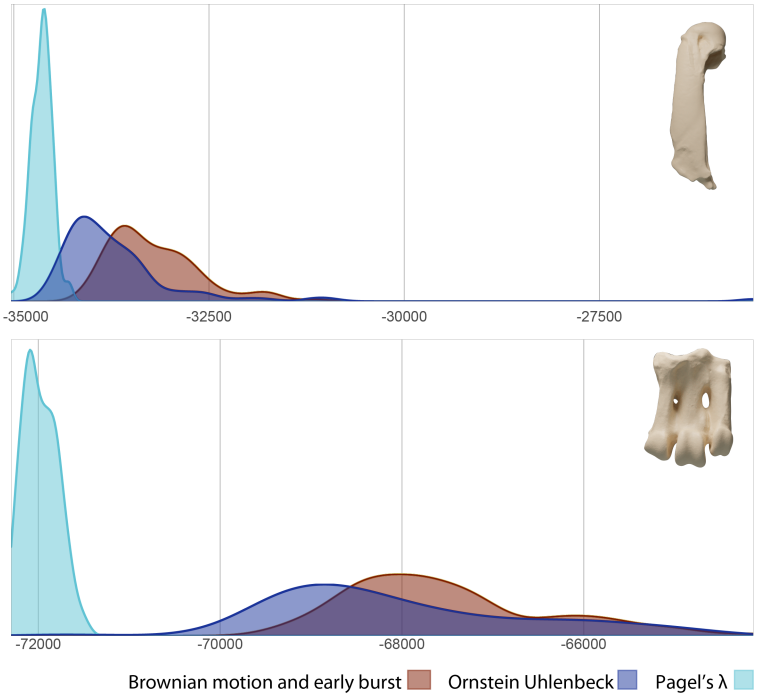


Figure 4.9: Generalised information criterion (GIC) distributions for each of the four tested models of evolution for the complete humerus comparative dataset (upper) and the tarsometatarsus (lower) comparative dataset. Densities show the distribution of the GIC value for all 100 trees sampled from the posterior. Note that Brownian motion and early burst (red) have identical GIC distributions.

The median for the λ parameter in the complete humerus comparative dataset was around 0.88 with 89% credible intervals (CI) between 0.57 and 0.95 (Fig. 4.10 red portion). The distribution of λ exhibited a relatively long tail toward lower values. The distribution of the same parameter for the tarsometatarsus dataset gives a median value of λ around 0.86 with 89% CI between 0.83 and 0.89 (Fig 4.10 blue portion). The distribution of λ for the tarsometatarsus dataset had a more symmetric and tight distribution of values.

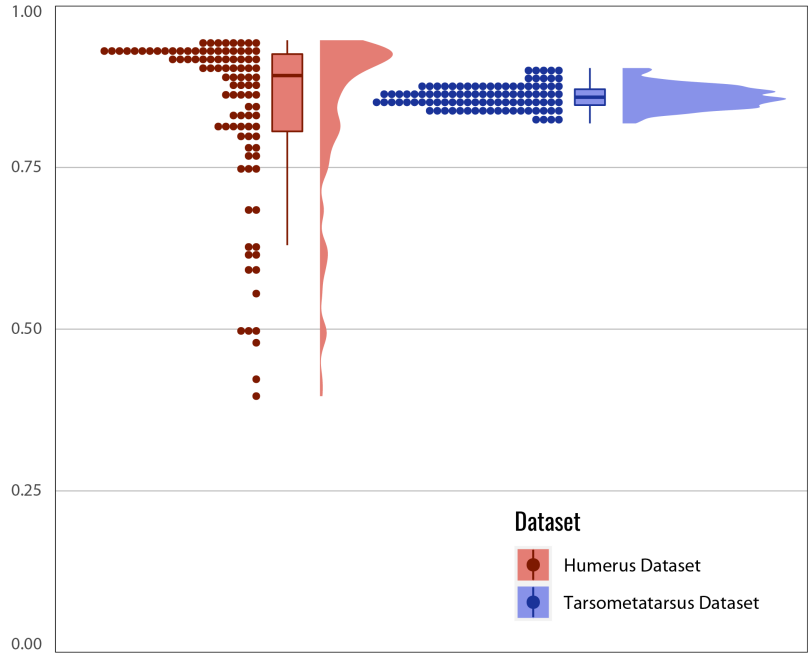


Figure 4.10: λ raincloud plot distributions. λ value distributions for all 100 trees on which the penalised likelihood method was used for both the complete humerus comparative dataset (red) and the tarsometatarsus comparative dataset (blue). Each dot represents the λ value for each phylogenetic tree to which the penalised likelihood method was applied.

Estimated rates of morphological change calculated for separate datasets cannot be directly compared because the Procrustes distances used to calculate these rates are only meaningful within the same dataset. However, comparing how the distributions of morphological rate values shifted in time from separate analyses may still provide some insight into the evolutionary rates of different bones. Rates of evolution along branches younger than the crown MRCA exhibit the lowest values in analyses of the humerus. In contrast, the shape of the humerus evolved at its fastest rates between the early penguin MRCA and the Oligocene MRCA. Rates of shape evolution were thus moderate between the time of the Oligocene MRCA and the crown MRCA. Taken altogether, the rate of morphological evolution of the humerus had a steady decrease from the origin of Sphenisciformes towards modern day (Fig. 4.11 left panel). The tarsometatarsus similarly shows an overall decrease in rate of morphological change but the change occurs more abruptly. The rate of morphological change between Oligocene MRCA to crown MRCA is almost identical to the rates observed between the crown MRCA to *Eudyptes* values (Fig. 4.11

right panel, green and red curves) and is substantially lower than the observed rates between the Paleocene MRCA and the Oligocene MRCA (Fig. 4.11 right panel, blue curves).

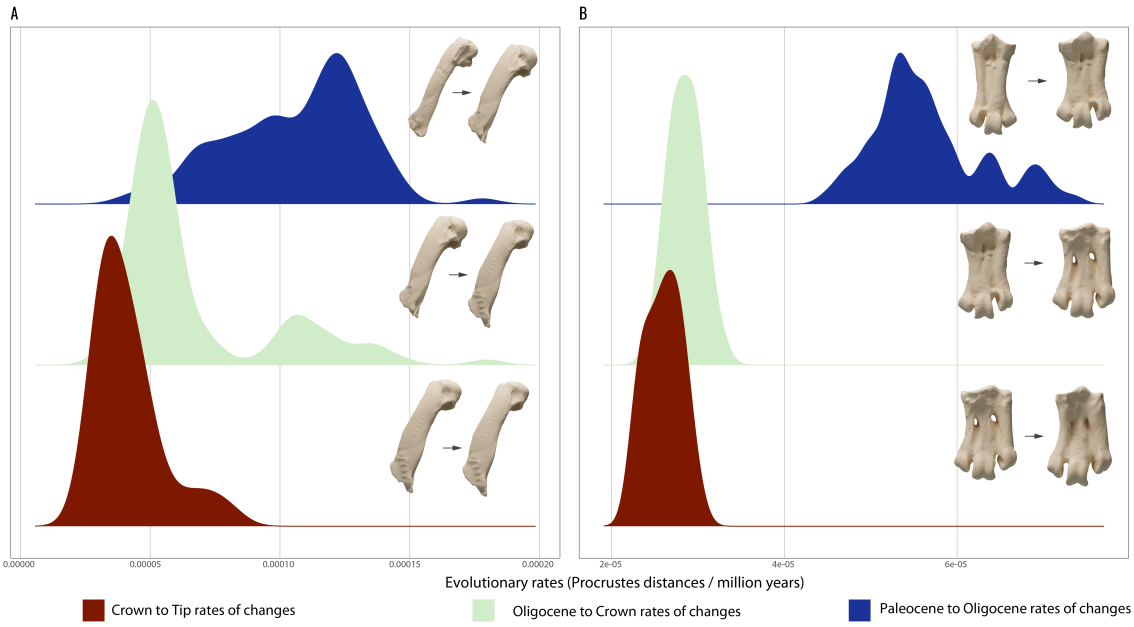


Figure 4.11: Comparison of evolutionary morphological rates for the A) humerus and B) tarsometatarsus. Each curve represents the distance between two homologous points on 100 phylogenetic trees. Blue curves represent distances between the Sphenisciformes most recent common ancestor (MRCA) and the Oligocene MRCA, green curves represent distances between the Oligocene MRCA and the crown MRCA, and red curves represent distances between the crown MRCA and the *Eudyptes* MRCA. Densities show estimated rates of morphological change per million years.

4.3.4 Morphospace

The first two axes of the principal component analysis (PCA) applied to the dataset of Procrustes-transformed landmarks from the complete humerus dataset explain 32.4% and 19.7% of the variance, respectively. The first principal component (PC1) is associated with the widening of the humeral shaft and head. Specimens with higher PC1 scores tend to be wider and more robust in overall shape, with proportionately enlarged heads and anteroposteriorly broadened shafts (e.g. *Pachydyptes ponderosus*). Specimens with lower PC1 scores are more gracile and have a comparatively more-slender shaft with a proportionately smaller humeral head (e.g.

Kuopoupou stilwelli). The second principal component (PC2) denotes instead the shape and the relative size of the margins of the fossa pneumotricipitalis. High PC2 scores are associated with smaller fossae with margins that look more horizontal in caudal view (e.g. *Muriwaimanu tuatahi*) whereas instead lower PC2 scores are associated with a more-parabolic margin in caudal view around a relatively wider fossa (e.g. *Aptenodytes forsteri*).

The PCA applied to the Procrustes-transformed landmarks from the tarsometatarsus explains 34.4% and 10.5% of the variance in that dataset along PC1 and PC2, respectively. For the tarsometatarsus, PC1 describes the relative elongation of the tarsal shaft. Specimens with higher PC1 values are more-slender and have an overall more rectangular shape in dorsal view (e.g. *Waimanu manneringi*). Consequently, lower PC1 values are associated with more squared shaped morphologies (e.g. *Aptenodytes* sp.). Higher PC2 scores are associated with more straight lateral profiles in dorsal view (e.g. *Eudyptula* spp., *Spheniscus* spp.,*Eudyptes* spp.,), whereas lower PC2 scores characterise specimens that have a more curved lateral profile (e.g. *Kairuku*).

Given that together PC1 and PC2 represent a considerable amount of the total morphological variation in both datasets, the morphospace was described using the first two components for both humerus and tarsometatarsus. The position of score values along PC1 and PC2 (i.e. morphospace occupation) through time for the humerus (Fig. 4.12) and for the tarsometatarsus (Fig. 4.13) show little overlap between the earliest penguins, the Oligocene giants, and the crown clade. The occupation of different regions of morphospace by these clades suggests that shape variation along PC1 and PC2 is heavily influenced by phylogeny. The pattern of morphospace occupation through time for the humerus thus suggests that, from the Paleocene to the Oligocene, the extinct species may have had a much broader distribution compared to extant species. For the tarsometatarsus the crown clade (Fig. 4.13) exhibits a much broader distribution along PC1 and PC2. For example, both extant species of *Aptenodytes* have PC1 and PC2 score values that are more

similar to the score values from the tarsometatarsi of *Kairuku* when compared with the PC1 and PC2 score values from tarsometatarsi of other crown taxa (Fig. 4.13B). The morphological affinity between *Aptenodytes* and *Kairuku* tarsometatarsus may be explained by gross outline and allometry being strongly weighted along both the first and second principal components. Consider that both *Aptenodytes* and *Kairuku* have the more box-shaped tarsometatarsus observed in larger penguin species.

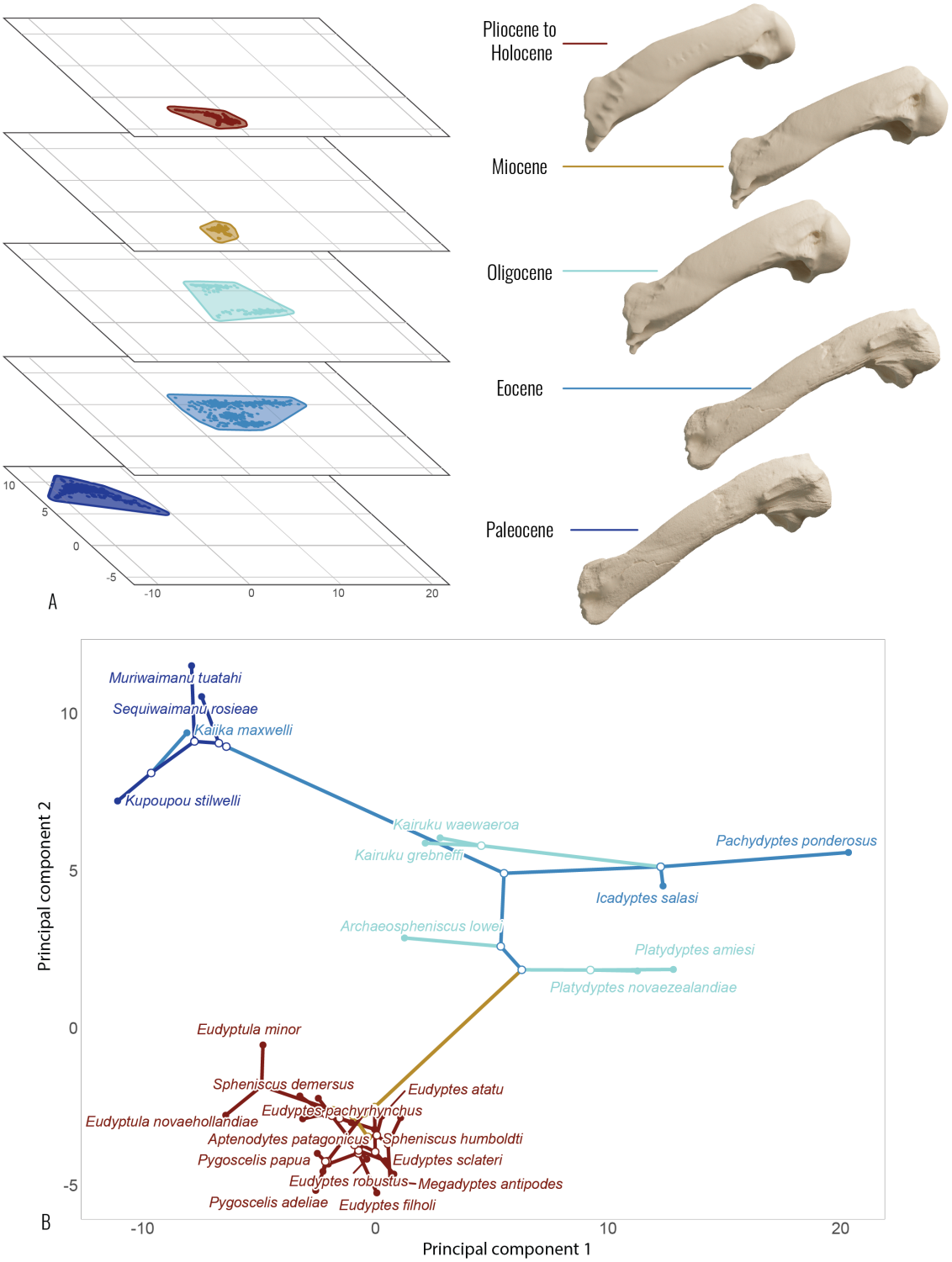


Figure 4.12: Morphospace of penguin humeral shape. A) Morphospace occupation through time. Each section is presented along with the average shape estimated for the corresponding time slice. B) Phylomorphospace of extinct and extant penguin humeri. Both A and B are shown with the same axis coordinates.

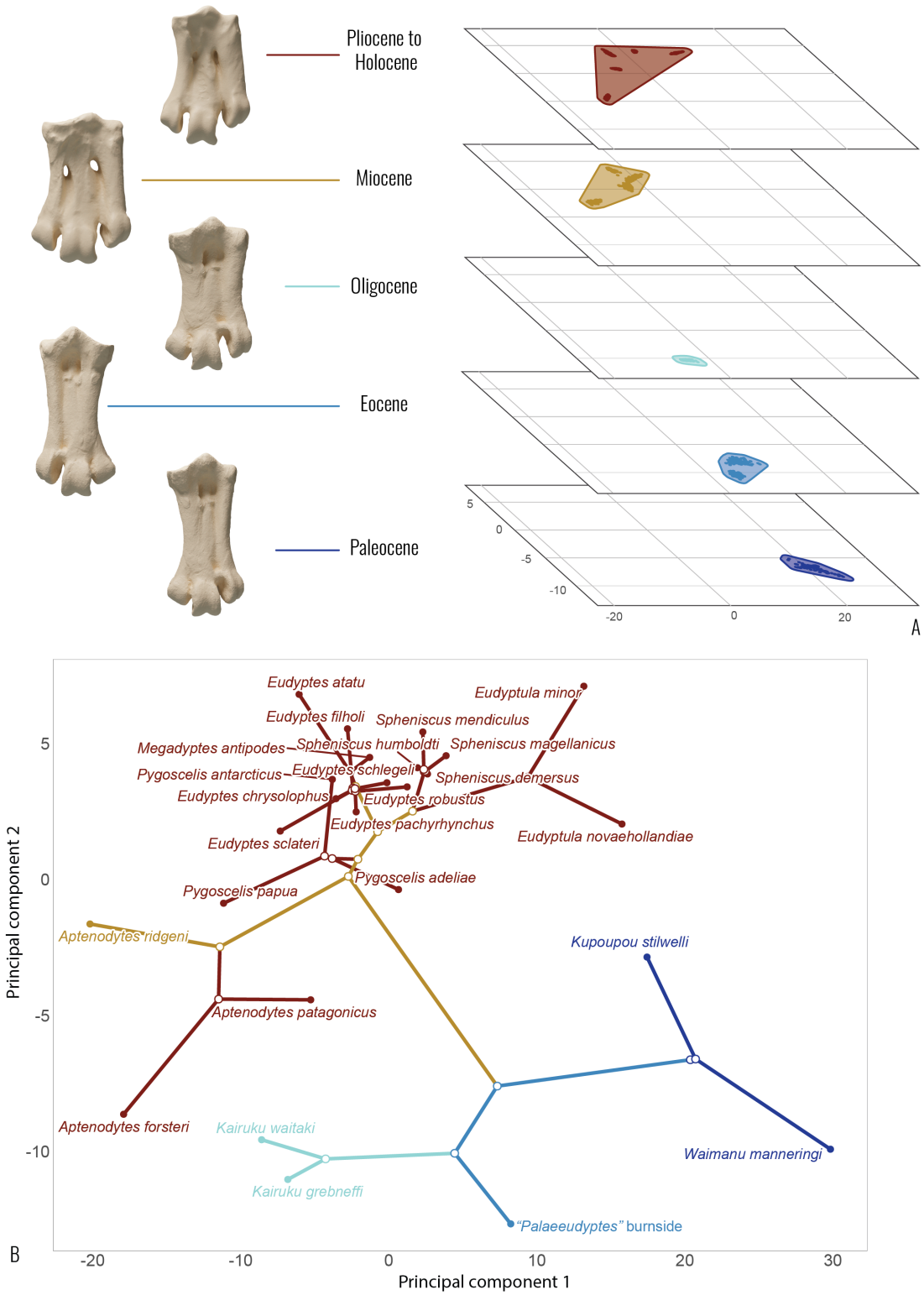


Figure 4.13: Morphospace of penguin tarsometatarsal shape. A) Morphospace occupation through time. Each section is presented along with the average shape estimated for the corresponding time slice. B) Phylomorphospace of extinct and extant penguin tarsometatarsi. Both A and B are shown with the same axis coordinates.

4.3.5 Morphospace prediction

The humeri from BMBH RMA43 *Anthropornis nordenskjoeldi*, NMNZ S47304 Sphenisciformes indet., OU22168 *Kairuku* sp. indet., and the DM1449 "Seal rock specimen" have shapes that project very near to the shapes of humeri from similarly-aged penguin species along PC1 and PC2 (Fig. 4.14 A-B). Notably, the shape of the humerus from BMBH RMA43 *Anthropornis nordenskjoeldi* projects within the convex hull created by the shapes of humeri from similarly-aged penguins. In contrast, the shape of the tarsometatarsi from OU21977 "*Pakudyptes*", OU22127 *Palaeeudyptes antarcticus*, and OU22181 *Palaeeudyptes gunnari* seem to be more distant from their expected morphospace area (Fig. 4.14D). Perhaps the tarsometatarsus morphospace is more "skewed" by allometry as mentioned above. Moreover, the PC1 and PC2 score values of "*Pakudyptes*" are relatively closer to the PC1 and PC2 values of crown species (Fig. 4.14C red area) compared to the PC1 and PC2 score values of more contemporaneous fossils (Fig. 4.14C blue area).

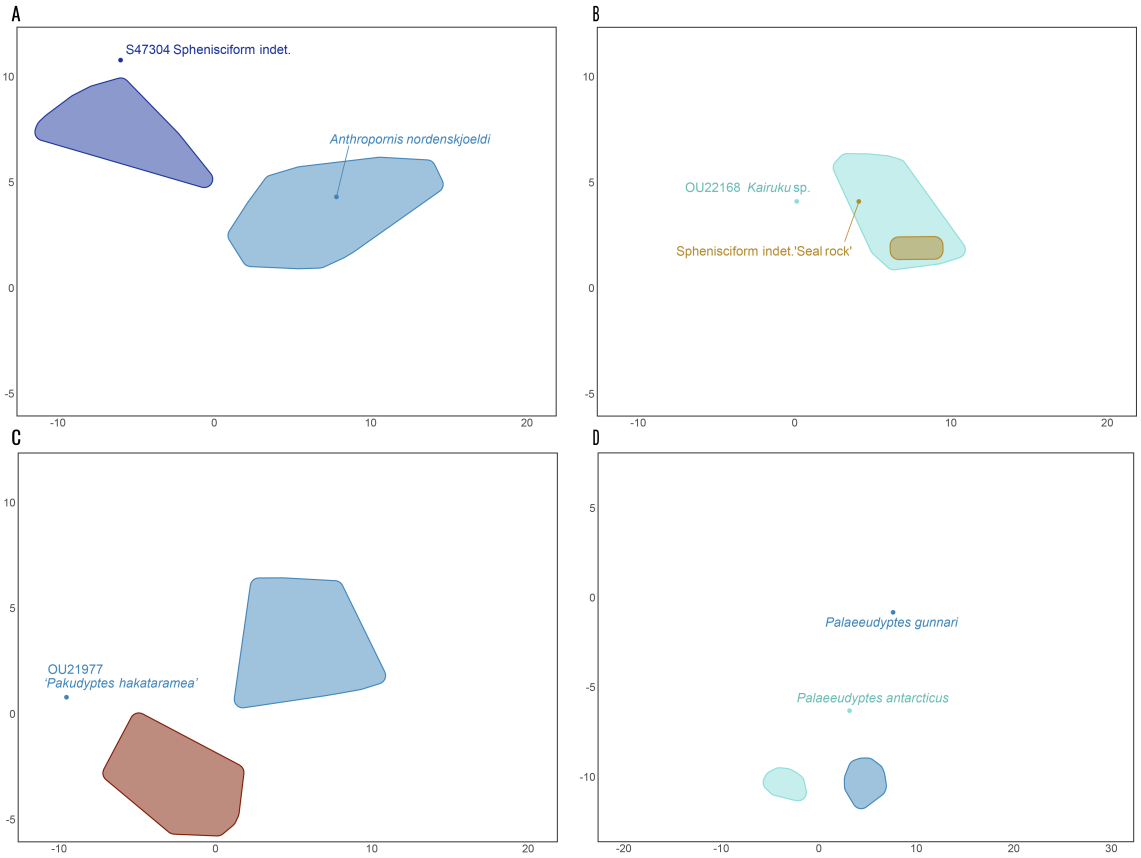


Figure 4.14: Projection of fossil A-C) humeri and D) tarsometatarsi that were not included in the comparative datasets into the principal component one (PC1) and PC2 morphospace calculated from the comparative datasets. Projected fossils were part of the Post-hoc dataset. Panels A to C represent the same humerus morphospace seen in figure 4.12 and panel D represents the same tarsometatarsus morphospace seen in figure 4.13. Coloured regions show convex hulls of morphospace occupation for bone shapes from hypothetical ancestors that were found inside the same time range as the sampled fossil shape.

4.4 Discussion

4.4.1 Morphological clock

The Bayes factor computation shows that the phylogenetic distribution of humerus shape and tarsometatarsus shape are best explained by a morphological clock model that has rates of evolution along each branch that are independent from one another. Hence, shape evolution for these two bones is best explained by heterogeneous evolutionary rates over the phylogenetic tree, and that such heterogeneity seems to not be influenced by parent branches. This is reminescent of a punctuated equilibrium-like model of evolution (Gould & Eldredge, 1977) for the penguin clade where rates of evolution may vary suddenly. Note though that the independence of rates may be a potential consequence of "outliers" like *Waimanu* and *Pachydyptes*. These taxa after `CONTML` analysis (Fig. 4.4) are found at the end of extremely long branches meaning that their morphologies are much more different than other penguins. It has previously been demonstrated that outliers may drive clock divergence analyses towards incorrect clade origin estimates (Hedges & Shah, 2003). Note though that the independent rates model was also found to provide the best explanation for the phylogenetic distribution of shape in the study by Álvarez-Carretero et al. (2019) when only the morphological alignment was analysed. Moreover, (Álvarez-Carretero et al., 2019) reported the presence of *Smilodon fatalis* as a potential outlier in their analysis.

The morphological clock analysis consistently predicted older clade origin dates when compared with evolutionary analyses from other studies (e.g. Blokland et al., 2019; Gavryushkina et al., 2017; Chapter 2 section 3.2). One interpretation of this finding is that the shape of the humerus and tarsometatarsus within any particular penguin taxon is often an ‘old shape’ that evolved in deeper time, and which is inherited with relatively little change during more recent speciation events. The shape of the humerus and tarsometatarsus either trace deeper into time than previously expected if morphological evolution had a uniform pace, and may show punctuated

bursts of shape change after long periods of morphological stasis. Nevertheless, wings and legs as separate locomotory modules have experienced different rates of evolutionary change. The evidence for these rates of morphological change can be seen in the age differences between homologous nodes in the morphological clock analyses of the humerus and tarsometatarsus comparative datasets (Fig. 4.7).

The clade origination dates from the morphological clock analysis on penguins highlights a pattern not too dissimilar from the morphological clock analysis performed by Álvarez-Carretero et al. (2019), although their analysis focused on order Carnivora within mammals. Firstly, crown Carnivora was estimated to have originated between 48 and 37 Ma in traditional phylogenetic studies (Heinrich et al., 2008; Tomiya, 2011). Álvarez-Carretero et al. (2019) performed several analyses on morphological alignments, molecular alignments and a combination of these two datasets; among these the analyses performed with only the morphological alignment returned overall the earliest node ages, dating the origins of Carnivora to at least 55 Ma. The penguin morphological clock results presented here likewise estimate earlier clade origins when compared with traditional phylogenetic methods, and often the penguin age estimates are displaced far earlier in time when compared with the age displacement in Álvarez-Carretero et al. (2019). It is thus possible that morphological clock analyses are methodologically-prone to pushing clade origins deeper into time. Although, compared to the Carnivora analysis, the penguin analysis exhibits a sampling of taxa that is more evenly distributed through time. Another reason for this discrepancy between the node ages from the morphological clock analyses and traditional phylogenetic analyses could be the absence of an outgroup. Although these tip-dated analyses require that an outgroup not be included (Álvarez-Carretero et al., 2019; Cau, 2017; Gavryushkina et al., 2017) it is possible that establishing a clade outside the focus group with a distinct overall morphology may help to "pin-point" more effectively the rates of change over time in absolute terms.

Although the estimated ages of nodes may be influenced by the number of taxa

in the analysis, the hypothesis that differences between the complete humerus and tarsometatarsus comparative datasets are exclusively due to undersampling can be ruled out for two main reasons: 1) node ages remained mostly unchanged when comparing the morphological clock analyses for the reduced and complete humerus datasets (Fig. 4.8); 2) the root of the tree is younger in the morphological clock analysis (using the independent rates model) performed on the tarsometatarsus dataset when compared with the complete humerus dataset (Fig. 4.6 and Fig. 4.7). The former point indicates that differences between the node ages in the analyses of the humerus and tarsometatarsus datasets are not due exclusively to differences in the number of taxa. The latter point is the exact opposite of what one would expect if sample size was influencing node age given that the tarsometatarsus dataset is more "undersampled" (Arcila et al., 2015). Note though that node ages did slightly increase (i.e. become older) in the reduced humerus dataset analyses, with a thickening of the distribution's tails. It is thus possible that taxon sampling has an impact on age uncertainty rather than affecting the median age of a node. By increasing the uncertainty around the age of each node there is a subsequent thickening in distribution tails, thus resulting in overall older estimates for all clade origins. This interpretation may be in line with one of the properties of Bayesian modelling that sees uncertainty increase over parameter estimates (i.e. clade origins) as the number of observations decrease (i.e. number of taxa included in a given dataset) (Gelman et al., 2013; Kruschke, 2015; McElreath, 2020).

4.4.2 Penalised likelihood results

The GIC distributions (Fig. 4.9) from the penalised likelihood analyses indicate that the Pagel's lambda model is the one that fits the best on all analysed posterior trees, for both the complete humerus and tarsometatarsus datasets. The Pagel's lambda model here suggests that there is a strong phylogenetic signal in the shape of the humerus and in the shape of the tarsometatarsus (Boettiger et al., 2012; Clavel et al., 2019), but that shape variation across the tree is not simply described by Brownian

motion (i.e. a random walk where the λ parameter is assumed to be 1). This result is supported by the median value of the inferred λ parameter which was between 0.85-0.86 for each dataset. Evolution under a Pagel's lambda model would suggest that the shape of the humerus and tarsometatarsus is subject to phylogenetic constraint, however, further analysis would be required to assess whether this signal is the product of a real macroevolutionary pattern or a consequence of taxon sampling (Boettiger et al., 2012). Moreover, the distribution of the λ for the humerus dataset is wider, more asymmetric, and has a longer tail (Fig. 4.10), meaning that for a subsample of the trees used for the humerus dataset the "phylogenetic signal" can become much lower than for the tarsometatarsus. At first it may seem that tarsometatarsus is more phylogenetically constrained compared with the humerus, but this conclusion needs to be reviewed cautiously. Given that there are more taxa in the humerus dataset the sampling of the posterior trees may explore more diverse phylogenetic topologies than in the tarsometatarsus dataset. Hence, the uncertainty around the λ parameter may be influenced by heteroscedasticity between datasets.

4.4.3 Evolutionary rates

The node ages estimated for penguins using humerus and tarsometatarsus shape suggest that the wing as a locomotory module experienced its most rapid evolutionary changes very early in the history of penguins, followed by a more steady and constant decrease in rates from the end of the Paleocene to modern day (Figs. 4.6 and 4.11). The shape of the lower leg experienced more dramatic changes at later stages of penguins history compared with the wing. When comparing how the shape of the humerus and the shape of the tarsometatarsus changed through time (Figs. 4.12A and 4.13 A), we observe that most of the morphological changes occurred between the Paleocene and Eocene for the humerus, with a straightening and widening of the humeral shaft and an enlargement of the humeral head. The major change to the gross shape of the tarsometatarsus involved the "shortening" of the tarsal shaft, which occurred from the Paleocene up to the Oligocene and thus

took more time to occur than the gross outline changes to the humerus.

If one compares humeri of early or modern penguins to those from other birds (Fig. 4.1) it is evident that, with the exception of auks, the morphology is quite distinct from most volant birds (Figs. 4.1A, C and E) as explained above in the introductory section of this chapter. Given the slower inferred rates of humerus morphological evolution with the penalised likelihood approach it is unsurprising to see that the root of the phylogeny estimated from the shape of the humerus with the morphological clock is pushed deeper to the past. With no specimens sampled outside from Sphenisciformes (the so called pan-Sphenisciforms in Ksepka and Ando, 2011) it is possible that the clock fails to estimate the correct rates of change and hence extend the same steady evolutionary rates to Mesozoic lineages. As seen above, given the functional selective pressure acting over the wing for underwater movement, it is possible that the humerus shape that fulfilled the locomotory needs of penguins was reached before the diversification of the whole clade during the Paleocene. After this major morphological shift, comparatively minor changes have occurred on the whole wing in order to increase efficiency in underwater locomotion. The same may not hold true for the tarsometatarsus given that the structure was not under the same strong selective pressure to optimise mobility on land. In this manner, most of the morphological change to the foot module was apparently "delayed" behind the changes to the humerus and so the most substantial changes to the shape of the tarsometatarsus occurred relatively rapidly from the Oligocene onwards (Fig. 4.7, blue curve).

With reference to the adaptive peaks metaphor, while the shape of the humerus may have already reached a high point (peak or ridge) for a non-volant marine diver in Paleocene southern hemisphere waters where just minor fluctuations in shape were possible, the tarsometatarsus was proportionally further away from it's own optimum. If the shape of the humerus was already close to the an adaptive high point (i.e. this shape contributed to an optimal fitness) then perhaps selection pressure acted earlier on this appendicular module that is crucial for locomotion in wing

propelled divers (Johansson & Aldrin, 2002; Mayr et al., 2021). Although finding evidence for adaptation is an extremely challenging task, one way to assess the pressure that the adaptive vertex of the constructional morphology triangle exerts is in being able to recover evidence for differential rates of evolution (Galen, 1996; Hoffmann & Ross, 2018).

4.4.4 Prediction insights

Incomplete and unnamed fossils provide a resource for testing hypotheses that involve a morphospace, including testing if the method used to generate a morphospace can be considered valid. Given that many giant and early penguins still fall within or nearby coeval areas of explored morphospace (Figure 4.14 A-B) means that the method used in this chapter to assess morphospace occupation is robust and produces meaningful results. More importantly, projecting specimens into a morphospace may provide further insight on the taxonomic identity of different specimens. For example, PC1 and PC2 scores from the humerus of OU22168 *Kairuku* sp. project close to the PC1 and PC2 scores of the two other sampled *Kairuku* species, reinforcing the attribution of this specimen to *Kairuku* in Ksepka et al. (2012) (Fig. 4.14B). Also, the humerus from S47304 Sphenisciformes indet. discussed in Blokland et al. (2019) has a morphological affinity to the humeri of *Muriwaimanu* and *Sequiwaimanu*, being closer to the latter in terms of absolute size (Figs. 4.14A). However this projecting method presents some limits. Consider the humerus of the Oligocene penguin OU21977 "*Pakudyptes*" that demonstrates that humeri with a more "crown-like" morphology may have originated earlier than previously expected. Consequently, if taxon sampling would be more extensive, perhaps we might discover that the morphospace of Oligocene taxa and crown taxa overlapped. Nevertheless, discovering that the shapes of modern bones also occur deeper in time is in line with what the morphological clock analysis is returning, with the morphological origin of humeri and tarsometatarsi for the crown clade dated between 22 and 46 Ma (89% credible intervals, Fig. 4.7, Table 4.3). More importantly, projecting fossils into

PC1 and PC2 is informative about the temporal areas of penguin evolution where shape diversity is well understood, and about the sections of time that would benefit from more input. Being able to describe and sample more specimens from these under-sampled time slices would enable us to generate more complete and precise models of morphospaces occupation.

4.4.5 Future directions for 3D Geometric morphometrics and phylogeny

Studies that aim to include geometric morphometrics as a tool to infer phylogenetic branching events may still be greatly improved. A critical point here is that `PamlX` (Yang, 2007; v 1.3.1) supports only Brownian motion models of trait evolution, and as has already seen in this thesis (i.e. Chapter 3, Section 3.2), this model may not provide the best explanation for a given dataset and phylogeny (Blomberg et al., 2020). Moreover, another limit of `PamlX` is the constrained topology that forces analyses to explore only one specific evolutionary hypothesis (i.e. the phylogenetic tree). Being able to sample from a distribution of trees would allow a great improvement in the field of phylogenetic inference. Some software packages are currently available that support the inclusion of continuous characters to infer phylogenetic trees, and these packages also have the capacity to analyse the resulting trees using multivariate Ornstein-Uhlenbeck models (other than Brownian motion) of trait evolution in a Bayesian context (e.g. `RevBayes` Höhna et al., 2016). Hence, after the dataset transformation with the $\mathbf{R}^*$ covariation matrix proposed in Álvarez-Carretero et al. (2019), it would be possible to perform the analysis also with models of trait evolution other than Brownian motion (Parins-Fukuchi, 2018). However, for high-dimensional datasets the computational time may represent a considerable hurdle.

4.5 Conclusion

As stated above, finding evidence for adaptations in deep time is an extremely challenging matter (Futuyma, 2010; Losos, 2011), however in the current chapter differences among rates of morphological evolution were used as a proxy to measure this evidence. The analysis of evolutionary rates was performed with two separate methods: the morphological divergence analysis (Álvarez-Carretero et al., 2019) and a penalised likelihood approach (Clavel et al., 2019). The former technique estimated rates of change in a similar way to how molecular clocks work. The second method enabled marginalisation over phylogenetic uncertainty to generate areas of morphospace occupation through time that contributed to highlight which sections of penguin evolutionary history were better understood, and which would benefit from greater taxon sampling. These two techniques both show that the humerus and tarsometatarsus changed through time at different rates. Whereas the rate of change in the shape of the humerus decreased steadily and was more uniform, the shape of the tarsometatarsus changed at a more varied rate which increased in more recent times. Adaptation may explain the differences between these rates, especially thanks to the adaptive landscape interpretation: The shape of the humerus was already well adapted to underwater locomotion before a post-Cretaceous clade diversification, whereas the shape of the tarsometatarsus was far more distant to an "adaptive peak". Given a less stringent pressure initially acting over the lower leg the morphology of the tarsometatarsus changed at a different pace compared with the humerus. If we look at crown penguins then the overall shape of the humerus was achieved earlier than the shape of the tarsometatarsus, providing evidence for a modular-type of trait evolution and suggesting that the humeral morphology is older than the tarsometatarsal morphology.

Figure 4.15: Unexpected encounter between a *Dinornis novaezelandiae* and *Eudyptes warhami* in the forest of North Island of New Zealand.

Chapter 5

Conclusion

5.1 Overview

The constructional morphology concept of Seilacher (1970) inspired the structure on which this thesis was conceived and written. By focusing on three major subjects in evolutionary biology (i.e. phylogeny, structural constraint and adaptation), fossil penguins were analysed with a linked-study approach. This thesis took advantage of recently developed methods in the field of of phenotypic character analysis like fossilised birth death tree (FBDT) models (Höhna et al., 2016), Bayesian modelling (Kruschke, 2015), 3D geometric morphometrics (3DGM) (Zelditch et al., 2012), morphological clock divergence analysis (Álvarez-Carretero et al., 2019) and penalised likelihood modeling (Clavel et al., 2019). Segmenting the analyses into three separate blocks allowed me to assess phylogeny, structural constraint and adaptation on three orthogonal perspectives. All of this was made in an attempt to follow one of the key assumptions of constructional morphology that acknowledges that all aspects of living organisms are the result of these three major forces together, each acting at different levels of magnitude.

The first step of my research aimed to reconstruct the phylogeny of penguins in order to have a better grasp of the evolutionary relationships among taxa and how these relationships may have impacted other traits. Chapter 2 describes a phylogenetic study using one of the most extensive penguin data matrices that

has been assembled (with important contributions from Blokland et al. (2019) and Thomas et al. (2020)), both in terms of taxon coverage and number of characters. Focusing on the phylogenetic aspect first allows the impact of the historical vertex of the aptive triangle to be quantified during the next phases of the research. Using both parsimony-based inference and Bayesian fossilised birth death tree (FBDT) models, the emerging phylogenetic hypotheses suggested a new perspective for the stem. Where instead of the traditional "pectinated" topology (e.g. Degrange et al., 2018; Gavryushkina et al., 2017; Ksepka et al., 2012; Thomas et al., 2020), several large clades were recovered. Key amongst these is a clade uniting the newly described *Kairuku waewaeroa* (Giovanardi et al., 2021) with other giant penguins.

The Bayesian phylogenetic analysis suggested a slightly earlier origin for Sphenisciformes than has traditionally been recovered, perhaps due to the inclusion of several Paleocene taxa that had not been included in previous FBDT analyses (i.e. Gavryushkina et al., 2017; Thomas et al., 2020). This earlier origin highlights the importance of extensive taxon sampling to understand the timing of major diversification events within Sphenisciformes. Moreover, the publication of *Kairuku waewaeroa* was an integral part of the research design of this chapter and meaningfully contributes to the understanding of extinct penguin diversity. *Kairuku waewaeroa* shows that the *Kairuku* clade was widespread across northern and southern Zealandia during the Oligocene, and was more morphologically diverse than previously recognised. These results suggest that the body plan of extinct penguins may have changed dramatically in response to environmental factors and are also significant from a regional perspective. Discovering that this fossil penguin is a new species is rewarding for the Hamilton Junior Naturalist Club, the group who found and recovered it, and it encourages other young people to connect with nature and make their own discoveries.

The second part of the research aimed to establish the importance of structural constraints on estimates of body mass of extinct penguins in the general context of avian diversity and is reported in Chapter 3. The historical constraints were modeled

from the distribution of phylogenetic trees that were estimated in Chapter 2 and the role of adaptation was modeled from published ecological bird classifications (Pigot et al., 2020). Bird body mass was estimated using Bayesian-informed models of growing complexity by taking advantage of the humeral articular facet of the coracoid (HAF, Field et al., 2013) and the entire femoral volume.

Phylogenetic information provided as a tree did not help explain more variation in the models of species-averaged body mass against the femur volume of individuals. However, including taxonomic order as a discrete variable did improve the models and ultimately represented the optimal trade-off between over-parameterisation and under-parameterisation. In addition, including ecological factors did not significantly improve body size estimation, suggesting that the relationship between HAF, femur volume and body mass represents a true functional constraint (McElreath, 2020; Pearl, 2009). Most importantly, the best performing model applied to femur volume and coracoid HAF provided concordant body mass estimates for extinct penguins, except in the case of early Paleocene taxa. These results suggested that many giant penguins were smaller than have previously been estimated (Jadwiszczak, 2009; Mayr, Scofield, et al., 2017). However, an important caveat to consider is that femur volume is total internal volume, and as described in Chapter 1, penguins have osteosclerotic bones. Future revisions to the models presented here may consider cross-sectional strength, or proportion of bone mineral vs. air and soft-tissue.

The aim of Chapter 4 was to evaluate the role of adaptation on the shape of penguins by estimating the rates of evolution for traits from different locomotory modules (i.e. humerus and tarsometatarsus) with two analytical approaches. These approaches followed the constructional morphology framework by accounting for the historical and structural constraints. The first method used a morphological clock analysis (Álvarez-Carretero et al., 2019) performed over the phylogeny produced in Chapter 2 (historical apex) and applied to a dataset for which the trait correlation matrix $\mathbf{R}$ was estimated (structural apex, as mentioned in Chapter 4 section 1.1;

Cheverud, 1982; Love et al., 2021. Results from this first method highlight that the humerus and tarsometatarsus experienced different rates of evolution. The humerus experienced more uniform and steady changes from the early Paleocene to present day, whereas the tarsometatarsus experienced more heterogeneous fluctuations in the rate of change.

The second method used the posterior distribution of trees resulting from the FBDT of Chapter 2 (historical apex) combined with a penalised likelihood (PL) approach for the estimation of character evolution (structural apex, as mentioned in Chapter 4 section 1.1). Using this second method it was possible to visualise and estimate the morphospace occupation through time for stem and crown penguins, revealing that changes in the morphology of the humerus decreased uniformly from the Paleocene to present day whereas the change of morphology of the tarsometatarsus was more abrupt in the earlier stages of penguin evolutionary history. The PL method applied to a distribution of phylogenetic trees to assess morphospace occupation was developed for this thesis and the results show the accuracy of the method: fossils that were not included in the phylogenetic analysis generally do not fall far outside their expected areas when projected into the morphospace.

5.2 A morphospace for penguin humeri and tarsometatarsi

An important synthesis about the evolution of penguin phenotypes can be visualised by adding the estimated body masses from Chapter 3 to the humerus and tarsometatarsus morphospaces generated in Chapter 4 (Fig. 5.1), and then drawing hypothetical evolutionary trajectories on these morphospaces (Fig. 5.2).

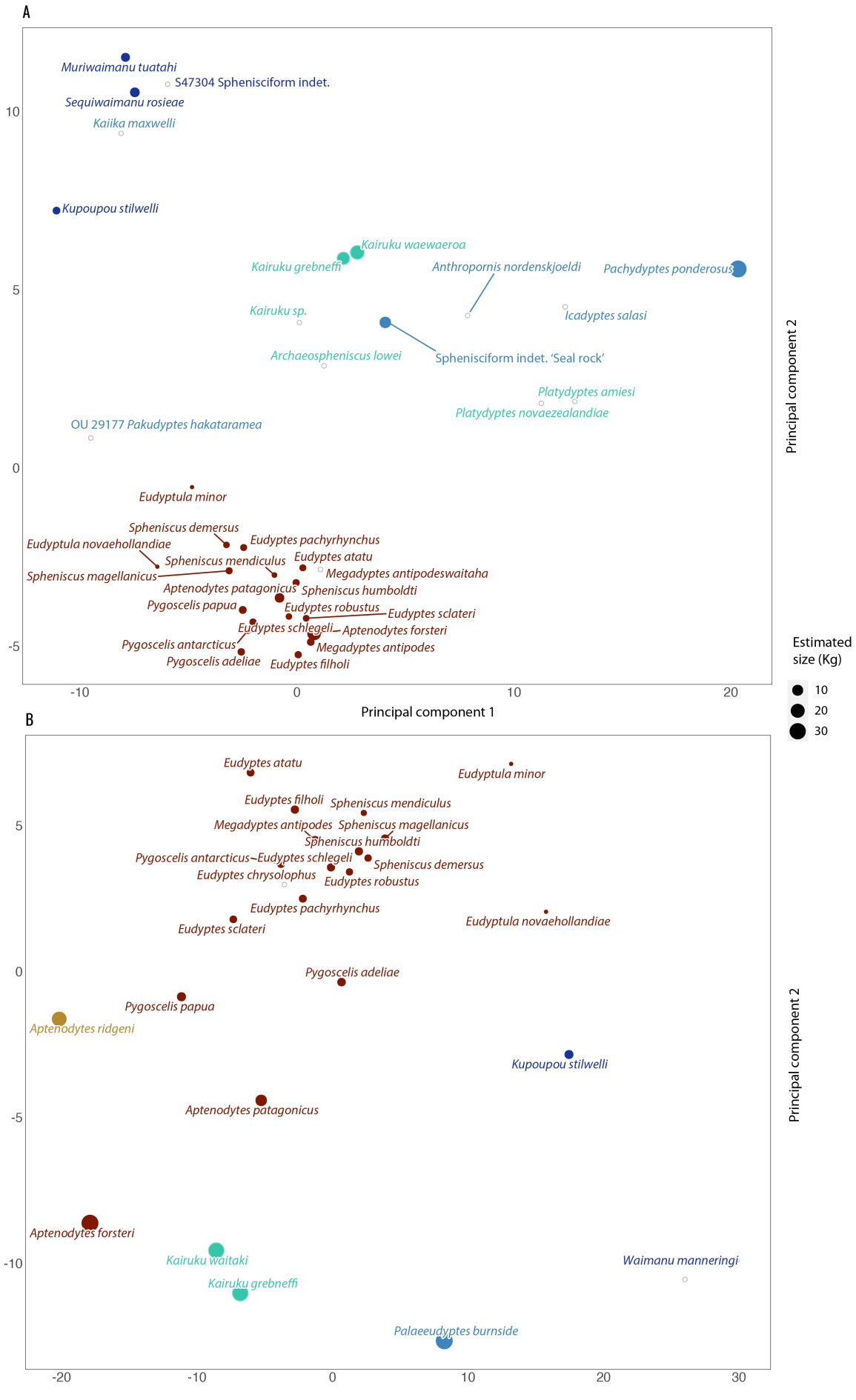


Figure 5.1: Body mass distribution for penguins shown along the principal component one (PC1) and PC2 morphospace for the humerus (A) and the tarsometatarsus (B) from Chapter 4. Body mass estimates from Table B.3 or from estimates in Table 3.3 if the taxon did not preserve the corresponding bone. Morphospace plots derived from Figs. 4.12 - 4.13. Size of the circles indicate the estimated or measured body mass of the penguin. Taxa with no size estimates are represented by smaller empty dots.

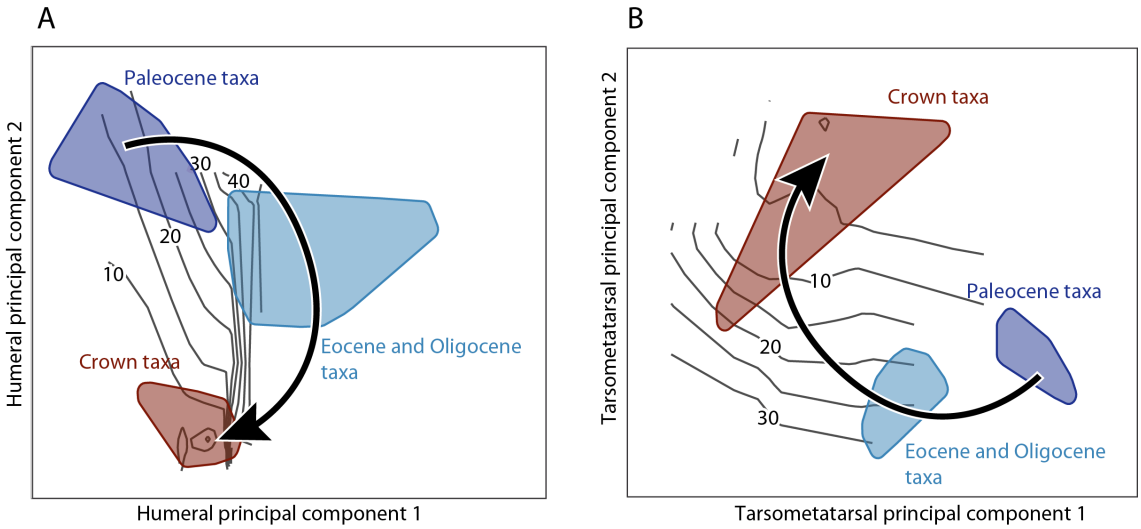


Figure 5.2: Evolutionary trajectory of the humerus (A) and the tarsometatarsus (B) for penguins in relation to body mass. Areas of morphospace occupation are defined from Chapter 4 whereas the arrows aims to describe the overall trajectory through time. Splines determine interpolated body mass in kilograms.

The humerus morphospace is based on principal component one (PC1) and PC2 which together explain approximately 52% of the shape variation in the dataset. Across the evolutionary history of penguins the humerus becomes initially wider and then narrow again whereas the margin of the tricipital fossa tends to uniformly flatten from Paleocene penguins towards modern penguins. The shape evolution of the humerus is represented in the PC1 and PC2 morphospace by an initial increase in PC1 values followed by a decrease in PC1 values, and a general decrease in PC2 values (Fig. 5.2A, black arrow). This evolutionary trajectory was traced by approximately following the backbone of the phylogenetic tree overlain on the morphospace (Fig. 4.12 - 4.13).

The PC1 and PC2 morphospace for the tarsometatarsus explains approximately 45% of the shape variation in the dataset. Across the evolutionary history of penguins the tarsometatarsus became more squared, and then again more elongated but just for smaller taxa, and the curvature of the lateral margin was gradually lost. The shape evolution of the tarsometatarsus is represented in the PC1 and PC2 morphospace by an initial decrease in both PC1 and PC2 values followed by a subsequent increase in both PC1 and PC2 values (Fig. 5.2B, black arrow).

Much of the insight into penguin evolution is gained here from the body mass and humerus shape relationship as opposed to evolutionary patterns in the shape of the tarsometatarsus. In contrast to the humerus, the two most important sources of shape variation in the tarsometatarsus dataset (i.e. PC1 and PC2) appear to be highly correlated with size: smaller crown penguins weighing less than 10 kg have more positive values along both axes (Fig. 5.1B) and larger genera like *Aptenodytes* and *Kairuku* that may reach up to 40 kg have more negative PC1 and PC2 values (Fig. 5.1B). This suggests that there is a strong allometric signal in the tarsometatarsus morphospace (Klingenberg, 2016). Given that broadly contemporaneous species show a diversity of body sizes for most of the history of penguins, the shape of the tarsometatarsus is here hypothesised to be changing as a consequence of body size and not be a driver of penguin diversification.

The PC1 and PC2 morphospace for the humerus shows different sized penguins sharing similar shapes (Fig. 5.1A). If the evolutionary trajectory for humerus shape (Fig. 5.2A, black arrow) is compared to estimated body mass (Fig. 5.1A) then a series of implications arise from the perspective of the adaptive landscape metaphor (i.e. sensu S. Wright, 1932). For example, early penguins including *Sequiwaimanu rosieae*, *Muriwaimanu tuatahi*, *Kaiika maxwelli* and *Kuopoupou stilwelli* have humeri (Fig. 5.1A) that exhibit a relatively small head, a less flattened shaft, smaller and more horizontal margins of the tricipital fossa, and have body masses ranging from below 10 kg up to 30 kg. Crown taxa instead have humeri with a wider dorsoventral surface, more straight anterodorsal margins, relatively wider tricipital fossae, and range in body mass from 1 kg to 45 kg. Lastly, species with the most positive PC1 score values (Fig. 5.2A, center right) include *Pachydyptes ponderosus*, *Platydyptes* spp. and *Icadyptes salasi*, which have humeri with extremely widened shafts and relatively enlarged heads. The body mass of *Pachydyptes ponderosus* probably surpassed 50 kg. Other extinct taxa like *Kairuku* spp., *Archaeosphenis-cus lowei*, DM1449 (the "Seal rock specimen") and *Anthropornis nordenskjoeldi* that were as large as living emperor penguin or larger tend to both fill the more central

morphospace (i.e. do not have extreme PC1 or PC2 values) (Fig. 5.1B, center) and fit the morphospace trajectory (Fig. 5.2A). The evolutionary pattern that emerges is that penguins diversify with a range of body sizes upon gaining a given humerus morphology. Eventually most of the species with a particular humerus shape become extinct, and a surviving lineage with a more-derived morphology becomes the progenitor for a subsequent diversification event (Fig. 5.2). As highlighted in Chapter 4 this pattern is reminiscent of a punctuated-equilibrium mode of evolution with sudden shifts in form that don't lead to the coexistence of plesiomorphic and apomorphic conditions (Gould & Eldredge, 1977). Further analysis with a more complete and unbiased dataset would be required in order to confirm this pattern.

While the evolutionary pattern for humerus shape and body size is influenced by the distribution of phylogenetic trees from Chapter 2, these trees did not include the small Oligocene humerus OU 29177 "*Pakudyptes hakataramea*" that looks comparatively modern. OU 29177 is important for showing that unexpected morphologies were explored by small taxa during the late Oligocene, given that OU 29177 has PC1 and PC2 score values that are unlike other humeri from other Oligocene penguins, but which are more similar to crown penguins (Figs.5.1A and 5.2A). The unexpected morphology of OU 29177 emphasises one of the most common paleontological sampling biases that favors the retrieval of larger specimens (Brown et al., 2021; Brown et al., 2013) and highlights that smaller taxa lived along to the "Oligocene giants". Moreover, the morphological similarity of OU 29177 to crown taxa may perhaps suggest that the lineage of penguins that gave rise to the crown went through a size reduction phase during the Oligocene. This hypothesis may be rejected if we could recover a considerably larger crown-like humerus from the Oligocene but such a specimen has yet to be found. Instead, if the small size OU 29177 is entirely due to early ontogenetic stages it may provide even further insights about crown penguin evolution. Consider that if early ontogenetic stages of the "Oligocene giants" are reminiscent of a crown-like morphology then it may even suggest that morphological novelties of the crown could be achieved through neoteny. To confirm such

hypothesis still more evidence would be required, especially from investigations in developmental studies.

Regardless of the way size changed through time, the size reduction hypothesis has some appealing implications especially given that the origin of penguins from a small-sized bird was originally proposed by Simpson (1946), and it has often been suggested that size may be a trait with substantial genetically-controlled variation (i.e. variation from growth rate or developmental timing that is subject to selection, Clarke et al., 2007; Ksepka et al., 2006; Mayr, Scofield, et al., 2017; Tambussi et al., 2005) as well as often representing a "line of least evolutionary resistance" (Marroig & Cheverud, 2005, 2010). Crucially, the humeral morphospace may even suggest that size variation has the potential to become a mechanism that contributes to cladogenesis by easing the trajectory of a given lineage over the hills of an adaptive landscape. Consider that the transition towards wing-propelled diving in an aerially-volant bird probably required the wing to be a locomotory module suited to flying and swimming at the same time. The tradeoff for flapping locomotion in air as well as in water is perhaps achievable for smaller birds but becomes more of a constraint for larger birds unless flight capability is sacrificed (Ksepka et al., 2006). The step of reducing and then increasing size perhaps did not occur just at the early radiation of the penguin clade, but as observed in the humerus morphospace, may have occurred multiple times during penguin evolution. Within more derived penguins the driver towards smaller body sizes may be foraging efficiency, echoing the initial driver for the initial transition into life in water. If body size changes through time help to navigate morphological ‘sinkholes’ in a fitness landscape, then size could be the driver by which penguins achieved novel wing morphologies (flatter forelimb bones, stiffened wing join) from the early Paleocene to modern day.

Lastly, the relatively short time to transition from land to sea in penguins is a macroevolutionary pattern not dissimilar to what has often been observed in the paleontological record during the land to sea transition for other tetrapods (Pyenson et al., 2014; Pyenson & Vermeij, 2016). The majority of the evolutionary history

of penguins involved a general morphological diversification whereas the functional transition that enabled underwater locomotion had already occurred at the beginning of Paleocene if not even during the Cretaceous. Another common aspect shared between penguins and other marine tetrapods may be a strong relationship between body size and the origin of morphological novelties. For example, body size may have been an important driving factor for filter feeding in baleen whales (Fordyce & Marx, 2018), as well as driving swimming efficiency in ichthyosaurs (Gutarra et al., 2019) and plesiosaurs (Troelsen et al., 2019). Size variation can be extremely constrained in aerially-volant birds (Vizcaíno & Fariña, 1999) especially compared to the size variation that marine vertebrates may achieve (McNeill Alexander, 1998). By transitioning toward marine niches penguins and other organisms may effectively relax selective pressures acting over size, allowing clades to diversify and evolutionary novelties to develop. These evolutionary hypotheses may be tested in future studies by comparing developmental growth and fluctuations in body size through the paleontological record.

5.3 The future of constructional morphology

Constructional morphology provides a general guideline for performing macroevolutionary research. By compartmentalising the total analysis into discrete stages it is possible to outline the key parts that are needed (here these key parts were phylogeny estimation, trait inference and morphospace reconstruction coupled with rates estimation). Although what was achieved in this thesis is mostly an attempt to incorporate a morphodynamics interpretation of constructional morphology from Seilacher (1970) into a structured workflow pipeline, future improvements might condense all steps into a better integrated method.

During my research into Bayesian statistical data analysis for this thesis I had the opportunity to investigate causal inference analysis (e.g. McElreath, 2020; Pearl, 2009; Pearl et al., 2016), a type of approach that aims to quantify the causal links between two separate factors. While we are aware that “correlation does not imply

causation", we also know that causation can be reliably measured, and traditionally this task has often been achieved with randomised control trials (Shrier, 2013). However, for evolutionary biologists, paleontologists, ecologists, and natural scientists in the broader sense, conducting a randomised control trial experiment is often not possible. Instead, causal inference can be achieved with a series of adjustments that aim to account for potential confounders that are acting between two variables of interest (Pearl et al., 2016). The method takes advantage of a directed acyclic graph (DAG)(Fig. 5.3) and has its origins in an approach developed by Sewall S. Wright (1921) to assess the impact of different factors on the development of Guinea pigs (Pearl & Mackenzie, 2017)). In general terms, the first step is tracing down in graph structure the possible confounders that act over our variables of interest (Fig. 5.3A). The next step is to hypothesise how variables interact with one another in causal terms (Fig. 5.3B). Finally, by excluding some variables following a set of graph rules (McElreath, 2020; Pearl et al., 2016), the impact of a specific variable over another can be estimated (Fig. 5.3C-D). By designing an experiment using a DAG it is possible to assess which variables need to be included and which need to be excluded in any given analysis.

Although Seilacher did not originally present constructional morphology to be read under a rigorous statistical light (Seilacher, 1970), the aptive triangle is reminiscent of the structure of a DAG, with three major variables interacting between each other. With a series of adjustments to the structure of the aptive triangle, the resulting DAG may incorporate the phylogeny and variables describing the phenotype and other aspects of the organism, along with potential confounders, and thus enable the quantification of cause and effect over complex macroevolutionary events. Causal inference in a comparative context has already been formalised by (Pagel, 1994) to assess the causational impact between two characters. However, recent developments in evolutionary psychology and anthropology have allowed the development of a latent variable analysis performed over a complex dataset of multivariate traits with an underlying phylogeny (Ringen et al., 2021). Given these

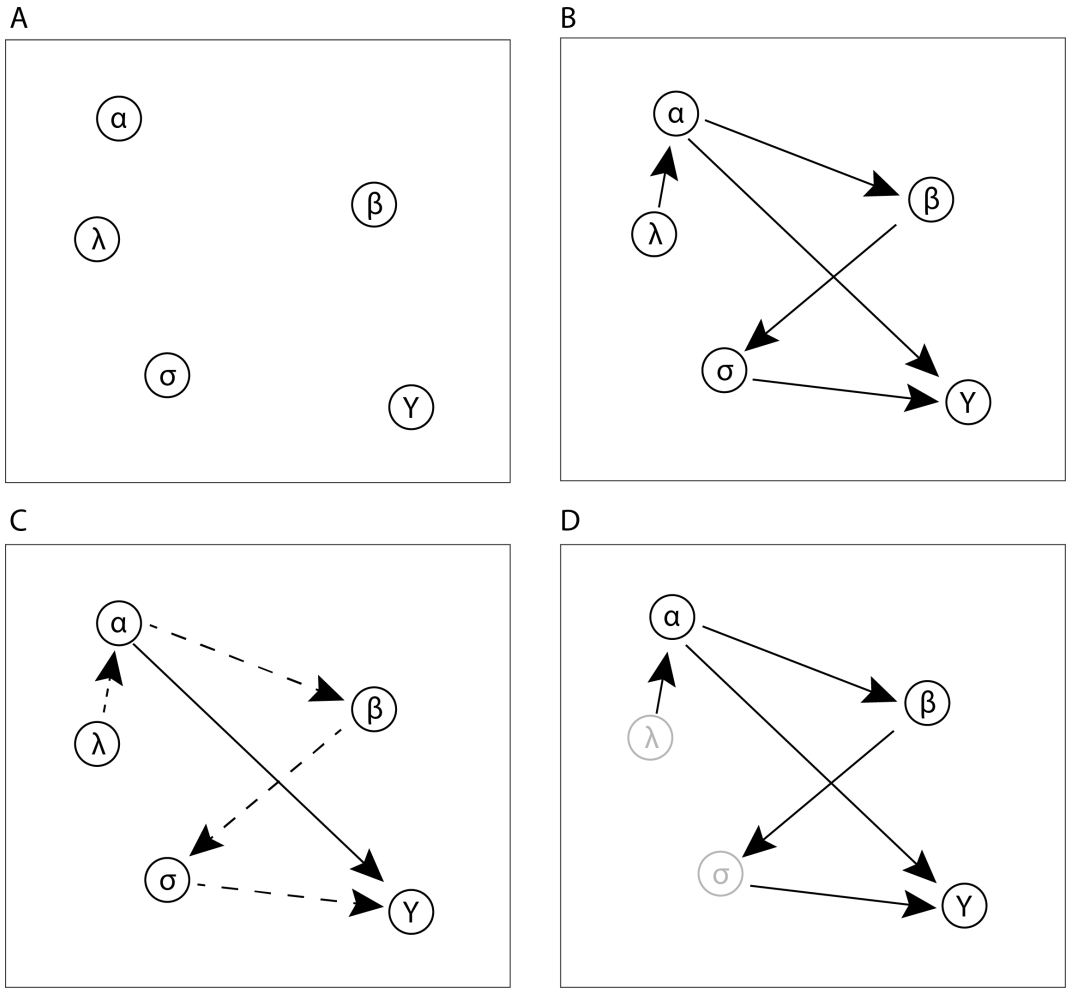


Figure 5.3: A directed acyclic graph (DAG) causal pipeline for the scenario of estimating the impact of variable α over variable Y. A) All potential confounds are first traced down on paper as nodes. B) Next, an hypothetical net of causal relations is drawn connecting each node. C) After deciding the causal relation to investigate (solid line), D) a series of rules (McElreath, 2020; Pearl et al., 2016) determines which variables to include (black) and which to exclude (light gray).

advancements it would be possible to extend the analysis to more than two traits, even to large multivariate shape datasets, like the ones deriving from geometric morphometrics (Bardua, Wilkinson, et al., 2019; Felice et al., 2021).

Exploring a DAG-informed experimental design is a possible future direction for research using the aptive triangle approach. Overall, however, I hope that this research may have contributed to highlight the strengths of the constructional morphology framework to study present and past forms of life.

Bibliography

Acosta Hospitaleche, C. (2016). Paleobiological remarks on a new partial skeleton of the Eocene antarctic penguin *Palaeeudyptes klekowskii*. *Ameghiniana*, *53*(3), 269–281. https://doi.org/10.5710/AMGH.27.08.2015.2890

Acosta Hospitaleche, C., & Alicia, C. I. (2004). *Los pingüinos (Aves: Sphenisciformes) fósiles de Patagonia* (Thesis). Universidad Nacional de La Plata. https://doi.org/10.35537/10915/4286

Acosta Hospitaleche, C., Haidr, N., Paulina-Carabajal, A., & Reguero, M. (2019). The first skull of *Anthropornis grandis* (Aves, Sphenisciformes) associated with postcranial elements. *Comptes Rendus Palevol*, *18*(6), 599–617. https://doi.org/10.1016/j.crpv.2019.06.003

Acosta Hospitaleche, C., Paulina-Carabajal, A., & Yury-Yáñez, R. (2021). The skull of the Miocene *Spheniscus urbinai* (Aves, Sphenisciformes): Osteology, brain morphology, and the cranial pneumatic systems. *Journal of Anatomy*, *239*(1), 151–166. https://doi.org/10.1111/joa.13403

Acosta Hospitaleche, C., & Tambussi, C. (2006). Skull morphometry of *Pygoscelis* (Sphenisciformes): Inter and intraspecific variations. *Polar Biology*, *29*(9), 728. https://doi.org/10.1007/s00300-006-0109-6

Adams, D. C. (2014). A method for assessing phylogenetic least squares models for shape and other high-dimensional multivariate data. *Evolution*, *68*(9), 2675–2688. https://doi.org/10.1111/evo.12463

Adams, D. C., & Collyer, M. L. (2019). Phylogenetic comparative methods and the evolution of multivariate phenotypes. *Annual Review of Ecology, Evolution,*

and Systematics, *50*(1), 405–425. https://doi.org/10.1146/annurev-ecolsys-110218-024555

Adams, D. C., & Otárola-Castillo, E. (2013). Geomorph: An R package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, *4*(4), 393–399. https://doi.org/10.1111/2041-210X.12035

Adams, D. C., Rohlf, F. J., & Slice, D. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix*, *24*(1), 7–14. https://doi.org/10.4404/hystrix-24.1-6283

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution'. *Italian Journal of Zoology*, *71*(1), 5–16. https://doi.org/10.1080/11250000409356545

Allmon, W. D., & Yacobucci, M. M. (2016). *Species and speciation in the fossil record*. University of Chicago Press.

Álvarez-Carretero, S., Goswami, A., Yang, Z., & Dos Reis, M. (2019). Bayesian estimation of species divergence times using correlated quantitative characters. *Systematic Biology*, *68*(6), 967–986. https://doi.org/10.1093/sysbio/syz015

Anderson, J. F., Hall-Martin, A., & Russell, D. A. (1985). Long-bone circumference and weight in mammals, birds and dinosaurs. *Journal of Zoology*, *207*(1), 53–61. https://doi.org/https://doi.org/10.1111/j.1469-7998.1985.tb04915.x

Ando, T. (2007). *New Zealand fossil penguins: Origin, pattern, and process* (PhD Thesis). University of Otago.

Ando, T., & Fordyce, R. E. (2014). Evolutionary drivers for flightless, wing-propelled divers in the Northern and Southern Hemispheres. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *400*, 50–61. https://doi.org/10.1016/j.palaeo.2013.08.002

Arbour, J. H., & Brown, C. M. (2014). Incomplete specimens in geometric morphometric analyses. *Methods in Ecology and Evolution*, *5*(1), 16–26. https://doi.org/10.1111/2041-210X.12128

Arcila, D., Alexander Pyron, R., Tyler, J. C., Ortí, G., & Betancur-R., R. (2015). An evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (Teleostei: Percomorphaceae). *Molecular Phylogenetics and Evolution*, *82*, 131–145. https://doi.org/10.1016/j.ympev.2014.10.011

Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2013). Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution*, *30*(2), 239–243. https://doi.org/10.1093/molbev/mss243

Bardua, C., Felice, R. N., Watanabe, A., Fabre, A. .-C., & Goswami, A. (2019). A practical guide to sliding and surface semilandmarks in morphometric analyses. *Integrative Organismal Biology*, *1*(1). https://doi.org/10.1093/iob/obz016

Bardua, C., Fabre, A.-C., Bon, M., Das, K., Stanley, E. L., Blackburn, D. C., & Goswami, A. (2020). Evolutionary integration of the frog cranium. *Evolution*, *74*(6), 1200–1215. https://doi.org/10.1111/evo.13984

Bardua, C., Wilkinson, M., Gower, D. J., Sherratt, E., & Goswami, A. (2019). Morphological evolution and modularity of the caecilian skull. *BMC Evolutionary Biology*, *19*(1), 30. https://doi.org/10.1186/s12862-018-1342-7

Baumel, J. J., & Witmer, L. M. (1993). Osteology Handbook of avian anatomy: Nomina Anatomica Avium. *Publications of the Nuttall Ornithological Club, Cambridge, Massachusetts*, 118–152.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, *53*, 370–418.

Beavan, A. J. S., Pisani, D., & Donoghue, P. C. J. (2021). Diversification dynamics of total-, stem-, and crown-groups are compatible with molecular clock estimates of divergence times. *Science Advances*, *7*(24), eabf2257. https://doi.org/10.1126/sciadv.abf2257

Benson, R. B. J., Hunt, G., Carrano, M. T., & Campione, N. (2018). Cope's rule and the adaptive landscape of dinosaur body size evolution. *Palaeontology*, *61*(1), 13–48. https://doi.org/https://doi.org/10.1111/pala.12329

Bertelli, S., & Giannini, N. P. (2005). A phylogeny of extant penguins (Aves: Sphenisciformes) combining morphology and mitochondrial sequences. *Cladistics*, *21*(3), 209–239. https://doi.org/https://doi.org/10.1111/j.1096-0031.2005.00065.x

Bessho, M., Ohnishi, I., Matsuyama, J., Matsumoto, T., Imai, K., & Nakamura, K. (2007). Prediction of strength and strain of the proximal femur by a CT-based finite element method. *Journal of Biomechanics*, *40*(8), 1745–1753. https://doi.org/10.1016/j.jbiomech.2006.08.003

Blackith, R. E., & Reyment, R. A. (1971). *Multivariate morphometrics*. Academic Press.

Blanckenhorn, W. U. (2000). The evolution of body size: What keeps organisms small? *The Quarterly Review of Biology*, *75*(4), 385–407. https://doi.org/10.1086/393620

Blender Online Community. (2020). Blender - a 3d modelling and rendering package. *Blender Foundation*. http://www.blender.org

Blokland, J. C., Reid, C. M., Worthy, T. H., Tennyson, A. J. D., Clarke, J. A., & Scofield, R. P. (2019). Chatham Island Paleocene fossils provide insight into the palaeobiology, evolution, and diversity of early penguins (Aves, Sphenisciformes). *Palaeontologia Electronica*, *22*(3), 1–92. https://doi.org/https://doi.org/10.2687/1009

Blomberg, S. P., Rathnayake, S. I., & Moreau, C. M. (2020). Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *The American Naturalist*, *195*(2), 145–165. https://doi.org/10.1086/706339

Boettiger, C., Coop, G., & Ralph, P. (2012). Is your phylogeny informative? Measuring the power of comparative methods. *Evolution*, *66*(7), 2240–2251. https://doi.org/10.1111/j.1558-5646.2011.01574.x

Bookstein, F. L. (1991). *Morphometric tools for landmark data: Geometry and biology*. Cambridge Univ. Press, New York.

Bookstein, F. L. (1996). Combining the tools of geometric morphometrics. In L. F. Marcus, M. Corti, A. Loy, G. J. P. Naylor, & D. E. Slice (Eds.), *Advances in morphometrics* (pp. 131–151). Springer US. https://doi.org/10.1007/978-1-4757-9083-2_12

Bookstein, F. L. (1998). A hundred years of morphometrics. *Acta Zoologica Academiae Scientiarum Hungaricae*, *44*(1), 7–59.

Bookstein, F. L., Chernoff, B., Elder, R. L., Humphries, J. J. M., Smith, G. R., & Strauss, R. F. (1985, July 1). *Morphometrics in evolutionary biology: The geometry of size and shape change, with examples from fishes* (1st Edition.). Academy of Natural Sciences of Philadelphia.

Bouckaert, R. R., Heled, J., Kühnert, D., Vaughan, T. G., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.*, *10*(4). http://dblp.uni-trier.de/db/journals/ploscb/ploscb10.html#BouckaertHKVWXSRD14

Brassey, C. A. (2017). Body mass estimation in paleontology: A review of volumetric techniques. *The Paleontological Society Papers*, *22*, 133–156.

Brazeau, M. D. (2011). Problematic character coding methods in morphology and their effects. *Biological Journal of the Linnean Society*, *104*(3), 489–498. https://doi.org/https://doi.org/10.1111/j.1095-8312.2011.01755.x

Briggs, D. E. G. (2017). Seilacher, Konstruktions-Morphologie, morphodynamics, and the evolution of form. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *328*(3), 197–206. https://doi.org/10.1002/jez.b.22725

Bright, J. A., Marugán-Lobón, J., Cobb, S. N., & Rayfield, E. J. (2016). The shapes of bird beaks are highly controlled by nondietary factors. *Proceedings of the National Academy of Sciences*, *113*(19), 5352–5357. https://doi.org/10.1073/pnas.1602683113

Bromham, L., & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics*, *4*(3), 216–224. https://doi.org/10.1038/nrg1020

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo* (1° edition). Chapman; Hall/CRC.

Brown, C. M., Campione, N. E., Wilson Mantilla, G. P., & Evans, D. C. (2021). Size-driven preservational and macroecological biases in the latest maastrichtian terrestrial vertebrate assemblages of north america. *Paleobiology*, 1–29. https://doi.org/10.1017/pab.2021.35

Brown, C. M., Evans, D. C., Campione, N. E., O'Brien, L. J., & Eberth, D. A. (2013). Evidence for taphonomic size bias in the Dinosaur Park Formation (Campanian, Alberta), a model Mesozoic terrestrial alluvial-paralic system. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *372*, 108–122. https://doi.org/https://doi.org/10.1016/j.palaeo.2012.06.027

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Buser, T. J., Boyd, O. F., Cortés, Á., Donatelli, C. M., Kolmann, M. A., Luparell, J. L., Pfeiffenberger, J. A., Sidlauskas, B. L., & Summers, A. P. (2020). The natural historian's guide to the CT galaxy: Step-by-step instructions for preparing and analyzing Computed Tomographic (CT) data using cross-platform, open access software. *Integrative Organismal Biology*, *2*(obaa009). https://doi.org/10.1093/iob/obaa009

Campione, N. E., & Campione, M. N. E. (2020). Package 'MASSTIMATE'. *Ecology and Evolution*, *5*(9), 913–923.

Campione, N. E., & Evans, D. C. (2020). The accuracy and precision of body mass estimation in non-avian dinosaurs. *Biological Reviews*, *95*(6), 1759–1797. https://doi.org/https://doi.org/10.1111/brv.12638

Campione, N. E., Evans, D. C., Brown, C. M., & Carrano, M. T. (2014). Body mass estimation in non-avian bipeds using a theoretical conversion to quadruped stylopodial proportions. *Methods in Ecology and Evolution*, *5*(9), 913–923. https://doi.org/https://doi.org/10.1111/2041-210X.12226

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). *Stan: A probabilistic programming language* (Vol. 76). https://doi.org/10.18637/jss.v076.i01

Cau, A. (2017). Specimen-level phylogenetics in paleontology using the Fossilized Birth-Death model with sampled ancestors. *PeerJ*, *5*, e3055. https://doi.org/10.7717/peerj.3055

Chambers, J. M. (1992). Linear models. *Statistical models in S* (pp. 95–144). Routledge.

Chávez Hoffmeister, M., Carrillo Briceño, J. D., & Nielsen, S. N. (2014). The evolution of seabirds in the Humboldt current: New clues from the Pliocene of Central Chile. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0090043

Chávez-Hoffmeister, M. (2014). Phylogenetic characters in the humerus and tarsometatarsus of penguins. *Polish Polar Research*, *35*(3), 469.

Chávez-Hoffmeister, M. (2020). Bill disparity and feeding strategies among fossil and modern penguins. *Paleobiology*, *46*(2), 176–192. https://doi.org/10.1017/pab.2020.10

Cheverud, J. M. (1982). Phenotypic, genetic, and environmental morphological integration in the cranium. *Evolution*, *36*(3), 499–516. https://doi.org/10.2307/2408096

Chown, S. L., & Gaston, K. J. (2010). Body size variation in insects: A macroecological perspective. *Biological Reviews*, *85*(1), 139–169. https://doi.org/https://doi.org/10.1111/j.1469-185X.2009.00097.x

Clark, B. D., & Bemis, W. (1979). Kinematics of swimming of penguins at the detroit zoo. *Journal of Zoology*, *188*(3), 411–428. https://doi.org/10.1111/j.1469-7998.1979.tb03424.x

Clarke, J. A., Ksepka, D. T., Salas-Gismondi, R., Altamirano, A. J., Shawkey, M. D., D'Alba, L., Vinther, J., DeVries, T. J., & Baby, P. (2010). Fossil evidence for evolution of the shape and color of penguin feathers. *Science*, *330*(6006), 954–957. https://doi.org/10.1126/science.1193604

Clarke, J. A., Ksepka, D. T., Stucchi, M., Urbina, M., Giannini, N., Bertelli, S., Narváez, Y., & Boyd, C. A. (2007). Paleogene equatorial penguins challenge the proposed relationship between biogeography, diversity, and Cenozoic climate change. *Proceedings of the National Academy of Sciences*, *104*(28), 11545–11550. https://doi.org/10.1073/pnas.0611099104

Clavel, J., Aristide, L., & Morlon, H. (2019). A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to New-World monkeys brain evolution. *Systematic Biology*, *68*(1), 93–116. https://doi.org/10.1093/sysbio/syy045

Clifton, G. T., Carr, J. A., & Biewener, A. A. (2018). Comparative hindlimb myology of foot-propelled swimming birds. *Journal of Anatomy*, *232*(1), 105–123. https://doi.org/10.1111/joa.12710

Cohen, K. M., Finney, S. C., Gibbard, P. L., & Fan, J.-X. (2013). The ics international chronostratigraphic chart. *Episodes Journal of International Geoscience*, *36*(3), 199–204.

Colbert, E. H. (1962). The weights of dinosaurs. *American Museum Novitates*, (2076), 1–16.

Cole, S. R., Chu, H., & Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, *179*(2), 252–260. https://doi.org/10.1093/aje/kwt245

Cole, T. L., Ksepka, D. T., Mitchell, K. J., Tennyson, A. J. D., Thomas, D. B., Pan, H., Zhang, G., Rawlence, N. J., Wood, J. R., Bover, P., Bouzat, J. L.,

Cooper, A., Fiddaman, S. R., Hart, T., Miller, G., Ryan, P. G., Shepherd, L. D., Wilmshurst, J. M., & Waters, J. M. (2019). Mitogenomes uncover extinct penguin taxa and reveal island formation as a key driver of speciation. *Molecular Biology and Evolution*, *36*(4), 784–797. https://doi.org/10.1093/molbev/msz017

Collyer, M. L., & Adams, D. C. (2021). Phylogenetically aligned component analysis. *Methods in Ecology and Evolution*, *12*(2), 359–372.

Collyer, M. L., Davis, M. A., & Adams, D. C. (2020). Making heads or tails of combined landmark configurations in geometric morphometric data. *Evolutionary Biology*, *47*(3), 193–205. https://doi.org/10.1007/s11692-020-09503-z

Costa, R., & Bisol, P. M. (1978). Genetic variability in deep-sea organisms. *The Biological Bulletin*, *155*(1), 125–133. https://doi.org/10.2307/1540870

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, *14*(1), 1–13. https://doi.org/10.1119/1.1990764

Croft, D. A., Engelman, R. K., Dolgushina, T., & Wesley, G. (2018). Diversity and disparity of sparassodonts (Metatheria) reveal non-analogue nature of ancient South American mammalian carnivore guilds. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1870), 20172012. https://doi.org/10.1098/rspb.2017.2012

Cubo, J. (2004). Pattern and process in constructional morphology. *Evolution & Development*, *6*(3), 131–133. https://doi.org/10.1111/j.1525-142X.2004.04018.x

Cubo, J., Legendre, P., Ricqlès, A. D., Montes, L., Margerie, E. D., Castanet, J., & Desdevises, Y. (2008). Phylogenetic, functional, and structural components of variation in bone growth rate of amniotes. *Evolution & Development*, *10*(2), 217–227. https://doi.org/10.1111/j.1525-142X.2008.00229.x

Darwin, C. (1872). *The origin of species: By means of natural selection or the preservation of favored races in the struggle for life* (Vol. 2). Modern library.

Davies, S., Bamford, M. J., & Loomes, D. (2002). *Ratites and tinamous: Tinamidae, Rheidae, Dromaiidae, Casuariidae, Apterygidae, Struthionidae*. Oxford University Press.

Davis, L. S., & Darby, J. T. (2012). *Penguin biology*. Elsevier.

de Finetti, B. (1970). *Theory of probability: A critical introductory treatment*. New York: John Wiley.

Degrange, F. J., Ksepka, D. T., & Tambussi, C. P. (2018). Redescription of the oldest crown clade penguin: Cranial osteology, jaw myology, neuroanatomy, and phylogenetic affinities of *Madrynornis mirandus*. *Journal of Vertebrate Paleontology*, *38*(2), e1445636. https://doi.org/10.1080/02724634.2018.1445636

Dellinger, A. S., Artuso, S., Pamperl, S., Michelangeli, F. A., Penneys, D. S., Fernández-Fernández, D. M., Alvear, M., Almeda, F., Scott Armbruster, W., Staedler, Y., & Schönenberger, J. (2019). Modularity increases rate of floral evolution and adaptive success for functionally specialized pollination systems. *Communications Biology*, *2*(1), 1–11. https://doi.org/10.1038/s42003-019-0697-7

Demery, A.-J. C., Burns, K. J., & Mason, N. A. (2021). Bill size, bill shape, and body size constrain bird song evolution on a macroevolutionary scale. *Ornithology*, *138*(2). https://doi.org/10.1093/ornithology/ukab011

Dickersin, K., Chan, S., Chalmersx, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials*, *8*(4), 343–353. https://doi.org/https://doi.org/10.1016/0197-2456(87)90155-3

dos Reis, M., Gunnell, G. F., Barba-Montoya, J., Wilkins, A., Yang, Z., & Yoder, A. D. (2018). Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. *Systematic Biology*, *67*(4), 594–615. https://doi.org/10.1093/sysbio/syy001

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C. J., & Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in

estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1742), 3491–3500. https://doi.org/10.1098/rspb.2012.0683

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLOS Biology*, *4*(5), e88. https://doi.org/10.1371/journal.pbio.0040088

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. https://doi.org/10.1186/1471-2148-7-214

Dryden, I. L., & Mardia, K. V. (1998). *Statistical analysis of shape*. Wiley.

Dunning, J. B. (2007). *CRC Handbook of avian body masses*. CRC Press.

Eliason, C. M., Maia, R., Parra, J. L., & Shawkey, M. D. (2020). Signal evolution and morphological complexity in hummingbirds (Aves: Trochilidae). *Evolution*, *74*(2), 447–458. https://doi.org/10.1111/evo.13893

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S., & Kikinis, R. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, *30*(9), 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001

Felice, R. N., Pol, D., & Goswami, A. (2021). Complex macroevolutionary dynamics underly the evolution of the crocodyliform skull. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1954), 20210919. https://doi.org/10.1098/rspb.2021.0919

Felice, R. N., Watanabe, A., Cuff, A. R., Hanson, M., Bhullar, B.-A. S., Rayfield, E. R., Witmer, L. M., Norell, M. A., & Goswami, A. (2020). Decelerated dinosaur skull evolution with the origin of birds. *PLOS Biology*, *18*(8), e3000801. https://doi.org/10.1371/journal.pbio.3000801

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, *22*(3), 240–249. https://doi.org/10.1093/sysbio/22.3.240

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1–15. https://doi.org/10.1086/284325

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, *19*, 445–471. Retrieved June 21, 2021, from https://www.jstor.org/stable/2097162

Felsenstein, J. (1993). *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein.

Felsenstein, J. (2002). Quantitative characters, phylogenies, and morphometrics. *Morphology, Shape and Phylogeny* (pp. 27–44). London; Chapman & Hall.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates.

Field, D. J., Lynner, C., Brown, C., & Darroch, S. A. F. (2013). Skeletal correlates for body mass estimation in modern and fossil flying birds. *PLOS ONE*, *8*(11), e82000. https://doi.org/10.1371/journal.pone.0082000

Finarelli, J. A. (2008). Hierarchy and the reconstruction of evolutionary trends: Evidence for constraints on the evolution of body size in terrestrial caniform carnivorans (Mammalia). *Paleobiology*, *34*(4), 553–562. https://doi.org/10.1666/07078.1

Fisher, R. A. (1973). *Statistical methods and scientific inference*. Oliver & Boyd.

Fordyce, R. E., & Marx, F. G. (2018). Gigantism precedes filter feeding in baleen whale evolution. *Current Biology*, *28*(10), 1670–1676.e2. https://doi.org/https://doi.org/10.1016/j.cub.2018.04.027

Freckleton, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, *3*(5), 940–947. https://doi.org/10.1111/j.2041-210X.2012.00220.x

Futuyma, D. J. (2010). Evolutionary constraint and ecological consequences. *Evolution*, *64*(7), 1865–1884. https://doi.org/https://doi.org/10.1111/j.1558-5646.2010.00960.x

Futuyma, D. J., & Kirkpatrick, M. (2017). *Evolution* (4° edizione). Sinauer Associates Inc.

Galen, C. (1996). Rates of floral evolution: Adaptation to bumblebee pollination in an alpine wildflower, *Polemonium viscosum*. *Evolution*, *50*(1), 120–125. https://doi.org/https://doi.org/10.1111/j.1558-5646.1996.tb04478.x

Galton, F. (1894). *Natural inheritance*. Macmillan & Company.

Garland, T., & Adolph, S. C. (1994). Why not to do two-species comparative studies: Limitations on inferring adaptation. *Physiological Zoology*, *67*(4), 797–828. https://doi.org/10.1086/physzool.67.4.30163866

Garland, T., & Ives, A. R. (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *The American Naturalist*, *155*(3), 346–364. https://doi.org/10.1086/303327

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., & Drummond, A. J. (2017). Bayesian Total-Evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, *66*(1), 57–73. https://doi.org/10.1093/sysbio/syw060

Gavryushkina, A., Welch, D., Stadler, T., & Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, *10*(12). https://doi.org/10.1371/journal.pcbi.1003919

Geisler, J. H., McGowen, M. R., Yang, G., & Gatesy, J. (2011). A supermatrix analysis of genomic, morphological, and paleontological data from crown Cetacea. *BMC Evolutionary Biology*, *11*(1), 112. https://doi.org/10.1186/1471-2148-11-112

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gelman, A., & Nolan, D. (2017). *Teaching statistics: A bag of tricks* (2nd ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198785699.001.0001

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Giovanardi, S., Ksepka, D. T., & Thomas, D. B. (2021). A giant Oligocene fossil penguin from the North Island of New Zealand. *Journal of Vertebrate Paleontology*, *0*(0), e1953047. https://doi.org/10.1080/02724634.2021.1953047

Goloboff, P. A., & Catalano, S. A. (2016). TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics*, *32*(3), 221–238. https://doi.org/10.1111/cla.12160

Goloboff, P. A., Torres, A., & Arias, J. S. (2018). Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics*, *34*(4), 407–437. https://doi.org/10.1111/cla.12205

Goolsby, E. W. (2016). Likelihood-based parameter estimation for high-dimensional phylogenetic comparative models: Overcoming the limitations of "distance-based" methods. *Systematic Biology*, *65*(5), 852–870. https://doi.org/10.1093/sysbio/syw051

Goswami, A., & Finarelli, J. A. (2016). EMMLi: A maximum likelihood approach to the analysis of modularity. *Evolution*, *70*(7), 1622–1637. https://doi.org/10.1111/evo.12956

Goswami, A., & Polly, P. D. (2010). Methods for studying morphological integration and modularity. *The Paleontological Society Papers*, *16*, 213–243. https://doi.org/10.1017/S1089332600001881

Gould, S. J., & Lewontin, R. C. (1979). The Spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *205*(1161), 581–598.

Gould, S. J. (2002). *The structure of evolutionary theory*. Belknap Press of Harvard University Press.

Gould, S. J., & Eldredge, N. (1977). Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, *3*(2), 115–151. https://www.jstor.org/stable/2400177

Griffin, T. M., & Kram, R. (2000). Penguin waddling is not wasteful. *Nature*, *408*(6815), 929–929. https://doi.org/10.1038/35050167

Grouw, H. v. (2017). The dark side of birds: Melanism—facts and fiction. *Bulletin of the British Ornithologists' Club*, *137*(1), 12–36. https://doi.org/10.25226/bboc.v137i1.2017.a9

Guillerme, T., & Healy, K. (2014). mulTree: A package for running MCMCglmm analysis on multiple trees. *Zenodo.(doi: 10.5281/zenodo. 12902)*.

Gutarra, S., Moon, B. C., Rahman, I. A., Palmer, C., Lautenschlager, S., Brimacombe, A. J., & Benton, M. J. (2019). Effects of body plan evolution on the hydrodynamic drag and energy requirements of swimming in ichthyosaurs. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1898), 20182786. https://doi.org/10.1098/rspb.2018.2786

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Heath, T. A., Huelsenbeck, J. P., & Stadler, T. (2014). The Fossilized Birth–Death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, *111*(29), E2957–E2966. https://doi.org/10.1073/pnas.1319091111

Hedges, S. B., & Shah, P. (2003). Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics*, *4*(1), 31. https://doi.org/10.1186/1471-2105-4-31

Heers, A. M., & Dial, K. P. (2015). Wings versus legs in the avian bauplan: Development and evolution of alternative locomotor strategies. *Evolution*, *69*(2), 305–320. https://doi.org/10.1111/evo.12576

Heinrich, R. E., Strait, S. G., & Houde, P. (2008). Earliest Eocene Miacidae (Mammalia: Carnivora) from northwestern Wyoming. *Journal of Paleontology*, *82*(1), 154–162. Retrieved October 27, 2021, from https://www.jstor.org/stable/20144177

Held, L. I., Jr. (2009). *Quirks of human anatomy: An evo-devo look at the human body*. Cambridge University Press. https://doi.org/10.1017/CBO9780511626890

Hennig, W., Davis, D. D., & Zangerl, R. (1966). *Phylogenetic Systematics*. University of Illinois Press.

Hinić-Frlog, S., & Motani, R. (2010). Relationship between osteology and aquatic locomotion in birds: Determining modes of locomotion in extinct ornithurae. *Journal of Evolutionary Biology*, *23*(2), 372–385. https://doi.org/10.1111/j.1420-9101.2009.01909.x

Ho, S. Y. W., & Duchêne, S. (2014). Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*, *23*(24), 5947–5965. https://doi.org/10.1111/mec.12953

Ho, S. Y. W., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., & Cooper, A. (2011). Time-dependent rates of molecular evolution. *Molecular Ecology*, *20*(15), 3087–3101. https://doi.org/10.1111/j.1365-294X.2011.05178.x

Hoffmann, A. A., & Ross, P. A. (2018). Rates and patterns of laboratory adaptation in (mostly) insects. *Journal of Economic Entomology*, *111*(2), 501–509. https://doi.org/10.1093/jee/toy024

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., & Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic Biology*, *63*(5), 753–771. https://doi.org/10.1093/sysbio/syu039

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, *65*(4), 726–736. https://doi.org/10.1093/sysbio/syw021

Huxley, J. S. (1932). *Problems of relative growth*. Methuen & Co.

Huxley, T. H. (1859). On a fossil bird and a fossil cetacean from New Zealand. *Quarterly Journal of the Geological Society*, *15*(1), 670–677. https://doi.org/10.1144/GSL.JGS.1859.015.01-02.73

Ibrahim, N., Maganuco, S., Dal Sasso, C., Fabbri, M., Auditore, M., Bindellini, G., Martill, D. M., Zouhri, S., Mattarelli, D. A., Unwin, D. M., Wiemann, J., Bonadonna, D., Amane, A., Jakubczak, J., Joger, U., Lauder, G. V., & Pierce, S. E. (2020). Tail-propelled aquatic locomotion in a theropod dinosaur. *Nature*, *581*(7806), 67–70. https://doi.org/10.1038/s41586-020-2190-3

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jadwiszczak, P. (2001). Body size of Eocene Antarctic penguins. *Polish Polar Research*, *22*(2), 147–158.

Jadwiszczak, P. (2009). Penguin past: The current state of knowledge. *Polish polar research*, *30*(1), 3–28.

Jadwiszczak, P. (2020). Outline shape analysis of penguin humeri: A robust approach to taxonomic classification. *Polar Research*, *39*. https://doi.org/10.33265/polar.v39.4370

Jadwiszczak, P., & Mörs, T. (2019). First partial skeleton of *Delphinornis larseni* Wiman, 1905, a slender-footed penguin from the Eocene of antarctic peninsula. *Palaeontologia Electronica*, *22*(2), 1–31. https://doi.org/10.26879/933

Jadwiszczak, P., Reguero, M., & Mörs, T. (2021). A new small-sized penguin from the late Eocene of Seymour Island with additional material of *Mesetaornis polaris*. *GFF*, *0*(0), 1–9. https://doi.org/10.1080/11035897.2021.1900385

Jenkins, P. F., & Veitch, C. R. (1991). Sexual dimorphism and age determination in the North Island saddleback (*Philesturnus carunculatus rufaster*). *New Zealand Journal of Zoology*, *18*(4), 445–450. https://doi.org/10.1080/03014223.1991.10422851

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., & Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature*, *491*(7424), 444–448. https://doi.org/10.1038/nature11631

Johansson, L. C., & Aldrin, B. S. W. (2002). Kinematics of diving Atlantic puffins (*Fratercula arctica*): Evidence for an active upstroke. *Journal of Experimental Biology*, *205*(3), 371–378. https://doi.org/10.1242/jeb.205.3.371

Jombart, T., Balloux, F., & Dray, S. (2010). adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics*, *26*(15), 1907–1909. https://doi.org/10.1093/bioinformatics/btq292

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism: Volume III*. Elsevier.

Jungers, W. L., Falsetti, A. B., & Wall, C. E. (1995). Shape, relative size, and size-adjustments in morphometrics. *American Journal of Physical Anthropology*, *38*, 137–161. https://doi.org/10.1002/ajpa.1330380608

Kendall, D. G. (1977). The diffusion of shape. *Advances in Applied Probability*, *9*(3), 428–430. https://doi.org/10.2307/1426091

Kish, L. (1965). *Survey sampling*. Chichester : Wiley New York.

Klingenberg, C. P. (2005). Developmental constraints, modules, and evolvability. In B. Hallgrímsson & B. K. Hall (Eds.), *Variation* (pp. 219–247). Academic Press. https://doi.org/10.1016/B978-012088777-4/50013-2

Klingenberg, C. P. (2016). Size, shape, and form: Concepts of allometry in geometric morphometrics. *Development Genes and Evolution*, *226*(3), 113–137. https://doi.org/10.1007/s00427-016-0539-2

Klingenberg, C. P. (2020). Walking on Kendall's shape space: Understanding shape spaces and their coordinate systems. *Evolutionary Biology*, *47*(4), 334–352. https://doi.org/10.1007/s11692-020-09513-x

Klingenberg, C. P., & Marugán-Lobón, J. (2013). Evolutionary covariation in geometric morphometric data: Analyzing integration, modularity, and allometry in a phylogenetic context. *Systematic Biology*, *62*(4), 591–610. https://doi.org/10.1093/sysbio/syt025

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. https://doi.org/10.1007/BF02289694

Ksepka, D. T., & Ando, T. (2011). Penguins past, present, and future: Trends in the evolution of the Sphenisciformes. In G. Dyke & G. Kaiser (Eds.), *Living dinosaurs: The evolutionary history of modern birds* (pp. 155–186). John Wiley & Sons.

Ksepka, D. T., Bertelli, S., & Giannini, N. P. (2006). The phylogeny of the living and fossil Sphenisciformes (penguins). *Cladistics*, *22*(5), 412–441. https://doi.org/10.1111/j.1096-0031.2006.00116.x

Ksepka, D. T., Fordyce, R. E., Ando, T., & Jones, C. M. (2012). New fossil penguins (Aves, Sphenisciformes) from the Oligocene of New Zealand reveal the skeletal

plan of stem penguins. *Journal of Vertebrate Paleontology*, *32*(2), 235–254. https://doi.org/10.1080/02724634.2012.652051

Ksepka, D. T., & Thomas, D. B. (2012). Multiple cenozoic invasions of Africa by penguins (Aves, Sphenisciformes). *Proceedings of the Royal Society B: Biological Sciences*, *279*(1730), 1027–1032. https://doi.org/10.1098/rspb.2011.1592

Ksepka, D. T., Werning, S., Sclafani, M., & Boles, Z. M. (2015). Bone histology in extant and fossil penguins (Aves: Sphenisciformes). *Journal of Anatomy*, *227*(5), 611–630. https://doi.org/10.1111/joa.12367

Lande, R. (1976). Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution*, *30*(2), 314–334. https://doi.org/https://doi.org/10.1111/j.1558-5646.1976.tb00911.x

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution*, *33*(1), 402–416. https://doi.org/10.2307/2407630

Larramendi, A., Paul, G. S., & Hsu, S. (2020). A review and reappraisal of the specific gravities of present and past multicellular organisms, with an emphasis on tetrapods. *The Anatomical Record*, *4*(9). https://doi.org/https://doi.org/10.1002/ar.24574

Larson, D. W., Brown, C. M., & Evans, D. C. (2016). Dental disparity and ecological stability in bird-like dinosaurs prior to the end-Cretaceous mass extinction. *Current Biology*, *26*(10), 1325–1333. https://doi.org/10.1016/j.cub.2016.03.039

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*(2), 119–122. https://doi.org/10.1080/00031305.1999.10474443

Lee, M. S. Y., Cau, A., Naish, D., & Dyke, G. J. (2014). Morphological clocks in paleontology, and a Mid-Cretaceous origin of crown aves. *Systematic Biology*, *63*(3), 442–449. https://doi.org/10.1093/sysbio/syt110

Lee, M. S. Y., & Palci, A. (2015). Morphological phylogenetics in the genomic age. *Current Biology*, *25*(19), R922–R929. https://doi.org/10.1016/j.cub.2015.07.009

Lee, M. S. Y., & Worthy, T. H. (2012). Likelihood reinstates *Archaeopteryx* as a primitive bird. *Biology Letters*, *8*(2), 299–303. https://doi.org/10.1098/rsbl.2011.0884

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics* (1st edition). Wiley.

Levin, H. L., & King, D. T. J. (2016). *The earth through time*. John Wiley & Sons.

Lewis, P. O. (2001a). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, *50*(6), 913–925. https://doi.org/10.1080/106351501753462876

Lewis, P. O. (2001b). Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution*, *16*(1), 30–37. https://doi.org/10.1016/S0169-5347(00)02025-5

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(1), 1–18. https://doi.org/https://doi.org/10.1111/j.2517-6161.1972.tb00885.x

Livezey, B. C. (1989). Morphometric patterns in recent and fossil penguins (Aves, Sphenisciformes). *Journal of Zoology*, *219*(2), 269–307. https://doi.org/10.1111/j.1469-7998.1989.tb02582.x

Lloyd, G. T., Wang, S. C., & Brusatte, S. L. (2012). Identifying heterogeneity in rates of morphological evolution: Discrete character change in the evolution of lungfish (Sarcopterygii; Dipnoi). *Evolution*, *66*(2), 330–348. https://doi.org/10.1111/j.1558-5646.2011.01460.x

Losos, J. B. (1994). An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Systematic Biology*, *43*(1), 117–123. https://doi.org/10.2307/2413584

Losos, J. B. (2011). Convergence, adaptation, and constraint. *Evolution*, *65*(7), 1827–1840. https://doi.org/https://doi.org/10.1111/j.1558-5646.2011.01289.x

Love, A. C., Grabowski, M., Houle, D., Liow, L. H., Porto, A., Tsuboi, M., Voje, K. L., & Hunt, G. (2021). Evolvability in the fossil record. *Paleobiology*, 1–24. https://doi.org/10.1017/pab.2021.36

Lovvorn, J., Liggins, G. A., Borstad, M. H., Calisal, S. M., & Mikkelsen, J. (2001). Hydrodynamic drag of diving birds: Effects of body size, body shape and feathers at steady speeds. *Journal of Experimental Biology*, *204*(9), 1547–1557.

Lü, J., Unwin, D. M., Jin, X., Liu, Y., & Ji, Q. (2010). Evidence for modular evolution in a long-tailed pterosaur with a pterodactyloid skull. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1680), 383–389. https://doi.org/10.1098/rspb.2009.1603

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. https://doi.org/https://doi.org/10.1002/sim.3680

Maddison, W. P., & Maddison, D. R. (2019). Mesquite: A modular system for evolutionary analysis. Version 3.61. https://www.mesquiteproject.org/

Marcus, L. F., Hingst-Zaher, E., & Zaher, H. (2000). Application of landmark morphometrics to skulls representing the orders of living mammals. *Hystrix, the Italian Journal of Mammalogy*, *11*(1).

Marples, B. J., & Finlay, H. J. (1952). *Early tertiary penguins of new zealand*. Print. by J. McIndoe by authority RE Owen, Government Printer.

Marroig, G., & Cheverud, J. M. (2005). Size as a line of least evolutionary resistance: Diet and adaptive morphological radiation in new world monkeys. *Evolution*, *59*(5), 1128–1142. https://doi.org/https://doi.org/10.1111/j.0014-3820.2005.tb01049.x

Marroig, G., & Cheverud, J. M. (2010). Size as a line of least resistance II: Direct selection on size or correlated response due to constraints? *Evolution*, *64*(5), 1470–1488. https://doi.org/https://doi.org/10.1111/j.1558-5646.2009.00920.x

Martins, E. P. (1996). Conducting phylogenetic comparative studies when the phylogeny is not known. *Evolution*, *50*(1), 12–22. https://doi.org/https://doi.org/10.1111/j.1558-5646.1996.tb04468.x

Martins, E. P., & Hansen, T. F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, *149*(4), 646–667.

Matzke, N. J., & Irmis, R. B. (2018). Including autapomorphies is important for paleontological tip-dating with clocklike data, but not with non-clock data. *PeerJ*, *6*, e4553. https://doi.org/10.7717/peerj.4553

Maurer, B. A., Brown, J. H., & Rusler, R. D. (1992). The micro and macro in body size evolution. *Evolution*, *46*(4), 939–953. https://doi.org/https://doi.org/10.1111/j.1558-5646.1992.tb00611.x

Maurits, L., Forkel, R., Kaiping, G. A., & Atkinson, Q. D. (2017). BEASTling: A software tool for linguistic phylogenetics using BEAST 2. *PLOS ONE*, *12*(8), e0180908. https://doi.org/10.1371/journal.pone.0180908

Mayr, G. (2005). Tertiary plotopterids (Aves, Plotopteridae) and a novel hypothesis on the phylogenetic relationships of penguins (Spheniscidae). *Journal of Zoological Systematics and Evolutionary Research*, *43*(1), 61–71. https://doi.org/10.1111/j.1439-0469.2004.00291.x

Mayr, G. (2016). Aequornithes: Aquatic and semi-aquatic carnivores. *Avian Evolution* (pp. 161–188). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119020677.ch10

Mayr, G., De Pietri, V. L., Love, L., Mannering, A. A., Bevitt, J. J., & Scofield, R. P. (2020). First complete wing of a stem group Sphenisciform from the

Paleocene of New Zealand sheds light on the evolution of the penguin flipper. *Diversity*, *12*(2), 46. https://doi.org/10.3390/d12020046

Mayr, G., De Pietri, V. L., Love, L., Mannering, A. A., & Scofield, R. P. (2017). A well-preserved new mid-Paleocene penguin (Aves, Sphenisciformes) from the Waipara Greensand in New Zealand. *Journal of Vertebrate Paleontology*, *37*(6), e1398169.

Mayr, G., Goedert, J. L., De Pietri, V. L., & Scofield, R. P. (2021). Comparative osteology of the penguin-like mid-Cenozoic Plotopteridae and the earliest true fossil penguins, with comments on the origins of wing-propelled diving. *Journal of Zoological Systematics and Evolutionary Research*, *59*(1), 264–276.

Mayr, G., Pietri, V. L. D., Love, L., Mannering, A., & Scofield, R. P. (2019). Leg bones of a new penguin species from the Waipara Greensand add to the diversity of very large-sized Sphenisciformes in the Paleocene of New Zealand. *Alcheringa: An Australasian Journal of Palaeontology*, *0*(0), 1–8. https://doi.org/10.1080/03115518.2019.1641619

Mayr, G., Scofield, R. P., Pietri, V. L. D., & Tennyson, A. J. D. (2017). A Paleocene penguin from New Zealand substantiates multiple origins of gigantism in fossil Sphenisciformes. *Nature Communications*, *8*(1), 1–8. https://doi.org/10.1038/s41467-017-01959-6

McClain, C. R., & Boyer, A. G. (2009). Biodiversity and body size are linked across metazoans. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2209–2215. https://doi.org/10.1098/rspb.2009.0245

McElreath, R. (2020). *Statistical rethinking : A Bayesian course with examples in R and Stan* (2nd edition). Chapman; Hall/CRC. https://doi.org/10.1201/9780429029608

McGhee, G. R. (1999). *Theoretical morphology: The concept and its applications*. Columbia University Press.

McGrayne, S. B. (2011). *The theory that would not die : How Bayes' rule cracked the enigma code, hunted down russian submarines, and emerged triumphant from two centuries of controversy*. Yale University Press.

McNab, B. K., & Ellis, H. I. (2006). Flightless rails endemic to islands have lower energy expenditures and clutch sizes than flighted rails on islands and continents. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *145*(3), 295–311. https://doi.org/10.1016/j.cbpa.2006.02.025

McNeill Alexander, R. (1998). All-time giants: The largest animals and their problems. *Palaeontology*, *41*(6), 1231–1246.

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Mitteroecker, P., & Gunz, P. (2009). Advances in geometric morphometrics. *Evolutionary Biology*, *36*(2), 235–247. https://doi.org/10.1007/s11692-009-9055-x

Mongiardino Koch, N., & Thompson, J. R. (2021). A Total-Evidence dated phylogeny of Echinoidea combining phylogenomic and paleontological data. *Systematic Biology*, *70*(3), 421–439. https://doi.org/10.1093/sysbio/syaa069

Monteiro, L. R. (1999). Multivariate regression models and geometric morphometrics: The search for causal factors in the analysis of shape. *Systematic Biology*, *48*(1), 192–199. https://doi.org/10.1080/106351599260526

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Morlon, H., Lewitus, E., Condamine, F. L., Manceau, M., Clavel, J., & Drury, J. (2016). RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. *Methods in Ecology and Evolution*, *7*(5), 589–597. https://doi.org/10.1111/2041-210X.12526

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Paul-Christian, B. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Neal, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, *111*(1), 194–203. https://doi.org/10.1006/jcph.1994.1054

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, *20*(3), 263–294.

O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., Pisani, D., & Donoghue, P. C. J. (2016). Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biology Letters*, *12*(4), 20160081. https://doi.org/10.1098/rsbl.2016.0081

Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *255*(1342), 37–45.

Panchen, A. L. (1999). Homology–history of a concept. *Novartis Foundation Symposium*, *222*, 5–18, discussion 18–23.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Parham, J. F., Donoghue, P. C. J., Bell, C. J., Calway, T. D., Head, J. J., Holroyd, P. A., Inoue, J. G., Irmis, R. B., Joyce, W. G., Ksepka, D. T., Patané, J. S. L., Smith, N. D., Tarver, J. E., van Tuinen, M., Yang, Z., Angielczyk, K. D., Greenwood, J. M., Hipsley, C. A., Jacobs, L., ... Benton, M. J. (2011). Best practices for justifying fossil calibrations. *Systematic Biology*, *61*(2), 346–359. https://doi.org/10.1093/sysbio/syr107

Parins-Fukuchi, C. (2018). Use of continuous traits can improve morphological phylogenetics. *Systematic Biology*, *67*(2), 328–339. https://doi.org/10.1093/sysbio/syx072

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd edition). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2017). *The book of why: The new science of cause and effect* (1° edition). Basic Books.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pender, J. E., Hipp, A. L., Hahn, M., & Starr, J. R. (2021). Trait evolution rates shape continental patterns of species richness in north america's most diverse angiosperm genus (*Carex*, Cyperaceae). *Journal of Systematics and Evolution*, *59*(4), 763–775. https://doi.org/https://doi.org/10.1111/jse.12739

Pigot, A. L., Sheard, C., Miller, E. T., Bregman, T. P., Freeman, B. G., Roll, U., Seddon, N., Trisos, C. H., Weeks, B. C., & Tobias, J. A. (2020). Macroevolutionary convergence connects morphological form to ecological function in birds.

Nature Ecology & Evolution, *4*(2), 230–239. https://doi.org/10.1038/s41559-019-1070-4

Pimiento, C., Cantalapiedra, J. L., Shimada, K., Field, D. J., & Smaers, J. B. (2019). Evolutionary pathways toward gigantism in sharks and rays. *Evolution*, *73*(3), 588–599. https://doi.org/https://doi.org/10.1111/evo.13680

Pinshow, B., Fedak, M. A., & Schmidt-Nielsen, K. (1977). Terrestrial locomotion in penguins: It costs more to waddle. *Science*, *195*(4278), 592–594. https://doi.org/10.1126/science.835018

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, *124*, 1–10.

Pol, D., & Escapa, I. H. (2009). Unstable taxa in cladistic analysis: Identification and the assessment of relevant characters. *Cladistics*, *25*(5), 515–527. https://doi.org/10.1111/j.1096-0031.2009.00258.x

Puttick, M. N., Thomas, G. H., & Benton, M. J. (2014). High rates of evolution preceded the origin of birds. *Evolution*, *68*(5), 1497–1510. https://doi.org/10.1111/evo.12363

Pyenson, N. D., Kelley, N. P., & Parham, J. F. (2014). Marine tetrapod macroevolution: Physical and biological drivers on 250ma of invasions and evolution in ocean ecosystems. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *400*, 1–8. https://doi.org/https://doi.org/10.1016/j.palaeo.2014.02.018

Pyenson, N. D., & Vermeij, G. J. (2016). The rise of ocean giants: Maximum body size in Cenozoic marine mammals as an indicator for productivity in the pacific and atlantic oceans. *Biology Letters*, *12*(7), 20160186. https://doi.org/10.1098/rsbl.2016.0186

Quiroz, M., Kohn, R., Villani, M., & Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, *114*(526), 831–843. https://doi.org/10.1080/01621459.2018.1448827

R core Team. (2021). R: A language and environment for statistical computing. https://www.R-project.org

Rannala, B., & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, *43*(3), 304–311. https://doi.org/10.1007/BF02338839

Rannala, B., & Yang, Z. (2007). Inferring speciation times under an episodic molecular clock. *Systematic Biology*, *56*(3), 453–466. https://doi.org/10.1080/10635150701420643

Raup, D. M. (1967). Geometric analysis of shell coiling: Coiling in ammonoids. *Journal of Paleontology*, *41*(1), 43–65.

Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, *1*(4), 319–329. https://doi.org/https://doi.org/10.1111/j.2041-210X.2010.00044.x

Rhoda, D., Segall, M., Larouche, O., Evans, K., & Angielczyk, K. D. (2021). Local superimpositions facilitate morphometric analysis of complex articulating structures. *Integrative and Comparative Biology*, (icab031). https://doi.org/10.1093/icb/icab031

Richards, M. (2019). *Two giant penguins from the Eocene-Oligocene of Otago, New Zealand.* (Master's thesis). University of Otago. Dunedin, New Zealand.

Ringen, E., Martin, J. S., & Jaeggi, A. (2021). Novel phylogenetic methods reveal that resource-use intensification drives the evolution of "complex" societies.

Rittmeyer, E. N., Allison, A., Gründler, M. C., Thompson, D. K., & Austin, C. C. (2012). Ecological guild evolution and the discovery of the world's smallest vertebrate. *PLOS ONE*, *7*(1), e29797. https://doi.org/10.1371/journal.pone.0029797

Robinson, W. R., Peters, R. H., & Zimmermann, J. (1983). The effects of body size and temperature on metabolic rate of organisms. *Canadian Journal of Zoology*, *61*(2), 281–288.

Rohlf, F. J. (1999). Shape statistics: Procrustes superimpositions and tangent spaces. *Journal of Classification*, *16*(2), 197–223. https://doi.org/10.1007/s003579900054

Rohlf, F. J., & Marcus, L. F. (1993). A revolution morphometrics. *Trends in Ecology & Evolution*, *8*(4), 129–132. https://doi.org/10.1016/0169-5347(93)90024-J

Rohlf, F. J., & Slice, D. (1990). Extensions of the procrustes method for the optimal superimposition of landmarks. *Systematic Biology*, *39*(1), 40–59. https://doi.org/10.2307/2992207

Rolfe, S., Pieper, S., Porto, A., Diamond, K., Winchester, J., Shan, S., Kirveslahti, H., Boyer, D., Summers, A., & Maga, A. M. (2021). SlicerMorph: An open and extensible platform to retrieve, visualize and analyze 3D morphology. *Methods in Ecology and Evolution*, *12*(10), 1816–1825. https://doi.org/10.1111/2041-210X.13669

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., & Rasnitsyn, A. P. (2012). A Total-Evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, *61*(6), 973–999. https://doi.org/10.1093/sysbio/sys058

Rüber, L., & Adams, D. C. (2001). Evolutionary convergence of body shape and trophic morphology in cichlids from lake tanganyika. *Journal of Evolutionary Biology*, *14*(2), 325–332. https://doi.org/10.1046/j.1420-9101.2001.00269.x

Rudwick, M. J. S. (1964). The function of zigzag deflexions in the commissures of fossil brachiopods. *Palaeontology*, *7*(1), 135–171.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and Implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1175

Schlager, S. (2017). Morpho and Rvcg – shape analysis in R: R-packages for geometric morphometrics, shape analysis and surface manipulations. In G. Zheng, S. Li, & G. Székely (Eds.), *Statistical Shape and Deformation Analysis* (pp. 217–256). Academic Press. https://doi.org/10.1016/B978-0-12-810493-4.00011-0

Schreiweis, D. O. (1982). A comparative study of the appendicular musculature of penguins (Aves, Sphenisciformes). *Smithsonian Contributions to Zoology*.

Sears, R., & Perrin, W. F. (2009). Blue whale: *Balaenoptera musculus*. In W. F. Perrin, B. Würsig, & J. G. M. Thewissen (Eds.), *Encyclopedia of Marine Mammals (Second Edition)* (pp. 120–124). Academic Press. https://doi.org/10.1016/B978-0-12-373553-9.00033-X

Seilacher, A. (1970). Arbeitskozept zur Konstruktions-Morphologie. *Lethaia*, *3*(4), 393–396. https://doi.org/https://doi.org/10.1111/j.1502-3931.1970.tb00830.x

Seilacher, A. (1984). Constructional morphology of bivalves: Evolutionary pathways in primary versus secondary soft-bottom dwellers. *Palaeontology*, *27*(2), 207–237.

Seilacher, A., & Gishlick, A. D. (2019). *Morphodynamics*. CRC Press. https://doi.org/10.1201/b17557

Shea, B. T. (1977). Eskimo craniofacial morphology, cold stress and the maxillary sinus. *American Journal of Physical Anthropology*, *47*(2), 289–300. https://doi.org/https://doi.org/10.1002/ajpa.1330470209

Shrier, I. (2013). Estimating causal effect with randomized controlled trial. *Epidemiology*, *24*(5), 779–781. https://doi.org/10.1097/EDE.0b013e31829f6d21

Sibert, E., Friedman, M., Hull, P., Hunt, G., & Norris, R. (2018). Two pulses of morphological diversification in pacific pelagic fishes following the Cretaceous–Palaeogene mass extinction. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1888), 20181194. https://doi.org/10.1098/rspb.2018.1194

Simpson, G. G. (1944). *Tempo and Mode in Evolution*. Columbia University Press.

Simpson, G. G. (1946). Fossil penguins. *Bulletin of the American Museum of Natural History*, *87*(1), UR6.

Simpson, G. G. (1953). *The major features of evolution*. Columbia University Press.

Simpson, G. G. (1981). Notes on some fossil penguins, including a new genus from Patagonia. *Ameghiniana*, *18*(3-4), 266–272.

Slack, K. E., Jones, C. M., Ando, T., Harrison, G. L., Fordyce, R. E., Arnason, U., & Penny, D. (2006). Early penguin fossils, plus mitochondrial genomes, calibrate avian evolution. *Molecular Biology and Evolution*, *23*(6), 1144–1155. https://doi.org/10.1093/molbev/msj124

Slater, G. J., Goldbogen, J. A., & Pyenson, N. D. (2017). Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1855), 20170546. https://doi.org/10.1098/rspb.2017.0546

Šmíd, J., & Tolley, K. A. (2019). Calibrating the tree of vipers under the Fossilized Birth-Death model. *Scientific Reports*, *9*(1), 5510. https://doi.org/10.1038/s41598-019-41290-2

Smith, A. F. M. (1973). A general Bayesian linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, *35*(1), 67–75. https://doi.org/https://doi.org/10.1111/j.2517-6161.1973.tb00937.x

Stadler, T., Gavryushkina, A., Warnock, R. C. M., Drummond, A. J., & Heath, T. A. (2018). The Fossilized Birth-Death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology*, *447*, 41–55. https://doi.org/10.1016/j.jtbi.2018.03.005

Stone, E. A. (2011). Why the phylogenetic regression appears robust to tree misspecification. *Systematic Biology*, *60*(3), 245–260. https://doi.org/10.1093/sysbio/syq098

Stonehouse, B. (1975). *The Biology of Penguins* (B. Stonehouse, Ed.; 1st Edition). Macmillan/University Park Press.

Sundberg, P. (1989). Shape and size-constrained principal components analysis. *Systematic Zoology*, *38*(2), 166–168. http://www.jstor.org/stable/2992385

Sutton, M., Rahman, I., & Garwood, R. (2016). Virtual Paleontology — An overview. *The Paleontological Society Papers*, *22*, 1–20. https://doi.org/10.1017/scs.2017.5

Swofford, D., Olsen, G., Waddel, P., & Hillis, D. (1996). Phylogenetic inference. In D. Hillis, C. Moritz, & B. Mable (Eds.), *Molecular systematics*. Sinauer.

Symonds, M. R. E., & Blomberg, S. P. (2014). A primer on phylogenetic generalised least squares. In L. Z. Garamszegi (Ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice* (pp. 105–130). Springer. https://doi.org/10.1007/978-3-662-43550-2_5

Tambussi, C. P., Reguero, M. A., Marenssi, S. A., & Santillana, S. N. (2005). *Crossvallia unienwillia*, a new Spheniscidae (Sphenisciformes, Aves) from the Late Paleocene of Antarctica. *Geobios*, *38*(5), 667–675. https://doi.org/https://doi.org/10.1016/j.geobios.2004.02.003

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, *17*(2), 57–86.

Theobald, D. L., & Wuttke, D. S. (2006). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proceedings of the National Academy of Sciences*, *103*(49), 18521–18527. https://doi.org/10.1073/pnas.0508445103

Thomas, D. B., & Ksepka, D. T. (2016). The Glen Murray fossil penguin from the North Island of New Zealand extends the geographic range of *Kairuku*. *Journal of the Royal Society of New Zealand*, *46*(3-4), 200–213. https://doi.org/10.1080/03036758.2016.1211541

Thomas, D. B., Tennyson, A. J. D., Scofield, R. P., Heath, T. A., Pett, W., & Ksepka, D. T. (2020). Ancient crested penguin constrains timing of recruitment into seabird hotspot. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1932), 20201497. https://doi.org/10.1098/rspb.2020.1497

Thompson, D. W. (1945). *On growth and form*. Cambridge: University Press.

Thorne, J. L., Kishino, H., & Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, *15*(12), 1647–1657. https://doi.org/10.1093/oxfordjournals.molbev.a025892

Tobalske, B. W. (2007). Biomechanics of bird flight. *Journal of Experimental Biology*, *210*(18), 3135–3146. https://doi.org/10.1242/jeb.000273

Tocheri, M. W. (2009). Laser scanning: 3D analysis of biological surfaces. In C. W. Sensen & B. Hallgrímsson (Eds.), *Advanced Imaging in Biology and Medicine: Technology, Software Environments, Applications* (pp. 85–101). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68993-5_4

Tomiya, S. (2011). A new basal caniform (Mammalia: Carnivora) from the Middle Eocene of North America and remarks on the phylogeny of early Carnivorans. *PLOS ONE*, *6*(9), e24146. https://doi.org/10.1371/journal.pone.0024146

Troelsen, P. V., Wilkinson, D. M., Seddighi, M., Allanson, D. R., & Falkingham, P. L. (2019). Functional morphology and hydrodynamics of plesiosaur necks: Does size matter? *Journal of Vertebrate Paleontology*, *39*(2), e1594850. https://doi.org/10.1080/02724634.2019.1594850

Tschopp, E., Mateus, O., & Benson, R. B. J. (2015). A specimen-level phylogenetic analysis and taxonomic revision of Diplodocidae (Dinosauria, Sauropoda). *PeerJ*, *3*, e857. https://doi.org/10.7717/peerj.857

Tschopp, E., Napoli, J. G., Wencker, L. C. M., Delfino, M., & Upchurch, P. (2021). How to render species comparable taxonomic units through deep time: A case study on intraspecific osteological variability in extant and extinct lacertid lizards. *Systematic Biology*, syab078. https://doi.org/10.1093/sysbio/syab078

Uyeda, J. C., Caetano, D. S., & Pennell, M. W. (2015). Comparative analysis of Principal Components can be misleading. *Systematic Biology*, *64*(4), 677–689. https://doi.org/10.1093/sysbio/syv019

Valkenburgh, B. V. (1985). Locomotor diversity within past and present guilds of large predatory mammals. *Paleobiology*, *11*(4), 406–428. https://doi.org/10.1017/S0094837300011702

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1–26. https://doi.org/10.1038/s43586-020-00001-2

Vasilakis, D. P., Whitfield, D. P., Schindler, S., Poirazidis, K. S., & Kati, V. (2016). Reconciling endangered species conservation with wind farm development: Cinereous vultures (*Aegypius monachus*) in south-eastern Europe. *Biological Conservation*, *196*, 10–17. https://doi.org/10.1016/j.biocon.2016.01.014

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vernygora, O. V., Simões, T. R., & Campbell, E. O. (2020). Evaluating the performance of probabilistic algorithm for phylogenetic analysis of big morphological datasets: A simulation study. *Systematic biology*, *69*(6), 1088–1105.

Vizcaíno, S. F., & Fariña, R. A. (1999). On the flight capabilities and distribution of the giant Miocene bird *Argentavis magnificens* (Teratornithidae). *Lethaia*, *32*(4), 271–278. https://doi.org/https://doi.org/10.1111/j.1502-3931.1999.tb00546.x

Wagner, G. P. (1989). The biological homology concept. *Annual Review of Ecology and Systematics*, *20*(1), 51–69.

Walton, S. A., & Korn, D. (2018). An ecomorphospace for the ammonoidea. *Paleobiology*, *44*(2), 273–289. https://doi.org/10.1017/pab.2017.33

Wang, M., & Lloyd, G. T. (2016). Rates of morphological evolution are heterogeneous in Early Cretaceous birds. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1828), 20160214. https://doi.org/10.1098/rspb.2016.0214

Watanabe, A. (2018). How many landmarks are enough to characterize shape and size variation? *PloS one*, *13*(6), e0198341. https://doi.org/https://doi.org/10.1371/journal.pone.0198341

Watanabe, A., Fabre, A.-C., Felice, R. N., Maisano, J. A., Müller, J., Herrel, A., & Goswami, A. (2019). Ecomorphological diversification in squamates from conserved pattern of cranial integration. *Proceedings of the National Academy of Sciences*, *116*(29), 14688–14697.

Watanabe, J., Field, D. J., & Matsuoka, H. (2020). Wing musculature reconstruction in extinct flightless auks (*Pinguinus* and *Mancalla*) reveals incomplete convergence with penguins (Spheniscidae) due to differing ancestral states. *Integrative Organismal Biology*, *3*(1). https://doi.org/10.1093/iob/obaa040

Webster, M., & Sheets, H. D. (2010). A practical introduction to landmark-based geometric morphometrics. *The Paleontological Society Papers*, *16*, 163–188. https://doi.org/10.1017/S1089332600001868

Weir, J. T., & Schluter, D. (2008). Calibrating the avian molecular clock. *Molecular Ecology*, *17*(10), 2321–2328. https://doi.org/10.1111/j.1365-294X.2008.03742.x

Wiens, J. J., & Morrill, M. C. (2011). Missing data in phylogenetic analysis: Reconciling results from simulations and empirical data. *Systematic Biology*, *60*(5), 719–731. https://doi.org/10.1093/sysbio/syr025

Wiley, E. O., & Lieberman, B. S. (2011). *Phylogenetics: Theory and practice of phylogenetic systematics* (2nd edition). John Wiley & Sons.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *22*(3), 392–399. https://doi.org/https://doi.org/10.2307/2346786

Willmer, P., Stone, G., & Johnston, I. (2009). *Environmental physiology of animals*. John Wiley & Sons.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science*, *69*, S316–S330. https://doi.org/10.1086/341855

Wright, A. M., & Hillis, D. M. (2014). Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109210

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, *20*(7), 557–585.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the sixth international congress of Genetics*, *1*, 356–366.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, *60*(2), 150–160. https://doi.org/10.1093/sysbio/syq085

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results*. *The Quarterly Journal of Economics*, *134*(2), 557–598. https://doi.org/10.1093/qje/qjy029

Zelditch, M. L., Swiderski, D. L., & Sheets, H. D. (2012). *Geometric morphometrics for biologists: A primer*. Academic Press.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006

Zuckerkandl, E., & Pauling, L. (1965, January 1). Evolutionary divergence and convergence in proteins. In V. Bryson & H. J. Vogel (Eds.), *Evolving genes and proteins* (pp. 97–166). Academic Press. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6

Appendix A

Publication of *Kairuku waewaeroa*

The following section will include the paper published on the Journal of Vertebrate Paleontology that described *Kairuku waewaeroa*.

Journal of Vertebrate Paleontology e1953047 (13 pages)

DOI: 10.1080/02724634.2021.1953047

ARTICLE

A GIANT OLIGOCENE FOSSIL PENGUIN FROM THE NORTH ISLAND OF NEW ZEALAND

SIMONE GIOVANARDI, *[1] DANIEL T. KSEPKA, [2] and DANIEL B. THOMAS [3]
[1]School of Natural and Computational Sciences, Massey University, Auckland, 0632 New Zealand, giovanerd90@gmail.com;
[2]Bruce Museum, Greenwich, CT, 06830, USA, ksepka@gmail.com;
[3]School of Natural and Computational Sciences, Massey University, Auckland, 0632 New Zealand, d.b.thomas@massey.ac.nz

ABSTRACT—Penguins (Sphenisciformes) have arguably the most complete and continuous fossil record of any avian clade, offering an ever-improving understanding of penguin phylogeny, biogeography, and the evolution of wing-propelled diving. Yet, our knowledge of the precise body proportions of stem-group penguins remains poor due to a dearth of articulated specimens. Here, we describe *Kairuku waewaeroa* sp. nov., a new giant penguin species from the Glen Massey Formation (Whaingaroan stage, 34.6–27.3 Ma). The holotype skeleton, discovered in Kawhia Harbour, North Island, New Zealand, is one of the most complete skeletons of a giant penguin yet uncovered. Our phylogenetic analysis recovers a clade uniting the New Zealand endemics *Kairuku waewaeroa*, *Kairuku waitaki*, and *Kairuku grebneffi*, which is supported by synapomorphies including a stout femoral shaft and tibiotarsi with a distinctly convex medial condyle. *Kairuku waewaeroa* is unique among stem penguins in having elongate tibiotarsi, revealing a new long-legged stem penguin body plan. The discovery of *Kairuku waewaeroa* contributes yet another penguin species to an Oligocene avifauna for Zealandia that is replete with giant birds.

http://zoobank.org./urn:lsid:zoobank.org:pub:8B8EBB57-80CB-4720-920D-7DABC5E5B95B

SUPPLEMENTAL DATA—Supplemental materials are available for this article for free at www.tandfonline.com/UJVP

Citation for this article: Giovanardi, S., D. T. Ksepka, and D. B. Thomas. 2021. A giant Oligocene fossil penguin from the North Island of New Zealand. Journal of Vertebrate Paleontology. DOI: 10.1080/10.1080/02724634.2021.1953047

INTRODUCTION

Penguins have a broad austral distribution and a fossil record that spans almost the entire Cenozoic. Since the first report of a fossil penguin by Thomas Henry Huxley (1859), the fossil record for these diving birds has grown to more than 60 species (Jadwiszczak, 2009; Ksepka and Ando, 2011; Mayr, 2017). The rich fossil record of penguins has provided much insight into the evolution of adaptations to a secondarily aquatic existence in penguins and has provided a valuable case study for the evolution of adaptation itself (Simpson, 1953). While modern penguins appear to have a relatively conservative bauplan, the fossil record has revealed a wide range of body size and shape variation for penguins. In recent years, discoveries have improved our understanding of how penguins diversified shortly after the Cretaceous–Paleogene extinction and documented the presence of many "giant" Paleogene species (Jadwiszczak, 2001; Slack et al., 2006; Clarke et al., 2007; Ksepka et al., 2012; Acosta Hospitaleche, 2016; Mayr et al., 2017a; Mayr et al. 2019). Phylogenetic analyses suggest that giant penguins do not represent a clade, but that multiple lineages of penguins attained larger size independently and that these species may instead represent an example of parallel evolution (Clarke et al. 2007; Chávez Hoffmeister et al., 2014; Gavryushkina et al., 2017; Mayr et al., 2017b).

Despite the aforementioned abundance in specimens, many fossil penguins (including most 'giant' species) are incomplete and disarticulated. Among the earliest penguins the most notable taxa in term of completeness and size are respectively *Sequiwaimanu rosieae* Mayr, De Pietri, Love, Mannering, and Scofield, 2017 and *Kumimanu biceae* Mayr, Scofield, De Pietri, and Tennyson, 2017 whereas in regard to Oligocene giants some of the most complete specimens belong to the genus *Kairuku* Ksepka, Fordyce, Ando, and Jones, 2012. These latter specimens revealed a tall and thin body plan which differed substantially from even the largest modern species (Emperor Penguin *Aptenodytes forsteri* and King Penguin *Aptenodytes patagonicus*). Here, we describe a new species of *Kairuku* from Kawhia Harbour in the North Island of New Zealand that shows key similarities with the previously described *Kairuku* specimens, but also reveals a different overall body shape. The new fossil is also of regional importance: the South Island of New Zealand has historically been one of the most productive localities worldwide for fossil penguins (Fordyce and Jones, 1990), whereas records from the North Island have long been limited to a few fragmentary specimens (Marples and Fleming, 1963; Grant-Mackie and Simpson, 1973; Thomas and Ksepka, 2016).

Institutional Abbreviations— **AMNH**, American Museum of Natural History, New York, USA; **BMNH**, Natural History Museum, Tring, UK; **CM**, Canterbury Museum, Christchurch, New Zealand; **MUSM**, Museo de Historia Natural, Universidad Nacional Mayor de San Marcos, Lima, Peru; **NMNZ**, Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand; **OM**, Otago Museum, Dunedin, New Zealand; **OU**, Otago University, Dunedin, New Zealand; **UA**, The University of Auckland, Auckland, New Zealand; **UCMP**, University of California Museum of Paleontology, Berkeley, CA, USA; **WM**, Waikato Museum Te Whare Taonga o Waikato, Hamilton, New Zealand.

MATERIALS AND METHODS

Comparative Material

Specimens compared for this study included: *Archaeospheniscus lowei* Marples, 1952 OM GL407; *Anthropornis nordenskjoeldi*

*Corresponding author
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujvp.

DRC 16

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Simone Giovanardi
Name/title of Primary Supervisor:	Daniel Thomas
In which chapter is the manuscript /published work:	Chapter 2

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Giovanardi S, Ksepka DT & Thomas DB (2021) A giant Oligocene fossil penguin from the North Island of New Zealand, Journal of Vertebrate Paleontology, DOI: 10.1080/02724634.2021.1953047

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

DBT and DTK conceived the study. SG, DTK and DBT collected the data. SG conducted the majority of data analysis with input from DBT and DTK. SG, DBT and DTK wrote and edited the manuscript.

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Simone Giovanardi
Date:	29/11/2021
Primary Supervisor's Signature:	
Date:	29/11/2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

GRS Version 5 – 13 December 2019
DRC 19/09/10

Appendix B

Chapter 3 Additional material

The following section will include the parameter estimates from both HAF and Femur Bayesian models (Table B.1) along with the results from the body mass Re-estimation step (Table B.2) and the body mass estimates for fossil specimens that preserved femurs (Table B.3).

Table B.1: Parameter estimates for all models estimated in chapter 3. Mean represents the average from the posterior weights estimates, whereas upper and lower denote the 89% credible interval

	Model number	Parameter	Mean	Standard deviation	Lower	Upper
HAF	I	β	2.45	0.01	2.44	2.47
		α	1.99	0.01	1.97	2.01
		σ	0.20	0.00	0.19	0.21
	II	β	2.44	0.01	2.42	2.47
		α	2.35	0.06	2.26	2.44
		σ	0.17	0.00	0.16	0.17
Femur	I	β	0.84	0.01	0.82	0.86
		α	6.12	0.02	6.08	6.16
		σ	0.44	0.02	0.42	0.47
	II	β	0.84	0.01	0.82	0.86
		α	6.93	0.08	6.80	7.06
		σ	0.31	0.01	0.29	0.33
	III	β	0.84	0.01	0.81	0.86
		α	6.76	0.08	6.64	6.89
		σ	0.31	0.01	0.29	0.33
	IV	β	0.85	0.01	0.83	0.87
		α	6.50	0.03	6.45	6.54
		σ	0.38	0.02	0.35	0.40
	V	β	0.85	0.01	0.83	0.87
		α	6.49	0.03	6.44	6.54
		σ	0.38	0.02	0.35	0.40
	VI	β	0.36	0.03	0.31	0.41
		α	3.13	0.75	1.93	4.32
		σ	2.68	0.11	2.51	2.85
	VII	β	0.28	0.03	0.22	0.33
		α	5.38	0.87	3.97	6.76
		σ	2.43	0.10	2.27	2.59
	VIII	β	0.75	0.02	0.72	0.77
		α	6.04	0.16	5.79	6.30
		η^2	0.24	0.03	0.19	0.29
		ρ^2	3.06	0.24	2.67	3.44
	IX	β	0.76	0.02	0.74	0.79
		α	6.29	0.15	6.04	6.53
		η^2	0.21	0.03	0.17	0.25
		ρ^2	3.08	0.24	2.70	3.47

Table B.2: Body mass estimates from or the femur validation test set phase. Estimated average body mass reported for each model, as well as the lower and upper 89% credible intervals, along with the real body mass for each bird.

Species	Observed	Model	Mean	Lower	Upper
Aethia psittacula	0.27	I	0.16	0.15	0.17
		II	0.18	0.17	0.19
		III	0.18	0.17	0.19
		IV	0.2	0.19	0.21
		V	0.2	0.19	0.21
		VI	0.3	0.08	0.68
		VII	0.29	0.09	0.65
		VIII	0.25	0.17	0.35
		IX	0.26	0.17	0.36
Alca torda	0.73	I	0.32	0.31	0.33
		II	0.35	0.33	0.37
		III	0.35	0.33	0.37
		IV	0.39	0.37	0.41
		V	0.39	0.37	0.41
		VI	0.57	0.14	1.45
		VII	0.52	0.15	1.16
		VIII	0.57	0.38	0.8
		IX	0.57	0.39	0.8
Aptenodytes patagonicus	11.72	I	5.11	4.72	5.49
		II	10.26	8.98	11.62
		III	9.93	8.69	11.26
		IV	6.64	6.15	7.15
		V	6.56	6.1	7.06
		VI	26.21	3.84	75.5
		VII	25.3	4.76	67.05

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Aptenodytes patagonicus	11.72	VIII	14.69	9.16	22
		IX	14.74	9.46	21.99
Apteryx haastii	2.05	I	5.23	4.89	5.62
		II	4.05	3.33	4.85
		III	4.24	3.55	5.01
		IV	4.36	3.72	5.05
		V	4.36	3.79	4.98
		VI	4.3	1.03	10.83
		VII	4.16	1.26	9.19
		VIII	4.35	2.85	6.26
		IX	4.29	2.89	6.01
Bubulcus ibis	0.37	I	0.52	0.5	0.54
		II	0.36	0.31	0.4
		III	0.37	0.32	0.42
		IV	0.42	0.39	0.46
		V	0.42	0.4	0.45
		VI	1.46	0.08	4.96
		VII	1.37	0.1	4.29
		VIII	0.44	0.22	0.76
		IX	0.41	0.2	0.7
Coenocorypha pusilla	0.08	I	0.11	0.11	0.12
		II	0.1	0.09	0.12
		III	0.11	0.09	0.12
		IV	0.08	0.07	0.09
		V	0.08	0.08	0.09
		VI	0.09	0.06	0.13

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Coenocorypha pusilla	0.08	VII	0.1	0.06	0.14
		VIII	0.07	0.05	0.09
		IX	0.07	0.05	0.09
Columba vitiensis	0.35	I	0.34	0.33	0.35
		II	0.32	0.28	0.35
		III	0.32	0.29	0.35
		IV	0.27	0.25	0.29
		V	0.27	0.26	0.29
		VI	0.52	0.1	1.38
		VII	0.49	0.12	1.17
		VIII	0.36	0.23	0.52
		IX	0.37	0.23	0.53
Diomedea exulans	8.19	I	3.8	3.54	4.07
		II	6.28	5.55	7.09
		III	6.15	5.49	6.89
		IV	4.91	4.58	5.26
		V	4.86	4.54	5.21
		VI	4.88	3.58	6.53
		VII	4.64	3.53	5.98
		VIII	5.94	4.71	7.37
		IX	6	4.72	7.44
Egretta novaehollandiae	0.56	I	0.77	0.74	0.81
		II	0.53	0.47	0.6
		III	0.54	0.48	0.62
		IV	0.97	0.92	1.02
		V	0.97	0.92	1.01

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Egretta novaehollandiae	0.56	VI	0.73	0.13	1.78
		VII	0.7	0.15	1.77
		VIII	0.6	0.35	0.92
		IX	0.59	0.36	0.87
Eudyptes sclateri	3.45	I	2.59	2.43	2.74
		II	5.17	4.56	5.87
		III	5.02	4.42	5.68
		IV	3.32	3.11	3.53
		V	3.28	3.08	3.5
		VI	5.74	1.52	12.65
		VII	5.57	1.75	12.05
		VIII	5.28	3.58	7.54
		IX	5.3	3.61	7.32
Eudyptula minor	1.16	I	0.65	0.63	0.68
		II	1.29	1.13	1.48
		III	1.26	1.11	1.43
		IV	0.81	0.78	0.85
		V	0.81	0.77	0.85
		VI	4.77	0.43	14.83
		VII	4.55	0.56	14.2
		VIII	1.42	0.81	2.24
		IX	1.37	0.79	2.19
Falco cenchroides	0.16	I	0.34	0.32	0.35
		II	0.14	0.08	0.21
		III	0.21	0.13	0.3
		IV	0.26	0.23	0.29

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Falco cenchroides	0.16	V	0.26	0.23	0.29
		VI	0.57	0.05	1.72
		VII	0.99	0.1	3.26
		VIII	0.22	0.11	0.35
		IX	0.24	0.14	0.41
Larus dominicanus	0.99	I	0.98	0.94	1.03
		II	1.09	1.01	1.17
		III	1.08	1.01	1.16
		IV	0.81	0.75	0.87
		V	0.8	0.75	0.86
		VI	0.85	0.65	1.07
		VII	0.82	0.64	1.04
		VIII	1.02	0.84	1.24
		IX	1	0.83	1.2
Leucosarcia melanoleuca	0.43	I	0.4	0.39	0.42
		II	0.37	0.33	0.42
		III	0.38	0.34	0.42
		IV	0.32	0.3	0.35
		V	0.33	0.3	0.35
		VI	0.55	0.04	1.95
		VII	0.5	0.04	1.8
		VIII	0.38	0.2	0.63
		IX	0.4	0.22	0.66
Megapodius eremita	0.68	I	1.39	1.32	1.46
		II	1.01	0.88	1.16
		III	1.03	0.89	1.18

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Megapodius eremita	0.68	IV	1.16	1.08	1.25
		V	1.16	1.08	1.24
		VI	4.82	0.03	15.38
		VII	6.06	0.05	15.21
		VIII	1.24	0.54	2.32
		IX	1.29	0.57	2.4
Passer domesticus	0.03	I	0.03	0.03	0.04
		II	0.03	0.02	0.03
		III	0.03	0.02	0.03
		IV	0.03	0.03	0.04
		V	0.03	0.03	0.04
		VI	0.1	0	0.35
		VII	0.04	0	0.14
		VIII	0.03	0.02	0.05
		IX	0.03	0.01	0.05
Phalacrocorax chalconotus	2.27	I	2.26	2.13	2.38
		II	2.86	2.51	3.22
		III	2.83	2.5	3.18
		IV	2.89	2.72	3.08
		V	2.87	2.71	3.04
		VI	2.06	0.6	4.88
		VII	1.91	0.59	4.32
		VIII	2.7	1.85	3.79
		IX	2.65	1.84	3.68
Phalacrocorax ranfurlyi	2.5	I	1.78	1.7	1.88
		II	2.22	1.96	2.5

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Phalacrocorax ranfurlyi	2.5	III	2.21	1.93	2.5
		IV	2.27	2.15	2.41
		V	2.26	2.12	2.4
		VI	4.48	0.2	15.77
		VII	3.82	0.24	15.09
		VIII	2.3	1.17	3.98
		IX	2.31	1.29	3.87
Porphyrio hochstetteri	2.76	I	4.29	3.98	4.6
		II	2.74	2.38	3.17
		III	2.8	2.4	3.22
		IV	3.66	3.38	3.97
		V	3.67	3.37	3.98
		VI	1.49	0.21	4.3
		VII	1.16	0.18	3.16
		VIII	1.95	1.1	2.91
		IX	2.13	1.28	3.23
Ptychoramphus aleuticus	0.18	I	0.12	0.11	0.13
		II	0.13	0.12	0.14
		III	0.13	0.12	0.14
		IV	0.14	0.14	0.15
		V	0.14	0.14	0.15
		VI	0.31	0.1	0.74
		VII	0.33	0.09	0.72
		VIII	0.21	0.14	0.28
		IX	0.21	0.14	0.29
	0.01	I	0.01	0.01	0.01

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Rhipidura fuliginosa	0.01	II	0.01	0.01	0.01
		III	0.01	0.01	0.01
		IV	0.01	0.01	0.01
		V	0.01	0.01	0.01
		VI	0.1	0	0.36
		VII	0.19	0.01	0.59
		VIII	0.01	0	0.02
		IX	0.01	0	0.02
Sterna hirundo	0.12	I	0.08	0.07	0.08
		II	0.09	0.08	0.09
		III	0.09	0.08	0.09
		IV	0.09	0.09	0.1
		V	0.09	0.09	0.1
		VI	0.16	0.07	0.31
		VII	0.17	0.08	0.29
		VIII	0.12	0.09	0.15
		IX	0.12	0.08	0.15
Tryngites subruficollis	0.06	I	0.09	0.08	0.09
		II	0.08	0.07	0.1
		III	0.09	0.07	0.1
		IV	0.07	0.06	0.08
		V	0.07	0.06	0.08
		VI	0.1	0.02	0.29
		VII	0.09	0.02	0.24
		VIII	0.08	0.05	0.12
		IX	0.08	0.05	0.11

Continued on next page

Table B.2 – *Continued from previous page*

Species	Observed	Model	Mean	Lower	Upper
Turnix varius	0.09	I	0.15	0.14	0.16
		II	0.09	0.08	0.11
		III	0.1	0.08	0.11
		IV	0.12	0.11	0.13
		V	0.12	0.11	0.13
		VI	1.17	0.01	4.14
		VII	0.82	0.02	3.02
		VIII	0.18	0.09	0.33
		IX	0.19	0.08	0.34

Table B.3: Body mass estimates for the Femur models for a set of extinct fossil penguins. Measurements are expressed in kilograms (kg), mean represents the average from the posterior body mass estimates, whereas upper and lower denote the 89% credible interval.

Fossil	Model	Mean	Lower	Upper
Aptenodytes ridgeni	I	12.1	11.0	13.3
	II	23.7	20.6	27.2
	III	23.0	20.2	26.0
	IV	16.1	14.6	17.6
	V	15.8	14.3	17.3
	VI	17.7	6.8	35.2
	VII	16.1	6.9	31.8
	VIII	27.9	19.7	37.2
	IX	28.1	20.6	37.7
Eudyptes atatu	I	2.0	1.9	2.1
	II	3.9	3.4	4.4
	III	3.8	3.4	4.3
	IV	2.6	2.4	2.7
	V	2.5	2.4	2.7
	VI	4.7	1.7	9.9
	VII	4.6	1.9	8.8
	VIII	3.9	2.8	5.3
	IX	4.0	2.9	5.2
GL429 Burnside	I	17.9	16.1	19.9
	II	35.1	30.3	40.4
	III	33.9	29.7	38.6
	IV	23.9	21.6	26.4
	V	23.4	21.1	25.9

Continued on next page

Table B.3 – *Continued from previous page*

Fossil	Model	Mean	Lower	Upper
GL429 Burnside	VI	10.6	1.0	31.7
	VII	8.5	0.9	27.4
	VIII	23.2	11.9	39.3
	IX	22.2	11.7	36.1
Kairuku grebneffi	I	17.5	15.7	19.4
	II	34.2	29.6	39.4
	III	33.0	28.9	37.6
	IV	23.3	21.0	25.7
	V	22.8	20.6	25.3
	VI	16.5	0.5	59.1
	VII	11.5	0.6	36.5
	VIII	20.0	9.5	35.0
	IX	20.3	10.2	35.8
Kairuku waewaeroa	I	22.8	20.4	25.4
	II	44.7	38.6	51.6
	III	43.1	37.8	49.1
	IV	30.6	27.5	34.0
	V	29.9	26.9	33.3
	VI	19.3	0.7	65.5
	VII	12.4	0.6	46.3
	VIII	25.5	12.2	44.9
	IX	26.1	13.1	45.8
Kairuku waitaki	I	15.3	13.9	17.0
	II	30.0	26.0	34.5
	III	29.0	25.5	33.0

Continued on next page

Table B.3 – *Continued from previous page*

Fossil	Model	Mean	Lower	Upper
Kairuku waitaki	IV	20.4	18.5	22.5
	V	20.0	18.1	22.1
	VI	16.5	0.5	55.5
	VII	10.7	0.5	37.4
	VIII	17.6	8.2	31.7
	IX	17.9	9.1	31.7
Muriwaimanu tuatahi	I	4.9	4.6	5.3
	II	9.6	8.4	11.0
	III	9.4	8.3	10.6
	IV	6.4	6.0	6.9
	V	6.3	5.9	6.8
	VI	3.7	0.4	11.4
	VII	2.9	0.4	7.9
	VIII	6.0	3.0	10.3
	IX	5.3	2.8	8.6
Palaeeudyptes sealrock	I	12.2	11.1	13.4
	II	23.8	20.7	27.3
	III	23.0	20.3	26.1
	IV	16.1	14.7	17.7
	V	15.8	14.4	17.4
	VI	9.7	0.8	29.9
	VII	7.3	0.7	24.7
	VIII	16.5	8.5	27.7
	IX	15.6	8.6	24.9
Sequiwaimanu rosieae	I	6.5	6.0	7.0

Continued on next page

Table B.3 – *Continued from previous page*

Fossil	Model	Mean	Lower	Upper
Sequiwaimanu rosieae	II	12.6	11.0	14.4
	III	12.2	10.8	13.8
	IV	8.4	7.8	9.1
	V	8.3	7.7	9.0
	VI	4.1	0.5	11.9
	VII	3.1	0.5	9.0
	VIII	7.6	4.0	12.9
	IX	7.0	3.7	11.2
Seymour OU22195	I	14.5	13.1	16.0
	II	28.3	24.6	32.6
	III	27.4	24.0	31.1
	IV	19.2	17.4	21.2
	V	18.8	17.1	20.8
	VI	10.0	1.0	32.4
	VII	7.7	0.8	25.0
	VIII	19.3	9.6	33.7
	IX	18.3	9.8	29.6
Seymour Ref 11	I	2.8	2.6	3.0
	II	5.4	4.8	6.1
	III	5.3	4.7	6.0
	IV	3.6	3.4	3.8
	V	3.5	3.3	3.8
	VI	4.6	0.4	14.3
	VII	4.6	0.5	14.0
	VIII	4.4	2.2	7.3

Continued on next page

Table B.3 – *Continued from previous page*

Fossil	Model	Mean	Lower	Upper
Seymour Ref 11	IX	4.0	2.1	6.3
Seymour Ref 31	I	19.8	17.7	21.9
	II	38.6	33.4	44.5
	III	37.3	32.7	42.5
	IV	26.3	23.8	29.2
	V	25.8	23.3	28.6
	VI	11.1	1.1	38.0
	VII	8.2	1.0	24.6
	VIII	25.3	12.9	43.0
	IX	24.4	13.1	40.4

Appendix C

Chapter 4 Additional material

Landmark configuration

This section will describe the landmarks configuration for humerus and tarsometatarsus used in Chapter 4 (Fig. 4.2-4.3). The humerus will be oriented as in a "swimming" position for purposes of the description below. This follows the recent suggestion of Richards (2019), who considered that the vertical posture of penguins results in the terms "cranial" and "caudal" being applied to different faces of the flipper than when applied to the wings of a typical bird. Humerus curves were defined by a total of ten points including the fixed landmarks at the beginning and end of the curves. Likewise, tarsometatarsus curves one to nine were defined by a total of ten points, and tarsometatarsus curves ten to 16 were defined by 16 points, including the fixed landmarks at the start and end points.

Humerus

- **Fixed landmark 1**: Intersection point between the articular facet of the humeral head and the caudal margin of the scar of m. supracoracoideus.

- **Fixed landmark 2**: Intersection point between the articular facet of the humeral head and the secondary tricipital fossa caudal margin of the scar of m. supracoracoideus.

- **Fixed landmark 3**: Midpoint between previous and next landmark following the curve defined by the boundary between the incisura capitis and the articular facet of the humeral head.

- **Fixed landmark 4**: Point where the sulcus transversus connects with the capital incisure on the proximal side of the articular facet of the humeral head side. Given that in some penguins these two features do not connect in this case the point is the caudal most point of the sulcus transversus at the intersection of the articular facet of the humeral head side.

- **Fixed landmark 5**: Midpoint between previous and next landmark following the curve defined by the boundary between sulcus transversus and the articular facet of the humeral head.

- **Fixed landmark 6**: Cranial-most point at margin between sulcus transversus and the articular facet of the articular facet of the humeral head.

- **Fixed landmark 7**: Intersection point between the articular facet of the humeral head and the cranial margin of the scar of m. supracoracoideus.

- **Fixed landmark 8**: Apex of articular facet of the humeral head seen in caudal view.

- **Fixed landmark 9**: Distal-most point of the margin of the scar of m. supracoracoideus.

- **Fixed landmark 10**: Proximal-most point of the margin of the m. pectoralis fossa.

- **Fixed landmark 11**: Distal-most point of the margin of the m. pectoralis fossa.

- **Fixed landmark 12**: Base point of the posterior trochlear ridge facing the proximal part of the bone.

- **Fixed landmark 13**: Apex of the curve defined by the trochlear ridge seen in dorsal view.

- **Fixed landmark 14**: Base of the sulcus for m. scapulotricipitalis seen in distal view.

- **Fixed landmark 15**: Apex of articular facet of the dorsal condyle seen in ventral view.

- **Fixed landmark 16**: Apex of articular facet of the ventral condyle seen in ventral view.

- **Curve 1**: Curve defined by the margin of the fossa pneumotricipitalis. The starting point is defined by the distal-most point of the margin of the scar of m. scapulohumeralis. Curve end point is the distal-most point of the margin of the scar of m. coracobrachialis.

- **Curve 2**: Curve defined by the caudal margin of the humeral shaft. The starting point is defined by the distal-most point of the margin of the scar of m. latissimus dorsi. Curve end point is the base of the anterior trochlear ridge furrow seen in ventral view.

- **Curve 3**: Curve defined by the anterior margin of the humeral shaft. The starting point is defined by the apex of the tuberculum dorsale. Curve end point is the boundary articular facet of the dorsal condyle seen in anterior view.

Tarsometatarsus

- **Curve 1**: Curve defined by the margin of the medial cotyle. The starting point is defined by the dorsal projection on the rim of the lateral-most point of the cotyle. Curve end point is the plantar projection on the rim of the lateral-most point of the cotyle.

- **Curve 2**: Curve defined by the margin of the lateral cotyle. The starting point is defined by the plantar projection on the rim of the medial-most point of the cotyle. Curve end point is the apex of the eminentia intercotylaris.

- **Curve 3**: Curve defined by the medial margin of metatarsal II shaft. The starting point is defined by the base of the medial margin of the shaft at the boundary with the m. abductor digiti II. Curve end point is the proximal margin of the depression on the medial surface of the second trochlea.

- **Curve 4**: Curve defined by the cranial margin of metatarsal II shaft. The starting point is defined by the base of the cranial margin of the shaft at the boundary with the retinaculum extensorium tarsometatarsi. Curve end point is the base of the sulcus between the second and third trochleae.

- **Curve 5**: Curve defined by the cranial margin of metatarsal III shaft. The starting point is defined by the distal margin of the tuberositas m. tibialis cranialis. Curve end point is the proximal margin of the articulatory facet of trochlea III.

- **Curve 6**: Curve defined by the cranial margin of metatarsal IV shaft. The starting point is defined by the distal margin of the impression of ligamentum lateralis collateralis on the cranial side. Curve end point is the proximal margin of the articulatory facet of trochlea IV.

- **Curve 7**: Curve defined by the lateral margin of metatarsal IV shaft. The starting point is defined by the base of the medial margin of the shaft at the boundary with the m. abductor digiti IV. Curve end point is the proximal margin of the depression on the lateral surface of the fourth trochlea.

- **Curve 8**: Curve defined by the crista hypotarsi lateralis. The starting point is defined by the proximal base of crista hypotarsi lateralis. Curve end point is the proximal end of the foramen vascularis proximalis lateralis. Curve follows the margin given by the bony crest.

- **Curve 9**: Curve defined by the plantar depression between III and IV metatarsals. The starting point is defined by the distal end of the foramen vascularis proximalis lateralis. Curve end point is sulcus between third and fourth trochleae. Curve follows the depression between the metatarsals.

- **Curve 10**: Curve defined by the crista hypotarsus medialis. The starting point is defined by the proximal base of crista hypotarsi medialis. Curve end point is the proximal end of the foramen vascularis proximalis medialis at the distal base of the bony crest.

- **Curve 11**: Curve defined by the medial margin of the articular facet of trochlea II. On trochlea II the starting point is defined by the proximal end of the medial margin on the dorsal face. Curve end point is the proximal end of the medial margin on the plantar face.

- **Curve 12**: Curve defined by the lateral margin of the articular facet of trochlea II. On trochlea II the starting point is defined by the proximal end of the lateral margin on the dorsal face. Curve end point is the proximal end of the lateral margin on the plantar face.

- **Curve 13**: Curve defined by the medial margin of the articular facet of trochlea III. On trochlea III the starting point is defined by the proximal end of the medial margin on the dorsal face. Curve end point is the proximal end of the medial margin on the plantar face.

- **Curve 14**: Curve defined by the lateral margin of the articular facet of trochlea III. On trochlea III the starting point is defined by the proximal end of the lateral margin on the dorsal face. Curve end point is the proximal end of the lateral margin on the plantar face.

- **Curve 15**: Curve defined by the medial margin of the articular facet of trochlea IV. On trochlea IV the starting point is defined by the proximal

end of the medial margin on the dorsal face. Curve end point is the proximal end of the medial margin on the plantar face.

- **Curve 16**: Curve defined by the lateral margin of the articular facet of trochlea IV. On trochlea IV the starting point is defined by the proximal end of the lateral margin on the dorsal face. Curve end point is the proximal end of the lateral margin on the plantar face.

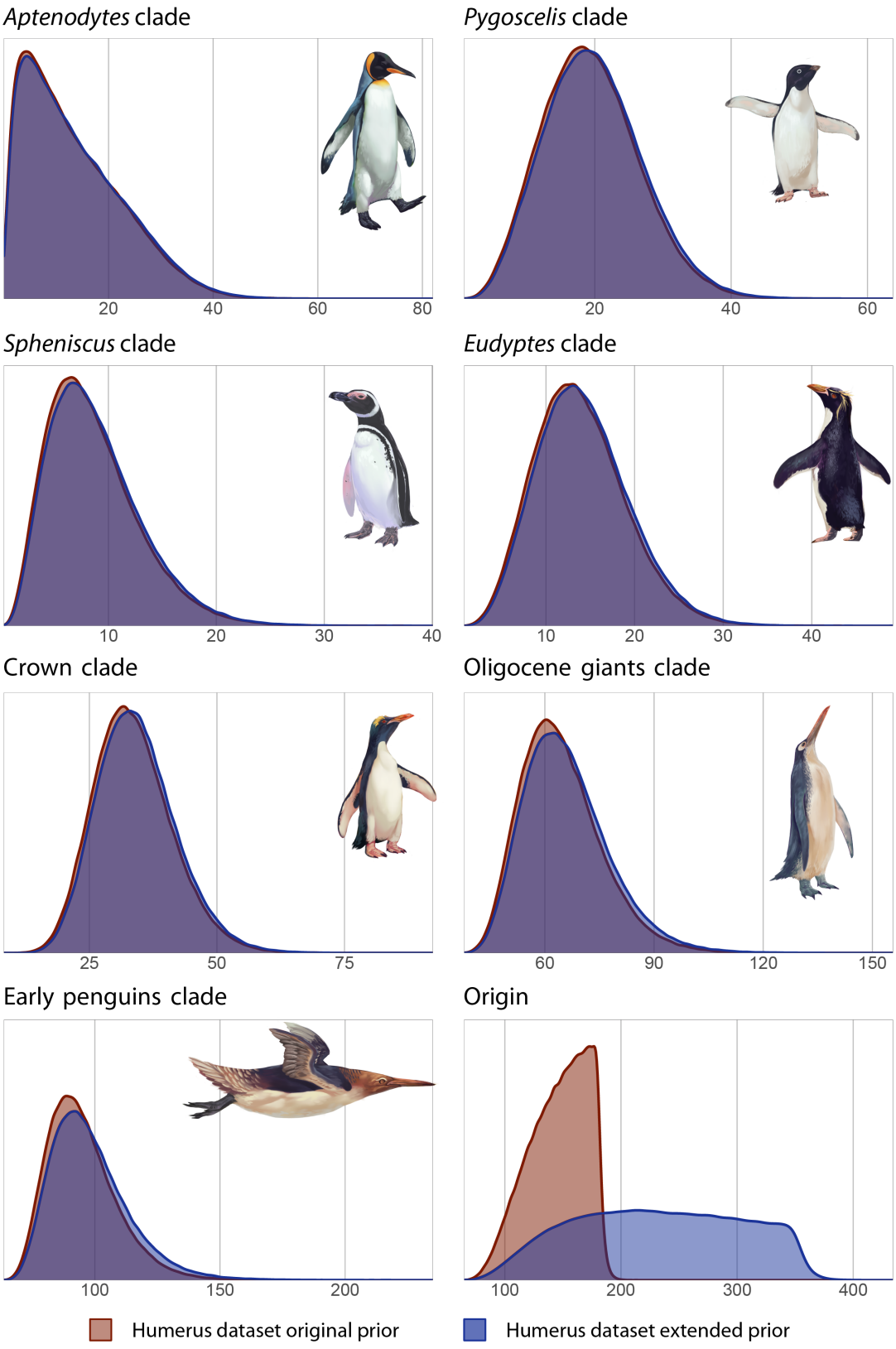


Figure C.1: Distributions of ages of selected clades in the morphological clock analyses performed on the humerus dataset with 61 to 180 Ma root origin prior (red) and tarsometatarsus 61 to 350 Ma (blue). Plots represent densities resulting from the Markov Chain Monte Carlo independent rates analyses. Numbers on the horizontal axes represent time in millions of years.Note that except from the root all clade origins rest unchanged.