Automatic identification and morphological comparison of bivalve and brachiopod fossils based on deep learning

Jiarui Sun [1], **Xiaokang Liu** [1,2], **Yunfei Huang** [3], **Fengyu Wang** [1], **Yongfang Sun** [1], **Jing Chen** [4], **Daoliang Chu** [1], **Haijun Song** [Corresp. 1]

[1] State Key Laboratory of Biogeology and Environmental Geology, School of Earth Sciences, China University of Geosciences, Wuhan, Hubei, China

[2] Department of Biology, University of Fribourg, Fribourg, Switzerland

[3] School of Geosciences, Yangtze University, Wuhan, Hubei, China

[4] Yifu Museum, China University of Geosciences, Wuhan, Hubei, China

Corresponding Author: Haijun Song
Email address: haijunsong@cug.edu.cn

Fossil identification is an essential and fundamental task for conducting palaeontological research. However, the artificial identification of fossils requires extensive experience and is time-consuming. The process is complex, tedious, and susceptible to subjective factors. In this study, an automatic identification model was established for bivalve and brachiopod fossils using deep learning. We built an available bivalve and brachiopod fossil image dataset (containing > 16,000 "image-label" data pairs) and completed the taxonomic determination to facilitate other researchers. We achieved > 80% identification accuracy at 22 genera and ~64% accuracy at 343 species using EfficientNetV2s architecture. We extracted the intermediate output of the model as fossil features and downscaled them to demonstrate the morphological feature space of fossils using t-distributed stochastic neighbor embedding (t-SNE). We found a distinctive boundary between the morphological feature points of bivalves and brachiopods. This study provides a possible method for studying the morphology evolution of fossil clades using computer vision in the future.

Automatic identification and morphological comparison of bivalve and brachiopod fossils based on deep learning

Jiarui Sun[1], Xiaokang Liu[1,2], Yunfei Huang[3], Fengyu Wang[1], Yongfang Sun[1], Jing Chen[4], Daoliang Chu[1], Haijun Song[1]*

[1] State Key Laboratory of Biogeology and Environmental Geology, School of Earth Sciences, China University of Geosciences, Wuhan, 430074, China

[2] Department of Biology, University of Fribourg, Fribourg, Switzerland

[3] School of Geosciences, Yangtze University, Wuhan 430100, China

[4] Yifu Museum, China University of Geosciences, Wuhan, 430074, China

*Corresponding author:

Haijun Song[1]

State Key Laboratory of Biogeology and Environmental Geology, School of Earth Sciences, China University of Geosciences, Wuhan, 430074, China

Email address: haijunsong@cug.edu.cn

Abstract

Fossil identification is an essential and fundamental task for conducting palaeontological research. However, the artificial identification of fossils requires extensive experience and is time-consuming. The process is complex, tedious, and susceptible to subjective factors. In this study, an automatic identification model was established for bivalve and brachiopod fossils using deep learning. We built an available bivalve and brachiopod fossil image dataset (containing > 16,000 "image-label" data pairs) and completed the taxonomic determination to facilitate other researchers. We achieved > 80% identification accuracy at 22 genera and ~64% accuracy at 343 species using EfficientNetV2s architecture. We extracted the intermediate output of the model as fossil features and downscaled them to demonstrate the morphological feature space of fossils using t-distributed stochastic neighbor embedding (t-SNE). We found a distinctive boundary between the morphological feature points of bivalves and brachiopods. This study provides a possible method for studying the morphology evolution of fossil clades using computer vision in the future.

Key words:

Fossil identification; Machine learning; Invertebrate; Morphology; Convolutional neural network.

Introduction

Fossil identification is a fundamental task in palaeontological research and has a wide range

of applications, including ~~stratigraphic~~ dating (Yin *et al.* 2001; Gradstein *et al.* 2012), biological evolution (Alroy *et al.* 2008; Fan *et al.* 2020; Song *et al.* 2021), palaeoenvironmental reconstruction (Flügel and Munnecke 2010; Scotese *et al.* 2021), and palaeoelevational estimation (Su *et al.* 2019). Because taxonomic identification requires a large amount of prior knowledge as a foundation, researchers need several years of training to accumulate enough experience to ensure the reliability of identification. However, the actual identification process still takes considerable time and is susceptible to subjective factors. The identification accuracy of some genera is even lower than 80% (Hsiang *et al.* 2019). In many fields of palaeontology, deep convolutional neural network (DCNN) has a significant advantage over humans, such as the identification of cut and trampling marks on bones (Byeon *et al.* 2019). To reduce the workload and work difficulty for researchers, automatic fossil identification methods relying on machine learning have been proposed extensively in recent years, among which models using convolutional neural networks (CNNs) [e.g., VGG-16 (Simonyan and Zisserman 2014), Inception-ResNet (Szegedy *et al.* 2017), GoogLeNet (Szegedy *et al.* 2015), etc.] have achieved good results (Dionisio *et al.* 2020; Niu and Xu 2022; Wang *et al.* 2022; Liu *et al.* 2022; Liu and Song 2020). This method can assist researchers in fossil identification, reduce the work stress of non-palaeontologists, and enable better identification and application of fossil materials in research. Furthermore, for identifying poorly preserved fossils, neural networks still maintain high identification accuracy (Bourel *et al.* 2020). Neural network in fossil identification is still at an early stage of development and cannot yet fully reach the identification level of professional palaeontologists. Neural network can provide a useful aid to manual identification rather than replace it, at least for now. It is still worth studying,

and as the level of the model improves and more training data are available, its accuracy will become higher.

The training of automatic taxonomy identification models (ATIM) requires a large dataset of labelled fossil images, which is still insufficient compared to general machine-learning datasets containing millions of items (Liu*, et al.* 2022). The lack of high-resolution (genus-level) fossil labels in the field of palaeontology is mainly due to the tedious and time-consuming process of dataset building. Machine learning has now achieved good results in fossil identification (above the genus level). Liu and Song (2020) achieved 95% accuracy for 22 fossil and abiotic grain groups during carbonate microfacies analysis. While 90% accuracy was achieved in the automatic identification of 50 fossil clades relying on web crawlers (Liu *et al.* 2022), genus- and species-level automatic identification focused mainly on a few taxa (mostly < 10). Dionisio *et al.* (2020) performed automatic identification of 9 radiolarian genera, obtaining 91.85% accuracy. Wang *et al.* (2022) used a Transpose Convolutional Neural Network to achieve 97% accuracy for 5 brachiopod species and was based on a small dataset. Niu and Xu (2022) performed automatic identification of fossils covering 113 graptolite species or subspecies. However, similar studies targeting a large number of taxa are less common. In addition, these studies all focus on the same fossil clade (e.g., radiolarians, brachiopods, etc.), and it is unclear whether mixed categories and large numbers of taxa can achieve automatic fossil identification.

Brachiopods and bivalves are the two most common invertebrate clades in the Phanerozoic (Sepkoski 1981; Clapham *et al.* 2006; Benton and Harper 2020). The similarities and differences between them in morphology and physiological mechanisms have long attracted the attention of

palaeontologists (Ballanti *et al.* 2012; Payne *et al.* 2014); however, the similar morphological features between them have caused problems for researchers to identify them accurately. Automatic identification of brachiopods has been carried out previously. Wang*, et al.* (2022) used the transposed convolutional neural network to realize the automatic identification of fossils with a relatively small dataset and they achieved 97% identification accuracy of five brachiopod species based on 630 training images. In this study, we built a "Bivalve and Brachiopod Fossil Image Dataset" (BBFID) (16,596 labelled fossil images covering 870 genera and 2033 species) for the first time by collecting and sorting a large amount of published literature. We built ATIMs using transfer learning in VGG-16 (Simonyan and Zisserman 2014), Inception-ResNet-v2 (Szegedy*, et al.* 2017), and EfficientNetV2s (Tan and Le 2021) architectures, which have performed well in general identifications. Furthermore, we extracted the process outputs of the model as fossil features and downscaled them to two-dimensional data using t-SNE (Van der Maaten and Hinton 2008). Plotting them in a two-dimensional space is an effective way to compare morphological differences between bivalves and brachiopods.

Materials and Data

The BBFID used for training ATIMs contains bivalve-part (BBFID-1) and brachiopod-part (BBFID-2), all collected from published literature and monographs (see Appendix). Detailed data on the number of each taxon are given in the Appendix (Table S1, S2).

We used Adobe Acrobat Pro DC to capture accurately named bivalve and brachiopod fossil images (mainly Permian and Triassic) from the collected literature and saved them as bmp, jpg, or

102 png images to minimize the quality loss of the images. Those that could not be saved due to the
103 encryption of PDF files in the literature were screenshotted as png files using Snipaste. The
104 majority of images collected from plates are single animal images, and the effect of plate
105 numbering was avoided as much as possible.

106 We obtained more than 16,000 fossil images from 188 publications and performed data
107 cleaning. During the data collection stage, we collected as many fossil images as possible. These
108 images were taken at any viewpoint and in any orientation. Different views of the same specimen
109 were treated as different instances and labelled separately. To ensure the reliability of the dataset,
110 we checked the bivalve and brachiopod images and corresponding labels. Because the taxonomic
111 system of bivalves and brachiopods is continuously improved (Konopleva *et al.* 2017; Sulser *et al.*
112 2010), we categorized the genera whose taxonomic names and positions had been changed in the
113 literature. Additionally, we removed poorly preserved fossil images. This contains two cases. The
114 first case is images with uncertain taxonomic names. The other discarded images are obtained from
115 scanned published documents (mostly monographs published in the last century) that are poorly
116 pixelated and difficult to identify even for palaeontologists. In both cases, the ambiguous images
117 are discarded based on whether the experts can distinguish the fossils or not. There is no filtering
118 based on deep learning preference, so this operation does not affect the utility of the deep learning
119 method.

120 Our dataset was divided into the training set (60%), validation set (20%), and test set (20%)
121 randomly to train, tune, and test the model. Because the validation set is used as a reference for
122 the tuning process, the identification accuracy of this part may have artificial bias and is not

universally meaningful. Thus, the final accuracy was calculated using a separate test set to evaluate model performance.

The final BBFID contains 870 genera, with 16,596 sets of "image-label" data pairs. All images have genus labels, with the 14,185 items having higher-resolution species labels. BBFID-1 contains 379 genera and 889 species, with 8,144 sets of image-label data pairs. BBFID-2 contains 491 genera and 1,144 species, with 8,452 sets of data pairs. Genus distributions of BBFID and examples of common categories are shown in Figure 1.

To meet the requirements of machine learning, each taxon should have at least three items. Therefore, we chose the categories with > 2 items of BBFID to perform the model training, which contains 16,389 sets of "image-label" data pairs.

Methods

Convolutional Neural Network

Convolutional neural networks (CNNs) perform well in general recognition and have been used in the automatic identification of palaeontological fossils (Dionisio, *et al.* 2020; Niu and Xu 2022; Wang, *et al.* 2022; Liu, *et al.* 2022; Liu and Song 2020; Kiel 2021). In this study, three pre-trained models of convolutional neural networks with good classification performance on the ImageNet dataset (Deng *et al.* 2009) namely VGG-16 (Simonyan and Zisserman 2014), Inception-ResNet-v2 (Szegedy, *et al.* 2017), and EfficientNetV2s (Tan and Le 2021) were selected and suitably modified (Fig. 2). VGG-16 and Inception-ResNet-v2 have been proven to automatically

identify fossils and perform well (Hsiang*, et al.* 2019; Liu*, et al.* 2022). We retained their main architecture, removed the top softmax layer and/or fully connected layer depending on fossil categories, and added a fully connected layer (with 256 output and Relu activation function), batch normalization layer (Ioffe and Szegedy 2015), dropout layer (with rate = 0.2), and fully connected layer (with output as fossil categories) (Fig. 2).

In fossil identification, CNNs first decode the fossil images to obtain the tensor that can be operated, and the model operates on these values to establish the correspondence between the image data and the fossil name. CNNs use convolutional operations to process image data and gradient descent to minimize the loss function to train the model (LeCun *et al.* 1998). The neural network can be divided into multiple network layers. More specifically, the convolutional, pooling, and fully connected layers play a crucial role in the automatic identification process. The convolutional layer reduces the data size and extracts high-dimensional information by operating on the image matrix with a certain size of the convolutional kernels. The pooling layer reduces the amount of computation, making the model easier to train (Giusti *et al.* 2013). The fully connected layer and activation function (Relu) fit the correspondence between fossils and labels (Nair and Hinton 2010) and output the predicted labels and probabilities we need at the top layer.

VGG-16 is a classic DCNN proposed by Simonyan and Zisserman (2014), which uses 16 layers and 3 × 3 convolutional kernels (convolution filters) to achieve good performance. And then, He *et al.* (2016) proposed a new residual connectivity method and applied it to Inception-ResNet-v2, which makes the network easier to optimize and allows the use of the deeper network to improve performance. EfficientNetV2 is currently a more advanced open-source image

163 classification model using the training-aware neural architecture search and scaling method to
164 improve training speed and parameter efficiency (Tan and Le 2021).

165 *Data preprocess*

166 Deep learning models have requirements for input data size. However, images in our dataset
167 were of different sizes and the labels were also inappropriate to model training. Thus, data needed
168 to be preprocessed. To match the model's requirement, all images were resized to a uniform size
169 (slightly different depending on the model in Fig. 2). To improve their generalization ability and
170 make the model easier to train, we randomly adjusted the image (training set and validation set)
171 brightness (within ± 0.5) and contrast (within 0 to + 10) to reduce the effect of noise. In addition,
172 the images were normalized and standardized. We conducted the discrete one-hot coding for image
173 labels. Finally, a one-to-one correspondence between the images and the labels was established,
174 and we obtained the processed machine-learning dataset.

175 *Training Methodology*

176 Achieving high accuracy in multiclass fossil identification using neural networks requires a
177 large dataset as a basis. Although we built the bivalve and brachiopod dataset manually, it was still
178 insufficient to train a model with random initialization of parameters to converge and achieve the
179 best results. Therefore, we applied transfer learning in the model training process, an effective way
180 to train a model on a small dataset (Tan *et al.* 2018; Brodzicki *et al.* 2020; Koeshidayatullah *et al.*
181 2020). Transfer learning uses parameters trained by general identification tasks for initialization

to accelerate the convergence of the new model. It is feasible to use this to reuse the general identification model parameters for palaeontological fossil identifications (Pires de Lima *et al.* 2020). This is why we only envision applying this method to the automatic identification of common fossils, while fossils with too few specimens will still need to rely on palaeontologists.

In this study, each model was loaded with pre-trained parameters that were originally trained on ImageNet. This method greatly reduces the amount of data required for automatic identification, greatly expanding their application scenarios.

We coded in Python and relied on the Tensorflow scientific computing library (Abadi *et al.* 2016) to train the model. The training process was performed using the Adam optimizer (Kingma and Ba 2014). The loss function uses the categorical cross-entropy loss function (Botev *et al.* 2013), and the accuracy is used as an evaluation metric. To facilitate training, the learning rate is adjusted with validation loss in training. Also, to prevent overfitting, EarlyStopping was set to ensure the good performance of the model in the test set. During the training process, the model saves architecture and parameters with the highest accuracy in the validation set in real-time for rapid deployment in subsequent applications. Because BBFID contains both the genus tags and species tags, we set the model to the genus mode (only read the genus tag) and species mode (read both genus tag and species tag) during model training and testing. Because of dataset size, model's architecture and hyperparameters significantly affect its performance; thus, we trained models and compared their performance under different scenarios (Table 1).

We chose the different sizes of the datasets to train models according to the taxonomic levels. At the genus-level, we set three scales to explore model performance using different volumes of

datasets. These three scales are the number of each genus > 100 images (scale A), > 50 images (scale B), and > 10 images (scale C) (Table 2). Among them, scale B/C contains all genera with more than 50/10 pictures, the same for other scales. The numbers of taxa in BBFID-1 are 13 (scale A), 34 (scale B), and 156 (scale C), respectively, whereas the numbers of items in BBFID-2 are 9 (scale A), 32 (scale B), and 223 (scale C). They display a clear gradient to match our research needs. For the selection of data adequacy (i.e., data scale) of the species-level, we selected scale B (number of each species > 50 images) and scale C (number of each species > 10 images) for training and testing, according to the performance of the genus mode. Furthermore, we also tried two larger scales: scale D (number of each species > 8 images) and scale E (number of each species > 6 images). There are four gradients in total to find the range that covers more genera with guaranteed accuracy. In addition, for BBFID, we added two larger scales (the number of each taxon > 4 images and > 2 images) to explore the model performance in small data sets. As mentioned earlier, all data (scales A, B, C, D, and E) were randomly divided into the training set, validation set, and test set in the ratios of 60%, 20%, and 20%, which is the ideal situation. In order to try a larger data scale, we discarded the requirement that the validation set cover all species. Therefore, the number of single-taxon images > 2 was the maximum data size we could try, because all taxonomic units shall be covered in the training set and test set.

Model architecture plays a pivotal role in models. Thus, we used BBFID-1 (scale A) to test model identification accuracy at the genus level under three different model architectures (i.e., VGG-16, Inception-ResNet-v2, and EfficientNetV2s). Subsequently, the best architecture was selected to build ATIM, trained and tested using different scales of BBFID-1, BBFID-2, and

BBFID, respectively, to obtain the corresponding model performance (Table 2).

Dimensionality reduction method

In this study, we employed a downscaling method of t-SNE that uses a probability measure of similarity and expresses probabilities as spatial distances (Van der Maaten and Hinton 2008). To compare fossil morphology, we extracted the output of the last maximum pooling layer as fossil features and downscaled the high-dimensional data of fossil features to a two-dimensional plane using t-SNE. Next, we visualized that to analyze easily the morphological differences and similarities between bivalves and brachiopods. The model training and downscaled visualization codes were referenced from some open-source projects.

Results

Model performance between different architectures and hyperparameters

Different architectures perform differently using BBFID-1 (scale A, genus level), with the best performance of 83.02% obtained with the EfficientNetV2s architecture and the corresponding hyperparameters (Table 1). The results of confusion matrix for this identification task are shown in Figures 3, 4, and 5. The identification recalls were > 79% for all categories except the genera *Pteria* (0.71), *Bakevellia* (0.72), and *Halobia* (0.72), ~~where~~ the accuracies of *Quemocuomegalodon,* and *Monotis* exceeded 90%.

Model performance using different data scale

We used EfficientNetV2s architecture that performed well on BBFID-1 and corresponding hyperparameters to build other models (genus mode), which performed as expected under different datasets (Table 2). The accuracy of BBFID-1 (scale A) was 82.10%, whereas those of scales B and C were 71.73% and 58.34% respectively, with the loss increasing by decreasing accuracy for all three. The accuracy of BBFID-2 was 85.43%, 71.35%, and 50.04% for the three dataset scales, whereas the identification accuracy of scale A exceeded 85%. Furthermore, in four categories, more than 90% of images were identified correctly (Fig. 4). The accuracy of model training by BBFID was 81.45%, 70.66%, and 53.71% at the three scales, and the performance of each scale was similar to the performance of the corresponding bivalve and brachiopod individual identifications. In species mode, the models also performed similarly (Table 2), with the accuracy of BBFID at scale C (148 categories for bivalves, 195 categories for brachiopods) of more than 60% (see Appendix S1 for confusion matrix and evaluation metrics). The accuracies of Scale D (bivalve 179 categories, brachiopod 265 categories) and scale E (bivalve 241 categories, brachiopod 396 categories) ranged from 51% to 59%.

Discussion

Identification accuracy

The ranking of automatic identification performance among three architectures trained by BBFID-1 (Table 1) is comparable to general task results (Simonyan and Zisserman 2014; Szegedy,

et al. 2017; Tan and Le 2021), indicating that the principle of fossil auto-identification is similar to that of general image classification. That corroborates the rationality of using transfer learning. The identification accuracy (> 80%) on genus mode is similar to ~~the previous study~~ (Romero *et al.* 2020).

All these models in EfficientNetV2s architecture met the early stopping condition and terminated training before 50 epochs, and the training set accuracy was close to 100% at this point. That indicates that the models completed fitting to the training set. The training process of BBFID (scale A) shows that the model basically converged about 20 epochs (Fig. 6), and its training set accuracy finally reached ~100%, while the maximum validation accuracy was over 80% (Table 2). The fossil images used in this study contain pictures of the whole shells and detailed pictures, such as structures of fossils. The identification accuracy was adversely affected by this factor. Thus, it is feasible to apply pre-trained parameters of the general model to the ATIM in the field of palaeontology using transfer learning. For the accuracy of different parts of the dataset, the accuracy of the validation set was comparable to that of the test set, but lower compared with the training set. Because the model was trained using the training set, the identification performance was better in this part. However, the data from the validation and test sets were not used to train models. Accordingly, the results were slightly worse compared with the training set. Furthermore, the validation set was purposefully optimized in the conditioning. Accordingly, the real performance of the ATIM is shown by the test set result, rather than that of the validation set.

The accuracy of the model using selected architecture and parameters (Table 1, Order 11) on genus mode exceeded 80% using BBFID-1 (scale A). In contrast, the accuracy decreases between

281 scale B and scale C stems from the single taxon images decrease and confusion caused by the
282 categories increase. Nevertheless, the identification accuracy of scale C (156 categories) was still
283 close to 60%. In addition, the model based on BBFID-2 achieved similar accuracy to the model
284 based on BBFID-1 at all scales. The identification accuracy at scale A exceeded 80%, which is
285 close to or even exceeds the identification level of palaeontologists (Hsiang*, et al.* 2019). Hsiang
286 *et al.* (2019) collected the accuracy of foraminiferal identification by palaeontologists and found
287 that human accuracy is only 71.4%, which is lower than automatic identification (87.4%). Another
288 study of planktonic foraminifera covering 300 specimens reported an average identification
289 accuracy of <78% for 21 experts (Al-Sabouni *et al.* 2018). In an automatic identification of
290 ~~dinoflagellates~~, the expert's accuracy was also only 72% (Culverhouse *et al.* 2003). Austen *et al.*
291 (2016) found that the accuracy of experts in bumblebees was even lower than 60%.

292 As mentioned previously, this study achieved automatic identification of fossils including 22
293 genera of bivalves and brachiopods, with a test set accuracy > 80%. The obtained model performed
294 relatively well considering the volume of categories and datasets in this task. Dionisio*, et al.* (2020)
295 also trained a model for identifying radiolarian fossils (containing only nine genera with 929
296 photographs) automatically. The accuracy of the CNN model is 91.85%, higher than ours. The
297 average number of images per genus used in this study was comparable to ours; however, they
298 used SEM photographs from the same source. Fewer extraneous factors and fewer categories
299 might have contributed to slightly higher accuracy. Models for the automatic identification of
300 pollen from 16 genera were also proposed with accuracies between 83% and 90%, also using
301 microscopic images (Romero*, et al.* 2020).

Moreover, models based on BBFID performed similarly to the models based on the corresponding scale of BBFID-1 or BBFID-2, which indicates that the ATIM is not easily affected by the similar morphology between bivalves and brachiopods with sufficient data volume (as further demonstrated by the confusion matrix). The models are highly reliable in bivalves and brachiopods identification at the genus level, which provides a basis for our subsequent comparison of their morphology. Moreover, the identification accuracy of BBFID (scale C, including 379 taxa) was 53.71%, which is understandable considering the large taxonomic unit number with the relatively limited training set. Large-scale automatic fossil identification based on a small dataset is feasible. However, it must be noted that the categories with fewer figures are more concentrated in the literature, which might have led to the similarity between the test set and the training set. Thus, these accuracies cannot objectively generalize the performance and ability of models.

Regarding species-level automatic identification performance, we achieved an accuracy of 82.83% for 16 species identification, with several species attributed to the same genus with relatively similar morphology. Although Kong *et al.* (2016) automatically identified three pollen species of the same genus in a confusing species classification task with 86.13% accuracy, it must be noted that their pollen task relied more on confusing information such as a texture for identification. Importantly, the identification accuracy of mixed data scale C in the species mode is similar to, or even slightly higher than, that in the genus mode. This implies that the number of taxonomic categories can have a greater impact on automatic identification performance relative to the differences between taxonomic units.

Although we independently built a dataset containing >16,000 images, it is still small for machine learning. Most studies in automatic fossil identification have focused on a few categories and large sample sizes (Liu*, et al.* 2022; Niu and Xu 2022; Wang*, et al.* 2022; Liu and Song 2020), which undoubtedly helps improve performance. Niu and Xu (2022) used a dataset of 34,000 graptolites to perform an automatic identification study of 41 genera, which resulted in 86% accuracy. In contrast, the identification accuracy of 47 genera in this study was 76.26%, which demonstrates the importance of larger data sets.

Analysis of identification results

We tested models in genus mode using BBFID-1, BBFID-2, and BBFID (scale A) and obtained a confusion matrix (Figs. 3, 4, 5), which truly reflects the model performance and misidentification. Example images of all 22 genera in this scenario are shown in the Appendix S2 for a better comparison of morphological differences. In the confusion matrix, the vertical axis represents the “true” genus name, whereas the horizontal axis represents the "predicted" genus name. The numbers in the matrix represent the proportion of "true" genera identified as "predicted" genera, and the larger the proportion, the darker the squares. The model performed well in the automatic identification of bivalves and brachiopods respectively, and misidentification was maintained at a low level.

In the hybrid auto-identification model (i.e., model based on BBFID), the overall performance was good although the accuracy (81.90%) decreased slightly compared to the separate auto-identification accuracies of bivalves and brachiopods (i.e., accuracies testing by BBFID-1 or

343 BBFID-2). Genus *Quemocuomegalodon* maintained a high identification recall (1.00) in the
344 bivalve categories, whereas the recall of *Proyalina* increased from 0.88 to 0.92. Other categories
345 decreased slightly. Most of the brachiopod categories showed significant or stable increases,
346 whereas only two genera exhibited recall decreases (*Araxathris* from 0.76 to 0.68 and *Paryphella*
347 from 0.77 to 0.72). The change in the recall may be related to the change in the distribution of the
348 training set. Among these misidentified categories, two cases were distinctive, each exceeding
349 0.20 of their respective categories in the test set. The bivalve *Pteria* was misidentified as *Bakevellia*
350 (0.25) and the brachiopod *Paryphella* was misidentified as *Fusichonetes* (0.24), with
351 morphological similarity being the main reason for misidentification. For example, the shells of
352 both *Pteria* and *Bakevellia* have similar outline and are anteriorly oblique. The posterior ear is
353 larger than the anterior ear. Distinctive concentric rings are visible on the shell surface. All these
354 features are very similar.

355 Importantly, the vast majority of misidentifications in the hybrid auto-identification model
356 occurred within categories (i.e., bivalves were misidentified as other bivalves and brachiopods
357 were misidentified as other brachiopods), whereas misidentifications between broad categories
358 were relatively rare. For example, only 0.04 of the brachiopod *Araxathris* were misidentified as
359 bivalve *Daonella* and 0.04 as bivalve *Eumorphotis*, which indicates that bivalves and brachiopods
360 have considerable morphological differences.

361 The above are all cases where the input fossil taxon is included in the training set, but in
362 reality, there are quite some fossil taxa that are not included in the training set. To deal with this
363 exception, we propose a new "Applicability Model" (AM) to identify such cases. We divide the

entire BBFID into "applicable" and "unapplicable", and perform binary classification training based on the Inception-ResNet-v2. The accuracy of AM (suitable for Order 22) is 85.54%. When the training is completed, the user can use the AM to verify whether the taxon of the input images is included in the training set and the usability of the genus/species identification model. If the result is “applicable”, the fossil will be identified automatically. If the result is “unapplicable”, the identification model will give the name of the fossil taxon that is most similar to it, and the user can continue the manual identification based on that taxon.

Morphological analysis of fossils

Fossils have complex and variable high-dimensional morphological features, which are difficult to visualize and analyze. Deep learning can extract features, downscale dimensions of data, and exclude the influence of human bias to fully reflect the fossil features. Neural networks can extract features more efficiently than manually selected features (Keceli *et al.* 2017). The accuracy of supervised classification of ammonoids using human-selected geometric features was only 70.4%-78.1% in 11 species (Foxon 2021), lower than the accuracy of > 80% for 22 species identifications in this study.

Machine learning can quantify morphological features and compare differences. Therefore, we extracted the process output from the ATIM (Order 22) and summed the same point data in each dimension to draw a feature map (Fig. 7). We can observe the identification features used by the convolutional neural network. However, the supervised deep learning used in this paper is a "result reason" approach that cannot verify the correctness of the taxonomic practice. Models may

use some features not used by experts to identify, which does not mean that the taxonomic practice is wrong. A possible scenario is that there are multiple differences between the two taxa, with experts and models choosing different perspectives. The model establishes a relationship between the input (fossil image, i.e., morphological features) and the output (taxon), and its ability to accurately identify fossil taxa indicates that taxonomic practice is well correlated with fossil morphology. Input-output relationships are established by feature extraction through convolutional neural networks. Automatic identification relies on these features that are similar with the working process of experts. The features extracted by the model are diverse, such as the umbilicus, ribs, and inner whorl of the ammonoid, spires and apices of gastropod, and growth lines and radial ribs of bivalve and brachiopod (Liu *et al*. 2022). For the identification results, there is no difference between the model's identification using images (actually fossil morphology) and the expert's identification using characterization. This is essentially determined by the prior knowledge, which is obtained by taxonomic practice. In the future, unsupervised learning may be able to provide unique insights to evaluate taxonomic practice.

We used the output of the top maximum pooling layer (this model is available at https://github.com/Jiarui-Sun/Automatic-fossil-identification) as fossil features and then used t-SNE (Van der Maaten and Hinton 2008) for dimension reduction, which achieved good results of morphology clustering and comparison (Fig. 8). The classification of each taxon in Figure 8 is clear, and the t-SNE results are similar between the training set (Fig. 8A) and the validation set and test set (Fig. 8B). However, the individual clusters obtained from the training set are more concentrated and the boundaries between different categories are clearer than the latter due to the

405 training process (Fig. 8).

406 In the downscaled visualization of this model for the validation and test sets, the brachiopods
407 and bivalves are clearly demarcated, but a few points are still mixed (Fig. 8B). A clear boundary
408 means that the brachiopod and bivalve fossils are sufficiently morphologically distinct, so that the
409 model can extract the differences well and represent them quantitatively. This demonstrates the
410 unique potential of deep learning models for fossil feature extraction. Without inputting any prior
411 knowledge other than the genus name (e.g., the model does not know which genus belongs to
412 bivalve or brachiopod), the model computationally obtains information on the morphological
413 differences between bivalve and brachiopod, which is compatible with the expert's classification.
414 In the future, it may be possible to use this feature to find similar classification boundaries relying
415 on models to perceive more detailed information about fossils (e.g., ornamental features), which
416 in turn could allow for quantitative differentiation of gradual features. That could not only provide
417 new possible perspectives for exploring fossil classification and biomorphological evolution, but
418 also try to explore whether there are important features that have been overlooked by experts. In
419 terms of the distribution area, the distribution of bivalve points is more extensive than that of
420 brachiopods, indicating that bivalves have greater morphological variability compared to
421 brachiopods in our dataset (but the effect of image context is not excluded here). Overall, the fossil
422 features extracted by CNNs can reflect the morphological characteristics of organisms to some
423 extent.

424 CNNs can complement existing methods for morphological studies such as morphological
425 matrix (Dai *et al.* 2021), landmark (Bazzi *et al.* 2018), and ornamentation index (Miao *et al.* 2022),

and provide new perspectives for studying the morphological evolution of fossils in the future. Geometric morphometry requires the extraction of fossil features by labelling manually and performing descending operations (e.g., principal component analysis), which has proven to be very effective (Topper *et al.* 2017; Aguirre *et al.* 2016). In this method, fossil features are selected by experts, with biological significance and better interpretation. However, it is also influenced by human factors, and some features may be missed (Villier and Korn 2004, Dai *et al.* 2021). Artificial intelligence differs in that it can obtain ~~the information~~ displayed in fossil images (not just a few dozen points). These obtained features are then downscaled (e.g., t-SNE used in this paper) to get the final fossil features. However, due to the black-box character of deep learning, the features obtained are poorly interpretable, and whether they are biologically meaningful needs further study in the future. Therefore, the advantage of artificial intelligence mainly lies in the feature extraction, which reduces the subjective influence and the time cost of manual marking. On the other hand, manual feature extraction is difficult to orient to a large number of specimens and is based only on some specific species. However, deep learning is capable of obtaining information from more specimens at the scale of big data, such as intraspecific differences, spatial and temporal differences, etc., due to its ability to automate the extraction of fossil features. Moreover, combining 3D information of fossils for palaeontological studies is also promising (Hou *et al.* 2020).

Conclusions

In this study, we used machine learning to automate fossil identification based on the practical

needs of palaeontological research. We built a bivalve and brachiopod fossil dataset by collecting open literature, with > 16,000 "image-label" data pairs. Using these data, we compared the performance of several convolutional neural network models based on VGG-16, Inception-ResNet-v2, and EfficientNetV2s, which are commonly used in the field of image classification and fossil identification. For this identification task, we found that EfficientNetV2s has the best performance.

We finally achieved automatic fossil identification including 22 fossil genera (genus mode, based on BBFID) and 16 fossil species (species mode, based on BBFID), both with > 80% accuracy. Furthermore, we conducted a study on the multiple categories' automatic fossil identification at the species level, and the test accuracy was ~64% based on BBFID (scale C, containing 343 bivalves and brachiopods). Models performed well in the automatic identification of multiple categories with a small dataset. These models can be deployed to a web platform [www.ai-fossil.com, (Liu, *et al.* 2022)] in the future to make them accessible more easily and usable by researchers. For the present, automatic fossil identification must be based on expert consensus, which is precisely why we emphasize the use of this model primarily for common fossil categories to aid in identification. With more taxa be included, we can use the output from deep learning models to accelerate the systematic palaeontology work during research rather than replace it. So, the researchers can focus on most challenging and ambiguous identification cases. When a new taxon is found, the AM output "unapplicable" and experts can perform further taxonomic studies on it. When experts decide to establish a new species, the fossil differences given by the algorithm can assist them in making determinations, which is what the model excels

at. But ultimately the establishment of new species still depends on how taxonomists apply the results of deep learning. We believe that there will be many palaeontologists working on fossil taxonomy and creating a steady stream of a priori knowledge to promote the interdisciplinary relationship between palaeontology and computer science together with AI researchers.

However, it must be noted that the model is an exploratory experiment and can ~~currently serve~~ as a useful assist to manual identification, not a complete replacement for it, at least for now. The current model still relies on a manually created taxonomy and uses it as a priori knowledge for model training. Current models are not able to combine all biological features (now only use morphological data) to build the taxonomy by themselves. However, when experts have completed the taxonomic criteria, researchers can use AI to identify fossils based on those criteria, reducing repetitive identification work and allowing palaeontologists to have more time and energy for ~~more~~ creative research work.

We also used machine learning to extract high-dimensional data of fossil morphology and downscaled them to obtain fossil morphological feature distribution maps, which present the similarity of fossil morphology in a visual way. It was found that the bivalve and brachiopod distribution regions have distinctive boundaries, and the morphological differences between the two are obvious enough from the neural network perspective. In this process, models based on deep learning are not absolutely objective. On the contrary, palaeontologists play a crucial role. This is precisely why we chose researcher consensus as a priori knowledge. Furthermore, we downscaled the fossil features to cast the map and observe their morphological distribution. Compared with the manually selected features, features based on the models are more objective

and can better reflect the morphological characteristics of fossils, which are still derived based on the consensus of researchers on fossil taxonomy to a certain extent. In the future, this can be used as a basis to quantify morphological information, analyze their morphological spatial distribution, and provide a new perspective for exploring biological evolution.

Data Availability Statement

BBFID is available from the Zenodo digital repository: https://doi.org/10.5281/zenodo.7248780. The main code and models of this study can be found at https://github.com/Jiarui-Sun/Automatic-fossil-identification.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. and Isard, M. 2016. TensorFlow: a system for Large-Scale machine learning. 265–283. *12th USENIX symposium on operating systems design and implementation (OSDI 16)*.

Aguirre, M. L., Richiano, S., Alvarez, A. and Farinati, E. A. 2016. Reading shell shape: implications for palaeoenvironmental reconstructions. A case study for bivalves from the marine Quaternary of Argentina (south-western Atlantic). *Historical Biology*, **28**, 753-773.

Al-Sabouni, N., Fenton, I. S., Telford, R. J. and Kučera, M. 2018. Reproducibility of species recognition in modern planktonic foraminifera and its implications for analyses of community structure. *Journal of Micropalaeontology*, **37**, 519-534.

Alroy, J., Aberhan, M., Bottjer, D. J., Foote, M., Fursich, F. T., Harries, P. J., Hendy, A. J. W., Holland, S. M., Ivany, L. C., Kiessling, W., Kosnik, M. A., Marshall, C. R., Mcgowan, A. J., Miller, A. I., Olszewski, T. D., Patzkowsky, M. E., Peters, S. E., Villier, L., Wagner, P. J., Bonuso, N., Borkow, P. S., Brenneis, B., Clapham, M. E., Fall, L. M., Ferguson, C. A., Hanson, V. L., Krug, A. Z., Layou, K. M., Leckey, E. H., Nurnberg, S., Powers, C. M., Sessa, J. A., Simpson, C., Tomasovych, A. and Visaggi, C. C. 2008. Phanerozoic trends in the global diversity of marine invertebrates. *Science*, **321**, 97–100.

Austen, G. E., Bindemann, M., Griffiths, R. A. and Roberts, D. L. 2016. Species identification by experts and non-experts: comparing images from field guides. *Scientific Reports*, **6**, 1-7.

Ballanti, L. A., Tullis, A. and Ward, P. D. 2012. Comparison of oxygen consumption by Terebratalia transversa

(Brachiopoda) and two species of pteriomorph bivalve molluscs: implications for surviving mass extinctions. *Paleobiology*, **38**, 525–537.

Bazzi, M., Kear, B. P., Blom, H., Ahlberg, P. E. and Campione, N. E. 2018. Static dental disparity and morphological turnover in sharks across the end-Cretaceous mass extinction. *Current Biology*, **28**, 2607–2615.

Benton, M. J. and Harper, D. A. 2020. *Introduction to paleobiology and the fossil record*. John Wiley & Sons, 642 pp.

Botev, Z. I., Kroese, D. P., Rubinstein, R. Y. and L'Ecuyer, P. 2013. The cross-entropy method for optimization. *Handbook of statistics*, **31**, 35–59.

Bourel, B., Marchant, R., De Garidel-Thoron, T., Tetard, M., Barboni, D., Gally, Y. and Beaufort, L. 2020. Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains. *Computers & Geosciences*, **140**, 104498.

Brodzicki, A., Piekarski, M., Kucharski, D., Jaworek-Korjakowska, J. and Gorgon, M. 2020. Transfer Learning Methods as a New Approach in Computer Vision Tasks with Small Datasets. *Foundations of Computing and Decision Sciences*, **45**, 179–193.

Byeon, W., Dominguez-Rodrigo, M., Arampatzis, G., Baquedano, E., Yravedra, J., Mate-Gonzalez, M. A. and Koumoutsakos, P. 2019. Automated identification and deep classification of cut marks on bones and its paleoanthropological implications. *Journal of Computational Science*, **32**, 36–43.

Chen, J. H. and Stiller, F. 2010. An early Daonella from the Middle Anisian of Guangxi, southwestern China, and its phylogenetical significance. *Swiss Journal of Geosciences*, **103**, 523–533.

Chen, Z. Q., Tong, J., Zhang, K., Yang, H., Liao, Z., Song, H. and Chen, J. 2009. Environmental and biotic turnover across the Permian-Triassic boundary on a shallow carbonate platform in western Zhejiang, South China. *Australian Journal of Earth Sciences*, **56**, 775–797.

Clapham, M. E., Bottjer, D. J., Powers, C. M., Bonuso, N., Fraiser, M. L., Marenco, P. J., Dornbos, S. Q. and Pruss, S. B. 2006. Assessing the ecological dominance of Phanerozoic marine invertebrates. *Palaios*, **21**, 431–441.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V. and González-Gil, S. 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, **247**, 17-25.

Dagys, A. 1965. Triassic brachiopods of Siberia. *Nauka, Moskva.*1–186.

Dai, X., Korn, D. and Song, H. 2021. Morphological selectivity of the Permian-Triassic ammonoid mass extinction. *Geology*, **49**, 1112–1116.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. and Li, F. F. 2009. ImageNet: A Large-Scale Hierarchical Image Database. 248–255. *IEEE-Computer-Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Miami Beach, FL.

Dionisio, A., Solano, G., Quisote, M. and Marquez, E. 2020. A Radiolarian Classifier using Convolutional Neural Networks. 1–5. *International Conference on Artificial Intelligence and Signal Processing (AISP)*. VIT AP Univ, Amaravati, India.

Fan, J. X., Shen, S. Z., Erwin, D. H., Sadler, P. M., Macleod, N., Cheng, Q. M., Hou, X. D., Yang, J., Wang, X. D., Wang, Y., Zhang, H., Chen, X., Li, G. X., Zhang, Y. C., Shi, Y. K., Yuan, D. X., Chen, Q., Zhang, L. N., Li, C. and Zhao, Y. Y. 2020. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. *Science*, **367**, 272–277.

Feng, R. and Jiang, Z. 1978. Phylum Brachiopoda. In, Geological and Palaeontological Team of Guizhou ed.,

Palaeontological Atlas of Southwest China; Guizhou, Part 2. Carboniferous to Quaternary Volume. *Carboniferous to Quaternary volume*.231–305.

FLÜGEL, E. and MUNNECKE, A. 2010. *Microfacies of carbonate rocks: analysis, interpretation and application*. Springer, Berlin, 924 pp.

Foxon, F. 2021. Ammonoid taxonomy with supervised and unsupervised machine learning algorithms. *PaleorXiv ewkx9, ver. 3*.https://doi.org/10.31233/osf.io/ewkx9.

Geyer, G., Hautmann, M., Hagdorn, H., Ockert, W. and Streng, M. 2005. Well-preserved mollusks from the lower Keuper (Ladinian) of Hohenlohe (Southwest Germany). *Paläontologische Zeitschrift*, **79**, 429–460.

Giusti, A., Ciresan, D. C., Masci, J., Gambardella, L. M. and Schmidhuber, J. 2013. Fast Image Scanning with Deep Max-Pooling Convolutional Neural Networks. 4034–4038. *20th IEEE International Conference on Image Processing (ICIP)*. IEEE, Melbourne, Australia.

Gradinaru, E. and Gaetani, M. 2019. Upper Spathian to Bithynian (Lower to Middle Triassic) Brachiopods from North Dobrogea (Romania). *Rivista Italiana Di Paleontologia E Stratigrafia*, **125**, 91–123.

Gradstein, F. M., Ogg, J. G., Schmitz, M. and Ogg, G. 2012. *The geologic time scale 2012*. Elsevier, 1144 pp.

He, K. M., Zhang, X. Y., Ren, S. Q. and Sun, J. 2016. Deep Residual Learning for Image Recognition. 770–778. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA.

Hofmann, R., Hautmann, M., Brayard, A., Nuetzel, A., Bylund, K. G., Jenks, J. F., Vennin, E., Olivier, N. and Bucher, H. 2014. Recovery of benthic marine communities from the end‐Permian mass extinction at the low latitudes of eastern Panthalassa. *Palaeontology*, **57**, 547–589.

Hou, Y. M., Cui, X. D., Canul-Ku, M., Jin, S. C., Hasimoto-Beltran, R., Guo, Q. H. and Zhu, M. 2020. ADMorph: A 3D Digital Microfossil Morphology Dataset for Deep Learning. *IEEE Access*, **8**, 148744–148756.

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., Baranowskils, U., Groeneveld, J., Edgar, K. M., Movellan, A., Aze, T., Dowsett, H. J., Miller, C. G., Rios, N. and 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.

Ioffe, S. and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 448–456. *International conference on machine learning*. PMLR,

Keceli, A. S., Kaya, A. and Keceli, S. U. 2017. Classification of radiolarian images with hand-crafted and deep features. *Computers & Geosciences*, **109**, 67–74.

Kiel, S. 2021. Assessing bivalve phylogeny using Deep Learning and computer vision approaches. *bioRxiv*.https://doi.org/10.1101/2021.04.08.438943.

Kingma, D. P. and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.https://doi.org/10.48550/arXiv.1412.6980.

Koeshidayatullah, A., Morsilli, M., Lehrmann, D. J., Al-Ramadan, K. and Payne, J. L. 2020. Fully automated carbonate petrography using deep convolutional neural networks. *Marine and Petroleum Geology*, **122**, 104687.

Komatsu, T. and Huyen, D. T. 2007. Lower Triassic bivalve fossils from the Song Da and An Chau Basins, North Vietnam. *Paleontological Research*, **11**, 135–144.

Kong, S., Punyasena, S. and Fowlkes, C. 2016. Spatially Aware Dictionary Learning and Coding for Fossil Pollen

Identification. 1305–1314. *29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV.

Konopleva, E. S., Bolotov, I. N., Vikhrev, I. V., Gofarov, M. Y. and Kondakov, A. V. 2017. An integrative approach underscores the taxonomic status of Lamellidens exolescens, a freshwater mussel from the Oriental tropics (Bivalvia: Unionidae). *Systematics and Biodiversity*, **15**, 204–217.

Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278–2324.

Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y. and Song, H. 2022. Automatic taxonomic identification based on the Fossil Image Dataset (> 415,000 images) and deep convolutional neural networks. *Paleobiology*.1–22.

Liu, X. K. and Song, H. J. 2020. Automatic identification of fossils and abiotic grains during carbonate microfacies analysis using deep convolutional neural networks. *Sedimentary Geology*, **410**, 105790.

Mcroberts, C. A. 2010. Biochronology of Triassic bivalves. *Geological Society, London, Special Publications*, **334**, 201–219.

Miao, L. Y., Dai, X., Song, H. C., Backes, A. R. and Song, H. J. 2022. A new index for quantifying the ornamentational complexity of animals with shells. *Ecology and Evolution*, **12**, e9247.

Nair, V. and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. *Icml*.

Niu, Z. B. and Xu, H.-H. 2022. AI-based graptolite identification improves shale gas exploration. *bioRxiv*.https://doi.org/10.1101/2022.01.17.476477.

Payne, J. L., Heim, N. A., Knope, M. L. and Mcclain, C. R. 2014. Metabolic dominance of bivalves predates brachiopod diversity decline by more than 150 million years. *Proceedings of the Royal Society B-Biological Sciences*, **281**, 20133122.

Pires De Lima, R., Welch, K. F., Barrick, J. E., Marfurt, K. J., Burkhalter, R., Cassel, M. and Soreghan, G. S. 2020. Convolutional Neural Networks as an aid to 131 biostratigraphy and micropaleontology: a test on late paleozoic microfossils. *Palaios*, **35**, 391–402.

Popov, A. M. and Zakharov, Y. D. 2017. Olenekian Brachiopods from the Kamenushka River Basin, South Primorye: New Data on the Brachiopod Recovery after the end-Permian Mass Extinction. *Paleontological Journal*, **51**, 735–745.

Romero, I. C., Kong, S., Fowlkes, C. C., Jaramillo, C., Urban, M. A., Oboh-Ikuenobe, F., D'apolito, C. and Punyasena, S. W. 2020. Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **117**, 28496–28505.

Scotese, C. R., Song, H. J., Mills, B. J. W. and Van Der Meer, D. G. 2021. Phanerozoic paleotemperatures: The earth's changing climate during the last 540 million years. *Earth-Science Reviews*, **215**, 103503.

Sepkoski, J. J. 1981. A factor analytic description of the Phanerozoic marine fossil record. *Paleobiology*, **7**, 36–53.

Silberling, N. J., Grant-Mackie, J. A. and Nichols, K. M. 1997. *The Late Triassic bivalve Monotis in accreted terranes of Alaska*. US Government Printing Office, 217 pp.

Simonyan, K. and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.https://doi.org/10.48550/arXiv.1409.1556.

Song, H. J., Kemp, D. B., Tian, L., Chu, D. L., Song, H. Y. and Dai, X. 2021. Thresholds of temperature change for mass extinctions. *Nature Communications*, **12**, 4694.

Song, T. 2018. Study on the Bivalve Faunas in Southwestern China during the Permian-Triassic Transitional Time. *China University of Geosciences*.1–182.

Su, T., Farnsworth, A., Spicer, R. A., Huang, J., Wus, F. X., Liu, J., Li, S. F., Xing, Y. W., Huang, Y. J., Deng, W. Y. D., Tang, H., Xu, C. L., Zhao, F., Srivastava, G., Valdes, P. J., Deng, T. and Zhou, Z. K. 2019. No high Tibetan Plateau until the Neogene. *Science Advances*, **5**, eaav2189.

Sulser, H., García-Ramos, D., Kürsteiner, P. and Menkveld-Gfeller, U. 2010. Taxonomy and palaeoecology of brachiopods from the South-Helvetic zone of the Fäneren region (Lutetian, Eocene, NE Switzerland). *Swiss Journal of Geosciences*, **103**, 257–272.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. *Thirty-first AAAI conference on artificial intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. 2015. Going deeper with convolutions. 1–9. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tan, C. Q., Sun, F. C., Kong, T., Zhang, W. C., Yang, C. and Liu, C. F. 2018. A Survey on Deep Transfer Learning. 270–279. *27th International Conference on Artificial Neural Networks (ICANN)*. Springer International Publishing Ag, Rhodes, Greece.

Tan, M. and Le, Q. 2021. Efficientnetv2: Smaller models and faster training. 10096–10106. *International Conference on Machine Learning*. PMLR.

Topper, T. P., Strotz, L. C., Skovsted, C. B. and Holmer, L. E. 2017. Do brachiopods show substrate‐related phenotypic variation? A case study from the Burgess Shale. *Palaeontology*, **60**, 269-279.

Van Der Maaten, L. and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Villier, L. and Korn, D. 2004. Morphological disparity of ammonoids and the mark of Permian mass extinctions. *Science*, **306**, 264-266.

Vörös, A. and BUDAI, T. 2003. *The Pelsonian Substage on the Balaton Highland (Middle Triassic, Hungary)*. Institutum Geologicum Hungaricum, 195 pp.

Wang, H. Z., Li, C. F., Zhang, Z. F., Kershaw, S., Holmer, L. E., Zhang, Y., Wei, K. Y. and Liu, P. 2022. Fossil brachiopod identification using a new deep convolutional neural network. *Gondwana Research*, **105**, 290–298.

Wu, H., He, W., Shi, G. R., Zhang, K., Yang, T., Zhang, Y., Xiao, Y., Chen, B. and Wu, S. 2018. A new Permian–Triassic boundary brachiopod fauna from the Xinmin section, southwestern Guizhou, south China and its extinction patterns. *Alcheringa: An Australasian Journal of Palaeontology*, **42**, 339–372.

Wu, H. T., Zhang, Y. and Sun, Y. L. 2019. A Mixed Permian-Triassic Boundary Brachiopod Fauna from Guizhou Province, South China. *Rivista Italiana Di Paleontologia E Stratigrafia*, **125**, 609–630.

Xu, G. and Grant, R. E. 1994. Brachiopods near the Permian-Triassic boundary in south China. *In Smithsonian Contributions to Paleobiology*, **76**, 1–68.

Yin, H. F., Zhang, K. X., Tong, J. N., Yang, Z. Y. and Wu, S. B. 2001. The Global Stratotype Section and Point (GSSP) of the Permian-Triassic Boundary. *Episodes*, **24**, 102–114.

Figures and Tables:

680 **FIG. 1.** Number of samples for each taxon at the genus level in (A) BBFID-1 and (B) BBFID-2
681 (scale B) and the distribution in subsets. The fossil images in the figure show examples of several
682 of the most common genera of the two datasets. Fossil images are from fourteen publications
683 (Komatsu and Huyen 2007; Silberling *et al.* 1997; Chen *et al.* 2009; McRoberts 2010; Vörös and
684 Budai 2003; Geyer *et al.* 2005; Chen and Stiller 2010; Hofmann *et al.* 2014; Gradinaru and Gaetani
685 2019; Wu *et al.* 2018; Xu and Grant 1994; Wu *et al.* 2019; Feng and Jiang 1978; Dagys 1965).
686 Fossil images are not to scale.
687
688 **FIG. 2.** DCNN architectures used in this study. Automatic identification model architectures of A, B, C are
689 modified from VGG-16 (Simonyan and Zisserman 2014), Inception-ResNet-v2 (Szegedy, *et al.* 2017),
690 and EfficientNetV2s (Tan and Le 2021) respectively.
691
692 **FIG. 3.** Confusion matrix and evaluation metrics of models trained by BBFID-1 (scale A) on genus
693 mode. The horizontal axis is the predicted label, and the vertical axis is the true label. Colors and
694 values represent the proportion of the corresponding taxon identified as the predicted label taxon.
695
696 **FIG. 4.** Confusion matrix and evaluation metrics of models trained by BBFID-2 (scale A) on genus
697 mode. Colors and values represent the proportion of the corresponding taxon identified as the
698 predicted label taxon.
699
700 **FIG. 5.** Confusion matrix and evaluation metrics of models trained by BBFID (scale A) on genus

701 mode. Colors and values represent the proportion of the corresponding taxon identified as the
702 predicted label taxon. The underlined categories are brachiopods, and the others are bivalves.
703
704 **FIG. 6**. The training process of ATIM on genus mode using BBFID (scale A) (Order 22).
705
706 **FIG. 7.** Feature maps of the most bivalve (*Claraia*) and brachiopod (*Piarorhynchella*) fossils in
707 BBFID, plotted by extracting model (Order 22) intermediate output. Fossil images are from
708 Song (2018), and Popov and Zakharov (2017).
709
710 **FIG. 8.** Fossil morphological feature distribution maps. (A) Training set data and (B) validation
711 set and test set data were fitted simultaneously using t-SNE. The accuracy of the original
712 identification model is 81.01%. The horizontal and vertical coordinates in the figure are the two
713 dimensions obtained by t-SNE (n_components=2, perplexity=10, init='pca', learning_rate=1,
714 n_iter= 6000, n_iter_without_progress=6000). The numbers represent different genera, where the
715 orange numbers represent the bivalves and the blue numbers represent the brachiopods. The
716 detailed correspondence is 0: *Pseudospiriferina*, 1: *Quemocuomegalodon*, 2: *Burmirhynchia*, 3:
717 *Promyalina*, 4: *Araxathyris*, 5: *Spiriferina*, 6: *Costatoria*, 7: *Fusichonetes*, 8: *Pteria*, 9: *Paryphella*,
718 10: *Neoschizodus*, 11: *Prelissorhynchia*, 12: *Juxathyris*, 13: *Piarorhynchella*, 14: *Leptochondria*,
719 15: *Daonella*, 16: *Unionites*, 17: *Bakevellia*, 18: *Halobia*, 19: *Eumorphotis*, 20: *Monotis*, 21:
720 *Claraia*.
721

722 **TABLE 1.** Identification accuracy training on BBFID-1 (scale A) at the genus level with different
723 model architectures and hyperparameters. Architectures in this table are shown in Fig. 2.
724 "Trainable layers of functional layers" represents the size of the parameters that can be trained.
725 "None" means that all layers of the backbone are frozen and the parameters involved in these
726 layers cannot be trained. These parameters maintain the values at the time of model initialization.
727 "Half layers" means that half of the backbone layer parameters are frozen, while "All layers"
728 means that all parameters of this model are not frozen and can be updated during the training
729 process. This setting has an impact on both the model training process and the model performance.

730

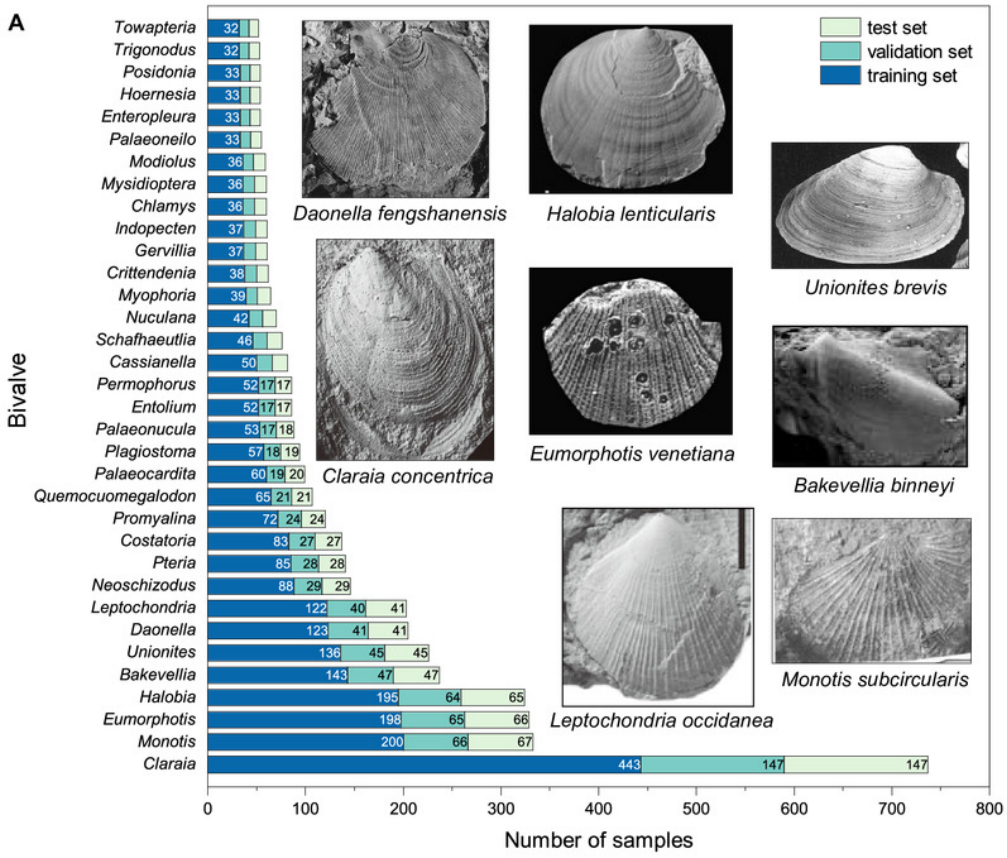
731 **TABLE 2.** Model performance using BBFID-1, BBFID-2 and BBFID in EfficientNetV2s
732 architecture. Learning rate starts from 1e-4 and the epoch is limited to less than 51. Test accuracy
733 / Test loss means the accuracy / loss of the saved model.

734

Figure 1

Number of samples for each taxon at the genus level in (A) BBFID-1 and (B) BBFID-2 (scale B) and the distribution in subsets.

The fossil images in the figure show examples of several of the most common genera of the two datasets. Fossil images are from fourteen publications (Komatsu and Huyen 2007; Silberling et al. 1997; Chen et al. 2009; McRoberts 2010; Vörös and Budai 2003; Geyer et al. 2005; Chen and Stiller 2010; Hofmann et al. 2014; Gradinaru and Gaetani 2019; Wu et al. 2018; Xu and Grant 1994; Wu et al. 2019; Feng and Jiang 1978; Dagys 1965). Fossil images are not to scale.


A
Towapteria 32
Trigonodus 32
Posidonia 33
Hoernesia 33
Enteropleura 33
Palaeoneilo 33
Modiolus 36
Mysidioptera 36
Chlamys 36
Indopecten 37
Gervillia 37
Crittendenia 38
Myophoria 39
Nuculana 42
Schafhaeutlia 46
Cassianella 50
Permophorus 52 17 17
Entolium 52 17 17
Palaeonucula 53 17 18
Plagiostoma 57 18 19
Palaeocardita 60 19 20
Quemocuomegalodon 65 21 21
Promyalina 72 24 24
Costatoria 83 27 27
Pteria 85 28 28
Neoschizodus 88 29 29
Leptochondria 122 40 41
Daonella 123 41 41
Unionites 136 45 45
Bakevellia 143 47 47
Halobia 195 64 65
Eumorphotis 198 65 66
Monotis 200 66 67
Claraia 443 147 147
Bivalve
0 100 200 300 400 500 600 700 800
Number of samples
test set
validation set
training set
Daonella fengshanensis
Halobia lenticularis
Unionites brevis
Claraia concentrica
Eumorphotis venetiana
Bakevellia binneyi
Leptochondria occidanea
Monotis subcircularis

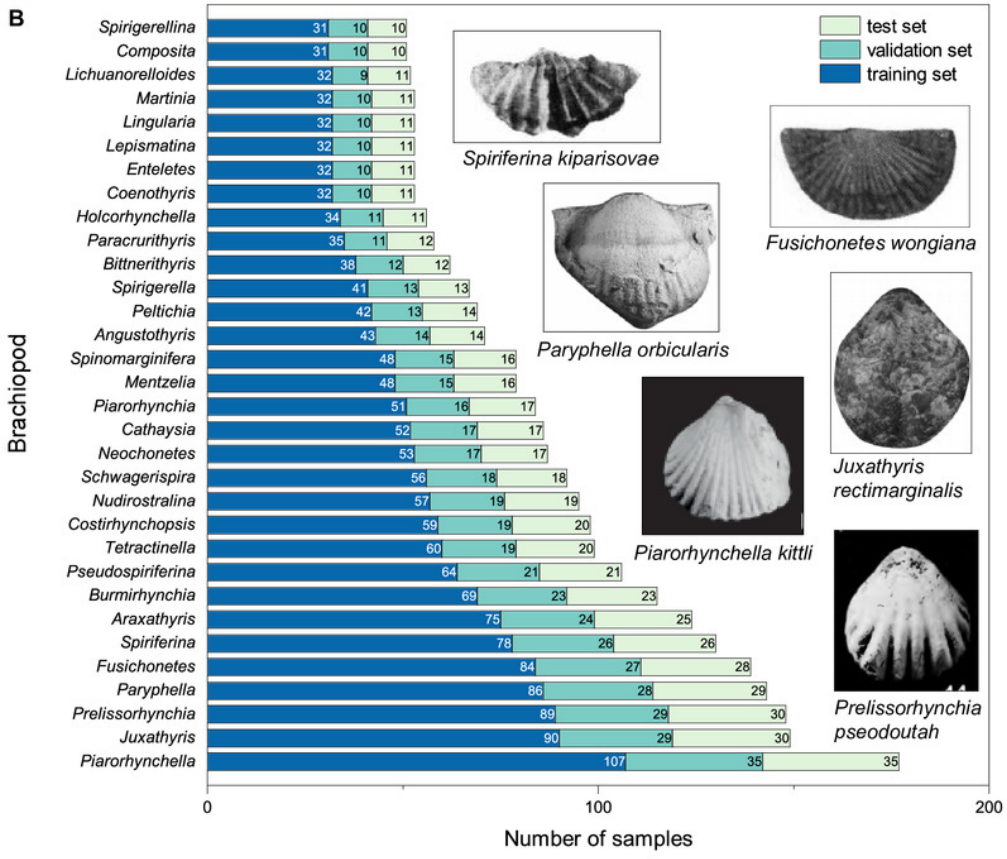

B
Spirigerellina 31 10 10
Composita 31 10 10
Lichuanorelloides 32 9 11
Martinia 32 10 11
Lingularia 32 10 11
Lepismatina 32 10 11
Enteletes 32 10 11
Coenothyris 32 10 11
Holcorhynchella 34 11 11
Paracrurithyris 35 11 12
Bittnerithyris 38 12 12
Spirigerella 41 13 13
Peltichia 42 13 14
Angustothyris 43 14 14
Spinomarginifera 48 15 16
Mentzelia 48 15 16
Piarorhynchia 51 16 17
Cathaysia 52 17 17
Neochonetes 53 17 17
Schwagerispira 56 18 18
Nudirostralina 57 19 19
Costirhynchopsis 59 19 20
Tetractinella 60 19 20
Pseudospiriferina 64 21 21
Burmirhynchia 69 23 23
Araxathyris 75 24 25
Spiriferina 78 26 26
Fusichonetes 84 27 28
Paryphella 86 28 29
Prelissorhynchia 89 29 30
Juxathyris 90 29 30
Piarorhynchella 107 35 35
Brachiopod
0 100 200
Number of samples
test set
validation set
training set
Spiriferina kiparisovae
Fusichonetes wongiana
Paryphella orbicularis
Juxathyris rectimarginalis
Piarorhynchella kittli
Prelissorhynchia pseodoutah

Figure 2

DCNN architectures used in this study.

Automatic identification model architectures of A, B, C are modified from VGG-16 (Simonyan and Zisserman 2014), Inception-ResNet-v2 (Szegedy, et al. 2017), and EfficientNetV2s (Tan and Le 2021) respectively.

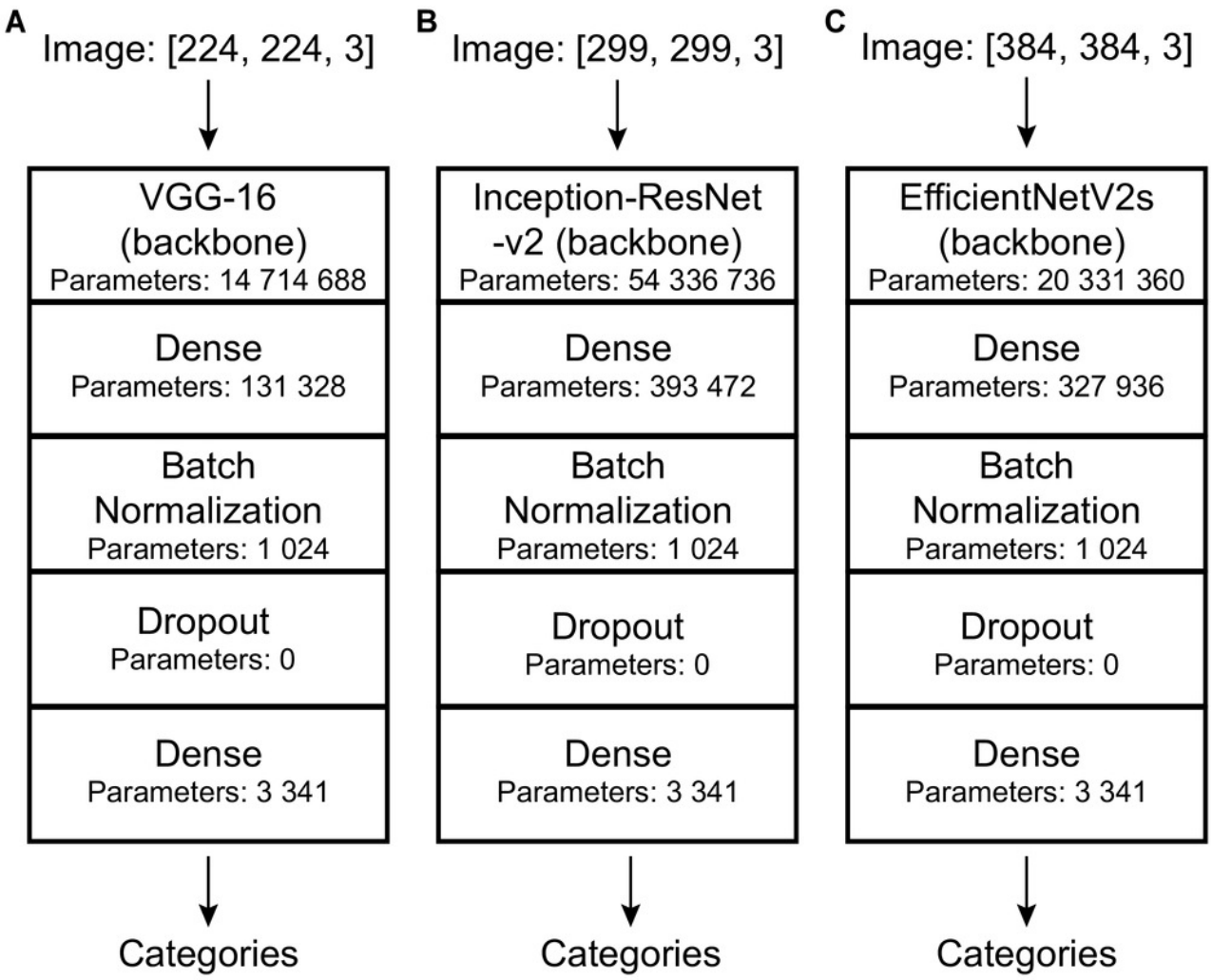


Figure 3

Confusion matrix and evaluation metrics of models trained by BBFID-1 (scale A) on genus mode.

The horizontal axis is the predicted label, and the vertical axis is the true label. Colors and values represent the proportion of the corresponding taxon identified as the predicted label taxon.

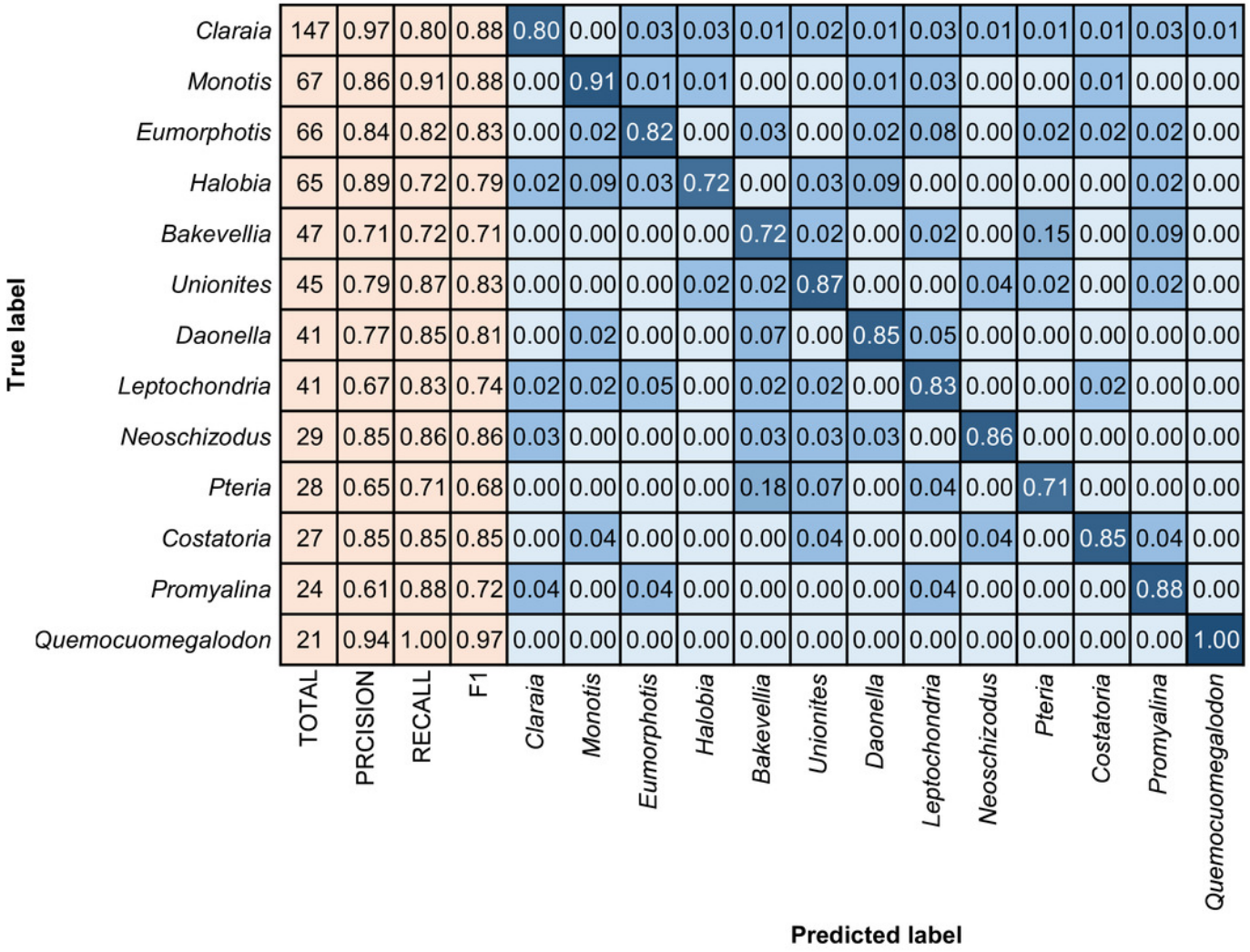


Figure 4

Confusion matrix and evaluation metrics of models trained by BBFID-2 (scale A) on genus mode.

Colors and values represent the proportion of the corresponding taxon identified as the predicted label taxon.

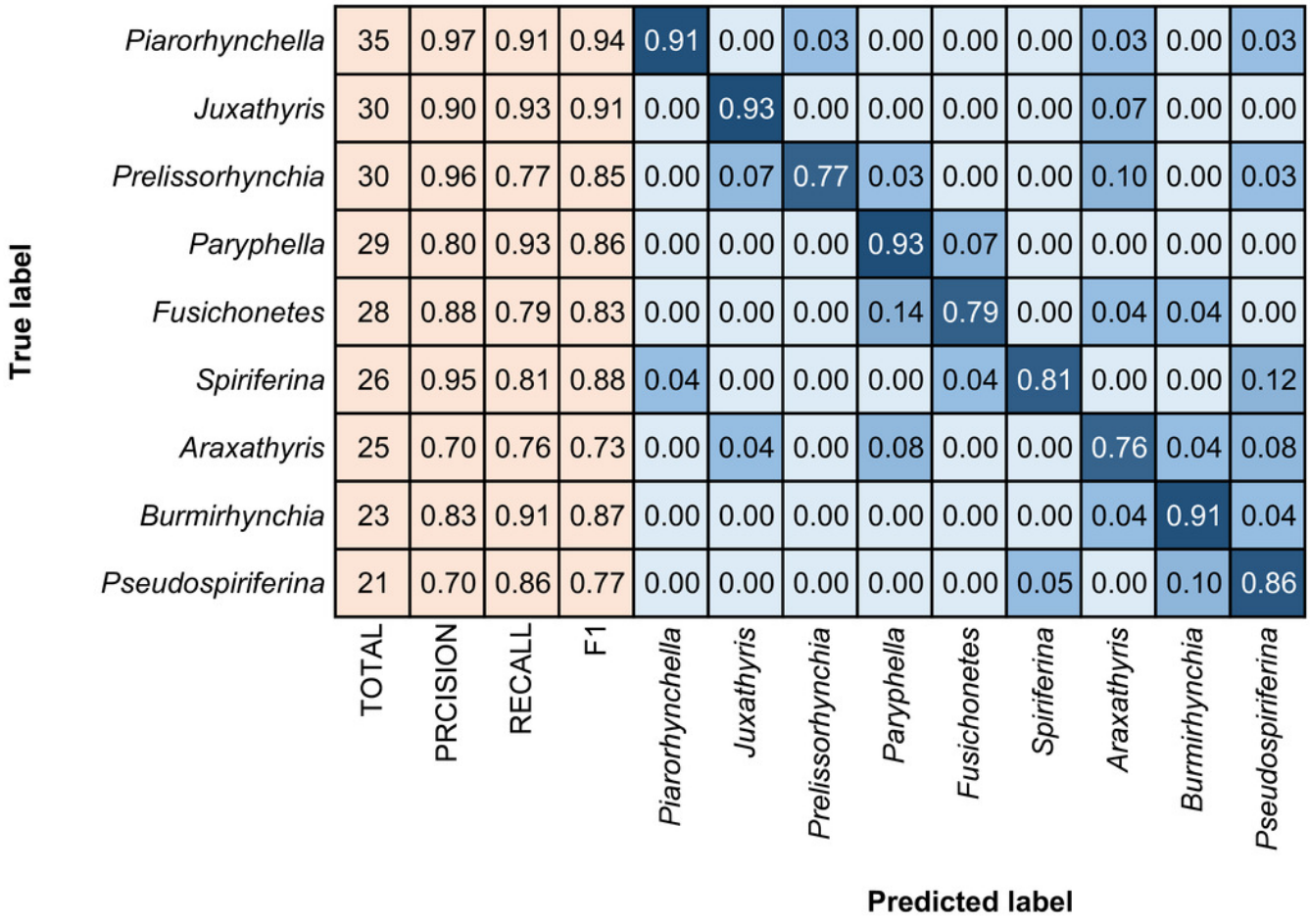


Figure 5

Confusion matrix and evaluation metrics of models trained by BBFID (scale A) on genus mode.

Colors and values represent the proportion of the corresponding taxon identified as the predicted label taxon. The underlined categories are brachiopods, and the others are bivalves.

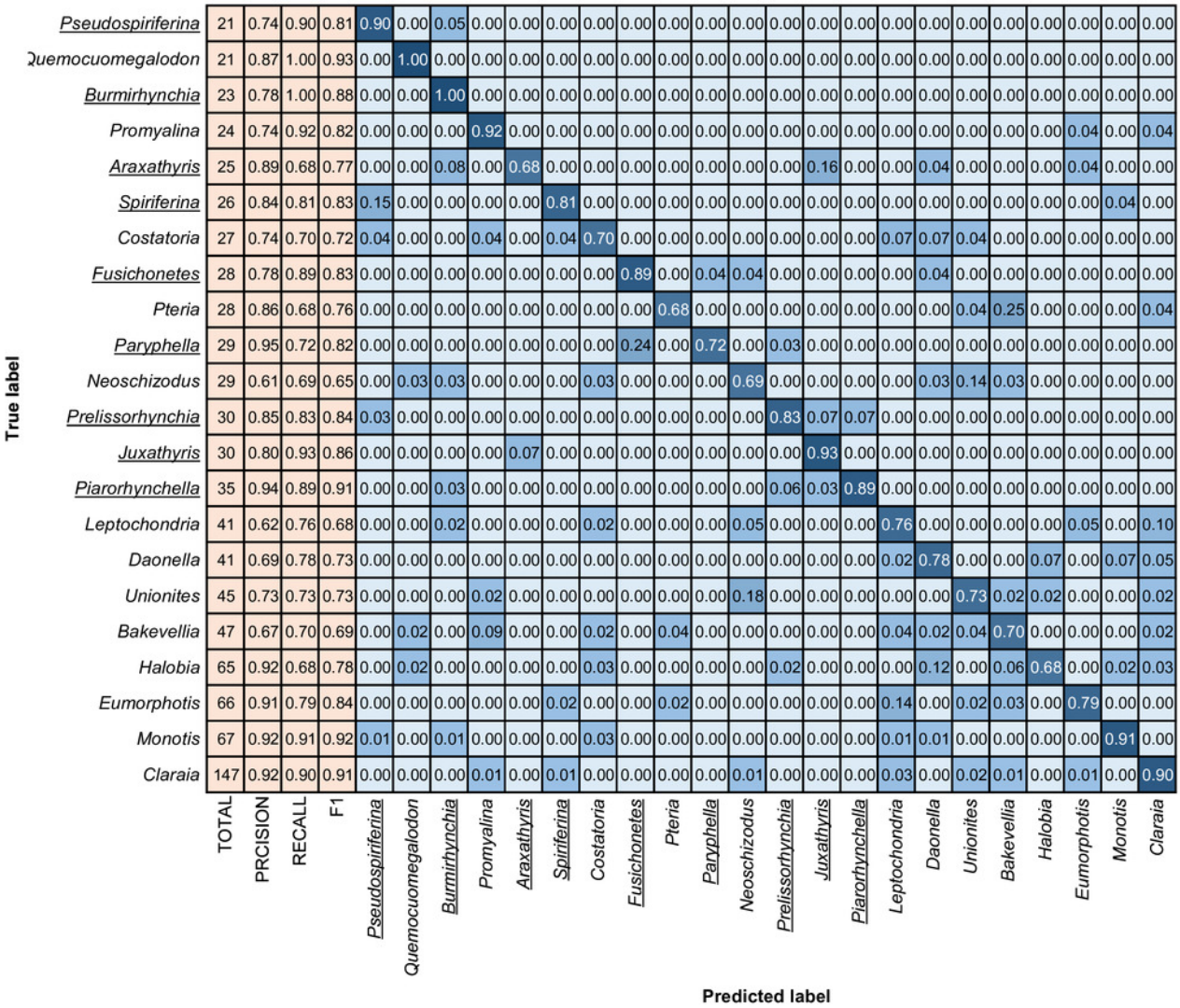


Figure 6

The training process of ATIM on genus mode using BBFID (scale A) (Order 22).

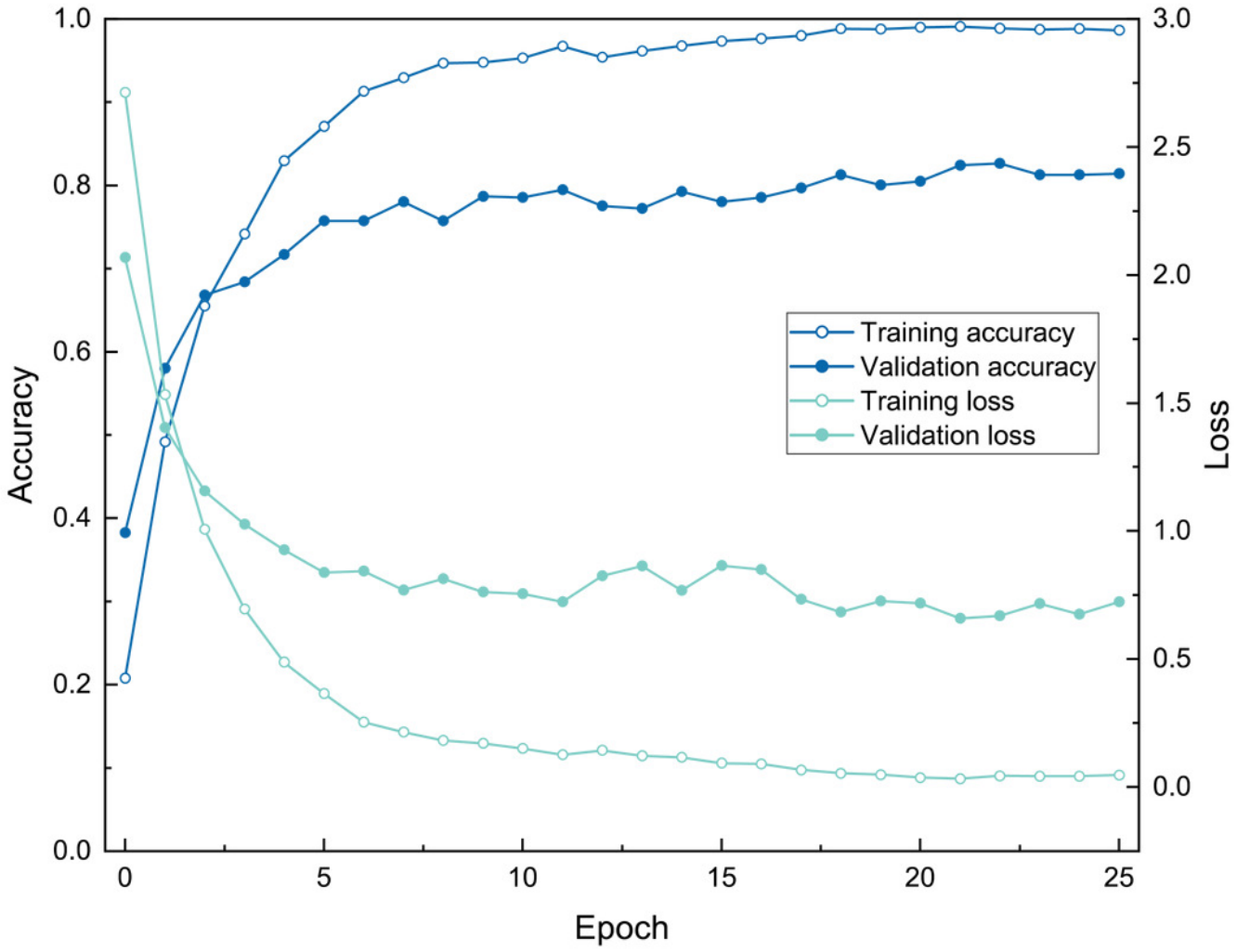


Figure 7

Feature maps of the most bivalve (Claraia) and brachiopod (Piarorhynchella) fossils in BBFID, plotted by extracting model (Order 22) intermediate output.

Fossil images are from Song (2018), and Popov and Zakharov (2017).

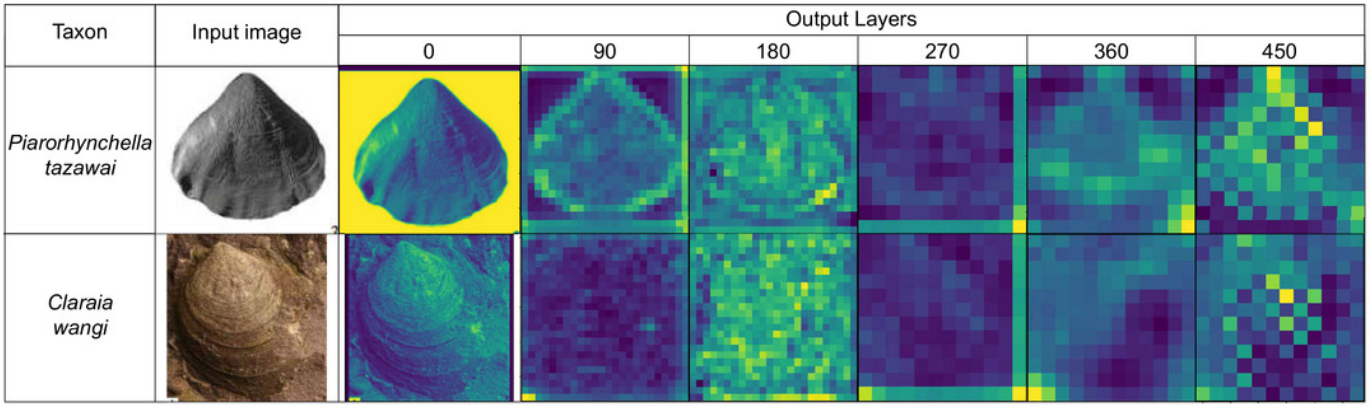


Figure 8

Fossil morphological feature distribution maps.

(A) Training set data and (B) validation set and test set data were fitted simultaneously using t-SNE. The accuracy of the original identification model is 81.01%. The horizontal and vertical coordinates in the figure are the two dimensions obtained by t-SNE (n_components=2, perplexity=10, init='pca', learning_rate=1, n_iter= 6000, n_iter_without_progress=6000). The numbers represent different genera, where the orange numbers represent the bivalves and the blue numbers represent the brachiopods. The detailed correspondence is 0: Pseudospiriferina, 1: Quemocuomegalodon, 2: Burmirhynchia, 3: Promyalina, 4: Araxathyris, 5: Spiriferina, 6: Costatoria, 7: Fusichonetes, 8: Pteria, 9: Paryphella, 10: Neoschizodus, 11: Prelissorhynchia, 12: Juxathyris, 13: Piarorhynchella, 14: Leptochondria, 15: Daonella, 16: Unionites, 17: Bakevellia, 18: Halobia, 19: Eumorphotis, 20: Monotis, 21: Claraia.

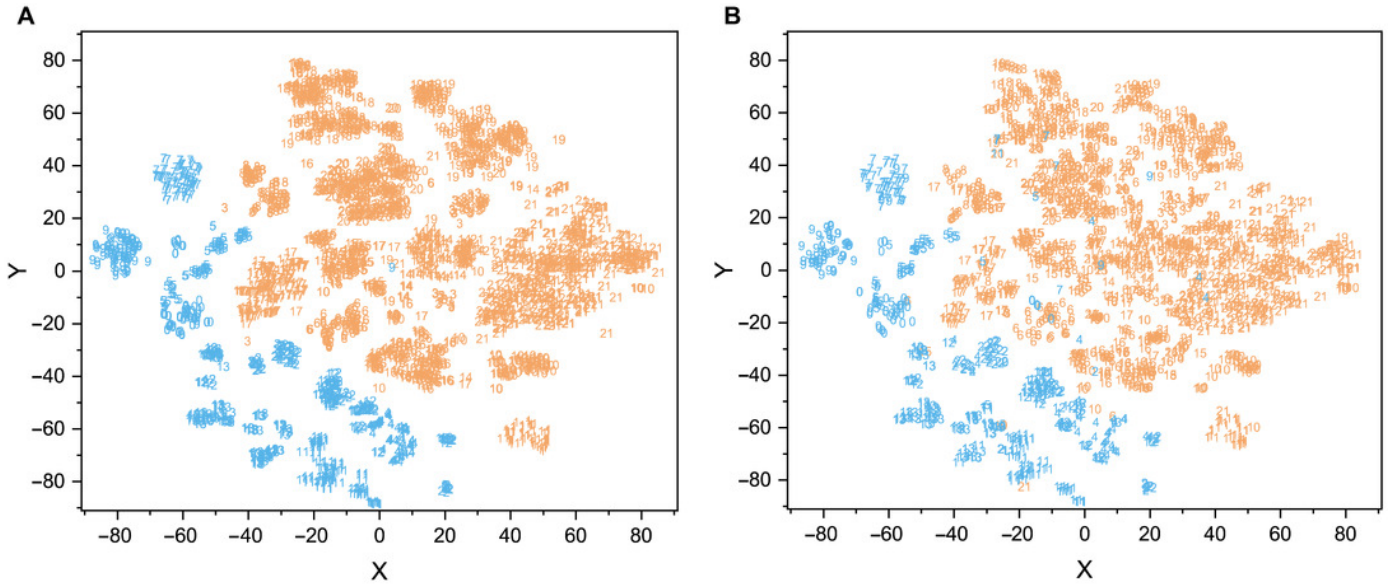

A
B
X
Y
80
60
40
20
0
−20
−40
−60
−80

Table 1(on next page)

Identification accuracy training on BBFID-1 (scale A) at the genus level with different model architectures and hyperparameters.

Architectures in this table are shown in Fig. 2. “Trainable layers of functional layers” represents the size of the parameters that can be trained. “None” means that all layers of the backbone are frozen and the parameters involved in these layers cannot be trained. These parameters maintain the values at the time of model initialization. “Half layers” means that half of the backbone layer parameters are frozen, while “All layers” means that all parameters of this model are not frozen and can be updated during the training process. This setting has an impact on both the model training process and the model performance.

TABLE 1. Identification accuracy training on BBFID-1 (scale A) at the genus level with different model architectures and hyperparameters. Architectures in this table are shown in Fig. 2. "Trainable layers of functional layers" represents the size of the parameters that can be trained. "None" means that all layers of the backbone are frozen and the parameters involved in these layers cannot be trained. These parameters maintain the values at the time of model initialization. "Half layers" means that half of the backbone layer parameters are frozen, while "All layers" means that all parameters of this model are not frozen and can be updated during the training process. This setting has an impact on both the model training process and the model performance.

Order	Backbone	Batch size	Trainable layers of functional layers	Reduce LR on plateau	Epochs	Max. training accuracy	Min. training loss	Max. validation accuracy	Min. validation loss	Test accuracy	Test loss
1	VGG-16	32	None	Yes	50	0.8648	0.4212	0.6444	1.1440	0.6281	1.2512
2	VGG-16	32	Half layers	Yes	40	0.9959	0.0181	0.7515	0.9126	0.7330	0.8444
3	VGG-16	32	All layers	Yes	50	0.7670	0.6080	0.5698	1.3465	0.5386	1.4802
4	VGG-16	32	All layers	No	36	0.3609	1.8002	0.3338	2.0523	0.0957	3.0871
5	Inception-ResNet-v2	8	None	Yes	50	0.3236	1.9945	0.3385	2.0345	0.3225	2.1000
6	Inception-ResNet-v2	8	Half layers	Yes	50	0.7363	0.7163	0.5263	1.4931	0.4877	1.5584
7	Inception-ResNet-v2	8	All layers	Yes	46	0.9959	0.0216	0.7934	1.2041	0.7778	2.5044
8	Inception-ResNet-v2	8	All layers	No	46	0.9805	0.0602	0.7981	0.8178	0.6590	1.2590
9	EfficientNetV2s	8	None	Yes	50	0.5693	1.2799	0.5419	1.4210	0.4923	1.5424
10	EfficientNetV2s	8	Half layers	Yes	50	0.9708	0.1013	0.7624	0.8314	0.7515	0.8633
11	EfficientNetV2s	8	All layers	Yes	44	0.9959	0.0139	0.8338	0.6130	0.8302	0.6807
12	EfficientNetV2s	8	All layers	No	37	0.9825	0.0578	0.8136	0.7905	0.7886	0.8122

Table 2(on next page)

Model performance using BBFID-1, BBFID-2 and BBFID in EfficientNetV2s architecture.

Learning rate starts from 1e-4 and the epoch is limited to less than 51. Test accuracy / Test loss means the accuracy / loss of the saved model.

TABLE 2. Model performance using BBFID-1, BBFID-2 and BBFID in EfficientNetV2s architecture. Learning rate starts from 1e-4 and the epoch is limited to less than 51. Test accuracy / Test loss means the accuracy / loss of the saved model.

Order	MODE	Dataset	Scale	> x images each taxon	Number of categories	Learning rate in the end	Epochs	Max. training accuracy	Min. training loss	Max. validation accuracy	Min. validation loss	Last epoch test accuracy	Last epoch test loss	Test accuracy	Test loss
13	Genus	BBFID-1	C	10	156	1.25E-05	49	0.9972	0.0080	0.5990	1.8758	0.5848	1.9234	0.5834	1.9320
14	Genus	BBFID-1	B	50	34	1.25E-05	34	0.9939	0.0281	0.7185	1.1308	0.6916	1.1420	0.7173	1.1142
15	Genus	BBFID-1	A	100	13	5.00E-05	29	0.9866	0.0446	0.8090	0.6661	0.8256	0.6719	0.8210	0.6650
16	Genus	BBFID-2	C	10	223	1.00E-04	22	0.9908	0.0848	0.5320	2.1067	0.4919	2.2493	0.5004	2.2021
17	Genus	BBFID-2	B	50	32	5.00E-05	21	0.9929	0.0483	0.7370	0.9765	0.7170	1.0273	0.7135	1.0625
18	Genus	BBFID-2	A	100	9	5.00E-05	25	0.9878	0.0486	0.8636	0.5007	0.8259	0.5409	0.8543	0.4904
19	Genus	BBFID	C	10	379	2.50E-05	35	0.9974	0.0134	0.5567	2.0772	0.5353	2.2279	0.5371	2.2333
20	Genus	BBFID	B	50	66	2.50E-05	27	0.9933	0.0299	0.7335	1.1080	0.7192	1.1866	0.7066	1.2000
21	Genus	BBFID	/	60	47	1.25E-05	34	0.9961	0.0177	0.7538	1.0721	0.7742	0.8506	0.7626	0.8921
22	Genus	BBFID	A	100	22	5.00E-05	26	0.9907	0.0335	0.8261	0.6590	0.8190	0.6615	0.8145	0.6759
23	Species	BBFID-1	E	6	241	5.00E-05	31	0.9949	0.0345	0.6117	1.8168	0.5971	1.9054	0.6080	1.9233
24	Species	BBFID-1	D	8	179	1.00E-04	28	0.9938	0.0645	0.6251	1.6484	0.5810	1.8759	0.6299	1.6987
25	Species	BBFID-1	C	10	148	2.50E-05	32	0.9975	0.0289	0.6629	1.4035	0.6642	1.4147	0.6790	1.4161
26	Species	BBFID-1	B	50	8	5.00E-05	27	0.9871	0.0789	0.7460	0.7560	0.7984	0.7489	0.8140	0.6747
27	Species	BBFID-2	E	6	396	1.00E-04	23	0.9950	0.0726	0.5128	2.3015	0.4677	2.5728	0.4813	2.5160
28	Species	BBFID-2	D	8	265	1.00E-04	28	0.9983	0.0492	0.5590	1.9957	0.5411	2.0768	0.5349	2.1075
29	Species	BBFID-2	C	10	195	1.00E-04	25	0.9969	0.0647	0.6162	1.6714	0.5540	1.9768	0.5791	1.8711
30	Species	BBFID-2	B	50	8	5.00E-05	24	0.9968	0.0472	0.9494	0.1308	0.9615	0.1806	0.9519	0.1610
31	Species	BBFID	/	2	1436	5.00E-05	41	0.9956	0.0271	0.4975	2.4540	0.4274	2.8980	0.4330	2.9233
32	Species	BBFID	/	4	914	1.00E-04	28	0.9920	0.0758	0.4958	2.4228	0.4707	2.5650	0.4899	2.5005
33	Species	BBFID	E	6	637	1.00E-04	25	0.9934	0.0677	0.5521	2.0340	0.5067	2.3438	0.5142	2.2276
34	Species	BBFID	D	8	444	5.00E-05	26	0.9975	0.0291	0.6148	1.6785	0.5752	1.8458	0.5957	1.8470
35	Species	BBFID	C	10	343	2.50E-05	34	0.9991	0.0143	0.6472	1.5119	0.6476	1.4888	0.6397	1.5602
36	Species	BBFID	B	50	16	1.00E-04	23	0.9787	0.1037	0.8399	0.5760	0.8283	0.5472	0.8283	0.5487