PeAPA | Publicación Electrónica
ASOCIACIÓN PALEONTOLÓGICA ARGENTINA

www.peapaleontologica.org.ar

Body mass prediction in biped dinosaurs using supervised learning

ALEJANDRO OTERO[1,*]
AILEN BORYA[2]

1. CONICET - División Paleontología de Vertebrados (Anexo Laboratorios), Facultad de Ciencias Naturales y Museo (UNLP). La Plata, Argentina.
2. Especialización en Geociencias de Exploración y Desarrollo de Hidrocarburos, Facultad de Ciencias Naturales y Museo (UNLP). La Plata, Argentina.
*Corresponding author

Recibido: 23 de abril de 2026 - **Aceptado:** 09 de noviembre de 2026 - **Publicado:** 24 de noviembre de 2026

Para citar este artículo: Alejandro Otero, & Ailen Borya (2026). Body mass prediction in biped dinosaurs using supervised learning. *Publicación Electrónica de la Asociación Paleontológica Argentina* 26(3): 44–86.

Link a este artículo: http://dx.doi.org/10.5710/PEAPA.09.11.2026.576

©**2026** Otero, Borya

ISSN 2469-0228

Asociación Paleontológica Argentina
Maipú 645 1° piso, C1006ACG, Buenos Aires
República Argentina
Tel/Fax (54-11) 4326-7563
Web: www.apaleontologica.org.ar

Acceso Abierto
Open Access

This work is licensed under
CC BY 4.0

ARTÍCULO Año 2026 - 26(3): 44–86 PeAPA Publicación Electrónica ASOCIACIÓN PALEONTOLÓGICA ARGENTINA

BODY MASS PREDICTION IN BIPED DINOSAURS USING SUPERVISED LEARNING

ALEJANDRO OTERO[1,*] AND AILEN BORYA[2]

[1]CONICET - División Paleontología de Vertebrados (Anexo Laboratorios), Facultad de Ciencias Naturales y Museo (UNLP). La Plata, Argentina. *alexandros.otero@gmail.com*
[2]Especialización en Geociencias de Exploración y Desarrollo de Hidrocarburos, Facultad de Ciencias Naturales y Museo (UNLP). La Plata, Argentina. *ailenborya@gmail.com*
*Corresponding author

AO: https://orcid.org/0000-0002-4766-7086; **AB:** https://orcid.org/0000-0001-9992-2512

Abstract. As machine learning (ML) becomes increasingly prevalent in paleontological research, establishing structured, transparent, and reproducible analytical workflows is essential to ensure scientific rigor, validate predictive performance, and extract biologically meaningful insights from fossil data. Here we present a machine learning workflow for body mass prediction, illustrated through a case study on bird-line theropods and other biped dinosaurs using femoral length and circumference as predictors. We implement and compare two supervised regression algorithms (random forests and neural networks) across different training dataset configurations. We demonstrate that both algorithms achieve high predictive accuracy when trained on datasets preserving broad morphological and body mass variance, with femoral circumference consistently emerging as the dominant predictor across all configurations. However, algorithmic behavior diverges sharply under morphospace restriction. Feature importance analysis reveals that predictive collapse arises from training distribution truncation rather than algorithmic instability. Furthermore, benchmarking ML models against traditional allometric equations highlights the necessity of using independent or hybrid volumetric targets when femoral measurements are used as predictors, breaking formulaic tautologies while capturing non-linear interactions across giant size regimes. Unsupervised clustering independently recovers structure aligned with body size gradients and taxonomic composition. Beyond paleobiological results, we provide explicit guidelines for hyperparameter optimization, learning curve diagnostics, interpretability, and reproducibility, advocating for machine learning as a rigorous, scalable complement to inference-based methods in paleontological research.

Key words. Body mass. Theropod dinosaurs. Machine learning. Random forests. Neural networks. Feature importance. Hyperparameters.

Resumen. PREDICCIÓN DE MASA CORPORAL EN DINOSAURIOS BÍPEDOS A TRAVÉS DE APRENDIZAJE SUPERVISADO. A medida que avanza la adopción del aprendizaje automático (*machine learning* o ML) en la investigación paleontológica, establecer flujos de trabajo analíticos estructurados, transparentes y reproducibles es esencial para garantizar el rigor científico, validar el rendimiento predictivo y extraer conocimientos biológicamente significativos a partir de los datos fósiles. Presentamos un flujo de trabajo de aprendizaje automático para la predicción de la masa corporal, a través de un caso de estudio en dinosaurios bípedos, utilizando longitud y circunferencia femoral como predictores. Comparamos dos algoritmos de regresión (bosque aleatorio y redes neuronales) a través de diferentes configuraciones de datos de entrenamiento. Ambos algoritmos logran una alta precisión predictiva cuando se entrenan con datos que preservan una amplia varianza en sus variables, siendo la circunferencia femoral el predictor dominante en todas las configuraciones. Sin embargo, el comportamiento algorítmico diverge bruscamente bajo la restricción del morfoespacio. La importancia de variables revela que el colapso predictivo surge del truncamiento de la distribución de entrenamiento en lugar de una inestabilidad algorítmica. Además, si se comparan los modelos de ML frente a las ecuaciones alométricas tradicionales utilizando los mismos predictores, se recomienda utilizar estimaciones volumétricas independientes o híbridas en los datos de entrenamiento. El agrupamiento no supervisado revela una estructura alineada con los gradientes de tamaño corporal y la composición taxonómica. Se presentan pautas explícitas para la optimización de hiperparámetros, diagnósticos de curvas de aprendizaje, interpretabilidad y reproducibilidad, resaltando al aprendizaje automático como un complemento riguroso y escalable para los métodos basados en inferencia en la investigación paleontológica.

Palabras clave. Masa corporal. Dinosaurios terópodos. Aprendizaje automático. Bosque aleatorio. Redes neuronales. Importancia de variables. Hiperparámetros.

INTRODUCTION

In earth sciences, two practices are central to scientific research: inference and prediction. Inference involves concluding data to understand the mechanisms and processes that generate observed phenomena (*e.g.*, Rudwick, 1964; Radinsky, 1987; Vizcaíno et al., 2016). It emphasizes uncovering patterns, testing hypotheses, and establishing cause/effect relationships. Prediction, by contrast, focuses on forecasting future or unobserved outcomes, often prioritizing accuracy even when the

©2026 Otero, Borya | CC BY 4.0 | Acceso Abierto - Open Access

underlying mechanisms are not fully understood (Wu & Vos, 2018; Tredennick et al., 2021). The primary goal of inference is explanation and generalization, whereas the goal of prediction is reliable forecasting (Shmueli, 2010; Bzdok et al., 2018). These two practices complement one another. Inference provides insight into the data-generating process, allowing scientists to generalize from samples to populations and ground models in theoretical understanding (*e.g.*, Campione, 2017; Campione & Evans, 2020). Prediction, meanwhile, evaluates whether models perform well under new or unseen conditions, serving as a practical test of their robustness (*e.g.*, Raschka & Mirjalili, 2019). Accurate predictions can validate inferred mechanisms, while unexpected predictive failures may challenge existing theories and motivate further investigation. Thus, inference and prediction form a feedback loop in scientific progress.

A key distinction between inference and prediction lies in the timeframe and purpose. Inference is primarily retrospective or explanatory, concerned with understanding past or current phenomena. For example, Whitlock (2011) provided inferences on feeding ecology of diplodocoid sauropods based on observable evidence. In this regard, by analyzing snout shape and tooth wear patterns, the author infers causal relationships between anatomical traits and feeding behaviors (*e.g.*, square snouts and fine scratches associated with non-selective ground-height browsing). The study further extends these findings to broader paleoecological contexts, connecting inferred feeding strategies to specific environments. Rather than attempting to predict new or unknown values for additional specimens, the focus of Whitlock's study is on understanding the mechanisms underlying sauropod ecology and their role in Mesozoic ecosystems, which is the essence of scientific inference.

Prediction is prospective, aimed at anticipating future states or outcomes. In paleontology, where both understanding processes (*e.g.*, macroevolutionary processes, extinction dynamics, fossilization processes) and anticipating events (*e.g.*, fossil occurrence, taxon diversity) are vital, inference and prediction are not mutually exclusive but interdependent. Campione et al. (2014) developed and tested mathematical models specifically designed to improve body mass prediction accuracy for extinct bipeds (unseen data) based on measurable variables (femoral circumference). Campione et al.'s model performance is quantitatively evaluated using validation against extant birds, highlighting prediction errors and comparing them to other equations. The ultimate application is forward-looking: predicting the unknown body masses of extinct dinosaurs using the derived equation, with the goal of producing reliable estimates that can be generalized to other taxa.

Traditional statistical approaches have historically emphasized inference. Researchers specify relationships using pre-defined models and rules, such as regression, time series, or hypothesis-testing frameworks, to test hypotheses and quantify uncertainty (*e.g.*, through confidence intervals or *p*-values; Sokal & Rohlf, 1987; Zelditch et al., 2012; Casella & Berger, 2024). In this context, the researcher programs the rules (*i.e.*, pre-defined logical pathways, physical assumptions, or mathematical structures that the researcher imposes on the data), and the data are processed according to those rules to generate outputs (Chollet, 2018; Ketkar & Moolayil, 2021) (Fig. 1). This approach represents the 'data modeling culture', in which emphasis is placed on describing the stochastic mechanisms presumed to generate the observations (Breiman, 2001a).

Quantitative research in paleontology has traditionally relied heavily on inference-based statistical methods, such as regression analyses and analysis of variance (ANOVA), to uncover relationships in relatively small and structured datasets (*e.g.*, Miller, 1949; MacTavish et al., 1968; Cassini et al., 2012; Benson et al., 2014; Mannion et al., 2015; Vizcaíno et al., 2016; Hammer & Harper, 2024). With a focus on explanation, hypothesis testing, and building mechanistic models of natural systems, these methodological approaches were foundational for establishing theoretical understanding regarding, for example, environmental reconstructions (Lamboy & Lesnikowska, 1988; Telford & Birks, 2011; Liu et al., 2020), paleoecology of fossil assemblages (Sepkoski, 1978, 1979, 1984; Monchot & Léchelle, 2002), and evolutionary processes (Simpson, 1944; Raup & Sepkoski, 1982; Benton, 1995; Benson et al., 2014; Voje et al., 2023), among others. The explanatory and hypothesis-driven nature of these methods remains indispensable for building mechanistic understanding of natural systems.

Over the last decade, however, advances in computational power, data availability, and algorithmic development have fostered the rapid expansion of artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), across the earth sciences (Yu et al., 2024; Borya et al., 2025; Marques et al., 2025a; Silvestro & Pimiento, 2025; Yaqoob et al., 2025). Unlike traditional statistical approaches, machine learning methods (in which deep learning is included) are primarily prediction-oriented (Bzdok et al., 2018; Rajula et al., 2020). Rather than relying on explicitly programmed rules, machine learning models are trained on labeled or unlabeled examples and learn statistical structure (*e.g.*, hidden patterns, correlations, non-linear relationships, and groupings that naturally exist within a dataset) directly from the data (Chollet, 2018; Ketkar & Moolayil, 2021; Ballell et al., 2022; Hartmann et al., 2026) (Fig. 1). This approach corresponds to what Breiman (2001a) described as the 'algorithmic modeling culture', in which predictive performance and generalization to unseen data are central objectives.

In geosciences, especially in hydrocarbon exploration and reservoir characterization, AI methods have been widely adopted to optimize decision-making, reduce costs, and automate interpretation workflows (Maxwell et al., 2018; Karpatne et al., 2019; Kong et al., 2019; Pires de Lima & Marfurt, 2020; Chen et al., 2023; Soulaimani et al., 2024; Borya et al., 2025). Bibliometric analyses suggest that, although AI applications in paleontology are increasing, the field still lags behind other geoscience sectors in methodological integration (Sadeqi-Arani & Kadkhodaie, 2023; Yu et al., 2024). This disparity likely reflects both the exploratory nature of paleobiological research and the historical emphasis on inference rather than prediction.

This historical underemphasis on prediction is rooted in fundamental theoretical and philosophical constraints, chief among them the uniformitarian assumption. Predictive modeling in deep time inherently relies on the premise that the biological, biomechanical, and ecological processes governing relationships today operated in the same way in the past. Because extrapolating models calibrated on living animals to extinct taxa carries unavoidable biological uncertainty, many researchers naturally prefer inference-based approaches that describe observable data rather than predict unknown states, a challenge particularly evident in body mass estimation (Campione, 2017; Campione & Evans, 2020).

Nevertheless, the uptake of machine learning in paleontology has accelerated in recent years (Dillon et al., 2023; Yu et al., 2024; Marques et al., 2025a; Yaqoob et al., 2025). Applications have focused predominantly on classification and pattern-recognition tasks, including fossil identification (Lallensack et al., 2022; Ha & Kim, 2023; Jones et al., 2024; Esperança Júnior et al., 2026; Hartmann et al., 2026; Wills & Poon, in press), dental morphology (Wills et al., 2021; Ballell et al., 2022; Hendrickx et al., 2023; Barucci et al., 2024; Bahn et al., 2025; Marques et al., 2025b, 2026a), automated taxonomic recognition (Pires de Lima et al., 2020; Fan et al., 2022; Wang et al., 2022; Liu et al., 2023a; Forcellati et al., 2026), taphonomic (Cifuentes-Alcobendas & Domínguez-Rodrigo, 2019), behavioral (Millar, 2019) and mass extinction patterns (Foster et al., 2022), as well as fossil segmentation (Yu et al., 2022; Knutsen & Konovalov, 2024). These studies demonstrate that machine learning can help mitigate long-standing challenges in paleontology, including incomplete preservation, high-dimensional morphological data, and observer bias.

AI frameworks should not be viewed as a replacement for inference-based statistics, but rather as a complementary approach. In paleontology, where datasets are often heterogeneous, partially incomplete, and influenced by deep-time biases, prediction-oriented approaches can offer practical advantages, provided that they are applied with methodological rigor and interpretative transparency. Thus, machine learning methods possess the potential to occupy a methodological middle ground between inference and prediction: they offer robust predictive power while remaining acutely sensitive to the complex, non-linear structure of the underlying data. However, it must be explicitly acknowledged that machine learning does not resolve the uniformitarian problem. Because these learned relationships are still fundamentally grounded in the extant data used to train the models, the exact same theoretical assumptions apply when extrapolating these patterns to fossil taxa (Campione, 2017; Campione & Evans, 2020).

Despite this growth in the paleontological field, many researchers remain unfamiliar with machine learning prin-

ciples. In addition, over 60% of the published studies do not share reproducible research, with just 37% making their code available and 56,5% providing public access to their data and most of the studies lack any pre-processing explicit methodology (for a comprehensive bibliometric study see Marques et al., 2025a).

In line with this, two components are particularly critical for responsible machine learning practice: hyperparameter optimization and model interpretability. Hyperparameters are configuration settings chosen before training a machine learning model. Hyperparameter optimization refers to the systematic selection of algorithm settings that control model complexity and learning behavior (Yang & Shami, 2020; Elgeldawi et al., 2021). Proper tuning improves predictive performance while balancing bias and variance, thereby reducing the risk of overfitting. Equally important, *post-hoc* interpretability methods, such as feature importance analyses, enable researchers to identify which predictors drive model outputs (Casalicchio et al., 2019; Saarela & Jauhiainen, 2021).

In scientific research, where interpretability is not optional but fundamental, a predictive model that delivers high accuracy without providing insight into the drivers of its predictions risks functioning as a 'black box', offering numerical performance without contributing to explanatory understanding. Recent advances in interpretable machine learning stress that transparency, explainability, and alignment with domain knowledge are essential requirements in research contexts (Murdoch et al., 2019; Rudin, 2019; Confalonieri et al., 2021; Ning et al., 2022). Rather than viewing interpretability as secondary to performance, these contributions argue that explanatory clarity enhances scientific validity, fosters reproducibility, and allows predictive modeling to meaningfully inform theory-building.

In disciplines such as paleontology, where the ultimate goal is to infer biological processes, evolutionary mechanisms, and functional relationships, predictive accuracy alone is insufficient. Models must also be interpretable, allowing researchers to understand the specific anatomical and statistical rules driving their predictions. Despite this need to explain exactly why a model predicts as it does, techniques such as feature importance analysis and hyperparameter optimization remain scarcely implemented in the field, with a few recent exceptions (Foster et al., 2022; Liu et al., 2023a; Barucci et al., 2024; Jones et al., 2024).

The primary aim of this contribution is to present a

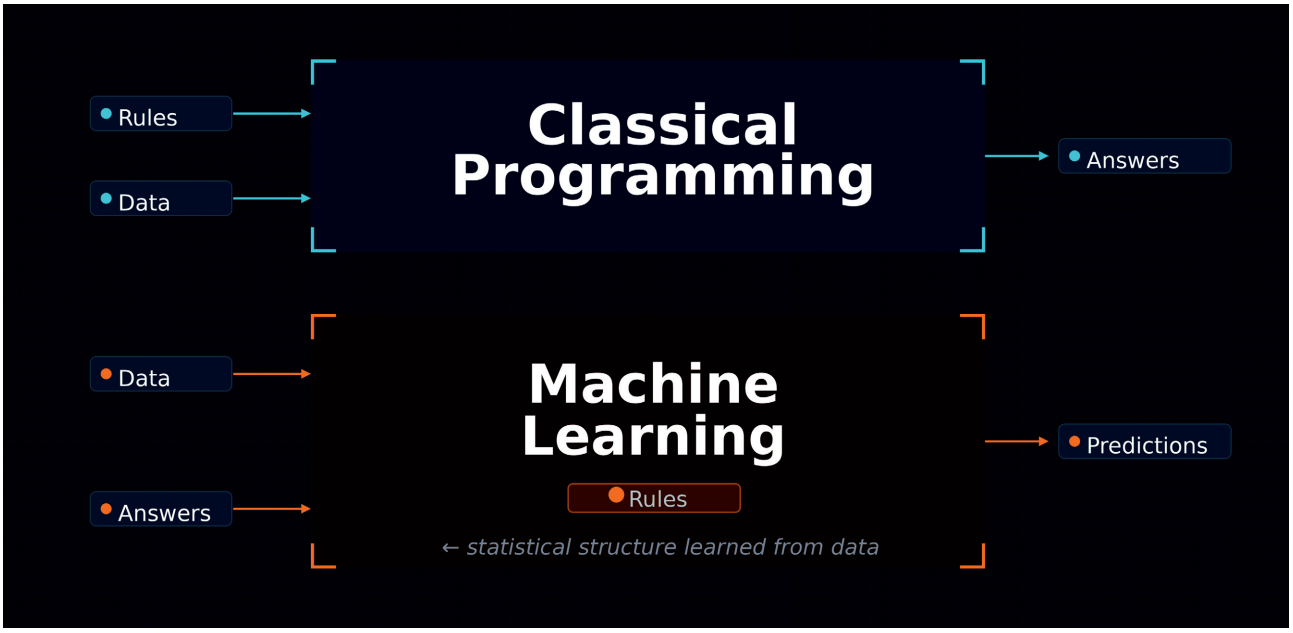

Figure 1. Classical programming versus machine learning. Systems designed in classical programming are fed by pre-established rules and data, retrieving answers as output. In contrast, supervised machine learning systems need data and answers (= target labels) as input; thus, the system is able to decipher the structure behind those to retrieve the rules that govern such relationships, providing prediction as output. Modified from Chollet (2018).

structured, transparent, and reproducible workflow for conducting machine learning analyses in paleontology (for an intuition of standard data science workflows see Müller & Guido, 2017; Chollet, 2018; Kelleher & Tierney, 2018, figs. 4, 5). The workflow proposed herein (Fig. 2) is intended not merely as a technical demonstration, or as a descriptive guide (see Sun et al., 2022; Marques et al., 2025a) but as a methodological guide using a practical and reproducible example. Rather than focusing solely on algorithmic performance, we explicitly detail each stage of the data science pipeline, from data curation and preprocessing to model selection, hyperparameter optimization, validation, and *post-hoc* interpretability, emphasizing methodological decisions that influence robustness and generalization.

Our workflow is illustrated through a practical case study on body mass prediction in bird-line theropod dinosaurs, using supervised regression algorithms. Body mass is one of the most critical biological traits influencing the physiology (*e.g.*, Bakker, 1972; LaBarbera, 1989; Grabowski et al., 2005), functional morphology (*e.g.*, Christiansen, 1999; Carrano, 2005; Otero et al., 2019), and ecology (*e.g.*, Farlow, 1976; McClain & Boyer, 2009) of

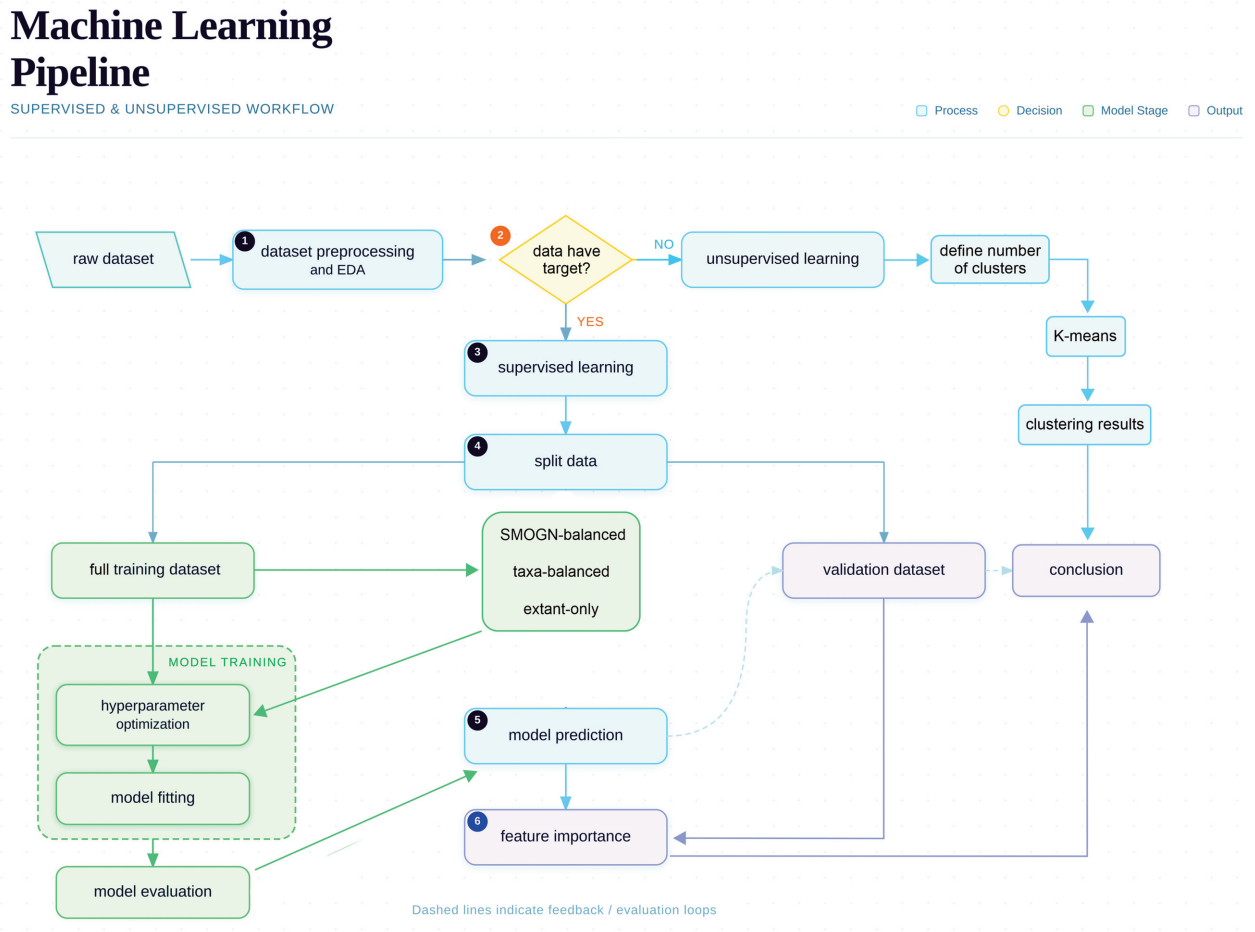

Figure 2. Workflow implemented in this contribution. Data preprocessing and exploratory data analysis (EDA) (1) are the first task of the workflow, in which the dataset is curated and standardized via preprocessing, whereas EDA helps with visualization of the relationships among variables. The decision of whether choosing an unsupervised or supervised algorithm (2) depends on the presence of a target variable or not. In supervised learning (3) the dataset is split (4) into training and validation subsets. For each algorithm, four training instances were implemented: full training dataset, SMOGN dataset, taxa-balanced dataset and extant-only dataset. Model training includes the hyperparameter optimization and model fitting, after which the model is evaluated through performance metrics. Predictions (5) and feature importance assessment (6) are made on the testing and validation datasets. On the other hand, unsupervised learning implemented in this contribution starts with the identification of the most suitable number of clusters (via silhouette and elbow methods), followed by the k-means clustering algorithm. 'Conclusions' box acts as a standard conceptual endpoint, indicating that every analytical branch ultimately requires a concluding synthesis step.

animals, both extant and extinct. In paleontology, where direct observation is impossible, estimating body mass plays a central role in reconstructing life histories, ecological niches, and evolutionary processes (Burness et al., 2001; Dahl et al., 2010; Harrison et al., 2010; Sookias et al., 2012a, b; Brown et al., 2013, 2022; Benson et al., 2014, 2018; Dempsey et al., 2025). In contrast to the widespread application of classification methods in paleontology, regression-based machine learning approaches remain comparatively underexplored (Marques et al., 2025a). Although machine learning approaches for body mass prediction have been applied to humans and farm animals (Alonso et al., 2015; Fujihara et al., 2023; Liu et al., 2023b), this is the first contribution attempted on fossil tetrapods. To complement the supervised analyses, we also implement unsupervised clustering techniques to investigate intrinsic structure in the dataset independent of the target variable.

Our broader goal is to help make machine learning both accessible and methodologically rigorous within paleontology. Although the case study centers on a specific paleobiological problem, the framework proposed herein is broadly applicable to quantitative questions across paleontology and related earth sciences. At each stage, we advocate current best practices designed to maximize transparency, reproducibility, and biological interpretability, thereby promoting a responsible and scientifically grounded integration of machine learning into paleontological research. By presenting a complete and explainable workflow, we seek to bridge the gap between traditional inference-driven approaches and modern predictive modeling, fostering applications that not only achieve strong predictive performance but also contribute meaningfully to paleobiological understanding.

The AI matryoshka

Artificial intelligence is not a method itself, but more like an umbrella that refers to computational systems designed to perform tasks or processes that resemble human perceptual, cognitive and decision-making functions, such as classification and prediction, or the creation of completely new information (Müller & Guido, 2017; Corea, 2019; Xu et al., 2021; Yu et al., 2024; Silvestro & Pimiento, 2025).

Machine learning, a core branch of artificial intelligence (Fig. 3), focuses on algorithms that can extract statistical patterns from data and use them to make predictions or decisions without explicit programming or rules (Fig. 1). Within it, algorithms are commonly grouped into two major categories: supervised and unsupervised learning. Semi-supervised and reinforcement learning are not covered here (for a review in paleontology see Yu et al., 2024).

In unsupervised learning, the goal is to uncover hidden structures in datasets that contain only predictor variables, but no labels or target values. The algorithm organizes the data into clusters or groups (*e.g.*, K-means, hierarchical clustering, autoencoders) according to similarity criteria, without external guidance (Chollet, 2018; Prince, 2023). Unsupervised learning also encompasses dimensionality reduction, a technique for simplifying complex, high-dimensional data by reducing the number of features or variables while preserving the data's most important patterns and relationships (*e.g.*, principal component analysis, PCA; autoencoders; uniform manifold approximation and projection, UMAP) (Aggarwal, 2018; Hartmann et al., 2026; Marques et al., 2026a).

In contrast, supervised learning is aimed at prediction, where datasets include both predictors and a target variable. The target can be categorical (classification) or continuous (regression). Typically, most of the dataset is used to train the model, while the remainder is reserved for testing. If the trained model can deliver accurate predictions on unseen data, it is considered robust and capable of generalization (Raschka & Mirjalili, 2019). Supervised methods have proven particularly valuable in fields like paleontology, where data are often incomplete or fragmentary (*e.g.*, Apostol et al., 2016; Wills et al., 2021, 2023; da Conceição et al., 2023; Ha & Kim, 2023).

Building on machine learning, deep learning represents a more advanced subset based mostly on artificial neural networks (ANN or simply NN) with many layers (Fig. 3). Deep learning models are especially powerful when dealing with complex, high-dimensional, or unstructured data (such as images, 3D models, or large fossil datasets), where they excel at automatically extracting hierarchical features that may be challenging to define manually (Nielsen, 2015; Keçeli et al., 2017; Zhong et al., 2017; Lallensack et al., 2022; Yu et al., 2024).

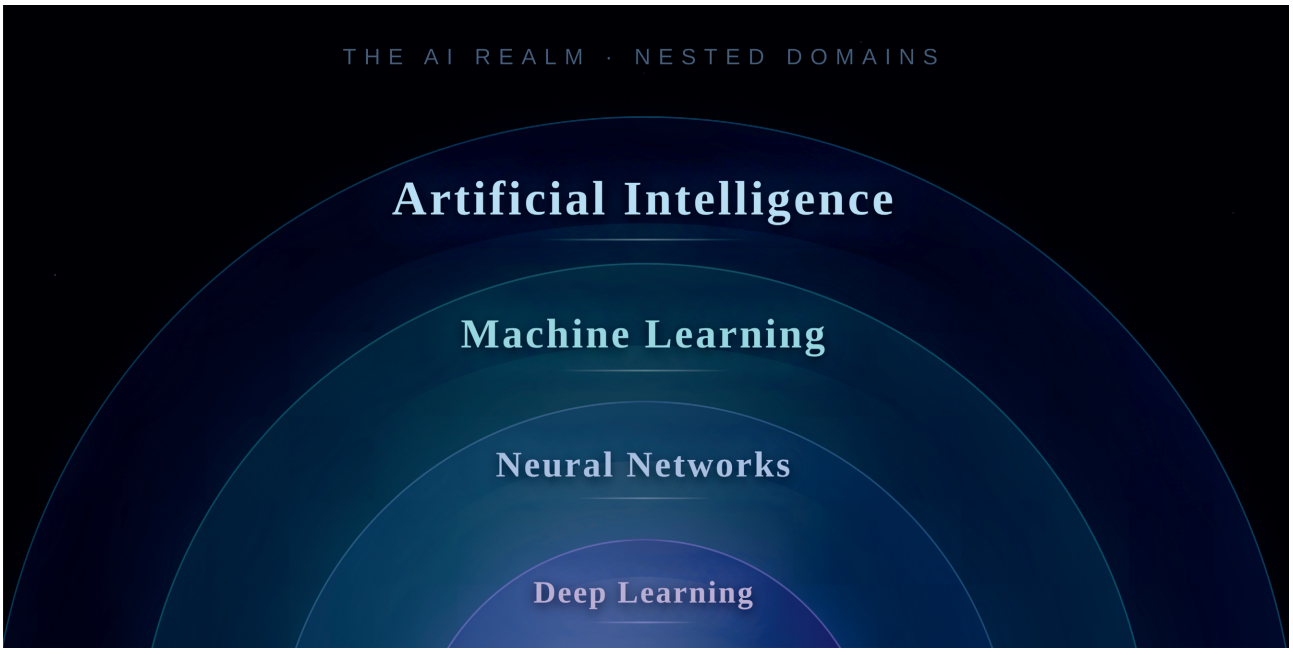

Figure 3. The artificial intelligence realm. AI includes systems that are able to automate intellectual tasks. Within it, machine learning captures algorithms that 'learn' from data. Neural networks go beyond through the inclusion of systems that automatically update the learning parameters. Finally, deep learning adds sophistication by multiple internal layers in the network. Adapted from Chollet (2018).

A fundamental (although not the only) distinction between less complex machine learning algorithms and neural networks lies in their training procedures. Training, fitting, and learning are all synonyms of the process of finding parameters that minimize the error. Machine learning algorithms (*e.g.*, decision trees, random forests, naïve Bayes, support vector machines, among others) learn the statistical structure of a training dataset by modeling the relationships between input features and the target variable, and subsequently generate predictions on unseen data. Model outputs are then evaluated against observed values using error (*e.g.*, mean absolute or squared error) and performance metrics (*e.g.*, R^2, accuracy, recall). These evaluations guide adjustments to the model's parameters by the researcher, or expanding the training dataset, after which the model is retrained to improve accuracy and generalization (Aggarwal, 2018; Chollet, 2018).

In contrast, neural networks initialize their parameters (weights) with arbitrary values at the start of training and generate predictions through a feedforward process (Fig. 4). These predictions are then compared with the true values to compute an error, or loss. Using this loss, the model automatically adjusts its parameters through backpropagation, which updates weights to minimize the loss function. The updated parameters are then passed through another feedforward step, and this cycle of prediction, evaluation, and adjustment is repeated until the loss converges toward a minimum. At this stage, the neural networks have achieved convergence. A complete pass of feedforward prediction followed by backpropagation-based adjustment across the training data is called an 'epoch' (Aggarwal, 2018; Chollet, 2018; Ketkar & Moolayil, 2021; Buduma et al., 2022; Prince, 2023). This process can be compared to tuning a musical instrument: the network starts with arbitrary 'notes' (initial parameters), produces a 'sound' (prediction), and checks how far off it is from the correct pitch (loss). Backpropagation acts like the musician adjusting the 'strings' (parameters) until the 'sound' aligns with the intended 'note'. Repeating this trial-and-error tuning gradually produces a well-adjusted, 'in-tune' network capable of accurate predictions.

The choice between implementing a machine learning algorithm and a deep learning approach using neural networks depends on several factors, including the research question, the type and size of the dataset, and the computational resources available. As deep learning models

require substantially higher computational resources and longer training times than traditional ML algorithms, the use of DL is not always justified: if the problem can be adequately addressed with a simpler, faster, and less resource-intensive ML method, then adopting such an approach is preferable.

Neural networks are particularly well-suited to medium-to-large datasets (thousands to millions of observations), as their performance typically improves with greater data availability. Although paleontological datasets are often limited in size, including extant analogs can substantially increase sample numbers (Deng et al., 2009; Hsiang et al., 2019), making neural networks training feasible and potentially more informative than simpler machine learning algorithms, which often reach a performance plateau once variance has stabilized. Furthermore, neural networks offer considerable flexibility through multiple hidden layers, diverse activation functions, and regularization strategies, allowing them to be tailored to specific data structures (Ketkar & Moolayil, 2021; Prince, 2023).

MATERIALS AND METHODS

Dataset

The original dataset is composed of 1121 specimens, including 1024 living and 97 extinct theropods already published (Field et al., 2013, tab. S1; Benson et al., 2014,

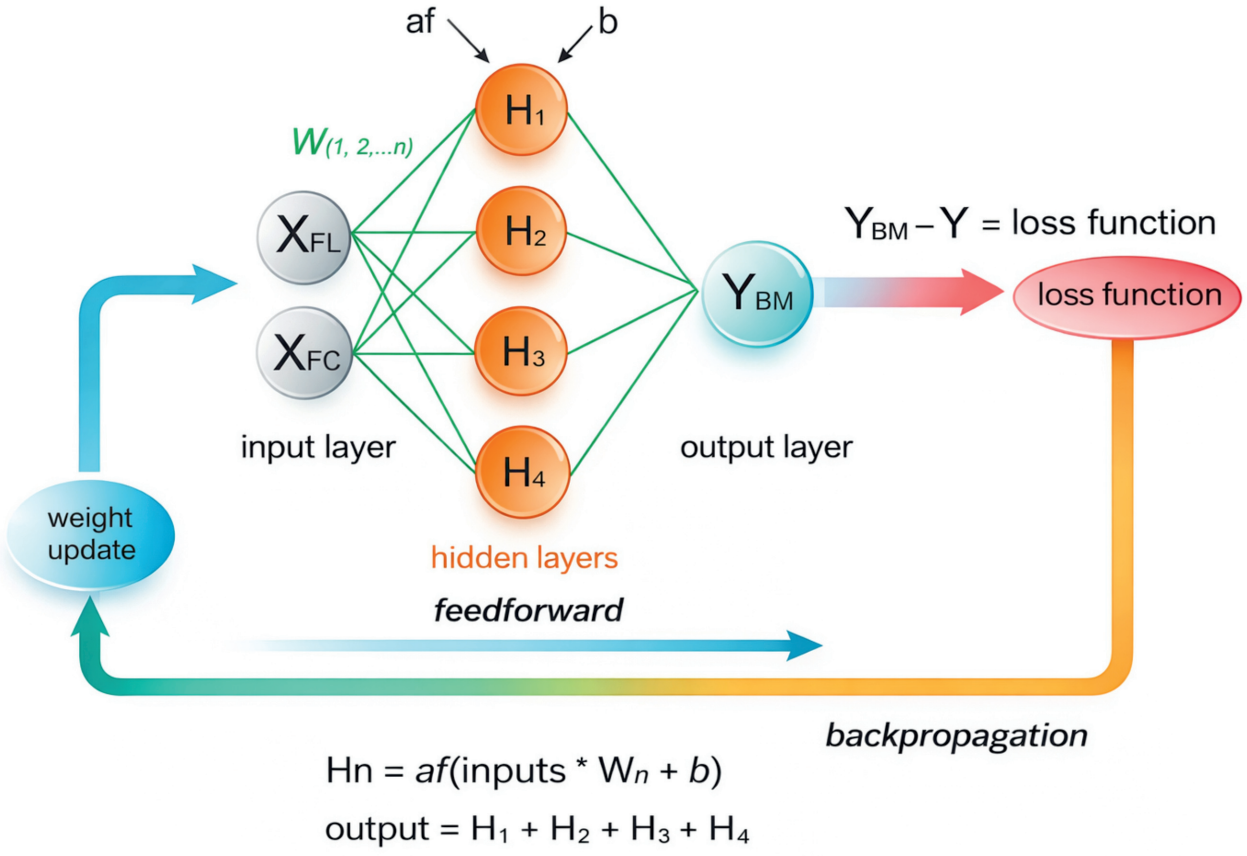

Figure 4. Basic architecture of a neural network with one hidden layer for simplification. The input layer (denoted by X) includes the established values for the variables in the dataset (in this case, FL = femoral length and FC = femoral circumference). The hidden layer(s) as well as the number of neurons per layer (denoted by H) are set by the researcher, depending on how complex the model is intended to be. The weight (W) and bias (b) are internal parameters that are automatically set by the algorithms. The activation function (af) is chosen by the researcher, depending on the structure of the dataset and the problem to be solved (*e.g.*, classification versus regression, image versus video versus text processing, etc.). The output layer (single or multiple, denoted by Y) presents the results (as BM = body mass), either as absolute values, or a probability, depending on the research problem. The feedforward culminates with a loss value, which is the difference between the predicted (Y_{BM}) versus the actual (Y) value. The backpropagation pass updates the weights in the direction of minimizing the loss and the cycle (=epoch) starts over. Iteration culminates when the loss reaches a minimum. Although the mathematical framework is outside the scope of this contribution, it is worth noting that such a minimum is reached when the derivative of the loss reaches a slope of zero.

dataset S1; Campione et al., 2014, data S1). The current dataset presents 429 species of living forms distributed across 24 orders. The maximum number of specimens per species is 10. Extinct forms are distributed across 88 species of non-avian Theropoda and represent 8,6% of the sample, with a maximum of three specimens per species.

The dataset includes linear measurements of limbs, body mass and taxonomic data. Here, we take femoral length (FL) and minimum femoral midshaft circumference (FC) as predictor variables and body mass (BM) as target variable. These variables are present in all selected specimens, with no missing values. More predictor variables could have been included, but it would have reduced drastically the final number of specimens included due to the presence of missing data in the extinct forms. Femur measurements have been proved useful for predicting body mass in bipeds (*e.g.*, Field et al., 2013; Campione et al., 2014; Serrano et al., 2015). For living specimens, body mass was directly taken (see Field et al., 2013; Campione et al., 2014 for details). For extinct forms, body mass was calculated via a universal scaling relationship formula by Campione et al. (2014; see Benson et al., 2014; corrected quadrupedal equation, cQE). The original dataset used in this contribution is presented in Table S1. In addition, we also incorporated a training set using non-avian theropods with body mass estimated using other methods than scaling formula (*i.e.*, volumetric) to compare results with the former (see below and Discussion).

Data preprocessing and exploratory data analysis

The preprocessing stage involves scaling, balancing the data and quality control for each variable, including checks for null and missing values, verification of correct data types and confirmation of appropriate measurement units (Fig. 2, stage 1). Data preprocessing is an essential step in data science, which converts the raw data into a high-quality, structured format that machine learning algorithms can effectively understand and learn from, improving model accuracy (Müller & Guido, 2017). Exploratory data analysis (EDA) is then performed to evaluate variable balance, distributions and interrelationships, identifying general patterns, detecting anomalies, understanding relationships between variables, and uncovering hidden insights using visual and statistical methods (Chatfield, 1986; Sahoo et al., 2019; Mukhiya & Ahmed, 2020; Singh et al., 2022).

Linear correlation between predictors (*i.e.*, collinearity) can negatively affect the model's ability to generalize, increasing the risk of overfitting (Garg & Tai, 2013; Chan et al., 2022; Kyriazos & Poga, 2023). A commonly used statistical approach to detect multicollinearity is the Variance Inflation Factor (VIF), which quantifies how much the variance of an estimated regression coefficient is increased due to collinearity. However, VIF is derived from linear regression models and relies on the R^2 value to assess predictor relationships. Given that the machine learning algorithms employed in this study (see below) are non-linear, the applicability of VIF in such contexts has been questioned. Specifically, Drobnič et al. (2020) highlighted the limitations of using VIF to assess multicollinearity in non-linear models, suggesting it may not adequately reflect their internal structure or behavior.

Scaling encompasses a broad class of preprocessing techniques that aim to render input features more comparable in magnitude, thereby facilitating learning and improving generalization to unseen data. Two commonly applied approaches are normalization and standardization. Normalization rescales variables to a fixed range (typically between 0 and 1) by subtracting the minimum value and dividing by the range of the feature (implemented as 'MinMaxScaler' in the Scikit-learn library). In contrast, standardization transforms features by subtracting the mean and scaling by the standard deviation, resulting in variables with zero mean and unit variance ('StandardScaler' in Scikit-learn; Müller & Guido, 2017).

The dataset utilized here encompasses body masses spanning several orders of magnitude, from small extant birds to multi-ton extinct dinosaurs. If the models were trained on raw arithmetic values, standard loss functions (*e.g.*, mean squared error) would be overwhelmingly dominated by the absolute errors associated with the largest taxa. Under such conditions, an algorithm would heavily penalize minor percentage errors in giant animals while virtually ignoring massive relative errors in smaller taxa. Log-transforming the variables ensures that the algorithms optimize for proportional (relative) error across the entire size spectrum, preventing giant taxa from unduly biasing

the models' predictive weights. Hence, prior to model training, all morphological variables and body mass targets were log-transformed. While certain tree-based algorithms, such as random forests, are natively scale-invariant and do not strictly require feature scaling, neural networks are highly sensitive to input scales (Chollet, 2018; Yaqoob et al., 2025). Because neural networks update their internal weights via gradient descent, unstandardized inputs spanning vastly different magnitudes can cause severe gradient instability and prevent model convergence. Although neural networks can, in principle, adapt to heterogeneous feature scales, learning is substantially more efficient and reliable when inputs are appropriately rescaled (Yaqoob et al., 2025). Therefore, standardization was applied to neural network models.

In addition to input-level scaling, batch normalization was evaluated as an internal normalization strategy within the network architecture. Batch normalization is implemented as a dedicated layer ('BatchNormalization' in Keras library) that adaptively normalizes activations during training by maintaining an exponential moving average of batch-wise means and variances. This procedure reduces internal covariate shift, stabilizes gradient propagation, and can improve training efficiency, often allowing for faster convergence and improved numerical stability (Ioffe & Szegedy, 2015; Chollet, 2018; Ketkar & Moolayil, 2021; Ozer et al., 2023a).

In this study, neural network models were trained using predictor variables that were both log-transformed and subsequently standardized across most regularized iterations (see below). Although the implemented network architectures are relatively shallow, one additional iteration incorporating batch normalization was performed for comparative purposes. In this configuration, input variables were log-transformed but not externally standardized, as normalization was handled internally by the batch normalization layers.

Python v. 3.12.11 (Python Software Foundation, http://www.python.org) libraries used for data preprocessing are Pandas v. 2.3.2 (McKinney, 2010), NumPy v. 2.0.1 (Harris et al., 2020) and Scikit-learn v. 1.7.1 (Pedregosa et al., 2011), whereas visualization was performed using Matplotlib v. 3.10 (Hunter, 2007), Seaborn v. 0.13.2 (Waskom, 2021), and Plotly v. 7.1.0 (Plotly Technologies Inc., Montréal, https://plot.ly).

Dataset modifications and resampling strategy. The original raw dataset (Tab. S1) is imbalanced both taxonomically and in the distribution of the target variable (body mass), spanning a wide and uneven range of values (see next section for quantitative assessment). In regression settings, imbalance in the response variable may lead to biased model behavior, particularly when extreme or high-magnitude values are underrepresented (Branco et al., 2016; Krawczyk, 2016). Models trained on such data may exhibit reduced predictive accuracy in sparsely sampled regions of the response distribution and may disproportionately favor densely populated ranges (Johnson & Khoshgoftaar, 2019).

To assess the robustness and generalization capacity of the models under different data configurations, we implemented a series of controlled modifications to the full dataset. These modifications were designed to evaluate the sensitivity of model performance to taxonomic representation, response-variable distribution, and potential label uncertainty.

First, we addressed imbalance in the body mass values distribution using the SMOGN algorithm (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise; Branco et al., 2017), implemented in the Python package SMOGN v. 0.1.2, an adaptation of SMOTE (Synthetic Minority Over-sampling Technique; Chawla et al., 2002; for paleontological examples see also Wills et al., 2021; Marques et al., 2025b) for regression problems. SMOGN combines under-sampling of overrepresented regions with synthetic generation of samples in underrepresented ranges of the response variable (see Supplementary Information). This approach allows a more homogeneous coverage of body mass values across the full spectrum. In practice, this ensures the model learns from a balanced representation of all size categories, preventing predictive bias toward the most abundant body sizes, while preserving the multivariate structure of the dataset (Tab. S2). Second, to test the model's sensitivity to phylogenetic bias, we evaluated the performance by explicitly balancing the 'State' column, distinguishing between extant and extinct forms. These two groups represent the main body plans within the dataset and differ not only phylogenetically but also in the certainty of body mass estimates. This resampling strategy enables assessment of whether model performance is dispropor-

tionately influenced by the dominant taxonomic component (Tab. S3: 85 extant; 85 extinct). Finally, given that body mass in extinct taxa is inferred rather than directly measured, we tested model performance after excluding extinct forms from the dataset through a leave-one-clade-out approach (LOCO; see Tahir et al., 2025 for an analog study) by excluding all non-avian theropods from the training phase. This modification allows evaluation of the extent to which inferred labels may introduce noise or bias into model training and provides a comparative framework for assessing predictive stability under reduced uncertainty (Tab. S4). Together, these dataset iterations allow a systematic evaluation of model robustness under realistic paleobiological constraints, rather than relying on a single data configuration. See Figure 5 for a summary of datasets configurations implemented.

It should be noted that because data resampling procedures (such as SMOGN and taxonomic balancing) were applied prior to the internal 75:25 partitioning, minor information leakage could potentially occur across internal training and testing folds; however, the globally withheld 10% validation set (Tab. S6) remained strictly protected and was never exposed to any preprocessing, resampling, feature selection, or tuning steps.

Finally, to evaluate the predictive accuracy and generalizability of supervised machine learning algorithms relative to traditional parametric scaling equations, we benchmarked our models against the corrected quadrupedal equation (cQE) of Campione et al. (2014, tab. 3) across a sample of 30 bipedal dinosaurs (acting as a validation or holdout sample). We conducted this comparative evaluation across three distinct taxa-balanced training paradigms to isolate the impact of statistical circularity, target leakage, and training set composition on deep-time body mass inference: (1) a cQE-derived fossil training setup in which target body masses for extinct taxa were calculated directly from cQE formulas (Tab. S3: 85 extant; 85 extinct); (2) an independent non-circular training setup in which extinct target masses were derived from 3D volumetric-density (VD) reconstructions compiled by Campione & Evans (2020,

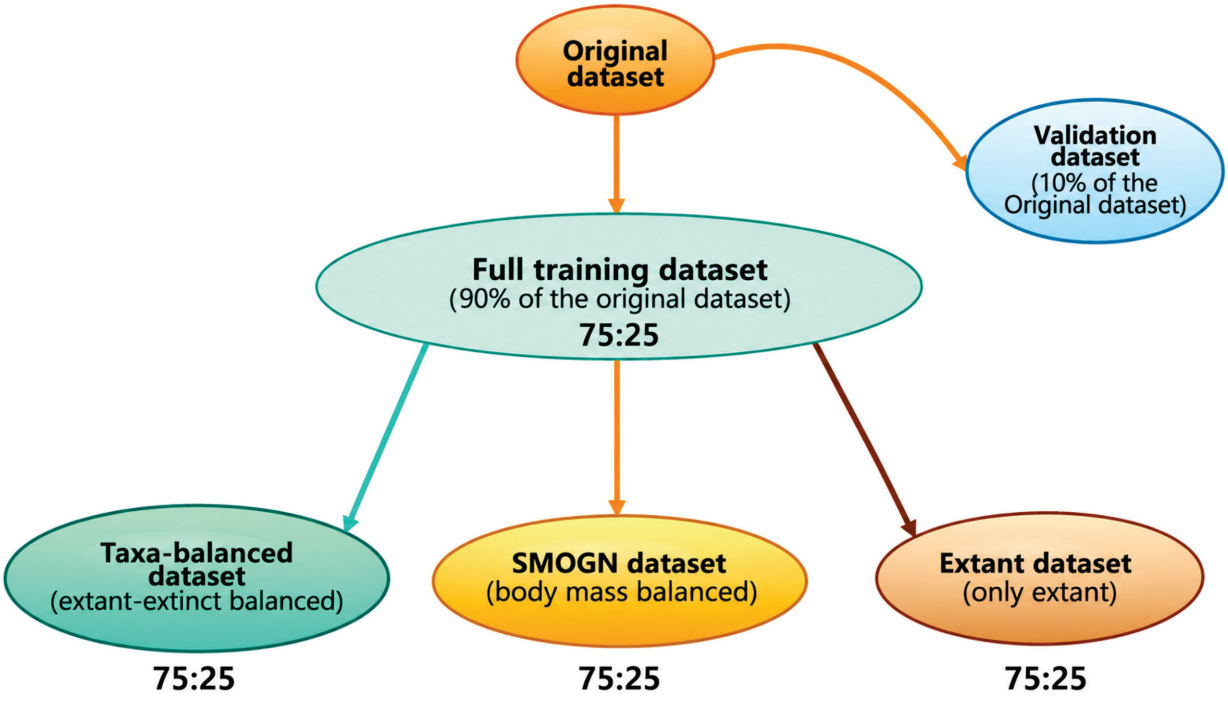

Figure 5. Dataset structure used in this contribution. The original raw dataset contains the whole samples before any balancing. From the latter, 10% of the samples are held-out for validation. The remaining dataset (the full training dataset) is the base to further rebalancing: taxa-balanced, SMOGN-balanced and extant-only. Each of these four datasets (full, taxa-balanced, SMOGN-balanced and extant-only) were split in a 75:25 (training:testing) ratio by the algorithm before model training.

app. S1) (Tab. S11: 85 extant; 85 extinct); and (3) a hybrid training set combining both cQE- and VD-derived targets to evaluate model behavior under mixed label generation regimes (Tab. S12: 85 extant; 170 extinct). Preprocessing, splitting and tuning configurations are the same that are implemented for our primary training datasets detailed above. Results are presented in Tables S7 for the validation set and S13 for the biped dinosaur estimation.

Exploratory data analysis (EDA). The study considers predictor variables that scale linearly (FL, FC) as well as a target variable that scales cubically (BM). Basic visualization among variables can reveal useful information (Fig. 2, stage 1). In this regard, we used boxplots to illustrate the spread and skewness of predictor and target variables through their quartiles, grouped by higher clade parameter (Fig. 6), whereas histograms were used to examine the distribution of each variable, allowing assessment of whether the data follow a normal or skewed distribution. Bar charts were used to display class proportions for higher clade and state, enabling the identification of class imbalances that may require resampling techniques (Fig. 7; Supplementary Information, Fig. S1) (Note: here and throughout the text, 'class' is applied as standard machine learning jargon to describe categorical groupings within the data, rather than the Linnaean taxonomic rank). Pairplots were used to show the relationships of predictors with body mass, highlighting the non-linear relationship of the raw data. After log-transformation, data present a more linear relationship, also smoothing the impact of extremely high body mass values and mitigating heteroscedasticity (Quinn & Keough, 2002; Serrano et al., 2015; Vizcaíno et al., 2016; Campione & Evans, 2020) (Fig. 8). A practical way to compare the data before and after log-transformation is using the 'describe' method in the Pandas library (McKinney, 2010), which generates descriptive statistics of the dataset, summarizing the central tendency, dispersion, and extreme values, among others (Supplementary Information, Tab. S5).

Model selection

Supervised learning. In machine learning, the selection of an appropriate algorithm is guided by the nature of the research question and the structure of the dataset (Fig. 2, stage 2). In this study, our objective was to predict body mass from linear anatomical measurements, which constitutes a supervised regression task (Fig. 2, stage 3). In this regard, we implemented and compared two complementary approaches: random forests (RF) and neural networks (NN).

It is important to consider that machine learning performance is not determined solely by sample size but by the relationship between model complexity and available data. Ensemble methods such as random forests are particularly robust in moderate sample-size settings due to bootstrap aggregation and implicit variance reduction (Breiman, 2001b; Biau & Scornet, 2016; Rigatti, 2017). Likewise, shallow neural networks with appropriate regularization and cross-validation can generalize effectively in tabular datasets below 2000 observations (Goodfellow et al., 2016; Ballell et al., 2022; Lallensack et al., 2022).

Random forests is an ensemble learning method that builds and aggregates multiple decision trees to enhance predictive accuracy and reduce overfitting (Dietterich, 2000; Breiman, 2001b), being largely insensitive to data normalization, making them well-suited for heterogeneous datasets typical of paleontological studies, where predictor variables and body masses often span vastly different numerical scales and magnitudes (*e.g.*, Keçeli et al., 2017; Ballell et al., 2022). In regression tasks, each tree is trained on a bootstrap dataset sample, and predictions are averaged across trees to produce the final output. This strategy reduces variance and bias, resulting in more stable and robust models than individual decision trees (*e.g.*, Breiman, 2001b; Madhav & HimanshuRawat, 2021). Although random forests for classification tasks have been applied consistently in paleontology during last years (*e.g.*, Keçeli et al., 2017; Ballell et al., 2022; Wills et al., 2023; Jones et al., 2024; Marques et al., 2025b), regression studies under random forests are less abundant (but see Otero, 2025); hence, this contribution constitutes an instance to comprehensively demonstrate how random forests can effectively handle a regression problem applied to paleontology.

We also implemented a neural networks model for algorithmic and performance comparisons (see Tab. 1 for algorithms setup). Unlike random forests, which aggregate

decision rules, neural networks rely on a fundamentally different learning paradigm: parameters (weights) are initialized and iteratively adjusted through backpropagation, where errors between predicted and observed values are used to update the network to minimize the loss function. This iterative process enables the model to progressively refine its internal representation of the relationships within the data (Aggarwal, 2018; Chollet, 2018).

The neural network was constructed using the Keras Sequential API (see below), a framework that allows models to be built as a linear stack of layers, from input to output. In our implementation, the input layer corresponds to the set of above-mentioned anatomical measurements (predictors), which were propagated through one or more hidden layers of densely connected neurons, each applying non-linear activation functions to capture complex dependencies. The final output layer consisted of a single neuron with a linear activation function, suitable for continuous regression tasks such as body mass estimation (see below).

By comparing neural networks with random forests in terms of setup complexity, error and performance, we can evaluate the relative strengths of these two approaches. Neural networks offer high flexibility in modeling intricate non-linearities and interactions, while random forests provide greater interpretability and robustness on smaller datasets. Including both methods allows for a balanced assessment of predictive performance, generalizability to unseen data, and potential trade-offs between accuracy and interpretability.

Three main instances of regression were trained for each dataset: two for random forests (RFdh: random forests with default hyperparameters; RFoh: random forests with optimized hyperparameters) and one for neural networks. For each instance, three iterations were performed, based on predictor combinations (*i.e.*, FC, FL and FC + FL). We started with both predictors (femoral length and minimum shaft circumference) and then fitted the model with one predictor at a time. Finally, these instances were tested in each selected dataset. The goal is to evaluate the combination of predictors that yield the best performance and lowest error metrics.

Random forests and neural networks algorithms used in this contribution were implemented using Scikit-learn v. 1.7.1 (https://scikit-learn.org/stable/) and Keras v. 3.11.2 (https://keras.io), respectively, both open-source tools. Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems (Pedregosa et al., 2011). Using the Scikit-learn interface, the method used for random forest regression is 'class sklearn.ensemble. RandomForestRegressor'. Keras is a deep learning API written in Python, whose core data structures are layers and models (Chollet, 2015). The scripts for each algorithm can be found in Supplementary Information as notebook Python files (.ipynb files) as well as in the GitHub repository (see Data archiving statement). In addition, a user-friendly body mass calculator for biped dinosaurs is provided (see Supplementary Information).

Unsupervised learning. In this study, we evaluated whether specimens can be classified into distinct groups based on morphological similarity using cluster analysis under femur length and femur circumference. To address this objective, we implemented K-means clustering using Scikit-learn (Fig. 2). K-means partitions observations into an optimal number of clusters (k) by minimizing within-cluster variance, thereby grouping specimens according to morphological similarity (Arthur & Vassilvitskii, 2007; Pedregosa et al., 2011). The optimal number of clusters was determined using the elbow and silhouette methods. The elbow method examines the total within-cluster sum of squares (WCSS) as a function of k, identifying the point at which increasing the number of clusters yields diminishing reductions in variance. Graphically, this point appears as an 'elbow' or inflection point in the WCSS curve (Supplementary Information, Fig. S2). The silhouette method assesses clustering quality by quantifying how similar each observation is to its assigned cluster compared to other clusters, resulting in a score from -1 to 1 (Shutaywi & Kachouie, 2021; Januzaj et al., 2023). Average silhouette widths greater than 0,7 are considered strong, values above 0,5 indicate reasonable structure, and values above 0,25 suggest weak clustering (Rousseeuw, 1987; Shutaywi & Kachouie, 2021).

The final K-means model was fitted with the following hyperparameters: k = 3, initialization method = *k-means++*,

TABLE 1. Computational settings combination for machine learning algorithms selected for this contribution. The average training time for neural networks iterations was 15 seconds. Training was performed in two different computers: 1) Intel®(R) Core (TM) i7-10750H CPU @ 2,60GHz (2,59 GHz), 32,0 GB RAM, NVIDIA® GeForce® GTX® 1660 Ti (6 GB); and 2) Intel®(R) Core (TM) i7-4790 CPU @ 3,60GHz (3,60 GHz), 32,0 GB RAM, AMD® Radeon® (TM) R7 370 Series (4 GB).

Algorithm	Settings	Default values	After optimization	After sensitivity analysis (testing)	
Random Forests	max depth*	None	None	10	
	bootstrap*	True	True	True	
	max features*	1	0,5	0,5	
	max samples*	None	0,5	0,5	
	min samples leaf*	1	1	3	
	min samples split*	2	2	6	
	n estimators*	100	300	300	
	random state	42	42	-	
	oob score	False	True	-	
	n jobs	None	-1	-	
Neural networks	**Settings**	**L2 regularization**	**Batch normalization**	**Dropout**	**Plain layers**
	hidden layers	2	2	2	2
	neurons	20	20	20	20
	activation function on hidden layers	relu	relu	relu	relu
	activation function on output layer	none	none	none	none
	learning rate	0,001	0,001	0,001	0,001
	regularization penalty	0,001	NA	NA	NA
	use bias	NA	True	NA	NA
	rate	NA	NA	0.1	NA
	seed	NA	NA	42	NA
	optimizer	Adam	Adam	Adam	Adam
	Total parameters	481	645	481	481
	epochs	100	100	100	100

NA= not applicable; * indicate hyperparameters (*i.e.*, those that affect how trees are built and therefore the fitted model). See main text and Supplementary Information for explanations.

and a maximum number of iterations set to the default value implemented in Scikit-learn. Model training was conducted using predictor variables that were treated differently prior to analysis: 1) log transformed and 2) standardized (*i.e.*, it centers and rescales each feature independently so that it has a mean of 0 and a standard deviation of 1). The resulting clusters were subsequently compared and visualized in relation to their corresponding body mass values (target variable), higher clade affiliation, and whether each cluster is predominantly composed of extinct or extant organisms (Fig. 9).

Model setup

Data split (Fig. 2, stage 4)**.** To evaluate model performance on unseen data, 10% of both extant and extinct specimens were randomly withheld from the raw original dataset (Tab. S1) and reserved as a validation set (Smith, 2002; Mendoza et al., 2006; Serrano et al., 2015; Otero, 2025) (Tab. S6). It is important to clarify that while some machine learning literature refers to this final holdout as a 'test set', we consistently refer to it as the 'validation set' throughout this manuscript, as defined in Scikit-learn documentation.

Random sampling across the full original dataset to generate the holdout/validation dataset was selected to ensure that both the validation and training partitions mirror the empirical joint distribution, range, and morphospace structure of the original sample, thereby preventing structural bias in model evaluation. By preserving the natural proportional representation of extant and extinct taxa across both partitions, this approach ensures that the primary holdout set provides an un-skewed benchmark of out-of-sample performance across the full size spectrum. Consequently, because holdout specimens were randomly sampled from the full dataset, some species may be represented in both training and holdout partitions; model performance on entirely unseen species missing from the training distribution may therefore exhibit greater error or variance than reported here. In spite of this limitation, randomly partitioning the full dataset ensures that the primary validation set maintains an unbiased representation of overall morphological variance and body mass spectrum. This baseline partition directly complements our non-random stress-tests (*e.g.*, extant-only dataset), which explicitly evaluate model performance under severe domain shift and morphospace restriction.

Predictions obtained on the validation set were generated in log-transformed space and subsequently back-transformed to arithmetic values. Because back-transformation can introduce bias, a correction factor was applied following the ratio estimator method of Snowdon (1991) (see validation results in Tab. S7). After setting aside the validation subset, the training dataset comprised 1009 specimens, including 924 extant and 85 extinct specimens (Tab. S8). Subsequent modifications on the training dataset include class balance (see above Dataset modifications and resampling strategy). Figure 5 shows the datasets structure used in this contribution.

Training and testing a model on the same samples introduces a risk of overfitting, where the model performs well on training data but fails to generalize to new observations (Dietterich, 1995; Ying, 2019). To mitigate this issue, datasets used for training the models (full, SMOGN-balanced, taxa-balanced, extant-only) were randomly partitioned under a 75:25 ratio using the 'train_test_split' function from Scikit-learn. The 'random_state' and 'keras.utils.set_random_seed' parameters were fixed to ensure reproducibility, as they set the seed for the pseudo-random number generator used to shuffle and split the data for Scikit-learn and Keras, respectively.

To avoid terminological confusion, it is important to explicitly distinguish between the training dataset configurations and the train-test split. The dataset configurations (*i.e.*, full, SMOGN-balanced, taxa-balanced, and extant-only) refer to data pre-processing strategies designed to alter the demographic composition and variance profile of the sample to address taxonomic imbalances. In contrast, the train-test split is a procedural partitioning step applied to each of these configurations (Fig. 5). While the configuration dictates what data the algorithm is exposed to, the 75:25 split dictates how that data is functionally divided: one portion (75%) is used strictly for algorithmic learning (*i.e.*, training), while the remainder (25%) is used for internal evaluation to tune the model and monitor for overfitting, independent of the globally withheld 10% validation set.

Hyperparameter optimization. Hyperparameter optimization automates the search for optimal model settings prior to training, reducing manual intervention and improving reproducibility (Fig. 2; Yang & Shami, 2020; Elgeldawi et al., 2021; Ballell et al., 2022; Foster et al., 2022; Linchamps et al., 2023; Liu et al., 2023a; see Supplementary Information for a complete description of the procedure). We utilized Grid Search, an exhaustive evaluation of user-specified combinations that is particularly well-suited for small datasets (Bergstra et al., 2011; Syarif et al., 2016; Wicaksono & Supianto, 2018; Liashchynskyi & Liashchynskyi, 2019; Elgeldawi et al., 2021; Madhav & HimanshuRawat, 2021; Ballell et al., 2022; Linchamps et al., 2023). Model evaluation was performed using 5-fold cross-validation

via Scikit-learn's 'GridSearchCV' to minimize underfitting and overfitting.

For the random forests models, we conducted a sensitivity analysis on key parameters (*e.g.*, 'max. depth', 'n. estimators'; see Supplementary Information) to identify configurations that minimize the mean squared error (MSE). The models were then re-trained using these refined configurations to ensure optimal stability (Tab. 1). Conversely, hyperparameter optimization for the neural networks was kept intentionally conservative. Extensive tuning of highly flexible NNs on limited datasets (~1000 observations) heavily increases the risk of overfitting (Aggarwal, 2018; Chollet, 2018). To ensure a fair, unbiased comparison against the inherently less-sensitive RF models, we constrained the NN to a parsimonious architecture, prioritizing robust generalization over artificially inflated training performance.

For comparison, models were also trained using default hyperparameters. Table 1 shows default values for random forests, as well as values after optimization and sensitivity analysis; in addition, parameters for neural networks are also presented therein. Training time and CPU/GPU features are also included in Table 1 description.

Regularization. Regularization techniques constrain model complexity during training to prevent overfitting, a scenario where a model memorizes noise rather than the underlying biological signal (Dietterich, 1995; Srivastava et al., 2014; Goodfellow et al., 2016; Aggarwal, 2018; Chollet, 2018; Ying, 2019; Ketkar & Moolayil, 2021). Because high model complexity relative to dataset size frequently exacerbates overfitting in paleontology (Cifuentes-Alcobendas & Domínguez-Rodrigo, 2019; Ballell et al., 2022; Marques et al., 2025a), regularization is essential for achieving accurate predictions on unseen data (see Supplementary Information for a complete description of the procedure).

To stabilize our neural network models, we implemented two established strategies: L2 regularization (weight decay) and Dropout. The former adds a penalty for excessively large network weights, promoting smoother solutions and reducing variance while preserving the ability to learn nonlinear relationships (Goodfellow et al., 2016; Aggarwal, 2018; Chollet, 2018; Hsiang et al., 2019; Ying, 2019; Ketkar & Moolayil, 2021; Ha & Kim, 2023). Since the extant-only dataset does not have extreme variance, L2 regularization was excluded from this dataset. Dropout, on the other hand, introduces stochasticity by randomly deactivating a fraction of neurons during training. It prevents the network from over-relying on specific pathways, effectively mimicking an ensemble approach to improve robustness and generalization (Srivastava et al., 2014; Goodfellow et al., 2016; Aggarwal, 2018; Cifuentes-Alcobendas & Domínguez-Rodrigo, 2019; Ying, 2019; Ozer et al., 2023b).

Model evaluation

Metrics. Model performance is evaluated here via traditional statistical modeling and machine learning regression tasks as well (Fig. 2, stage 5), such as the coefficient of determination (R^2) and the percent prediction error (PPE) (*e.g.*, Field et al., 2013; Campione et al., 2014; Serrano et al., 2015). PPE is calculated as follows:

$$PPE = (|\text{true body mass} - \text{estimated body mass}| \,/\, \text{estimated body mass}) \times 100$$

PPE is especially useful when comparing models across datasets with different scales or units, although care must be taken with near-zero target values, which can inflate error estimates (for application in biological and paleontological data see Field et al., 2013; Campione et al., 2014; Serrano et al., 2015; Campione & Evans, 2020). In addition, we evaluate the models' feature importance, which quantifies the contribution of individual features to a model's predictions, offering crucial insight into how models make decisions, regardless of whether the relationships between features and outcomes are linear or non-linear (see above; Casalicchio et al., 2019; Saarela & Jauhiainen, 2021).

To assess learning dynamics, convergence behavior, and potential overfitting in neural networks, we compared the training and testing error trajectories through learning curves (Pires de Lima et al., 2020; Marques et al., 2026b), in which epochs are displayed on the x-axis and error (mean squared error, MSE) on the y-axis.

Feature importance analysis. To ensure comparability between random forest and neural network models, feature importance was assessed using model-agnostic approaches that can be applied consistently across both algorithmic

families (Fig. 2, stage 6). Among these, SHAP (SHapley Additive exPlanations) represents one of the most robust and widely adopted techniques. Grounded in cooperative game theory, SHAP provides consistent and locally accurate explanations by assigning each predictor a contribution value relative to a baseline, typically defined as the model's expected (mean) prediction. Unlike simple importance metrics (*e.g.*, Gini importance in random forests), SHAP accounts for feature interactions and conditional dependencies, offering a consistent, additive explanation framework (Lundberg & Lee, 2017; see below).

SHAP provides essential insight into how predictions are generated, regardless of whether the underlying relationships between predictors and responses are linear or non-linear (Casalicchio et al., 2019; Saarela & Jauhiainen, 2021). It is worth noting that while some machine learning algorithms do not inherently require data linearization (for example, random forests), the log-transformation of our dataset remains critical. Log-space stabilizes the variance across variables that span several orders of magnitude, ensuring the algorithms minimize proportional rather than absolute error and preventing giant taxa from dominating the training phase. In the context of biological prediction, such as estimating body mass from morphological proxies, feature importance serves a dual purpose: it supports model transparency while also enabling evaluation of the biological relevance of individual anatomical variables (*e.g.*, skeletal dimensions) (Foster et al., 2022; Otero, 2025).

By design, SHAP values are additive: the sum of all feature contributions equals the difference between the predicted body mass and the baseline estimate (Ning et al., 2022; Song et al., 2024). This property allows direct interpretation of how individual morphological predictors increase or decrease predicted body mass, both for single specimens (local explanations) and across the full dataset (global importance). SHAP facilitates biological interpretability by quantifying not only the relative importance of predictors, but also the direction and magnitude of their effects. In addition, it enables assessment of whether highly ranked predictors align with established biomechanical or allometric expectations for body mass estimation, thereby strengthening the link between statistical performance and biological plausibility.

RESULTS

Exploratory data analysis

Femur length and circumference (predictor variables), along with body mass (target variable), exhibit mean values of 82,87 mm, 26,01 mm, and 6130 g, respectively. Both predictor and target variables show high standard deviations, approximately twice their respective means, indicating substantial variability within the dataset (Tab. S5). Additionally, several potential outliers are identified in both predictor and target variables (Fig. 6). For the predictors, those are primarily observed at values exceeding 100 mm for femur length and 40 mm for femur circumference. These extreme values are largely associated with Struthioniformes and non-avian theropods. In the case of body mass, potential outliers occur above 2500 g, and are associated with extant taxa such as Pelecaniformes, Phoenicopteriformes, and Struthioniformes, as well as non-avian theropods.

The distributions of all variables are positive-skewed, indicating a high frequency of relatively low body masses and a comparatively small number of extreme high values (Fig. 7; Supplementary Information, Fig. S1). The most frequent values of femur length, femur circumference, and body mass correspond to extant taxa belonging primarily to smaller-bodied clades that dominate the dataset. Within this group, Passeriformes, Charadriiformes, and Anseriformes each contribute more than one hundred records (Fig. 7). In contrast, the less frequent and higher values of these variables are mostly associated with extinct larger-bodied groups (non-avian theropods), as well as Struthioniformes. Together, these groups account for fewer than one hundred records, representing approximately 8,6% of the total records in the dataset (Fig. 7). These patterns indicate a pronounced group imbalance among higher clades and between extant and extinct taxa in both predictor and target variables.

Regarding the interrelationships among variables, the predictor variables exhibit a linear relationship with each other. On the other hand, the relationships between the predictor variables and the target variable (body mass) are non-linear in the raw data (Fig. 8). Following log-transformation, these relationships become linear, with both predictors and the target variable displaying clear linear trends (Fig. 8; Supplementary Information, Tab. S5).

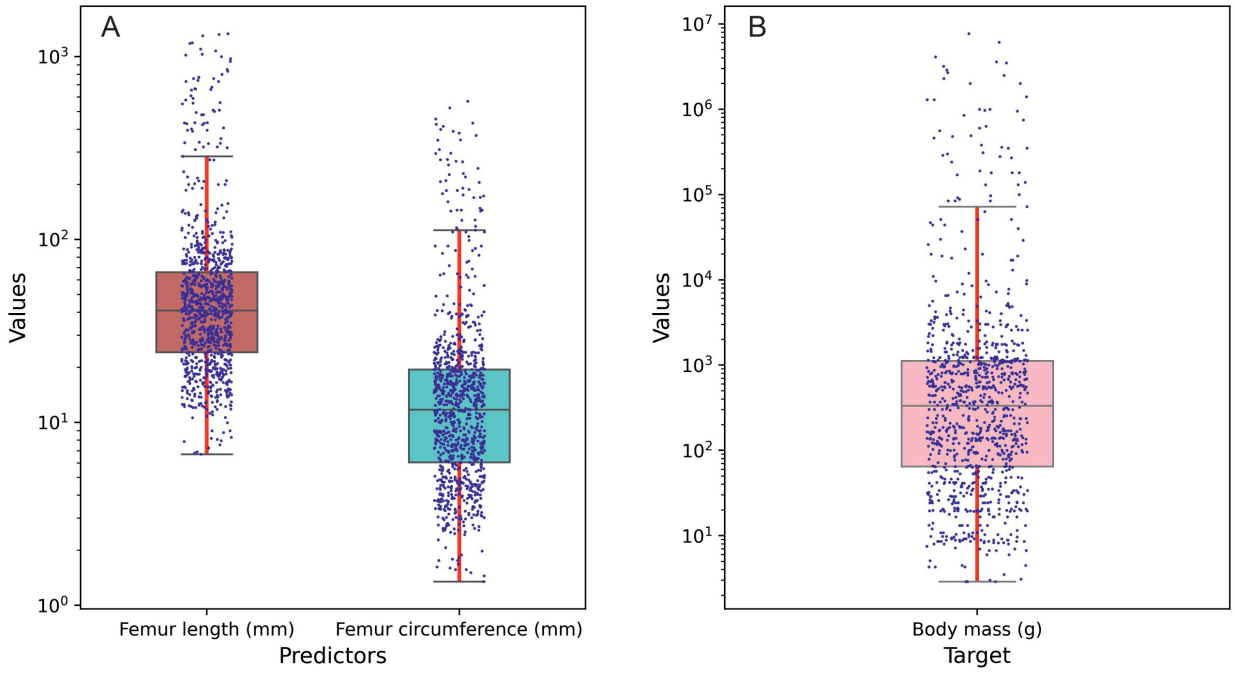

Figure 6. Boxplots depicting data statistics (based on Supplementary Information, Tab. S5, before log-transformation) of the full training-testing dataset. The bottom and top of the whiskers are the minimum values and those considered normal (1,5 x interquartile range), respectively, whereas the top and bottom margins of the box represent the upper boundary of quartile 3 (Q3) and the lower boundary of quartile 1 (Q1), respectively. The median value is denoted as the horizontal line inside the box. The interquartile range is the space between Q1 and Q3, containing the 50% of all occurrences. Dots placed outside min-max range are potential outliers. The y-axis is presented in log 10 scale for ease of visualization.

Unsupervised learning

The optimal number of clusters (k) was determined using both the elbow and silhouette methods. The silhouette analysis suggested a range between 2 and 5 clusters, with coefficients varying narrowly between 0,51 and 0,62 for the log-transformed space and between 0,61 and 0,91 for the standardized space. However, the elbow method indicated plausible solutions at k = 3 and k = 5 (Supplementary Information, Fig. S2). To further constrain the optimal solution, K-means models were fitted across the range k = 2–5. A final value of k = 3 was selected, supported by anatomical criteria and its consistency with patterns observed during exploratory data analysis (EDA). The number of specimens, along with the mean and standard deviation of femur length, femur circumference, and corresponding body mass for each cluster, are summarized in Supplementary Information (Tab. S9).

For the log-transformed space, most extant species are assigned to groups 1 and 2. Group 1 comprises 416 specimens and is characterized by mean femur length and circumference values of 21,97 mm and 5,42 mm, respectively, with standard deviations of 7,02 mm and 2,02 mm. This group corresponds to the lowest body mass values and includes mostly Passeriformes and Charadriiformes (Fig. 9A). Group 2 is the most represented cluster, containing 515 specimens and encompassing higher clades such as Anseriformes, Accipitriformes, and Ciconiiformes, among others (Fig. 9A). This group exhibits mean femur length and circumference values of 62,19 mm and 18,54 mm, respectively, with standard deviations of 24,31 mm and 7,69 mm. Both groups display relatively low variability, as indicated by standard deviation values lower than their respective means, suggesting moderate internal homogeneity. These results allow the identification of two distinct groups of extant species that were not clearly differentiated during EDA. Nevertheless, a subset of extant species is distributed across both groups 1 and 2 such as Charadriiformes and Procellariiformes (Fig. 9A).

Group 3, in contrast, is the least represented cluster, comprising 78 specimens, predominantly extinct species, and associated with the highest body mass values (Fig. 9A). It exhibits the largest mean femur length and circumference

(544,24 mm and 185,16 mm, respectively), along with substantial variability (standard deviations of 314,75 mm and 124,42 mm). This group is primarily composed of non-avian theropods, with a smaller representation of Struthioniformes. These patterns are consistent with those identified during EDA.

Clusterization after standardization shows a different cluster scheme for k=3, with most living forms nested in group 1, except for Struthioniformes, which also are included in group 2 (Fig. 9B). Non-avian theropods are distributed along the three groups. When scaling only to the raw data, the composition of the group 1 when k=2, 3 and 4 is very inclusive, covering a wide range of body masses (Supplementary Information, Fig. S3). Hence, caution should be taken when applying standardization to datasets with values along a continuous range and the presence of extreme values. In those cases, log-transformation might be a better approach.

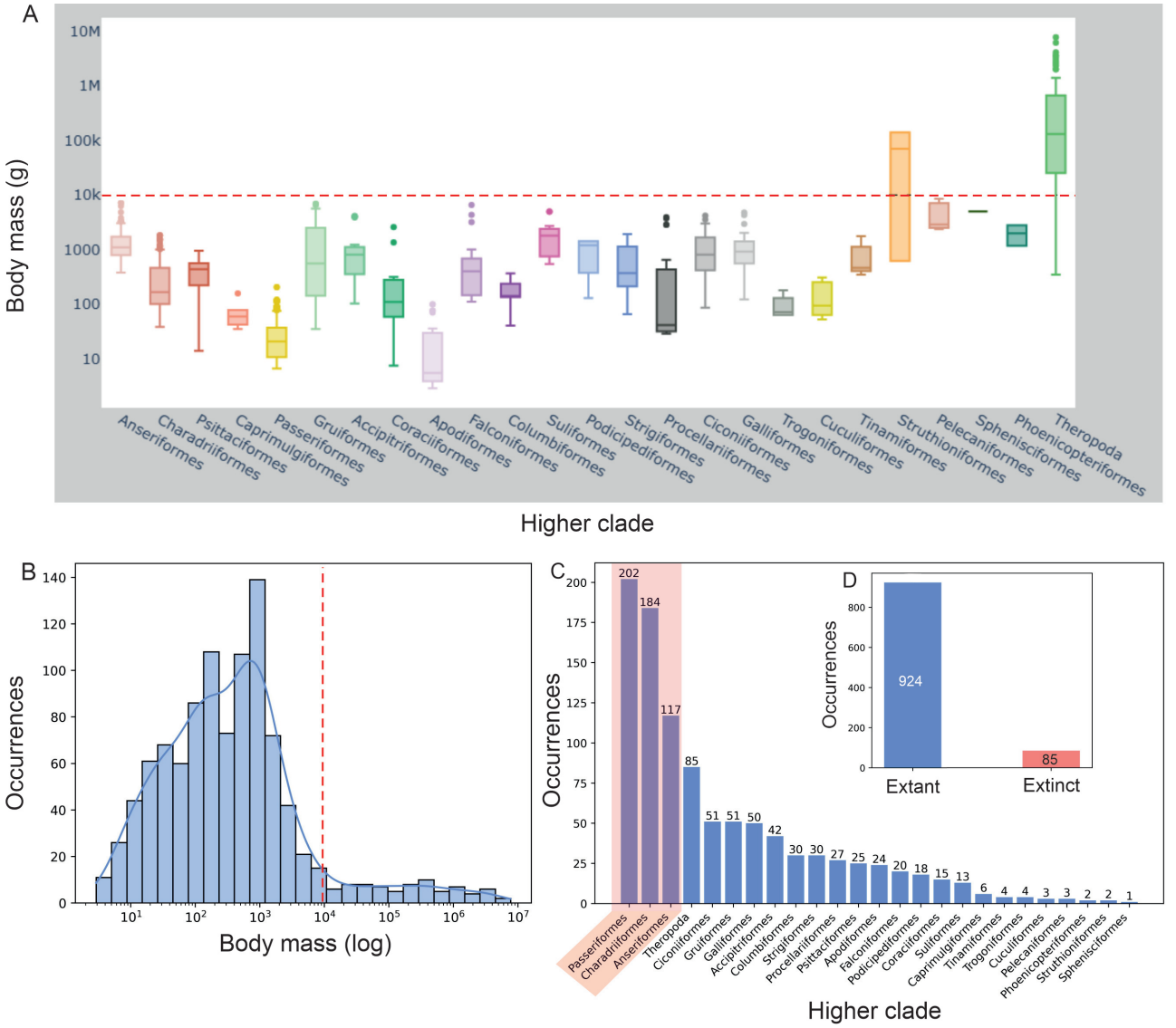

Figure 7. Visual representation of the structure of the full training dataset (Tab. S8). A, The boxplot shows the disaggregated structure, highlighting body mass at 10000 g (horizontal dotted line). Most taxa are placed below this threshold. The y-axis is presented in log10 scale for ease of visualization. B, The histogram shows the skewed distribution of the frequency of body mass values below 10000 g (denoted by a vertical dotted line). C, D, Bar charts depict the taxon imbalance of the dataset, either by Higher clade (C) and by State (D). In C, Passeriformes, Charadriiformes and Anseriformes account for about 50% of the total sampling of the full training dataset, whereas living forms represent 92% of the total sample (D). The clade name 'Theropoda' refers to non-avian theropods.

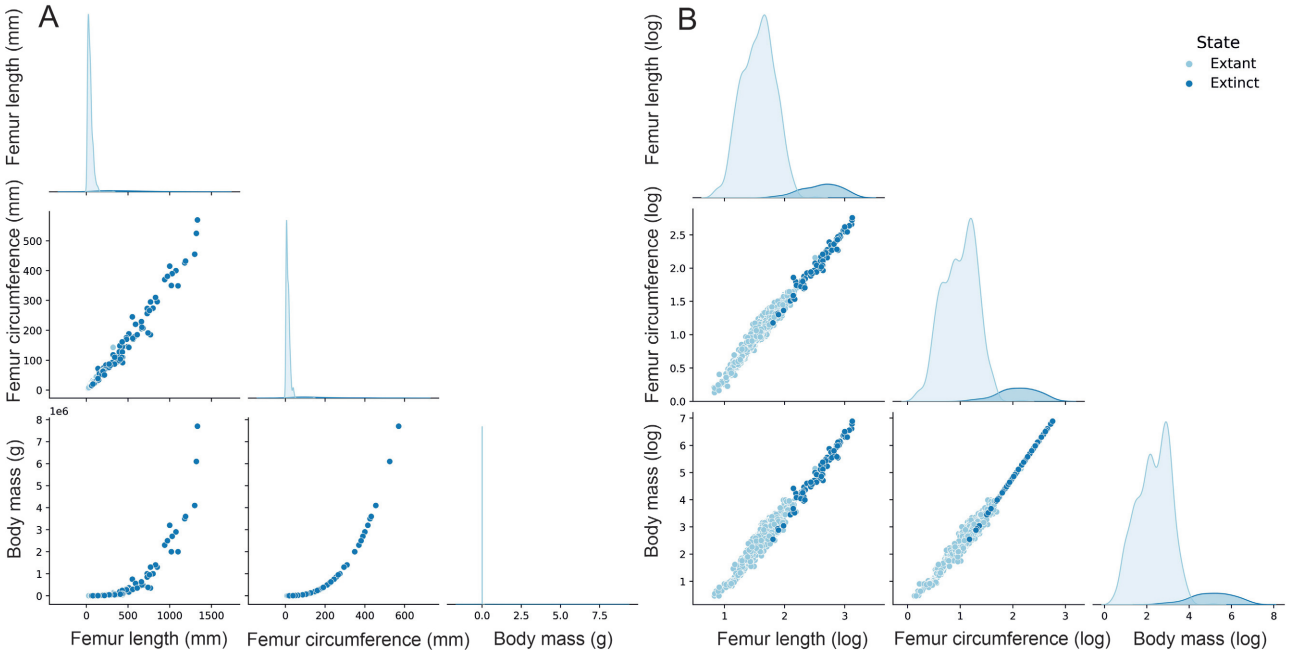

Figure 8. A, B, Variable pairplots showing the non-linear relationship between predictors (femoral length and circumference) and target (body mass) before log-transformation (A), and the linear relationship after log-transformation (B). Body mass parameter in pairplot A is presented in exponential scale for ease of visualization. Note in B the contrasting relationship between femoral circumference and body mass across groups: while extant taxa display natural biological scatter around the scaling trend, extinct taxa form a deterministic, perfectly straight line, visually highlighting the statistical circularity (target leakage) introduced when proxy-derived body mass estimates calculated directly from femoral circumference are used as target labels. This potential issue is addressed in the Discussion section.

Hyperparameter optimization ('tuning') and sensitivity analysis

Following our initial grid search optimization, several parameters for random forests deviated from their defaults, indicating that tuning meaningfully altered the model's behavior (Tab. 1). We subsequently used a sensitivity analysis to refine these results. This involved examining how the prediction error changed when we adjusted one hyperparameter at a time while holding the rest constant.

Comparing the fully optimized model to the default software settings, three of the seven evaluated hyperparameters ('max. features', 'max. samples', 'n. estimators') deviated from their default settings. This indicates that the default configuration does not necessarily provide optimal performance for the present dataset and that systematic tuning meaningfully alters model behavior.

Sensitivity analysis (Fig. 10; Tab. 1) further refined these results by examining model error across a range of hyperparameter values while holding others constant. Three hyperparameters ('max. depth', 'min. samples leaf' and 'min. samples split') reached their minimum test error at values different from those selected by the grid search procedure. In practical terms, these parameters act as biological 'reality checks' that control the growth of the decision trees, preventing them from becoming too complex and simply memorizing the training data (*i.e.*, overfitting). For example, rather than learning the general allometric scaling relationship between femoral circumference and body mass, an unrestricted model might create a highly specific, idiosyncratic rule just to accommodate a single anomalous specimen or taphonomic outlier. By constraining tree growth, we force the model to capture broader evolutionary and biomechanical patterns. This suggests that interactions among hyperparameters during joint optimization may mask alternative local minima that become evident when parameters are evaluated individually (Fig. 10B, E).

The hyperparameter values combination after sensitivity analysis reported in Table 1 corresponds to the configuration yielding the lowest error on the testing set (Fig. 10). When compared to the default settings, five out of

seven key hyperparameters differ, underscoring the importance of tuning these algorithms specifically for paleontological datasets rather than relying on out-of-the-box computer science defaults. Only two parameters ('bootstrap' and 'min. samples leaf'), retained their default value after optimization. It was also associated with the minimum error during sensitivity analysis, indicating that this baseline grouping size is robust under the morphometric variance of our current sample.

Additionally, the sensitivity analysis reveals that 'max. features' and 'n. estimators' exhibit relatively stable error values across a broad range of tested configurations for both training and testing sets (Fig. 10C, D). This pattern suggests that model performance is comparatively insensitive to moderate variations in these parameters, indicating a plateau region in the error landscape and reduced risk of performance instability due to minor parameter adjustments.

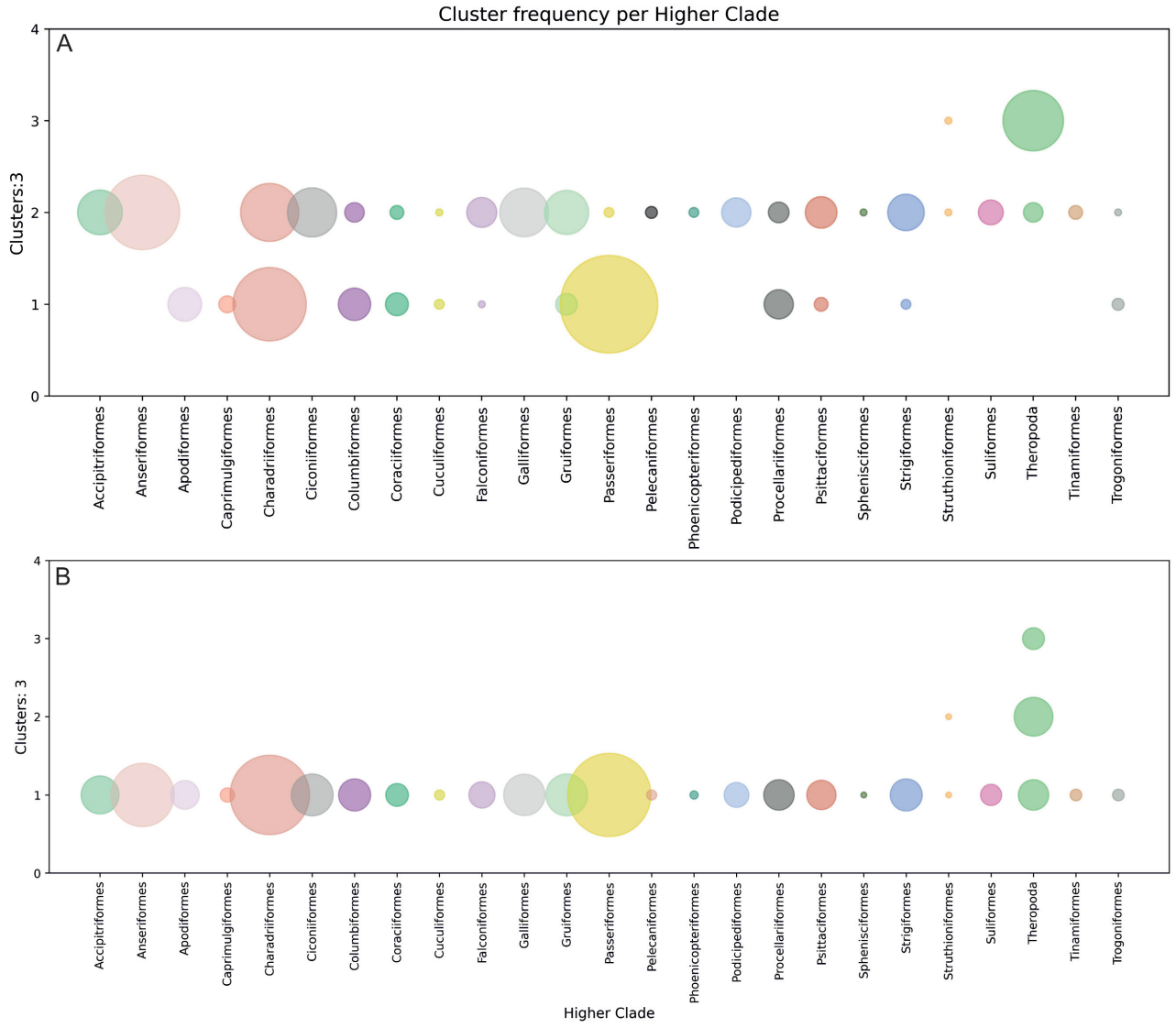

Figure 9. Data clustering. Three clusters were selected after analyzing the inflection curve via the silhouette and elbow methods (Supplementary Information, Fig. S2). For the log-transformed space (A), the disaggregated bubble plot shows the distribution of the groups among the three clusters, with most living forms distributed across clusters 1 and 2 (except for Struthioniformes, which are distributed in clusters 2 and 3). In the standardized space (B) most living forms fall in cluster 1, which covers a wide range of body masses (from 2,5 to 44000 g). The size of the bubble indicates the frequency of occurrences per cluster, whereas the color and tone indicate the taxon rank. The clade name 'Theropoda' refers to non-avian theropods.

These results demonstrate that systematic tuning substantially improves model configuration. The divergence between default, grid-optimized, and sensitivity-derived values highlights why researchers should explicitly evaluate and tune model complexity rather than treating machine learning algorithms as 'black boxes' with pre-defined parameters. These findings reinforce the necessity of combining grid search with sensitivity analysis as complementary steps to ensure both optimal predictive performance and a deeper understanding of model behavior.

Model performance

In this section we present the results yielded by iterations of the algorithms across different datasets for the combined variables (femur length and circumference) as well as for variables in isolation. We are interested primarily in analyzing the results on the validation set (*i.e.*, data entirely unseen during training and testing), since it is composed of data that was not included on the training set (see Tab. S7 for body mass prediction on the validation set along with corresponding error boundaries based on PPE). Performance (via R^2 and PPE) was evaluated across the four dataset variants: full, SMOGN-balanced, taxa-balanced and extant-only. In addition, mean squared error (MSE) was used in the learning curves for neural networks (for performance results on the test split for all training datasets see Supplementary Information, Fig. S4, Tab. S10). For NN, results under L2 regularization retrieved the better performance, outperforming dropout and batch normalization; hence, the former will be exposed below.

Model performance on the validation set reveals strong predictive accuracy (>0,9) under full, SMOGN-balanced, and taxa-balanced configurations, but a pronounced breakdown under extant-only training.

Across the full, SMOGN-balanced, and taxa-balanced datasets, both random forest and neural network models achieved consistently high explanatory power when trained with combined predictors (FL + FC) or femoral circumference (FC) alone. R^2 values frequently exceeded 0,95 and reached up to 0,97 in random forests with default hyperparameters trained only with femur circumference (RFdh, FC, Fig. 11A), while PPE values typically ranged between 27–31% for

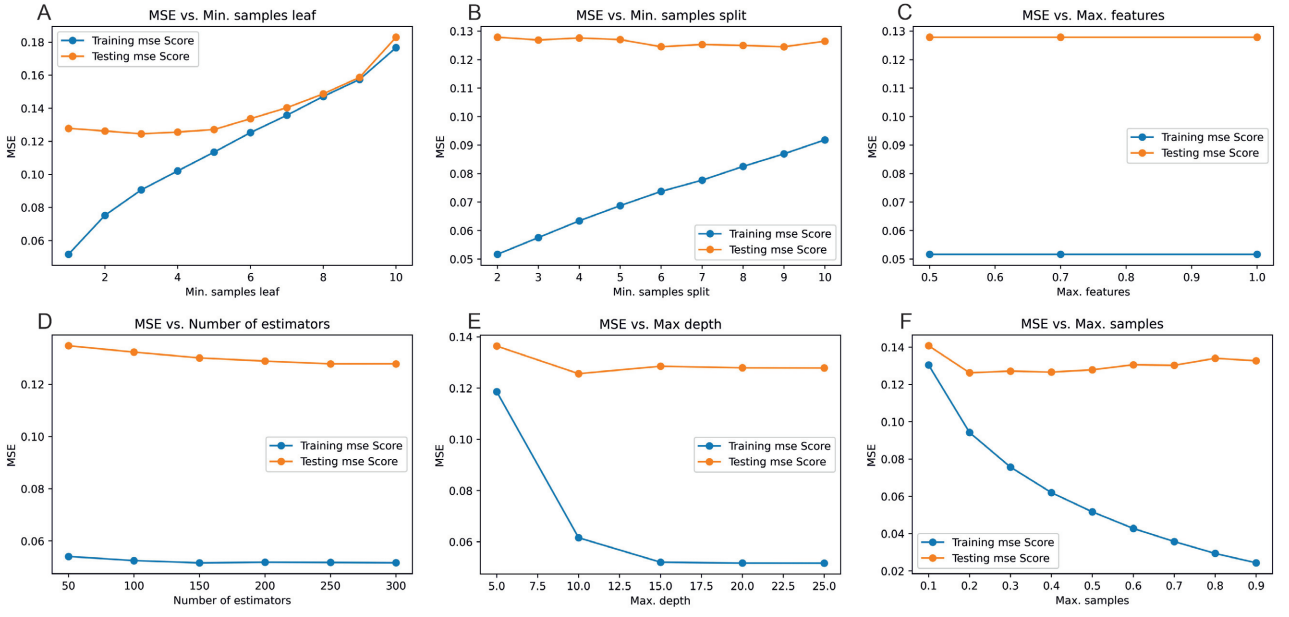

Figure 10. Sensitivity analysis of hyperparameters for random forests in terms of hyperparameter values (on x-axis) versus error (MSE; y-axis) for the full training set. For each hyperparameter, a set of values was evaluated, keeping fixed remaining ones, addressing the error for each value selected. This way, for each hyperparameter, we can see which value retrieves the lowest error, and adjust the model accordingly. Apart from being a useful tool for adjusting the model, sensitivity analysis also helps visualize hyperparameters that are more prone to influence the model output. In this case, 'Min. samples leaf' shows great variation in both training and testing sets as its value changes. Y-axis scales are optimized independently for each panel to clearly display the local minimum of the error landscape for every hyperparameter.

combined predictors and slightly higher for FC alone (Fig. 11B). Femoral circumference consistently outperformed femoral length (FL), reinforcing its biomechanical relevance as a primary body mass proxy (*e.g.*, Anderson et al., 1985; Campione et al., 2014). For this reason, interpretation focuses on combined predictors and FC-only models.

Balancing procedures produced moderate but interpretable effects. Results retrieved in SMOGN-balanced dataset slightly improved or preserved R^2 in several random forests configurations, while the full and taxa-balanced datasets maintained high explanatory power but introduced modest increases in PPE, particularly in neural network models. These modifications did not disrupt the overall scaling structure, indicating that model performance remains stable when morphological variance breadth is preserved. Analyzing extant versus extinct errors it can be seen that extinct forms present the overall lowest error values, most occurrences falling below 20%, with the exception of two extreme values, corresponding to *Bambiraptor* and *Rahonavis*, the theropod occurrences with the lowest predicted body mass in the validation set (Fig. 12). This low prediction error in extinct taxa arises because the target body masses in these cQE-derived training configurations were generated directly from minimum femoral circumference via scaling equations, creating a deterministic mathematical relationship between the predictor feature and the target variable. To evaluate model behavior under a scenario of more empirical independence and break this mathematical tautology, these cQE-derived training runs were evaluated alongside alternative model configurations trained on independent, volumetric-density (VD) body mass targets (see below and Discussion).

A fundamentally different pattern emerges under extant-only training. When models were trained exclusively on living birds and evaluated on the validation set (which includes both extant and extinct taxa spanning a substantially broader body mass spectrum) random forests performance collapsed, yielding negative R^2 values (approx. −0,04) across all configurations and extremely high PPE (>100%). Neural networks exhibited partial retention of explanatory power (R^2 up to 0,35 for combined predictors), but performance remained markedly reduced relative to full-dataset training (Supplementary Information, Fig. S4).

The negative R^2 values observed for random forests reflect a structural failure of generalization rather than stochastic instability. When confronted with values outside the training variance regime, predictions regress toward the training mean, inflating residual error such that it exceeds the total variance of the validation set, thereby producing negative R^2 (see Discussion).

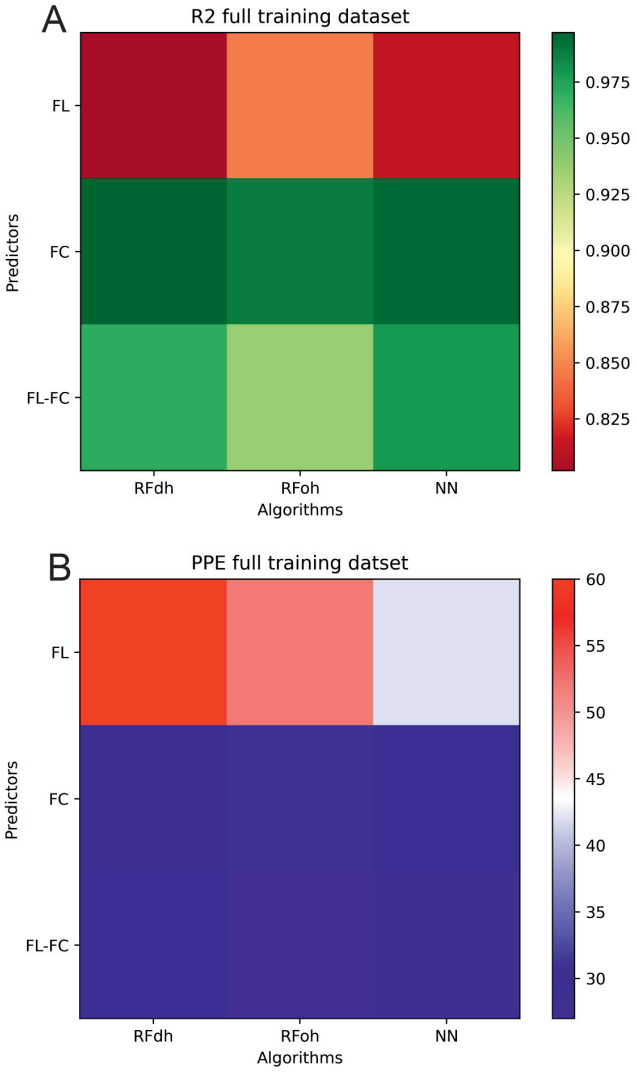

Figure 11. Heatmaps showing overall performance of algorithms (random forests and neural networks) versus predictors (FL-FC combined, FL alone, FC alone) in terms of R^2 (A) and percent prediction error (PPE, B) for the validation set after training with the full training dataset. Performance is similar for all algorithms, showing high R^2 (> 0,9) for combined predictors and FC-only, decreasing sharply under FL-only. A similar scenario is displayed for PPE. Heatmaps for remaining training datasets are presented in online supplementary information (Supplementary Information, Fig. S4). Abbreviations: NN, neural networks; RFdh, random forests with default hyperparameters; RFoh, random forests with optimized hyperparameters.

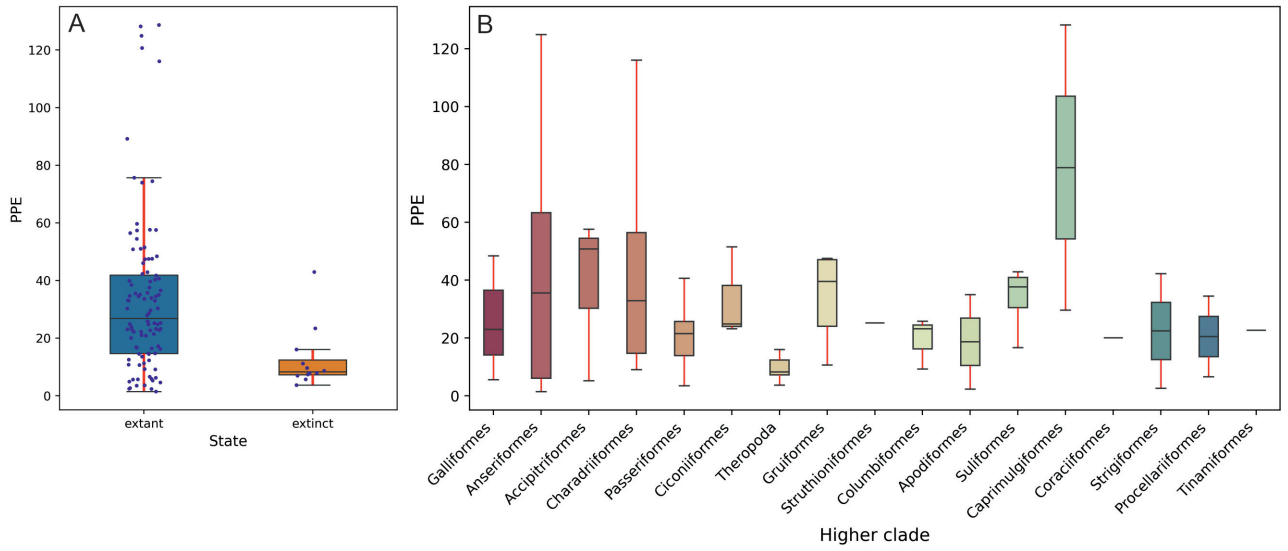

Figure 12. Boxplots showing the percent prediction error (PPE) for random forests with optimized hyperparameters for the validation set after training with the full training dataset, displaying extant versus extinct forms (A) and by higher clade (B). Overall, non-avian theropod dinosaurs display the lowest error values. The clade name 'Theropoda' refers to non-avian theropods.

Learning curves for the neural network models show rapid convergence across all predictor configurations for the full training dataset (Fig. 13) and also for remaining datasets used (Fig. S5). In all cases, both training and testing mean squared error (MSE) decreases sharply during the first few epochs and reaches a stable plateau after approximately 8–10 epochs (vertical dashed line in Fig. S5), with the exception in the taxa-balanced dataset, in which stabilization occurs soon after epoch 20 (Supplementary Information, Fig. S5). Such rapid convergence is typical in low-dimensional regression problems, where the underlying functional relationship is relatively smooth and the number of predictors is small (Schmidt-Hieber, 2020). In these cases, the gradient descent optimization quickly moves toward a stable region of the loss surface.

Beyond this point, additional training produces negligible reductions in error, indicating that the network quickly captures the dominant relationship between the predictors and body mass. The training and testing curves closely track one another throughout the optimization process, with only minimal divergence, suggesting limited overfitting and good generalization performance. If the network were overfitting, the training error would continue decreasing while the testing error increased or diverged (*e.g.*, Chollet, 2018).

Minor differences among predictor configurations are reflected in the final plateau values: models trained with combined predictors (FL + FC) and with femoral circumference alone converge to slightly lower MSE values than those using femoral length alone. This pattern is consistent with the stronger biomechanical and allometric association between femoral circumference and body mass documented in previous studies (*e.g.*, Anderson et al., 1985; Carrano, 2001; Christiansen & Fariña, 2004; Campione et al., 2014). Overall, the learning curves indicate stable training dynamics and confirm that the neural networks efficiently learn the underlying scaling relationship within a relatively small number of epochs.

Feature importance

In our analyses, SHAP-based feature importance was nearly identical between training and validation datasets, indicating stable attribution patterns and suggesting that the learned scaling relationships generalize consistently across partitions. For this reason, interpretation focuses on the validation results.

Across algorithms and dataset configurations, femoral circumference emerges as the dominant predictor of body mass (Fig. 14; Supplementary Information, Figs. S6–S11). This effect is particularly pronounced in random forests with default hyperparameters and neural networks (Fig. 14A, B,

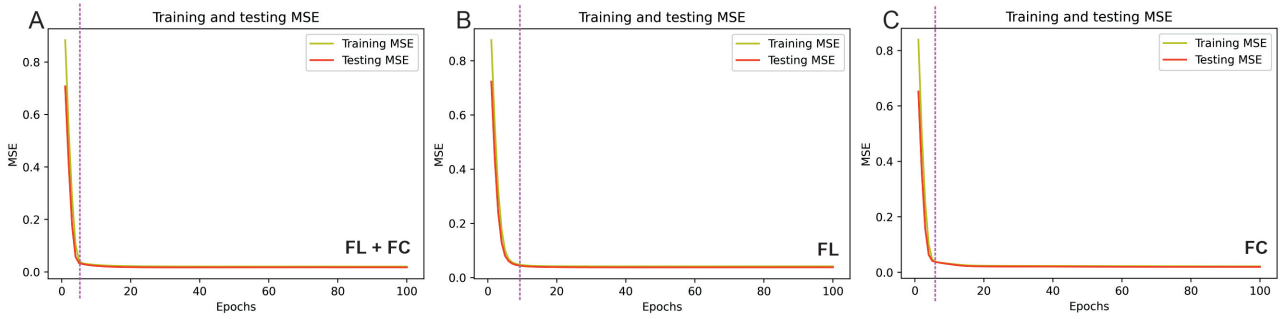

Figure 13. A–C, Neural networks' performance curves (epoch versus mean squared error—MSE—) for predictor combinations under the full training dataset. Two clear patterns can be highlighted. First, a sudden drop in error before reaching epoch 10; and second, training and testing curves converge (*i.e.*, both curves touch each other) very fast. These patterns indicate that low-complexity neural networks are enough to predict with good performance using current datasets.

E, F), where SHAP values for femoral circumference consistently exceed those of femoral length in the four training regimes (full, SMOGN-balanced, taxa-balanced and extant-only datasets; Supplementary Information, Figs. S6–S11). In these configurations, SHAP distributions for femoral circumference are farther from zero and exhibit greater spread, indicating stronger and more variable contribution to model output. This pattern suggests that random forests with default hyperparameters and neural networks architecture assign greater functional weight to circumference when constructing its nonlinear mapping, likely reflecting the stronger biomechanical and cross-sectional scaling signal embedded in femoral circumference relative to femoral length.

Random forests with optimized hyperparameters models also identify femoral circumference as the principal predictor, but importance is comparatively more balanced relative to femoral length (Fig. 14C, D). In the latter algorithm, SHAP values for femoral circumference exceed those of femoral length in three of the four training regimes (full, SMOGN-balanced and extant-only datasets; Supplementary Information, Figs. S7, S10). This difference likely reflects architectural distinctions. In tree-based ensembles, partition feature space locally; when predictors are correlated, splits can alternate between features depending on local variance structure (see Discussion). As a result, importance may be distributed more evenly when both predictors contribute similarly to impurity reduction (Breiman, 2001b). In contrast, neural networks approximate a single continuous function defined over the entire predictor space, tuning all model weights simultaneously through gradient-based optimization. Because this process systematically amplifies the contribution of whichever predictor reduces prediction error most effectively across the whole training dataset, femoral circumference (which carries a stronger biomechanical signal for body mass) receives disproportionately larger weights, resulting in a more pronounced and consistent dominance over femoral length in SHAP-based attribution (*e.g.*, Aggarwal, 2018).

An exception arises under the taxa-balanced dataset, where femoral length exhibits slightly higher SHAP values in the random forest model with optimized hyperparameters (Supplementary Information, Fig. S10C, D). In the latter, most of the specimens present in the validation set have femoral length as the most relevant predictor, from taxa below 100 g (*e.g.*, Passeriformes) to those exceeding 100000 g (non-avian theropods). While previous studies (*e.g.*, Campione & Evans, 2012) have established femur circumference as the primary correlate for body mass due to its direct relationship with static weight-bearing constraints, in our framework, FC serves as the predictive 'baseline' (the average of the dataset), while secondary skeletal features provide critical biological context. This process is explicitly visualized in the force plots for individual taxa (Fig. 15). For instance, in the extant *Gallus* (Fig. 15A), femur circumference provides the dominant positive pressure (red) to increase the mass estimate from the base value, while femur length acts as a slight corrective factor, pushing the estimate in the opposite direction (blue).

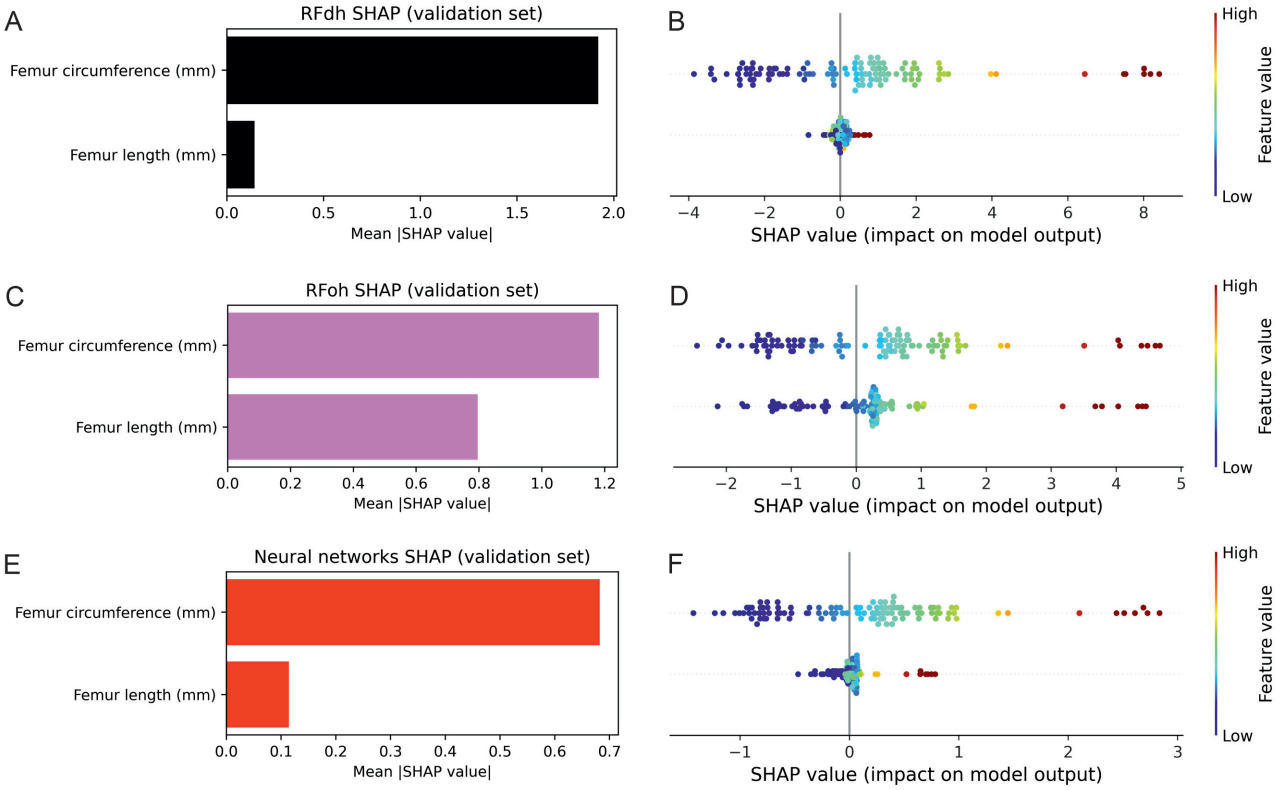

Figure 14. Feature importance output via SHAP (Shapley additive explanations) for the validation set after training with the full training dataset. A, B, E, F, Femur circumference emerges as the most influential predictor across algorithms, with major influence in random forests with default hyperparameters (RFdh, A, B) and neural networks (E, F). C, D, Random forests with optimized hyperparameters (RFoh), however, present more influence of femoral length (see Discussion for explanation). The beeswarm plots, on the right, combine global feature importance and directional influence, expressing the samples (as dots) in terms of the impact on the model output. Samples with SHAP values close to zero means the predictor has low impact on prediction and vice-versa. High feature values (y-axis, red dots) increase prediction, whereas low feature values (y-axis, blue dots) decrease prediction.

This shift likely reflects redistribution of morphological variance across clades. Because random forests importance depends on recursive partitioning driven by local heterogeneity, changes in variance distribution (as occur in the extant-extinct evenly balanced dataset) can modify feature usage frequency without fundamentally altering the underlying biological scaling relationship (see Discussion).

Machine learning predictions against benchmark equations in biped dinosaurs

Based on the preceding analysis and the performance and error metrics obtained, the machine learning algorithms evaluated here yielded sufficiently reliable results to warrant their consideration in future studies, particularly when applied to avian datasets. However, a fundamental methodological question remains: what is the precise impact of target leakage, specifically, introducing an explicit circularity where proxy-derived estimates calculated from the cQE formula appear as both predictor features and target variables, on algorithmic performance, morphospace extrapolation, and model generalizability?

When trained on the cQE-derived dataset configuration (1), the machine learning predictions exhibited exceptional, near-deterministic alignment with the cQE benchmark (Fig. 16; Tab. S13). The random forest model closely reproduced the biomechanical scaling relationships captured by the cQE equation, yielding a mean proportional difference of 7,15% across the sample, with 60% of taxa (n = 18) falling within 5% of the cQE estimate and an additional 20% (n = 6) showing minor discrepancies of 5–10%. Mid-to-large-bodied bipeds, ranging from ornithischians like *Jeholosaurus* (10 kg), *Stegoceras* (10 kg), and *Homalocephale* (40 kg), to massive

theropods like *Allosaurus* (2690 kg) and *Giganotosaurus* (6000 kg), fell tightly along the 1:1 parity line and well within the asymmetric 25% prediction error envelope of the cQE model. However, because the fossil target labels in this configuration were mathematically calculated from minimum femoral circumference using the cQE formula itself, incorporating FC as an input feature caused the algorithms to re-learn the underlying deterministic power-law equation. While this demonstrates that random forests and neural networks can successfully reconstruct continuous mathematical functions across multi-order size scales, Figure 16 visually exposes a possible target leakage through the unvarying linear alignment of fossil points in log-transformed space.

To break this circularity and test model behavior under true empirical independence, we re-trained the algorithms

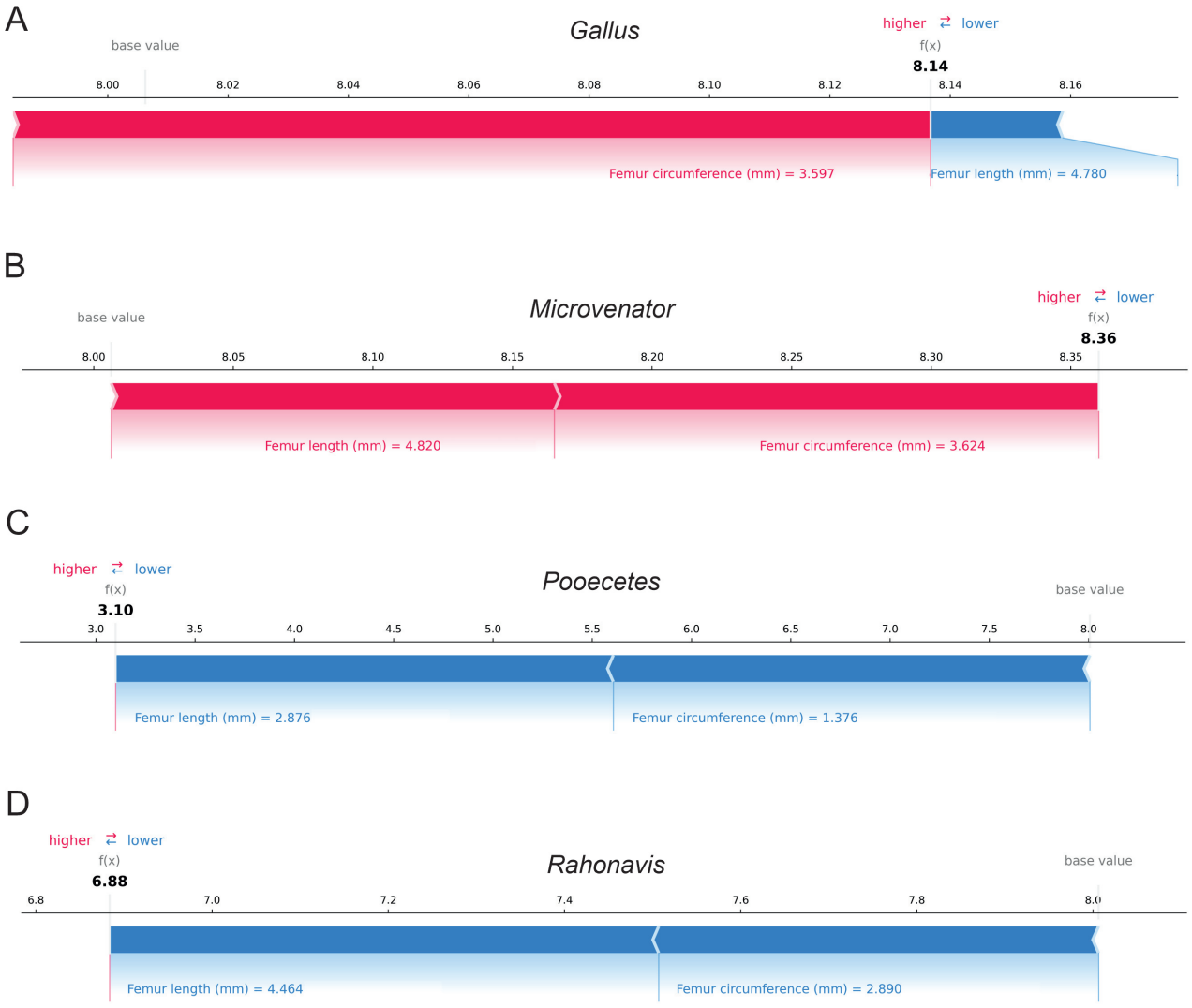

Figure 15. A–D, Example of force plot showing a detailed breakdown of the features' contributions for a specific prediction under the taxa-balanced dataset (random forests with optimized hyperparameters). In this case, the plot denotes the contribution of each predictor to prediction across some extant (*Gallus*, *Pooecetes*) and extinct (*Microvenator, Rahonavis*) theropods. In this case, *Gallus* and *Microvenator* present a higher femur circumference contribution to body mass prediction (note the longer femur circumference bar relative to femur length bar in A and B). Conversely, *Pooecetes* and *Rahonavis* have a body mass prediction with femoral length as the most influential predictor (note the longer femur length bar relative to femur circumference bar in C and D). The Relative Shapley additive explanations (SHAP) value attributes show how the different variable combinations vary in the model prediction. The x-axis (log-scale) is the model output value and f(X) is the predicted body mass for this particular sample. Features pushing the prediction higher relative to the mean or base value are shown in red, while those pushing it lower are in blue.

using 3D volumetric-density (2) body mass targets derived from soft-tissue volume reconstructions (Campione & Evans, 2020, app. S1) (Supplementary Information, Fig. S12; Tab. S13). Because VD mass estimates are calculated from 3D digital envelope volumes (*e.g.*, minimum convex hulls or photogrammetry) independent of long-bone shaft dimensions, using VD targets alongside femoral predictors removes the direct formulaic dependence on femoral circumference used in cQE-derived labels. The artificial linearity seen in Figure 16 dissolves into realistic biological scatter that reflects true variation in body proportions, soft-tissue distribution, and digital reconstruction methodologies, although not drastically different in results from Figure 16. In general, neural network predictions are placed within the error envelope of cQE estimations, with large-bodied theropods (*e.g.*, *Tyrannosaurus*, *Giganotosaurus*, *Mapusaurus*,

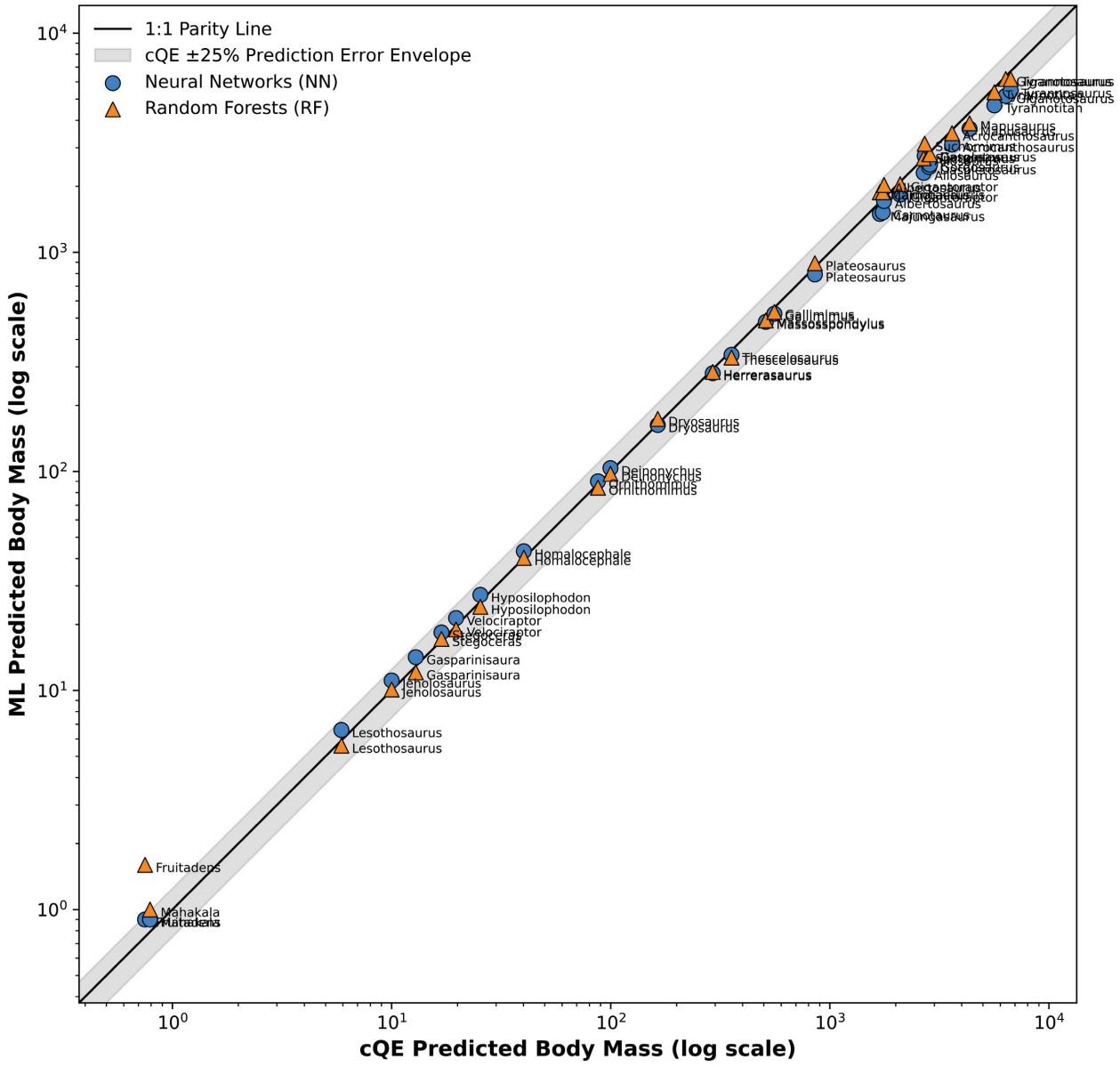

Figure 16. Parity plots comparing machine learning body mass predictions (random forests, orange triangles; neural networks, blue circles) against benchmark estimates derived from the corrected quadrupedal equation for bipeds (cQE; Campione et al., 2014) across 30 bipedal dinosaur taxa in log-log scale. The solid diagonal line indicates 1:1 parity, and the grey shaded envelope represents the 25% cQE prediction error interval, projected as asymmetric bounds in logarithmic space. Predictors and target variables share a formulaic dependency, causing models to re-learn the deterministic power-law scaling curve and producing linear alignment along the parity line (target leakage).

Acrocanthosaurus) positioned on the parity line, confirming that even when target leakage is completely removed, supervised machine learning algorithms trained on morphologically expanded datasets successfully capture deep-time allometric scaling trends without collapsing. Conversely, most predictions derived from random forest algorithms depart from the parity line (*e.g.*, *Velociraptor*, *Jeholosaurus*) and some of them fall outside the cQE error ranges, particularly at extreme body size (*e.g.*, *Gasparinisaura*, *Fruitadens*, *Mahakala*, *Suchomimus*, *Acrocanthosaurus*).

Evaluating predictions under the hybrid training configuration, which combines both cQE-derived and VD-derived targets (3) demonstrates a transitional, mediating effect between the deterministic constraints of Figure 16 and the empirical scatter of Supplementary Information, Figure S12 (Tab. S13). Predictions under random forest exhibit an overall mean difference with cQE estimations of 13,5%, in which 60% of the sample present less than 10% of deviation from the cQE parity line and only 15% fall outside the cQE error boundaries, whereas predictions under neural networks remain closely aligned with those inferred with cQE, departing an average of only 5% relative to cQE. In addition, 80% of the predictions differ less than 10% relative to cQE and 100% are within the cQE error margins, highlighting neural networks under hybrid configuration training as the best-performing among the evaluated configurations, according to the selected benchmark metrics for biped estimation.

Under hybrid training, smallest taxa like *Fruitadens* and *Mahakala* remain overestimated relative to cQE for random forests. Those taxa have femoral circumferences of 19,6 mm and 19,9 mm, respectively, showing the largest deviations from the 1:1 parity line, with random forests overestimating relative to cQE by approximately 55% and 50%, respectively. These discrepancies likely reflect the challenge of interpolating into extremely small body size ranges with both methods and are in line with the results presented for the validation set above, in which smallest theropods exhibit the greatest error (Fig. 12). For the random forests, the training distribution may include few analogous sub-kilogram bipeds (such as Psittaciformes, Passeriformes, Columbiformes), reducing interpolative accuracy in this region of morphospace. The cQE equation, while calibrated on a broad extant dataset that includes small-bodied taxa (Campione et al., 2014), applies a scaling relationship that may behave differently near the lower boundary of its calibration range.

Large theropods, in general, cluster tightly along the parity line or the upper boundary of the cQE error envelope for random forests. Among the largest theropods, several taxa showed random forests estimates that exceeded cQE predictions by more than 10%. Specifically, *Acrocanthosaurus* (>30%), *Tyrannotitan* (>20%), and *Giganotosaurus* (>12%) all showed random forests estimates above cQE values. In contrast, smaller taxa, such as *Gallimimus*, *Deinonychus* and *Herrerasaurus* showed the opposite pattern, with cQE exceeding the random forests estimate. These bidirectional discrepancies suggest that deviations are not driven by a systematic algorithmic bias but rather reflect taxon-specific factors.

Interpreting the above-mentioned patterns using overall body length as a metric must be approached with extreme caution, as reconstructed body lengths in large extinct dinosaurs are notoriously prone to significant error and preservation artifacts (Hone, 2012). Consequently, rather than relying on global length metrics, these deviations are better understood through localized skeletal proportions, such as differential femoral robusticity. In this regard, abelisaurids and some tyrannosauroids are characterized by relatively robust femoral shafts relative to their overall scale (Pintore et al., 2024), a pattern that reflects the well-established positive relationship between femoral robusticity and body mass in large-bodied theropods (Farlow et al., 1995; Hutchinson & Garcia, 2002). Conversely, *Suchomimus*, a spinosaurid with elongate, somewhat gracile hindlimb morphology relative to other large theropods, may exhibit a circumference signal that the random forests place towards a slightly higher mass than the strictly circumference-based cQE.

The estimate for *Tyrannosaurus rex* warrants specific commentary given the volume of independent estimates available for this taxon. Our random forests yielded near 7000 kg, while cQE returned 6688 kg (error interval: 5014–8361 kg). Both values fall well within the broad range reported in the literature, which spans approximately 5000–9500 kg depending on specimen and method (*e.g.*, Farlow et

al., 1995; Hutchinson et al., 2011; Campione et al., 2014). The neural networks estimate is somewhat conservative relative to the cQE central estimate (approx. 3% lower), yet remains within the cQE confidence interval, suggesting that the two methods are statistically concordant for this taxon.

These comparisons support the validity of the random forests and neural networks as a body mass estimation tool for non-avian dinosaurs, particularly across the intermediate size range where both methods converge consistently. Where the methods diverge, principally at extreme body sizes or in taxa with atypical femoral proportions, the discrepancies provide interpretively useful signals rather than simple error, pointing to biological variation in hindlimb scaling that neither method fully captures through circumference alone. This reinforces the utility of combining machine learning approaches with established allometric equations as complementary rather than competing tools for body mass inference in fossil taxa.

DISCUSSION

General algorithmic behavior under morphospace restriction

The results presented above for the unseen data on the validation set (Tab. S7) demonstrate that both random forest and neural network models achieve high predictive accuracy when trained on datasets that preserve broad morphological and body mass variance, particularly when femoral circumference is included as a predictor. Under training datasets with mixture of extant and extinct forms, R^2 values remain high and PPE remains low, suggesting that machine learning approaches implemented here effectively capture allometric scaling when variance structure is maintained (Fig. 11; Supplementary Information, Fig. S4). However, algorithmic behavior diverges markedly when dataset structure is restricted, revealing fundamental differences in architectural properties and generalization mechanisms.

Random forests exhibit strong robustness when broad variance is present, but collapses under variance truncation. Its ensemble structure, based on bootstrap aggregation and random feature subsampling, reduces variance and stabilizes predictions across heterogeneous datasets (Breiman, 2001b; Hastie et al., 2009; Biau & Scornet, 2016). Because individual trees partition the predictor space locally and predictions are averaged, random forests are typically resistant to moderate distributional perturbations and localized noise. This explains its stability under SMOGN rebalancing and taxa redistribution, as well as its strong performance in full-variance datasets. However, when training is restricted to extant taxa, random forests performance deteriorates dramatically (Supplementary Information, Fig. S4). This is not stochastic noise but a structural limitation. Tree-based methods are intrinsically interpolative: terminal node predictions are bounded by observed training responses and cannot extrapolate beyond the empirical range (Breiman, 2001b; Hastie et al., 2009; Hengl et al., 2018; Sekulić et al., 2020). Because the validation set contains large-bodied extinct taxa (>500 kg; *e.g.*, *Aucasaurus*, *Tyrannosaurus*, *Giganotosaurus*) absent from extant-only training, random forests predictions regress toward the smaller-bodied mean of the training distribution. This phenomenon mirrors the behavior observed in complex parametric frameworks when predicting body masses beyond the scope of living taxa; as Campione (2017) demonstrated, highly flexible or complex models (such as higher-order polynomials) become exceptionally unpredictable and prone to severe error during deep-time extrapolation. Residual error consequently exceeds total variance in the validation data, producing poor R^2 values. The latter highlights a fundamental limitation of decision-tree models when they are forced to make predictions for giant extinct animals that fall completely outside the size range and body shapes seen in living species (Meyer & Pebesma, 2021).

Neural networks display a distinct but complementary pattern. In full-variance datasets, neural networks performance rivals or exceeds random forests, consistent with their capacity as universal function approximators capable of modeling complex nonlinear scaling relationships (Goodfellow et al., 2016; Aggarwal, 2018; Yaqoob et al., 2025) (Fig. 11; Supplementary Information, Fig. S4). Rather than collapsing abruptly, predictive accuracy degrades more gradually under extant-only training, with R^2 values declining but remaining positive and PPE increasing moderately. This relative resilience reflects the capacity of sufficiently

parameterized feedforward networks to learn smooth, continuous functional mappings between predictors and the target variable, rather than locally bounded partitions (Hornik et al., 1989; Goodfellow et al., 2016). Nevertheless, neural networks performance also degrades under severe variance restriction, consistent with the general principle in statistical learning theory that generalization error increases as the divergence between training and test distributions widens (Ben-David et al., 2010). Neural networks optimize a global loss function via gradient descent, making them more sensitive to leverage points and uneven data distributions (Yaqoob et al., 2025). In the extant-only configuration, R^2 still declines substantially, reflecting reduced variance and diminished ability to infer generalized scaling structure. The partial retention of predictive capacity under extant-only training, contrasted with random forests' near-total collapse, therefore reflects a structural advantage in extrapolation for smooth function approximators over partition-based ensembles, rather than immunity to distributional shift. However, while neural networks demonstrate mathematical continuity during deep-time extrapolation, maintaining numerical estimates beyond the extant training range, this mathematical output must be interpreted strictly as an algorithmic projection rather than inherent biological validation. Because these models remain fundamentally constrained by the uniformitarian assumption that modern scaling laws govern extinct clades, and because clade-specific variations (such as skeletal pneumatization or atypical limb proportions) can alter real-world mass distribution, performance metrics at extreme size regimes must be evaluated alongside known anatomical and phylogenetic limitations.

The marked fluctuation in R^2 observed under different train-test partitions (Tab. S7) highlights the instability inherent to variance-restricted extant-only dataset. When a fixed random state is imposed (as we did here to ensure reproducibility), a single deterministic split is evaluated; if that split produces an uneven distribution of body mass between training and validation subsets, particularly underrepresentation of large-bodied taxa in the training set, model performance declines sharply. In contrast, alternative random splits may distribute morphological variance more evenly, yielding substantially higher R^2 values. This sensitivity reflects the limited variance structure of the extant-only dataset and the reliance of regression-based machine learning models, especially tree-based methods such as random forests, on adequate coverage of the predictor space.

This phenomenon represents a clear case of covariate shift (Sugiyama et al., 2007; Quiñonero-Candela et al., 2009), in which the joint distribution of predictors and target differs between training and validation phases. The extant-only dataset truncates the upper tail of body mass and compresses morphospace coverage. As a result, the learned regression surface is biased toward small- to medium-bodied taxa and cannot recover variance across deep-time size regimes. Neural networks, while capable of limited functional extrapolation due to their parametric structure (Goodfellow et al., 2016), remain constrained by the reduced variance structure of the training data.

Across all configurations, the consistent superiority of femoral circumference reinforces its biomechanical and allometric relevance as a primary proxy for body mass (Anderson et al., 1985; Carrano, 2001; Christiansen & Fariña, 2004; Campione & Evans, 2012, 2020). The strong performance of combined predictors further suggests that multivariate models capture complementary information embedded in skeletal geometry, enhancing regression stability when variance breadth is preserved (see also Christiansen & Fariña, 2004; Serrano et al., 2015). Nevertheless, applying multivariate models across deep time requires caution due to inherent theoretical constraints. As highlighted by Campione & Evans (2012) and Campione (2017) in studies of body mass scaling, extrapolating models derived largely from extant species to phylogenetically distant or hyper-sized extinct taxa involves fundamental assumptions regarding evolutionary constancy in limb proportion and design. When multivariate techniques combine multiple skeletal dimensions, they can inadvertently compound errors if individual predictors diverge due to clade-specific structural adaptations or specialized locomotor habits. Consequently, while algorithms can successfully exploit multi-variable geometric covariance in datasets with broad variance, their biological reliability depends entirely on whether the uniformitarian assumption holds, namely, that the functional relationships governing living reference

taxa remain valid when projected into the novel morphospace of deep-time, giant extinct clades.

These findings carry important methodological implications for machine learning applications in paleobiology. Preserving broad morphological and body size diversity within training datasets is not simply advantageous, but it is also essential for stable generalization. This remains true even when extinct taxa are associated with estimated rather than directly measured body masses. The way of estimating the body mass in extinct forms is another topic to deal with caution, as we will explore later in the discussion. Although estimated labels introduce uncertainty, their inclusion expands morphospace coverage and reduces extrapolative demands. Excluding such taxa in pursuit of label purity paradoxically narrows variance structure, increases model bias, and may induce complete predictive collapse, as observed for random forests under extant-only training. In modeling continuous biological traits across evolutionary timescales, variance breadth and morphospace representation are therefore more critical to generalization than strict restriction to directly observed values.

Neural networks intrinsic behavior: deep dive into learning curves

In neural networks, the behavior observed in the learning curves provides insight into the learning dynamics of the network architecture used in this study. As we highlighted in the results section, the rapid decline in error during the initial training epochs indicates that the network quickly captures the mapping between femoral measurements and body mass (Fig. 13; Supplementary Information, Fig. S5). Such fast convergence is rather common in relatively low-dimensional regression problems, where the predictor space contains strong scaling signals and the network can efficiently approximate the underlying function during early gradient updates (Goodfellow et al., 2016; Beltrán-Royo et al., 2024; MacFadden et al., 2026; Tong & Zhang, 2026).

The training and testing MSE curves remain closely aligned throughout optimization. In contrast to the unregularized case, the near-identical trajectories of the training and testing MSE curves throughout optimization could possibly be attributed to the effect of L2 regularization (although not exclusively, see below), which penalizes large weights in the network and thereby constrains model complexity. By shrinking parameter magnitudes toward zero, L2 regularization effectively smooths the learned regression surface and reduces variance in model predictions (Hastie et al., 2009; Hsiang et al., 2019; Ha & Kim, 2023). This mechanism discourages the network from fitting noise or local fluctuations in the training data, promoting solutions that generalize more consistently to unseen observations. As a result, even though neural networks are highly flexible universal function approximators, the presence of L2 regularization limits overfitting and stabilizes the optimization trajectory, thus contributing to explaining the close alignment between training and testing error across epochs.

The taxa-balanced dataset provides an additional perspective on network learning dynamics. In this configuration, the error plateau occurs slightly later, shortly after epoch 20, suggesting that the network requires additional optimization steps to accommodate the modified covariance structure introduced by equalizing extant and extinct representation (Supplementary Information, Fig. S5D–F). Balancing taxa effectively redistributes sampling density in morphospace, altering the joint distribution between femoral length, circumference, and body mass. As a result, the network must adjust its internal weight configuration to approximate a slightly different regression surface before reaching convergence. From a statistical learning standpoint, this behavior is expected since neural networks learn by iteratively minimizing a global loss function (*i.e.*, an error proxy), and changes in the predictor/response distribution can shift the location of the optimal parameter minimum in weight space (Chollet, 2018; Ketkar & Moolayil, 2021). The modest delay in reaching the error plateau therefore likely reflects the need to reconcile a more heterogeneous scaling structure rather than instability in the learning process itself.

The fact that the extant-only dataset, trained without L2 regularization, produces learning curves that are highly similar to those obtained from the regularized models (Supplementary Information, Fig. S5) can be explained by properties of the data distribution and the intrinsic dynamics of neural networks optimization. On the one hand, the extant-only dataset likely exhibits a more compact variance structure in both predictor and response space.

Because large-bodied extinct taxa are absent, the range of femoral circumference, femoral length, and body mass is substantially narrower. In regression problems with limited variance and relatively smooth relationships between predictors and the target variable, neural networks can converge to stable solutions even in the absence of explicit regularization. In such cases, the effective complexity of the learned function is naturally constrained by the restricted support of the training distribution, reducing the risk of overfitting (Gidel et al., 2019; Blanc et al., 2020; Wei et al., 2020). In other words, the data distribution itself acts as an implicit regularizer. Even when no explicit penalty term such as L2 is included, stochastic gradient descent (SGD) tends to favor solutions with relatively small weights and smoother functional mappings (Nielsen, 2015; Murphy, 2022). This optimization bias arises from the geometry of the loss landscape and the trajectory followed by gradient descent, which often leads the model toward low-complexity solutions that generalize reasonably well to unseen data. Consequently, the learning dynamics can resemble those of explicitly regularized models, particularly in small or moderately sized datasets with strong signal-to-noise ratios.

On the other hand, a key factor is the strength of the underlying biological signal linking femoral dimensions to body mass. Because both femoral length and circumference scale predictably with body mass under well-established biomechanical principles (*e.g.*, weight-bearing scaling relationships; Anderson et al., 1985; Carrano, 2001; Campione et al., 2014), the network is effectively learning a relatively smooth allometric function. When the signal-to-noise ratio is high and the dimensionality of the predictor space is low, as in the present case with only two predictors, overfitting becomes inherently less likely, even without formal regularization (Murphy, 2022).

These factors could explain why neural networks performance under the extant-only model behaves similarly to the regularized configurations. The restricted morphospace of extant taxa, combined with implicit regularization from gradient descent and the strong biological scaling relationship between femoral measurements and body mass, effectively limits model complexity and produces learning dynamics comparable to those obtained under explicit L2 regularization. This result further reinforces the interpretation that differences observed among dataset configurations primarily reflect variance structure and morphospace coverage, rather than the presence or absence of explicit regularization mechanisms.

SHAP-based feature attribution across dataset configurations

Femoral length and circumference exhibit a positive association with body mass, consistent with their direct relationship to overall skeletal size (see EDA section). Such correlations are expected in linear morphometric variables scaling with mass under allometric principles (Anderson et al., 1985; Damuth & MacFadden, 1990; Campione & Evans, 2012, 2020; Field et al., 2013). However, correlation alone does not clarify how each predictor contributes to model output in nonlinear learning frameworks. In multivariate settings, especially when predictors are collinear, as is typical for limb skeletal dimensions (Campione & Evans, 2012; Dormann et al., 2013), feature importance cannot be reliably inferred from pairwise associations, since standard importance metrics are known to be biased under collinearity (Strobl et al., 2008; Nicodemus & Malley, 2009). SHAP-based attribution addresses this by quantifying each predictor's conditional marginal contribution across the joint feature distribution (Lundberg & Lee, 2017; Lundberg et al., 2020). In line with this, *post-hoc* interpretability methods therefore provide a principled approach to disentangle feature contributions within complex models, making results more interpretable (Ekanayake et al., 2022; Foster et al., 2022; Nohara et al., 2022; Gebreyesus et al., 2023).

From a machine learning perspective, the stability of SHAP rankings (with femoral circumference predominance as the best predictor) across training and validation sets further supports the robustness of the learned relationships (Fig. 14, Supplementary Information, Figs. S6–S11). If feature attribution had shifted dramatically between datasets, this would indicate overfitting or unstable functional mapping. Instead, the near-identical SHAP patterns suggest that the models capture a biologically meaningful and generalizable scaling signal rather than spurious correlations.

The observed shift in feature importance between random forests configurations (default versus optimized

hyperparameters) can be explained by the dependence of tree-based importance measures on model architecture and predictor correlation structure. Feature importance in random forests is derived from the cumulative reduction in prediction error across tree splits, and is therefore sensitive to hyperparameter settings that control tree depth, feature subsampling, and ensemble size (Breiman, 2001b; Hastie et al., 2009). Under default hyperparameters, the model preferentially exploits the strongest global predictor, in this case, femoral circumference, resulting in a clear dominance in SHAP values. In contrast, the optimized configuration sets 'max. features' = 0,5, which, in a two-predictor setting, restricts each split to a single randomly selected predictor. Under default settings ('max. features' = 1,0), every split can choose the more informative variable, so femoral circumference is selected almost systematically. With random feature subsampling, femoral length is necessarily used in a substantial proportion of splits, and predictive contribution is therefore shared between the two correlated predictors. This redistribution is a known consequence of feature subsampling in tree-based ensembles, where correlated variables can share predictive contribution depending on model structure (Breiman, 2001b; Strobl et al., 2008; Hastie et al., 2009; Biau & Scornet, 2016). It reflects how the model uses the predictors, not a reversal of their underlying biomechanical relevance.

A further insight emerges when SHAP patterns are interpreted in light of the extant-only dataset performance collapse. Under variance-restricted training, SHAP values remain directionally consistent (femoral circumference continues to dominate) but their magnitude contracts and becomes less dispersed in random forests, reflecting a reduced functional range over which predictors influence model output (Supplementary Information, Figs. S6, S9). This contraction parallels the sharp decline in R^2 observed for random forests and the partial degradation in neural network performance (Fig. 11). From an interpretability perspective, this indicates that the models are not fundamentally re-weighting predictors; rather, the learned mapping is compressed due to truncated variance in the training distribution. In random forests, this manifests as terminal nodes spanning a narrower response interval, limiting marginal contribution even when a predictor is structurally important (Biau & Scornet, 2016), whereas in neural networks, reduced variance constrains gradient updates and diminishes sensitivity to large-scale nonlinear trends (Goodfellow et al., 2016). This pattern is consistent with statistical learning theory: when the support of the training distribution is truncated, both predictive variance and feature attribution shrink accordingly (Hastie et al., 2009). Thus, SHAP behavior under extant-only training provides mechanistic confirmation that the observed predictive collapse arises from morphospace restriction rather than algorithmic instability. In this sense, interpretability analysis reinforces the conclusion that variance breadth governs both generalization performance and the functional expressivity of learned scaling relationships (Lundberg & Lee, 2017; Molnar, 2022). The mechanistic insights provided by SHAP feature attribution under extant-only training must be disentangled from the statistical circularity inherent to the cQE-derived fossil configurations. Under extant-only training, target body masses consist exclusively of directly measured live weights from modern birds, completely free from proxy-derived target leakage or formulaic circularity. When models trained on this unconfounded extant baseline are tasked with predicting large-bodied extinct taxa, the contraction and collapse of SHAP attributions reflect a true structural limitation: tree-based algorithms cannot extrapolate beyond the physical boundaries of their training morphospace.

Although femoral circumference consistently emerges as the dominant predictor at the global level, individual-level SHAP analyses reveal localized instances in which femoral length exerts slightly greater influence, particularly among certain extant avian clades (*e.g.*, Galliformes, Columbiformes and Tinamiformes). Those clades are not collectively restricted to any particular locomotor habit (*i.e.*, cursorial versus flying); hence, those occurrences likely reflect clade-specific proportional scaling rather than a reversal of biomechanical relevance. Many small-bodied birds exhibit relatively gracile femora with constrained cross-sectional variance (Field et al., 2013; Serrano et al., 2015), whereas longitudinal proportions may retain greater discriminative variability within narrow size regimes. Because SHAP values quantify conditional marginal contributions, reduced variance in femoral circumference within these subpopula-

tions increases the relative importance of femoral length for individual predictions (Field et al., 2013; Lundberg & Lee, 2017). The modest magnitude of these differences indicates local redistribution of predictive contribution rather than structural dominance, underscoring the capacity of the models to capture both global allometric scaling and fine-scale morphological variation.

The SHAP analyses complement predictive performance metrics by revealing how explanatory power is distributed across morphometric variables. The results demonstrate that predictive success is not merely statistical but biologically coherent, with feature attribution aligning closely with established biomechanical expectations. The consistent dominance of femoral circumference across algorithms and dataset configurations reinforces its relevance as a primary body mass proxy: circumference directly captures weight-bearing cross-sectional geometry, which scales predictably with body mass under elastic similarity and structural support models (Anderson et al., 1985; Carrano, 2001; Campione & Evans, 2020). Femoral length, while correlated with mass (Hone et al., 2008; Erickson et al., 2009) encodes longitudinal scaling and may be more sensitive to ecological or phylogenetic variation (Field et al., 2013; Campione & Evans, 2020) resulting in slightly lower predictive centrality across most configurations.

Predictive machine learning versus classical statistical inference: data integrity, circularity, and morphospace expansion

Evaluating machine learning performance across paleontological datasets requires establishing a clear distinction between empirical target variables and target values derived from proxy formulas. In our primary workflow, the training dataset presents a composite target variable combining directly measured live weights from modern birds with proxy-derived body mass estimates for fossil taxa generated via scaling equations (Tabs. S1, S2, S3, S8). When minimum femoral circumference is simultaneously included as an input feature, this dual usage introduces partial target leakage, causing the algorithm to re-learn deterministic power-law relationships alongside empirical biological scaling. As visually demonstrated in log-transformed feature space (Fig. 8B), the extant portion of the training dataset exhibits natural biological scatter free from circularity, whereas the fossil portion displays strict deterministic alignment, reflecting the underlying equation used to generate the target labels.

This structural dichotomy highlights an unavoidable trade-off in deep-time predictive modeling between statistical independence and morphospace variance. Hence, is it better to prioritize a low-variance dataset without circularity or a high-variance dataset with circularity? We believe that the answer depends on the focus of the methodological approach. What is true is that the extant-only dataset is the only true benchmark for empirical biological scaling, but drastically sacrificing predictive performance (Supplementary Information, Fig. S4). The extant + extinct dataset is valuable not as an empirical ground-truth model, but as a pseudo-synthetic augmentation or domain expansion test to demonstrate how ML models perform across size morphospaces. The benchmark comparisons (Fig. 16; Supplementary Information, Fig. S12) illustrate this trade-off: circular targets reproduce the cQE formula, whereas independent volumetric targets yield biologically realistic scatter without model collapse.

Can machine learning and classical statistical frameworks be evaluated the same way? We believe not. Although both share the same constraints on data integrity, they differ in aim: classical inference seeks to recover the data-generating process, whereas machine learning optimizes out-of-sample prediction (Breiman, 2001a). However, ML is not immune to data leakage. If an input feature 'x' is mathematically used to construct target label 'y', the ML algorithm will learn a deterministic function rather than an empirical biological scaling relationship. Consequently, neither paradigm can bypass valid target construction or honest out-of-sample evaluation. Rather than viewing machine learning as less constrained by these assumptions, performance metrics on mixed datasets must be interpreted as a diagnostic evaluation of algorithmic behavior across extended size ranges rather than an uncompromised measure of biological prediction accuracy. In multivariate configurations, supervised algorithms (such as random forests and neural networks) may partially offset, but do not eliminate, this circularity. When femoral length is included alongside femoral circumference, the models are not restricted to re-

producing the circumference-based formula used to generate the fossil labels: SHAP attributions show that femoral length modulates individual predictions, particularly within some extant clades with directly measured body masses (*e.g.*, Galliformes, Passeriformes; Fig. 15; Lundberg & Lee, 2017). However, because cQE-derived fossil targets depend on femoral circumference alone, any contribution of femoral length to fossil predictions reflects its covariation with circumference rather than independent information on body mass. Circularity is therefore fully removed only when fossil targets are derived independently of the predictors, as in the volumetric-density configuration (Supplementary Information, Fig. S12).

CONCLUSIONS AND NEXT HORIZON

The workflow presented here demonstrates that machine learning can be successfully integrated into paleobiological research when applied with methodological rigor, biological awareness, and full transparency at each analytical stage. The results of the body mass prediction case study yield several conclusions with both methodological and paleobiological relevance.

Both random forest and neural network algorithms achieve high predictive accuracy when trained on datasets that preserve broad morphological and body mass variance. Under these conditions, R^2 values remain consistently high and prediction percentage errors remain low, confirming that non-linear machine learning approaches can effectively recover allometric scaling relationships encoded in femoral dimensions. Femoral circumference consistently emerges as the dominant predictor across all dataset configurations and both algorithms, corroborating its established biomechanical relevance as a primary body mass proxy in bipedal archosaurs.

Algorithmic behavior diverges fundamentally under morphospace restriction. Random forests, being an intrinsically interpolative method, collapses almost entirely when training is restricted to extant taxa, as predictions are bounded by the observed response range and cannot extrapolate to the larger-bodied extinct taxa present in the validation set. Neural networks, by contrast, degrade more gradually under the same conditions, retaining partial predictive capacity due to their capacity to encode smooth continuous functional mappings. This contrast highlights that morphospace coverage of the training distribution, rather than algorithm choice alone, is the primary determinant of generalization performance in deep-time applications. Finally, a comparative analysis between the original and augmented datasets (via SMOGN resampling) revealed consistent performance metrics, indicating that the synthetic instances effectively populated under-represented regions of the morphospace without introducing artificial bias.

Systematic hyperparameter optimization substantially improves model configuration relative to default settings in both algorithms, with optimized parameters differing markedly from defaults in five of seven evaluated hyperparameters for random forests after sensitivity analysis. Neural networks learning curves confirm rapid convergence in low-dimensional regression problems with strong biological signals, and demonstrate that L2 regularization effectively prevents overfitting by aligning training and testing error trajectories throughout optimization.

SHAP-based feature attribution provides mechanistic confirmation of these patterns. Under variance-restricted training, SHAP magnitudes contract without reversing predictor rankings, demonstrating that predictive collapse reflects distributional truncation rather than algorithmic instability or erroneous feature weighting. This result reinforces the interpretive value of combining performance metrics with *post-hoc* explainability tools. Predictions for 30 biped dinosaurs show broad agreement with the circumference-based corrected equation, with a mean deviation of 7,2% for random forests and 9,7% for neural networks across the full sample (Tab. S13). Discrepancies are concentrated at the extremes of the size distribution, particularly among sub-kilogram taxa and among large-bodied forms with atypical femoral proportions, pointing to biologically meaningful rather than arbitrary sources of error. In addition, a promising avenue for future research involves exploring the influence of femoral pneumaticity on the relative robusticity of the femur and, by extension, its impact on the femoral circumference-body mass relationship. Data differentiating pneumatic and non-pneumatic femora were previously evaluated by Campione et al. (2014). Although their study did not recover statistically significant

differential scaling patterns between the two groups using traditional linear allometric models, the integration of such osteological traits into a machine learning framework offers a renewed perspective. Because algorithms such as neural networks and random forests excel at identifying complex, non-linear relationships and higher-order interactions among multiple variables, it is possible that incorporating pneumaticity as an additional categorical predictor could improve overall predictive performance. Exploring the interaction between bone pneumaticity and cross-sectional robusticity represents a logical next step in expanding the multivariate capacity of these models, potentially refining body mass estimations for highly derived or strongly pneumatized theropod lineages.

Unsupervised clustering independently recovers structure consistent with body size gradients and the extant/extinct compositional divide, demonstrating that intrinsic data geometry reflects paleobiologically interpretable signals.

A wide variance morphospace (in this case represented by the inclusion of extinct taxa in training datasets) is highly beneficial for reliable generalization across deep-time size regimes. Restricting training to extant analogs systematically narrows variance structure, increases model bias, and risks near-complete predictive collapse for the large-bodied extinct forms that are often of greatest paleobiological interest. Variance breadth and morphospace representation are therefore more critical to machine learning performance in paleontology than strict adherence to directly measured body masses. In addition, the subsequent decrease in predictive performance demonstrates that these extinct forms occupy a distinct functional morphospace not fully captured by extant analogs. This result justifies the inclusion of fossil data in the training pipeline, ensuring the model learns the unique biomechanical constraints of non-avian dinosaurs rather than merely extrapolating from modern birds.

Addressing statistical circularity through independent volumetric targets represents a crucial step toward establishing data integrity in paleontological machine learning. When fossil body mass labels are derived directly from univariate scaling formulas, algorithms inevitably learn deterministic mathematical relationships alongside biological scaling. Re-evaluating our models using independent 3D volumetric-density reconstructions breaks this target leakage while preserving morphospace expansion across giant size regimes. Combining empirical live weights from modern taxa with independent 3D volumetric reconstructions of extinct forms thus provides a structurally sound and empirically constrained framework for deep-time predictive modeling.

The framework presented here is not limited to body mass prediction or to theropod dinosaurs. The principles of hypothesis-driven design, transparent preprocessing, systematic hyperparameter tuning, cross-validated evaluation, and *post-hoc* interpretability are broadly applicable to any quantitative paleontological problem amenable to supervised regression or classification. We advocate for this integrated approach as a foundation for responsible, reproducible, and scientifically meaningful machine learning in paleontology.

Effective machine learning in paleontology demands that analyses remain hypothesis-driven: the research question must precede algorithm selection, not follow it. Selecting an algorithm first risks producing models that are technically sophisticated but scientifically meaningless (Chollet, 2018; Corea, 2019; Escudero, 2019). High-quality, rigorously curated data are foundational, requiring thorough preprocessing, careful handling of missing values and potential outliers, and exploratory data analysis to detect structural biases inherent to fossil records. Dataset splitting should prevent data leakage by ensuring statistical independence between training-testing, and validation sets, ideally designed at the specimen, locality, or clade level. Transparency and reproducibility require full reporting of data sources, preprocessing steps, hyperparameters, random seeds, and evaluation metrics, enabling future replication as datasets evolve with new discoveries. Model evaluation must be adapted to paleontological constraints, combining cross-validation, and metrics aligned with the research question, while controlling hyperparameter tuning to minimize overfitting. *Post-hoc* interpretability tools are valuable but must be applied cautiously and integrated with domain expertise in biomechanics, phylogenetics, or taphonomy to strengthen scientific validity.

Looking ahead, the rapid proliferation of AI has introduced conceptual confusion alongside genuine analytical

power. Algorithmic effectiveness depends fundamentally on data quality, methodological transparency, and domain knowledge, not on technological enthusiasm. Emerging paradigms such as quantum computing may eventually accelerate optimization and sampling workflows currently bottlenecked by classical hardware (Moret-Bonillo, 2015; Abdelgaber & Nikolopoulos, 2020; Ayoade et al., 2022; Pineda et al., 2025), but will amplify rather than eliminate the need for rigorous scientific design. The future of machine learning in paleontological research lies in the thoughtful integration of statistical reasoning, biological expertise, and transparent methodology, ensuring that predictive modeling contributes meaningfully to scientific discovery rather than merely reproducing it numerically.

Finally, the coexistence between classical inference and predictive machine learning represents a debate that paleontology must inevitably confront as data-driven methodologies continue to permeate deep-time biological research. While individual researchers may differ on whether machine learning frameworks or classical regressions represent the most appropriate philosophy for paleobiological reconstruction, dismissing algorithmic approaches or evaluating them purely through traditional statistical lenses risks stalling methodological innovation.

ACKNOWLEDGMENTS

We would like to express gratitude to Diego Pol (Museo Argentino de Ciencias Naturales) and Néstor Toledo (Museo de La Plata) for meaningful methodological discussions on machine learning and statistics. We also thank the guest editors of this volume for considering this contribution. Nicolás Campione and an anonymous reviewer greatly improved the quality of this contribution.

Data archiving statement

The code and datasets used in this study are available on GitHub (https://github.com/aleotero80/otero_borya_body_mass) and permanently archived in Zenodo (https://doi.org/10.5281/zenodo.24130098).

Disclosure statement

The authors declare no conflict of interest.

REFERENCES

Abdelgaber, N., & Nikolopoulos, C. (2020). Overview on quantum computing and its applications in artificial intelligence. *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 198–199). Laguna Hills.

Aggarwal, C. C. (2018). *Neural networks and deep learning*. Springer.

Alonso, J., Villa, A., & Bahamonde, A. (2015). Improved estimation of bovine weight trajectories using Support Vector Machine Classification. *Computers and Electronics in Agriculture*, *110*, 36–41.

Anderson, J. F., Hall-Martin, A., & Russell, D. A. (1985). Long-bone circumference and weight in mammals, birds and dinosaurs. *Journal of Zoology*, *207*, 53–61.

Apostol, L. A., Márquez, E., Gasmen, P., & Solano, G. (2016). Radss: A radiolarian classifier using support vector machines. *7th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–6). Chalkidiki.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, *7*, 1027–1035.

Ayoade, O., Rivas, P., & Orduz, J. (2022). Artificial intelligence computing at the quantum level. *Data*, *7*(3), 28.

Bahn, J., Alférez, G. H., & Snyder, K. (2025). Machine learning classification of fossilized *Pectinodon bakkeri* teeth images: insights into troodontid theropod dinosaur morphology. *Machine Learning and Knowledge Extraction*, *7*(2), 45.

Bakker, R. T. (1972). Anatomical and ecological evidence of endothermy in dinosaurs. *Nature*, *238*(5359), 81–85.

Ballell, A., Benton, M. J., & Rayfield, E. J. (2022). Dental form and function in the early feeding diversification of dinosaurs. *Science Advances*, *8*(50), eabq5201.

Barucci, A., Ciacci, G., Lio, P., Azevedo, T., Di Cencio, A., Merella, M., & Collareta, A. (2024). An explainable Convolutional Neural Network approach to fossil shark tooth identification. *Bollettino della Società Paleontologica Italiana*, *63*(3), 215–227.

Beltrán-Royo, C., Llopis-Ibor, L., Pantrigo, J. J., & Ramirez, I. (2024). DC Neural Networks avoid overfitting in one-dimensional nonlinear regression. *Knowledge-Based Systems*, *283*, 111154.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1), 151–175.

Benson, R. B., 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.

Benson, R. B., Campione, N. E., Carrano, M. T., Mannion, P. D., Sullivan, C., Upchurch, P., & Evans, D. C. (2014). Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. *PLoS Biology*, *12*(5), e1001853.

Benton, M. J. (1995). Diversification and extinction in the history of life. *Science*, *268*, 52–58.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, *24*, 1–9.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Blanc, G., Gupta, N., Valiant, G., & Valiant, P. (2020). Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. *Proceedings of Machine Learning Research*, *125*, 483–513.

Borya, A., Henao, K., & Ruiz-Graham, C. (2025). Automating Core Data Integration: Insights from the Volve Field Case Study. *Abu Dhabi International Petroleum Exhibition and Conference* (p. D041S138R004). Abu Dhabi.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, *49*(2), 1–50.

Branco, P., Torgo, L., & Ribeiro, R. (2017). SMOGN: A pre-processing approach for imbalanced regression. *Proceedings of Machine Learning Research*, *74*, 36–50.

Breiman, L. (2001a). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*, 5–32.

Brown, C. M., Campione, N. E., Mantilla, G. P. W., & Evans, D. C. (2022). Size-driven preservational and macroecological biases in the latest Maastrichtian terrestrial vertebrate assemblages of North America. *Paleobiology*, *48*(2), 210–238.

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.

Buduma, N., Buduma, N., & Papa, J. (2022). *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. O'Reilly Media.

Burness, G. P., Diamond, J., & Flannery, T. (2001). Dinosaurs, dragons and dwarfs: the evolution of maximal body size. *Proceedings of the National Academy of Sciences*, *98*(25), 14518–14523.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233.

Campione, N. E. (2017). Extrapolating body masses in large terrestrial vertebrates. *Paleobiology*, *43*(4), 693–699.

Campione, N. E., & Evans, D. C. (2012). A universal scaling relationship between body mass and proximal limb bone dimensions in quadrupedal terrestrial tetrapods. *BMC Biology*, *10*, 60.

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.

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.

Carrano, M. T. (2001). Implications of limb bone scaling, curvature and eccentricity in mammals and non-avian dinosaurs. *Journal of Zoology*, *254*(1), 41–55.

Carrano, M. T. (2005). The evolution of sauropod locomotion: morphological diversity of a secondarily quadrupedal radiation. In K. A. Curry Rogers & J. A. Wilson (Eds.), *The sauropods: Evolution and paleobiology* (pp. 229–249). University of California Press.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 655–670). Dublin.

Casella, G., & Berger, R. (2024). *Statistical inference*. Chapman and Hall/CRC.

Cassini, G. H., Flores, D. A., & Vizcaíno, S. F. (2012). Postnatal ontogenetic scaling of Nesodontine (Notoungulata, Toxodontidae) cranial morphology. *Acta Zoologica*, *93*(3), 249–259.

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, *10*(8), 1283.

Chatfield, C. (1986). Exploratory data analysis. *European Journal of Operational Research*, *23*(1), 5–13.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., & Yang, Z. (2023). Machine learning methods in weather and climate applications: A survey. *Applied Sciences*, *13*(21), 12019.

Chollet, F. (2015). Keras. GitHub. https://github.com/fchollet/keras

Chollet, F. (2018). *Deep learning with Python*. Manning Publications.

Christiansen, P. (1999). Long bone scaling and limb posture in non-avian theropods: evidence for differential allometry. *Journal of Vertebrate Paleontology*, *19*(4), 666–680.

Christiansen, P., & Fariña, R. A. (2004). Mass prediction in theropod dinosaurs. *Historical Biology*, *16*(2–4), 85–92.

Cifuentes-Alcobendas, G., & Domínguez-Rodrigo, M. (2019). Deep learning and taphonomy: high accuracy in the classification of cut marks made on fleshed and defleshed bones using convolutional neural networks. *Scientific Reports*, *9*(1), 18933.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(1), e1391.

Corea, F. (2019). Introduction to data. In *An Introduction to Data: Everything You Need to Know About AI, Big Data and Data Science* (pp. 1–5). Springer.

da Conceição, D. M., Júnior, M. G. E., Iannuzzi, R., Recamonde-Mendoza, M., & Malta, G. B. (2023). PaleoWood: A machine learning approach for determining the affinity of Paleozoic gymnosperm woods. *Journal of South American Earth Sciences*, *121*, 104125.

Dahl, T. W., Hammarlund, E. U., Anbar, A. D., Bond, D. P., Gill, B. C., Gordon, G. W., Knoll, A. H., Nielsen, A. T., Schovsbo, N. H., & Canfield, D. E. (2010). Devonian rise in atmospheric oxygen correlated to the radiations of terrestrial plants and large predatory fish. *Proceedings of the National Academy of Sciences*, *107*(42), 17911–17915.

Damuth, J. D., & MacFadden, B. J. (1990). *Body size in mammalian paleobiology: estimation and biological implications*. Cambridge University Press.

Dempsey, M., Cross, S. R., Maidment, S. C., Hutchinson, J. R., & Bates, K. T. (2025). New perspectives on body size and shape evolution in dinosaurs. *Biological Reviews*, *100*, 1829–1860.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). Miami.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems* (pp. 1–15). Cagliari.

Dillon, E. M., Dunne, E. M., Womack, T. M., Kouvari, M., Larina, E., Claytor, J. R., & Zill, M. E. (2023). Challenges and directions in analytical paleobiology. *Paleobiology*, *49*(3), 377–393.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marqués, J. R., Gruber, B., Lafourcade, B., Leitão, P., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46.

Drobnič, F., Kos, A., & Pustišek, M. (2020). On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics*, *9*(5), 761.

Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in

compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, *16*, e01059.

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. *Informatics*, *8*(4), 79.

Erickson, G. M., Rauhut, O. W., Zhou, Z., Turner, A. H., Inouye, B. D., Hu, D., & Norell, M. A. (2009). Was dinosaurian physiology inherited by birds? Reconciling slow growth in *Archaeopteryx*. *PLoS One*, *4*(10), e7390.

Escudero, W. S. (2019). *Big data: breve manual para conocer la ciencia de datos que ya invadió nuestras vidas*. Siglo XXI Editores.

Esperança Júnior, M. G. F., Iannuzzi, R., & da Conceição, D. M. (2026). Support vector machine-based classification of paleozoic gymnosperm woods. *Publicación Electrónica de la Asociación Paleontológica Argentina*, *26*(3), 1–14.

Fan, L., Xu, C., Jarzembowski, E. A., & Cui, X. (2022). Quantifying plant mimesis in fossil insects using deep learning. *Historical Biology*, *34*(5), 907–916.

Farlow, J. O. (1976). A consideration of the trophic dynamics of a Late Cretaceous large-dinosaur community (Oldman Formation). *Ecology*, *57*(5), 841–857.

Farlow, J. O., Smith, M. B., & Robinson, J. M. (1995). Body mass, bone «strength indicator,» and cursorial potential of *Tyrannosaurus rex*. *Journal of Vertebrate Paleontology*, *15*(4), 713–725.

Field, D. J., Lynner, C., Brown, C., & Darroch, S. A. (2013). Skeletal correlates for body mass estimation in modern and fossil flying birds. *PLoS One*, *8*(11), e82000.

Forcellati, M. R., Napoli, J. G., Meyer, D., Watanabe, A., Benson, R., & Raxworthy, C. J. (2026). Machine learning can accurately assign fossil and extant species to crown toxicoferan (Reptilia: Squamata) groups using inner ear shape data. *Zoological Journal of the Linnean Society*, *206*(3), zlaf188.

Foster, W. J., Ayzel, G., Münchmeyer, J., Rettelbach, T., Kitzmann, N. H., Isson, T. T., & Aberhan, M. (2022). Machine learning identifies ecological selectivity patterns across the end-Permian mass extinction. *Paleobiology*, *48*(3), 357–371.

Fujihara, K., Yamada Harada, M., Horikawa, C., Iwanaga, M., Tanaka, H., Nomura, H., Sui, Y., Tanabe, K., Yamada, T., Kodama, S., Kato, K., & Sone, H. (2023). Machine learning approach to predict body weight in adults. *Frontiers in Public Health, 11*, 1090146.

Garg, A., & Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *International Journal of Modelling, Identification and Control*, *18*(4), 295–312.

Gebreyesus, Y., Dalton, D., Nixon, S., De Chiara, D., & Chinnici, M. (2023). Machine learning for data center optimizations: feature selection using Shapley additive exPlanation (SHAP). *Future Internet*, *15*(3), 88.

Gidel, G., Bach, F., & Lacoste-Julien, S. (2019). Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, *32*, *1–10*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grabowski, A., Farley, C. T., & Kram, R. (2005). Independent metabolic costs of supporting body weight and accelerating body mass during walking. *Journal of Applied Physiology*, *98*(2), 579–583.

Ha, Y., & Kim, S. S. (2023). Classification of large ornithopod dinosaur footprints using Xception transfer learning. *PLoS One*, *18*(11), e0293020.

Hammer, Ø., & Harper, D. A. (2024). *Paleontological data analysis*. John Wiley & Sons.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., …, Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362.

Harrison, J. F., Kaiser, A., & VandenBrooks, J. M. (2010). Atmospheric oxygen level and the evolution of insect body size. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1690), 1937–1946.

Hartmann, G., Blakesley, T., dePolo, P. E., & Brusatte, S. L. (2026). Identifying variation in dinosaur footprints and classifying problematic specimens via unbiased unsupervised machine learning. *Proceedings of the National Academy of Sciences*, *123*(5), e2527222122.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Springer.

Hendrickx, C., Trapman, T. H., Wills, S., Holwerda, F. M., Stein, K. H., Rauhut, O. W., Melzer, R. R., Woensel, J. V., & Reumer, J. W. (2023). A combined approach to identify isolated theropod teeth from the Cenomanian Kem Kem Group of Morocco: cladistic, discriminant, and machine learning analyses. *Journal of Vertebrate Paleontology*, *43*(4), e2311791.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Hone, D. W. (2012). Variation in the tail length of non-avian dinosaurs. *Journal of Vertebrate Paleontology*, *32*(5), 1082–1089.

Hone, D. W. E., Dyke, G. J., Haden, M., & Benton, M. J. (2008). Body size evolution in Mesozoic birds. *Journal of Evolutionary Biology*, *21*(2), 618–624.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hsiang, A. Y., Brombacher, A., Rillo, M. C., Mleneck-Vautravers, M. J., Conn, S., Lordsmith, S., Jentzen, A., Henehan, M. J., Metcalfe, B., Fenton, I. S., Wade, B. S., Fox, L., Meilland, J., Davis, C. V., Baranowski, U., Groeneveld, J., Edgar, K. M., Movellan, A., Aze, T., …, Hull, P. M. (2019). Endless forams: >34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. *Paleoceanography and Paleoclimatology*, *34*, 1157–1177.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Hutchinson, J. R., & Garcia, M. (2002). *Tyrannosaurus* was not a fast runner. *Nature*, *415*, 1018–1021.

Hutchinson, J. R., Bates, K. T., Molnar, J., Allen, V., & Makovicky, P. J. (2011). A computational analysis of limb and body dimensions in *Tyrannosaurus rex* with implications for locomotion, ontogeny, and growth. *PLoS One*, *6*(10), e26037.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, *37*, 448–456.

Januzaj, Y., Beqiri, E., & Luma, A. (2023). Determining the optimal number of clusters using silhouette score as a data mining technique. *International Journal of Online & Biomedical Engineering*, *19*(4), 174–182.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

Jones, M., Lallensack, J. N., Jarman, I., Falkingham, P., & Siekmann, I.

(2024). Classification of dinosaur footprints using machine learning. *Journal of Vertebrate Paleontology*, *44*(6), e2493157.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2019). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on knowledge and data engineering*, *31*(8), 1544–1554.

Keçeli, A. S., Kaya, A., & Keçeli, S. U. (2017). Classification of radiolarian images with handcrafted and deep features. *Computers & Geosciences*, *109*, 67–74.

Kelleher, J. D., & Tierney, B. (2018). *Data Science*. MIT Press.

Ketkar, N., & Moolayil, J. (2021). *Deep learning with Python: learn best practices of deep learning models with PyTorch*. Apress.

Knutsen, E. M., & Konovalov, D. A. (2024). Accelerating segmentation of fossil CT scans through Deep Learning. *Scientific Reports*, *14*(1), 20943.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, *90*(1), 3–14.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232.

Kyriazos, T., & Poga, M. (2023). Dealing with multicollinearity in factor analysis: the problem, detections and solutions. *Open Journal of Statistics*, *13*(3), 404–424.

LaBarbera, M. (1989). Analyzing body size as a factor in ecology and evolution. *Annual Review of Ecology and Systematics*, *20*, 97–117.

Lallensack, J. N., Romilio, A., & Falkingham, P. L. (2022). A machine learning approach for the discrimination of theropod and ornithischian dinosaur tracks. *Journal of the Royal Society Interface*, *19*(196), 20220588.

Lamboy, W., & Lesnikowska, A. (1988). Some statistical methods useful in the analysis of plant paleoecological data. *Palaios*, *3*, 86–94.

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv*, 1912.06059.

Linchamps, P., Stoetzel, E., Robinet, F., Hanon, R., Latouche, P., & Cornette, R. (2023). Bioclimatic inference based on mammal community using machine learning regression models: perspectives for paleoecological studies. *Frontiers in Ecology and Evolution*, *11*, 1178379.

Liu, J., Xiao, D., Liu, Y., & Huang, Y. (2023b). A pig mass estimation model based on deep learning without constraint. *Animals, 13(8)*, 1376.

Liu, M., Prentice, I. C., Ter Braak, C. J., & Harrison, S. P. (2020). An improved statistical approach for reconstructing past climates from biotic assemblages. *Proceedings of the Royal Society A*, *476*(2243), 20200346.

Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y., & Song, H. (2023a). Automatic taxonomic identification based on the Fossil Image Dataset (>415,000 images) and deep convolutional neural networks. *Paleobiology*, *49*, 1–22.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 1–10.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*, 56–67.

MacFadden, B. J., Barberis, C. A., Vallejo-Pareja, M. C., Zbinden, S. P., Perez, V. J., Killingsworth, S. R., & Porto, A. (2026). AI and paleontology: effects of vertebrate fossil sample size on machine learning image classification. *Paleobiology*, *52*(3), 1–17.

MacTavish, J. N., Malone, P. G., & Wells, T. L. (1968). RMAR: a reduced major axis regression program designed for paleontologic data. *Journal of Paleontology*, *42*(4), 1076–1078.

Madhav, S. K., & HimanshuRawat, P. J. (2021). A brief study on random forests using Python. *International Journal of Advances in Engineering and Management (IJAEM)*, *3*(6), 2063–2069.

Mannion, P. D., Benson, R. B., Carrano, M. T., Tennant, J. P., Judd, J., & Butler, R. J. (2015). Climate constrains the evolutionary history and biodiversity of crocodylians. *Nature Communications*, *6*(1), 8438.

Marques, C. S., Malafaia, E., Pereira, S., Santos, V. F., & Dufourq, E. (2025a). A review of machine learning applications for identification and classification problems in paleontology. *Ecological Informatics*, *91*, 103329.

Marques, C. S., Dufourq, E., Pereira, S., Santos, V. F., & Malafaia, E. (2025b). Enhancing the classification of isolated theropod teeth using machine learning: a comparative study. *PeerJ*, *13*, e19116.

Marques, C. S., Dufourq, E., Pereira, S., Santos, V. F., & Malafaia, E. (2026a). Automated multi-view silhouette generation and morphometric clustering of isolated theropod teeth using machine learning. *Publicación Electrónica de la Asociación Paleontológica Argentina*, *26*(3), 15–43.

Marques, C. S., Mota, A., Belvedere, M., Castanera, D., Díaz-Martínez, I., Malafaia, E., & Dufourq, E. (2026b). Deep tracks: Using deep learning and procedurally simulated data for automated vertebrate footprints classification. *Ecological Informatics*, *93*, 103523.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817.

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

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, *445*(1), 51–56.

Mendoza, M., Janis, C. M., & Palmqvist, P. (2006). Estimating the body mass of extinct ungulates: a study on the use of multiple regression. *Journal of Zoology*, *270*(1), 90–101.

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, *12*(9), 1620–1633.

Millar, M. (2019). Predicting theropod hunting tactics using machine learning. *Open Science Journal*, *4*(1), 1–11.

Miller, R. L. (1949). An application of the analysis of variance to paleontology. *Journal of Paleontology*, *23*(6), 635–640.

Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models interpretable*. Lulu Press.

Monchot, H., & Léchelle, J. (2002). Statistical nonparametric methods for the study of fossil populations. *Paleobiology*, *28*(1), 55–69.

Moret-Bonillo, V. (2015). Can artificial intelligence benefit from quantum computing? *Progress in Artificial Intelligence*, *3*(2), 89–105.

Mukhiya, S. K., & Ahmed, U. (2020). *Hands-on exploratory data analysis with Python*. Packt Publishing.

Müller, A. C., & Guido, S. (2017). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Murphy, K. P. (2022). *Probabilistic machine learning: Advanced topics*. MIT Press.

Nicodemus, K. K., & Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, *25*(15), 1884–1890.

Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.

Ning, Y., Ong, M. E. H., Chakraborty, B., Goldstein, B. A., Ting, D. S. W., Vaughan, R., & Liu, N. (2022). Shapley variable importance cloud for interpretable machine learning. *Patterns*, *3*(4), 100452.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, *214*, 106584.

Otero, A. (2025). Machine learning reveals predictor versatility for assessing body mass in theropods (Dinosauria, Saurischia). *Publicación Electrónica de la Asociación Paleontológica Argentina*, *26*(R3), R163.

Otero, A., Cuff, A. R., Allen, V., Sumner-Rooney, L., Pol, D., & Hutchinson, J. R. (2019). Ontogenetic changes in the body plan of the sauropodomorph dinosaur *Mussaurus patagonicus* reveal shifts of locomotor stance during growth. *Scientific Reports*, *9*(1), 1–10.

Ozer, I., Kocak, I., Cetin, O., Karaca, A. C., Ozer, C. K., & Gorur, K. (2023a). Towards investigation of transfer learning framework for Globotruncanita genus and Globotruncana genus microfossils in genus-level and species-level prediction. *Engineering Science and Technology, an International Journal*, *48*, 101589.

Ozer, I., Ozer, C. K., Karaca, A. C., Gorur, K., Kocak, I., & Cetin, O. (2023b). Species-level microfossil identification for *Globotruncana* genus using hybrid deep learning algorithms from the scratch via a low-cost light microscope imaging. *Multimedia Tools and Applications*, *82*(9), 13689–13718.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pineda, V. G., Valencia-Arias, A., Giraldo, F. E. L., & Zapata-Ochoa, E. A. (2025). Integrating artificial intelligence and quantum computing: A systematic literature review of features and applications. *International Journal of Cognitive Computing in Engineering*, *7*, 26–39.

Pintore, R., Hutchinson, J. R., Bishop, P. J., Tsai, H. P., & Houssaye, A. (2024). The evolution of femoral morphology in giant non-avian theropod dinosaurs. *Paleobiology*, *50*(2), 308–330.

Pires de Lima, R., & Marfurt, K. (2020). Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sensing*, *12*(1), 86.

Pires de Lima, R., Welch, K. F., Barrick, J. E., Marfurt, K. J., Burkhalter, R., Cassel, M., & Soreghan, G. S. (2020). Convolutional neural networks as an aid to biostratigraphy and micropaleontology: a test on late Paleozoic microfossils. *Palaios*, *35*(9), 391–402.

Prince, S. J. (2023). *Understanding deep learning*. MIT Press.

Python Software Foundation. Python Language Reference, version 3.12. Available at http://www.python.org

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. MIT Press.

Radinsky, L. B. (1987). *The evolution of vertebrate design*. University of Chicago Press.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development and treatment. *Medicina*, *56*(9), 455.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing.

Raup, D. M., & Sepkoski, J. J. (1982). Mass extinctions in the marine fossil record. *Science*, *215*, 1501–1503.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine*, *47*(1), 31–39.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rudwick, M. J. S. (1964). The inference of function from structure in fossils. *The British Journal for the Philosophy of Science*, *15*(57), 27–40.

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*(2), 272.

Sadeqi-Arani, Z., & Kadkhodaie, A. (2023). A bibliometric analysis of the application of machine learning methods in the petroleum industry. *Results in Engineering*, *20*, 101518.

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 4727–4735.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, *48*, 1875–1897.

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, *12*(10), 1687.

Sepkoski, J. J. (1978). A kinetic model of Phanerozoic taxonomic diversity I. Analysis of marine orders. *Paleobiology*, *4*, 223–251.

Sepkoski, J. J. (1979). A kinetic model of Phanerozoic taxonomic diversity II. Early Phanerozoic families and multiple equilibria. *Paleobiology*, *5*, 222–251.

Sepkoski, J. J. (1984). A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. *Paleobiology*, *10*, 246–267.

Serrano, F. J., Palmqvist, P., & Sanz, J. L. (2015). Multivariate analysis of neognath skeletal measurements: implications for body mass estimation in Mesozoic birds. *Zoological Journal of the Linnean Society*, *173*(4), 929–955.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, *23*(6), 759.

Silvestro, D., & Pimiento, C. (2025). Emerging uses of artificial intelligence in deep time biodiversity research. *Nature Reviews Biodiversity*, *1*(10), 671–677.

Simpson, G. G. (1944). *Tempo and mode in evolution*. Columbia University Press.

Singh, J., Singh, J., Singh, G., & Kaur, N. (2022). Exploratory data analysis for interpreting model prediction using Python. In *2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)* (pp. 1–6). Bangalore.

Smith, R. J. (2002). Estimation of body mass in paleontology. *Journal of Human Evolution*, *43*(2), 271–287.

Snowdon, P. (1991). A ratio estimator for bias correction in logarithmic regressions. *Canadian Journal of Forest Research*, *21*(5), 720–724.

Sokal, R. R., & Rohlf, F. J. (1987). *Biometry*. Freeman.

Song, Z., Cao, S., & Yang, H. (2024). An interpretable framework for modeling global solar radiation using tree-based ensemble machine learning and Shapley additive explanations methods. *Applied Energy*, *364*, 123238.

Sookias, R. B., Butler, R. J., & Benson, R. B. (2012a). Rise of dinosaurs reveals major body-size transitions are driven by passive processes of trait evolution. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1736), 2180–2187.

Sookias, R. B., Benson, R. B., & Butler, R. J. (2012b). Biology, not environment, drives major patterns in maximum tetrapod body size through time. *Biology Letters*, *8*(4), 674–677.

Soulaimani, S., Soulaimani, A., Abdelrahman, K., Miftah, A., Fnais, M. S., & Mondal, B. K. (2024). Advanced machine learning artificial neural network classifier for lithology identification using Bayesian optimization. *Frontiers in Earth Science*, *12*, 1473325.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307.

Sugiyama, M., Krauledat, M., & Müller, K. R. (2007). Covariate shift adaptation by importance weighting. *Journal of Machine Learning Research*, *8*(5), 985–1005.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., & John, A. (2022). A review of earth artificial intelligence. *Computers & Geosciences*, *159*, 105034.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, *14*(4), 1502–1509.

Tahir, M., Khan, S. S., Davie, J., Yamanaka, S., & Ashraf, A. (2025). LOCO-EPI: Leave-one-chromosome-out (LOCO) as a benchmarking paradigm for deep learning based prediction of enhancer-promoter interactions. *Applied Intelligence*, *55*(1), 71.

Telford, R. J., & Birks, H. J. B. (2011). A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages. *Quaternary Science Reviews*, *30*(9–10), 1272–1278.

Tong, L., & Zhang, Z. (2026). Neural network analysis of psychological data: a step-by-step guide. *Multivariate Behavioral Research*, *61*(2), 399–419.

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*(6), e03336.

Vizcaíno, S. F., Bargo, M. S., Cassini, G. H., & Toledo, N. (2016). *Forma y función en paleobiología de vertebrados*. EDULP, Universidad Nacional de La Plata.

Voje, K. L., Saulsbury, J. G., Starrfelt, J., Latorre, D. V., Rojas, A., Kinneberg, V. B., & Grabowski, M. (2023). Measurement theory and paleobiology. *Trends in Ecology & Evolution*, *38*(12), 1165–1176.

Wang, B., Sun, R., Yang, X., Niu, B., Zhang, T., Zhao, Y., & Han, J. (2022). Recognition of rare microfossils using transfer learning and deep residual networks. *Biology*, *12*(1), 16.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wei, C., Kakade, S., & Ma, T. (2020). The implicit and explicit regularization effects of dropout. *Proceedings of Machine Learning Research*, *119*, 10181–10192.

Whitlock, J. A. (2011). Inferences of diplodocoid (Sauropoda: Dinosauria) feeding behavior from snout shape and microwear analyses. *PLoS One*, *6*(4), e18304.

Wicaksono, A. S., & Supianto, A. A. (2018). Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction. *International Journal of Advanced Computer Science and Applications*, *9*(12), 263–267.

Wills, S., & Poon, S. T. S. (In press). AI in Paleontology: overcoming data constraints using 2.5D CT deep learning models. *Publicación Electrónica de la Asociación Paleontológica Argentina*. https://doi.org/10.5710/PEAPA.31.08.2026.578

Wills, S., Underwood, C. J., & Barrett, P. M. (2021). Learning to see the wood for the trees: machine learning, decision trees, and the classification of isolated theropod teeth. *Palaeontology*, *64*(1), 75–99.

Wills, S., Underwood, C. J., & Barrett, P. M. (2023). Machine learning confirms new records of maniraptoran theropods in Middle Jurassic UK microvertebrate faunas. *Papers in Palaeontology*, *9*(2), e1487.

Wu, Q., & Vos, P. (2018). Inference and prediction. *Handbook of Statistics*, *38*, 111–172.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., & Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, *2*(4), 100179.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316.

Yaqoob, M., Ishaq, M., Ansari, M. Y., Qaiser, Y., Hussain, R., Rabbani, H. S., & Seers, T. D. (2025). Advancing paleontology: a survey on deep learning methodologies in fossil image analysis. *Artificial Intelligence Review*, *58*(3), 83.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*, 022022.

Yu, C., Qin, F., Li, Y., Qin, Z., & Norell, M. (2022). CT segmentation of dinosaur fossils by deep learning. *Frontiers in Earth Science*, *9*, 805271.

Yu, C., Qin, F., Watanabe, A., Yao, W., Li, Y., Qin, Z., & Xu, X. (2024). Artificial intelligence in paleontology. *Earth-Science Reviews*, *252*, 104765.

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

Zhong, B., Ge, Q., Kanakiya, B., Marchitto, R. M. T., & Lobaton, E. (2017). A comparative study of image classification algorithms for Foraminifera identification. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1–8). Honolulu.

doi: 10.5710/PEAPA.09.11.2026.576

Recibido: 23 de abril de 2026
Aceptado: 09 de noviembre de 2026
Publicado: 24 de noviembre de 2026

Acceso Abierto
Open Access
This work is licensed under
CC BY 4.0