Received: 25 July 2025 | Accepted: 12 January 2026

DOI: 10.1111/2041-210x.70264

Methods in Ecology and Evolution

RESEARCH ARTICLE

Multi-barcoding-based species identification using a hierarchical attention network with staged curriculum learning

Bin Ye[1,2] | Junfeng Xia[1,3] | Xia Wan[4] | Min Wu[5] | Satoshi Chiba[6,7]

[1]School of Life Sciences and Medical Engineering, Anhui University, Hefei, China; [2]School of Computer Science and Technology, Anhui University, Hefei, China; [3]Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, China; [4]School of Resources and Environmental Engineering, Anhui University, Hefei, China; [5]School of Life Sciences, Nanjing University, Nanjing, China; [6]Graduate School of Life Sciences, Tohoku University, Sendai, Japan and [7]Center for Northeast Asian Studies, Tohoku University, Sendai, Japan

Correspondence
Junfeng Xia
Email: jfxia@ahu.edu.cn

Xia Wan
Email: wanxia@ahu.edu.cn

Funding information
China Postdoctoral Science Foundation, Grant/Award Number: 2025M772813; National Natural Science Foundation of China, Grant/Award Number: 31872196; Japan Society for the Promotion of Science, Grant/Award Number: KAKENHI-17H04611

Handling Editor: Ai-bing Zhang

Abstract

1. DNA barcode-based species identification has become pivotal in biodiversity research. Current barcoding-based deep learning methods have advantages in species identification but exhibit significant limitations. By using COI barcoding alone, previous studies were unable to fully capture species identification features. Prior methods showed that high classification accuracy at the species level was high but reduced accuracy for higher taxonomic levels.
2. We constructed a multi-type barcoding dataset, including COI, 16S, H3, 18S, ITS1 and ITS2, for taxa of the two most species-rich subclasses within Gastropoda, Caenogastropoda and Heterobranchia. Multi-type barcoding was represented by RNA secondary structure embedding, and COI was additionally featured by DNA barcode embedding. To address uneven sample coverage among barcoding types, we designed a three-stage curriculum learning framework by training COI data, freezing parameters, fusing low-coverage barcoding data and then full tuning to obtain the final model, SnailBaLLsp. By employing a hierarchical attention network that propagates higher taxonomy-level classifications as features to lower levels, SnailBaLLsp improved the classification accuracy across all taxonomy levels.
3. SnailBaLLsp achieved 84.99% species-level and 92.93% genus-level accuracy for Gastropoda. Multi-barcode analysis outperformed single COI barcode at all taxonomy levels. The three-stage strategy boosted the unseen species prediction accuracy by 3.79%, while further fusion of the additional DNA barcoding embedding of COI improved it by 8.55%, surpassing state-of-the-art baselines across all levels. SnailBaLLsp effectively predicted label-unknown case data, with failure predictions in species-level linked to high intraspecific sequence divergence. It also exhibits strong cross-taxon transferability, as shown by its high performance after fine-tuning on multi-barcoding data of Bivalvia.
4. By successfully utilizing RNA secondary structure while also incorporating DNA barcoding features, we processed multi-type barcoding data with incomplete

coverage while providing multi-taxonomy-level species predictions from subclass to species, offering an advanced deep learning-based solution in the molecular barcoding aspect for integrated taxonomy.

KEYWORDS
deep learning, DNA barcoding, hierarchical taxonomy, RNA secondary structure, species identification

1 | INTRODUCTION

With the growing urgency of biodiversity conservation, species classification technology based on DNA barcodes has emerged as a focal point in bioinformatics research (Kartzinel et al., 2025; Li et al., 2024; Meier et al., 2024, 2025; Pollock et al., 2025; Romeijn et al., 2024; Sanchez et al., 2025). While current mainstream methods, such as BarcodeBERT (Arias et al., 2025), BarcodeMamba (Gao & Taylor, 2024) and BarcodeMAE (Safari et al., 2025), primarily relied on the single barcode of mitochondrial c oxidase subunit I (COI) gene sequence and have achieved significant progress in insect groups. However, these models exhibited limitations. On one hand, their training data was predominantly derived from insect barcode data, leading to a decline in performance when transferring across taxa. On the other hand, current methods considered only COI barcode data for training (Arias et al., 2025; Gao & Taylor, 2024; Li et al., 2024; Safari et al., 2025). However, a single COI barcode lacked sufficient classification robustness when confronted with incomplete lineage sorting (Ahmed et al., 2022; Antil et al., 2023; Liu et al., 2017; Moritz & Cicero, 2004; Yang & Rannala, 2010). These limitations constrained the application depth of DNA barcode technology in species identification and biodiversity studies.

Existing deep learning studies, such as Bioscan-5m (Gharaee et al., 2024) and CLIBD (Gong et al., 2024), have achieved high accuracy (approximately 99%) in species classification by using a large-scale dataset containing 5 million insect samples. Their success hinged on using joint information from DNA barcoding and taxonomic information to guide the learning of image features. These works effectively integrated both molecular and image modalities, providing good examples for the application of deep learning in integrated taxonomy. However, in terms of DNA barcoding modality, these studies relied solely on the single barcode of COI, without considering other barcode types. In the absence of taxonomic information and image data, species-level classification accuracy relied only on COI barcoding data dropped sharply from 99% to 69.4%, and genus-level accuracy also dropped to 47.03%–63.6% (Gharaee et al., 2024; Gong et al., 2024). This highlighted the insufficiency in DNA feature extraction capabilities in the previous models. In addition, previous models had high accuracy in predicting the final classification target, the species level, but their accuracy in predicting genera and higher taxonomy levels was not good enough. For instance, the specific foundation models for DNA barcoding has been reported to have high species-level prediction accuracy with 99.7% for BarcodeBERT (Arias et al., 2025), 99.2% for BarcodeMamba (Gao & Taylor, 2024) and 80.6% for BarcodeMAE in novel species identification (Safari et al., 2025), but there is also big room for improvement in genus-level accuracy with only 70.2% for BarcodeMamba (Gao & Taylor, 2024) and 74.2% for BarcodeMAE (Safari et al., 2025). This situation underscored two critical challenges. First, the hierarchical features of DNA sequences (from subclass to species) have not been fully captured, limiting the classification accuracy at higher taxonomy levels. Second, the information from a single barcode type of COI was insufficient to support fine-grained classification, necessitating the integration of multiple barcode types to provide complementary phylogenetic signals. However, multi-barcodes fusion faces challenges due to the varying coverage of different barcodes, which is a common problem and lack of systematic organization. The most pronounced example is found in the BOLD system, where the vast majority of data consists of COI sequences for metazoans, while other barcode types for animals are not currently included (Prosser et al., 2025). The incomplete coverage of multi-barcoding across samples results in more missing values and lower-quality data for some barcode modalities. Since the low-quality modalities could adversely affect the training outcomes, we propose developing a staged learning framework that integrates hierarchical consistency modelling with adaptive fusion of multiple barcoding methods, aimed at enhancing the classification accuracy at every taxonomy level.

Hierarchical attention networks and curriculum learning strategy offer innovative solutions to address the aforementioned bottlenecks in species classification. The biological taxonomy that was organized into a hierarchical structure—phylum, class, order, family, genus and species—inherently represents a multi-granularity cognitive process of species concepts (Qin et al., 2023). Hierarchical attention mechanisms are effective in capturing such hierarchical structures (Yang et al., 2016), allowing for the extraction and selection of common and distinguishing features at each granularity for modelling the relationship between local and global features across different taxonomy levels (Chen et al., 2025). Previous studies have successfully employed hierarchical attention mechanisms within hierarchical guided deep learning methods to classify species based on image data, including plants (Fan et al., 2015; Liu et al., 2025; Weinbach et al., 2025; Zhang et al., 2020), phytoplankton (Ito et al., 2023), insects (Bjerge et al., 2023), fish (Elhamod et al., 2022) and birds (Han et al., 2025), as well as using acoustic data for bird classification (Noumida & Rajan, 2022; Wang et al., 2024). However, existing studies focused on image and sound modalities, leaving the

application of hierarchical attention modelling to DNA barcodes largely unexplored. Compared to morphological and acoustic features, DNA sequences possess unique advantages in phylogenetics reflected by genetic variation and provide stable and quantifiable evolutionary distances. Moreover, curriculum learning could provide a suitable training strategy for multi-barcode data with different sample coverages, as this approach trains a machine learning model from easier data to harder data (Bengio et al., 2009; Wang et al., 2022). Therefore, integrating hierarchical attention mechanisms into staged curriculum learning with multi-barcode DNA analysis is expected to capture evolutionary constraints across taxonomy levels at the molecular level, thereby enhancing the species classification accuracy.

Here, we employed a dynamic hierarchical attention network and applied a staged curriculum training strategy to enhance the accuracy of barcode-based species classification, using Gastropoda as an empirical example. First, we designed a gated attention mechanism to explicitly capture dependency relationships among classification levels, from subclass to species. This mechanism ensures hierarchical consistency by transmitting knowledge from higher to lower levels through feature gating, aligning the classification results at lower taxonomy levels with outcomes at higher taxonomy levels. Second, we developed a three-stage training strategy to address the challenges of low sample coverage and high missing value proportion for multi-barcodes other than COI. We initially trained using only COI barcoding, which has the highest sample coverage and then fused the other types of DNA barcoding that have lower coverage, and finally performed full tuning on the complete barcoding data including COI and other barcode types. To enhance the generalization ability of the model, we used domain adaptive methods to transfer learned knowledge to the target domain, thereby improving its ability to predict new species barcodes. Our study confirms that introducing multi-level information of taxonomy and fusing multi-barcode types can improve the accuracy of species classification and alleviate incomplete lineage sorting. Our approach is lightweight and has low computational requirements, making it easy to deploy in practical species classification workflows and offers a novel approach for gastropod diversity research.

2 | MATERIALS AND METHODS

2.1 | Data processing

2.1.1 | Accession of DNA barcodes

The DNA barcodes of gastropods used in this study were primarily sourced from National Center for Biotechnology Information (NCBI) GenBank and the Barcode of Life Data Systems (BOLD) (Ratnasingham & Hebert, 2007, 2013). To unify the taxonomy of Gastropoda, we employed the accepted taxonomic system based on the MolluscaBase (https://www.molluscabase.org/, accessed by 2025/05/23) as the classification standard for Gastropoda species. Considering the variation in species numbers among different gastropod taxa and the availability of barcode data, we focused on species within the two subclasses of gastropods, Caenogastropoda and Heterobranchia, which have larger species numbers and higher barcode availability. We manually obtained barcode sequences from GenBank as the training dataset, instead of directly using the barcode data from BOLD. This choice was motivated by two main reasons. Firstly, this study aims to integrate COI and multi-barcodes to construct a multimodal deep learning species classification model; however, BOLD mainly stores COI and a small amount of ITS data, lacking other types of barcode data (Ratnasingham & Hebert, 2007, 2013). Secondly, as the taxonomic information of gastropods in BOLD has not been strictly matched with the MolluscaBase database, manually obtaining data based on the accepted species records in MolluscaBase ensures the reliability of the classification labels.

To access COI sequences from GenBank, we integrated accepted family names and their corresponding unaccepted family names and then used them as search key words to obtain COI/COX1 barcode sequences from GenBank (Tables S1 and S2). Then, we used the Entrez tool of biopython package (Cock et al., 2009; Maglott et al., 2007) to search for other barcodes typically used for Gastropoda identification, 16S rRNA, 18S rRNA, histone 3 (H3), internal transcribed spacer 1 (ITS1) and internal transcribed spacer 2 (ITS2) sequences in GenBank, based on the voucher and species of specimens corresponding to COI barcodes, thereby constructing a multimodal dataset composed of multi-types of DNA barcoding. We also obtained the barcoding data of Gastropoda and the corresponding classification labels from BOLDistilled (Prosser et al., 2025), the comprehensive DNA barcode reference libraries. However, the sample size of the barcoding data from BOLDistilled (n_sample = 42,714) and the number of species (n_species = 2722) were much smaller than those manually obtained from GenBank (n_sample = 76,577, n_species = 11,295). More importantly, a large portion of sequences in BOLDistilled (21,131/42,714 = 49.47%) overlapped with sequences from GenBank. This indicates that the barcoding data obtained from GenBank is more suitable as a training set than data from BOLDistilled. Therefore, after removing duplicates of sequences and performing necessary label cleaning, sequences from BOLDistilled were used as an independent testing dataset.

2.1.2 | Sequence filtering and label cleaning

This study employed RNA secondary structure embedding extracted by the RNA foundation model ERNIE-RNA (Yin et al., 2025) to represent features of DNA barcodes for species classification. This embedding approach enables representing of deep features from DNA barcodes while avoiding multiple sequence alignment across varying lengths. Due to the input sequence length constraint of ≤1022 bp for the ERNIE-RNA foundation model, we exclusively analysed barcode sequences with lengths ≤1022 bp. Among the six barcodes involved in this study, except for approximately 37.9% of 18S rRNA sequences (length > 1022 bp) were removed, the majority

of all other sequences were within the length limit (18S, 62.1%; COI, 100%; 16S, 98.6%; H3, 100%; ITS2, 85.5%; ITS1, 97.7%).

Based on the accepted names at various taxonomy levels from MolluscaBase, the taxonomic labels of subclass, order, superfamily, genus and species for COI/COX1 sequences obtained from GenBank were verified and cleaned. The main steps included removing sequences longer than 1022 bp and those with conflicting labels. For sequences from other sources (e.g. BOLDistilled), labels were confirmed according to the labels in the training set, with labels not appearing in the training set assigned as 'NA' and sequences longer than 1022 bp or with conflicting labels were also removed.

2.1.3 | Dataset construction

We manually downloaded sequences of multi-barcodes from GenBank released before December 31, 2024, to serve as the training and testing datasets for this study. A total of 76,577 COI/COX1 sequences from the Gastropoda were obtained, with the number of categories at each taxonomy level detailed in Table 1. For the training, testing and independent testing 1 datasets from GenBank (Table 1), we obtained sequence data for other barcoding types corresponding to the COI sequences based on their specimen voucher and species. If multiple sequences existed for the same specimen voucher within a single barcoding type, only one sequence was randomly selected. For records containing a complete ITS sequence that were combined ITS1 and ITS2, the first and second halves of these sequences were treated as ITS1 and ITS2 sequences, respectively. Consequently, we developed a multimodal barcoding dataset for Gastropoda composed of complete COI records and partial records for the other five barcoding types: 16S rRNA, H3, 18S rRNA, ITS1 and ITS2. The sequence counts for the six barcoding types were detailed in Table 2.

In addition, we accessed COI/COX1 sequences of Gastropoda from GenBank between January 1, 2025 and May 23, 2025, removed sequences overlapping with the training and testing sets and performed label cleaning and verification, resulting in 1927 sequences forming independent testing dataset 1 (Table 1). We also accessed sequence data of five other barcode types corresponding to COI sequences in independent testing dataset 1 (Table 2). Furthermore, we obtained 42,714 COI sequences from BOLDistilled (Prosser et al., 2025) as independent testing dataset 2. However, 49.47% of these sequences overlapped with those in the GenBank training and testing datasets. After performing sequence filtering and label cleaning, a total of 14,023 COI sequences were retained in Table 1. Finally, we combined the two independent testing sets to obtain an 'unseen' independent dataset containing 3146 full labelled samples consistent with the training labels (Table 2).

2.1.4 | Sequence augmentation

Due to the extreme imbalance of data across hierarchical taxonomy levels, for sequences belonging to categories with fewer samples,

TABLE 1 Total counts of original and augmented sequences, along with classification numbers of each taxonomy level in different datasets.

Dataset	$N_{original\_seq}$	$N_{agumented\_seq}$	$N_{subclass}$	N_{order}	$N_{superfamily}$	N_{family}	N_{genus}	$N_{species}$	Accession
Training-Validation-Testing	76,577	101,607	2	24	103	354	2753	11,295	GenBank, by 2024/12/31
Independent testing 1	1927	–	2 (0)[a]	18 (0)	57 (0)	105 (0)	224 (38)	302 (264)	GenBank, from 2025/01/01 to 2025/05/23
Independent testing 2	14,023	–	2 (0)	15 (0)	80 (0)	105 (0)	232 (38)	673 (541)	BOLDistilled_COI_Mar2025
Case Study (Gastropoda)	1056	–	2 (0)	15 (0)	50 (0)	87 (0)	163 (0)	256 (0)	GenBank, from 2025/06/01 to 2025/10/31
Model Transfer (Bivalvia)	35,934[b]	–	3	17	33	71	617	3060	GenBank (by 2025/11/20) and BOLDistilled_COI_Mar2025

[a]Values in parentheses represent the number of categories at the corresponding level where the classification label has not appeared in Training-Validation-Testing.
[b]The total number of COI sequences of Bivalvia includes 34,852 sequences accessed from GenBank and 1082 sequences from BOLDistilled_COI_Mar2025, which all match the NCBI taxonomy system of Bivalvia.

TABLE 2 Counts and length ranges of sequences for multiple barcode types in different datasets.

Dataset	COI	16S	H3	18S	ITS1	ITS2	Total
Training-validation-testing							
$N_{original}$	76,577	12,538	2822	885	2312	3910	76,577
$N_{augmented}$	25,030	7330	1940	725	560	1285	25,030
N_{total}	101,607	19,868	4762	1610	2872	5195	101,607
L_{min} (bp)	94	182	163	141	126	111	—
L_{max} (bp)	799	1021	400	1020	1022	1021	—
Independent testing 1							
N_{total}	1927	5	40	0	8	21	1927
L_{min} (bp)	154	426	226	—	484	343	—
L_{max} (bp)	686	439	327	—	631	837	—
Independent testing 2							
N_{total}	140,23	0	0	0	0	0	14,023
L_{min} (bp)	500	—	—	—	—	—	—
L_{max} (bp)	1006	—	—	—	—	—	—
Unseen[a]							
N_{total}	3146	3	37	0	0	12	3146
L_{min} (bp)	154	428	291	—	—	777	—
L_{max} (bp)	1000	439	327	—	—	837	—
Case study (Gastropoda)							
N_{total}	1056	12	104	78	0	0	1056
L_{min} (bp)	179	436	270	407	—	—	—
L_{max} (bp)	707	676	369	466	—	—-	—
Model transfer (Bivalvia)							
N_{total}	35,934	1951	230	40	1479	142	35,934
L_{min} (bp)	103	221	148	196	109	242	—
L_{max} (bp)	1011	892	760	1001	1013	934	—

[a]An independent testing dataset formed by merging independent testing 1 and independent testing 2, where all samples have complete labels across every taxonomy level.

we randomly masked a certain proportion of nucleotides as 'N' in the sequence, thereby generating new samples for these categories to reduce imbalance to some extent in both the training and testing datasets. Specifically, we dynamically set the masked nucleotide ratio and the maximum number of augmented sequences per original sequence according to the number of existing true sequences at the genus level. This dynamic enhancement increased the number of sequences per genus, while the generated sequences retained the species labels of their original true sequences (Table S3). Following this approach, we first generated 25,053 new sequences for COI. Subsequently, we augmented sequences for the other five barcoding types, ensuring that the generated sequences for these five barcoding types were as matched as possible with the generated COI samples, thereby obtaining generated data with modalities/barcoding types as complete as possible. To achieve this, based on the already generated COI sequence data, if a species has been enhanced in the COI barcode and also has true sequences in other barcode types, we enhanced the sequences of this species in those barcoding types using the same dynamic masking and augmentation method as that in COI type. Finally, we obtained a training and testing dataset comprising 101,607 multimodal barcoding sequences, each containing complete COI records and partial records of the other five barcoding types (Table 2). No data enhancement was performed for the sequences of the independent testing dataset.

2.1.5 | Embedding extraction using RNA foundation model

The RNA secondary structure features of DNA barcoding have been proven to be effective features for species classification tasks (Acharya et al., 2022; Coleman, 2015; Mohanty et al., 2023; Pere et al., 2024). In this study, we used ERNIE-RNA (Yin et al., 2025), a foundation model for predicting RNA secondary structure, to represent the RNA secondary structure of DNA barcoding. The biggest advantage of using this tool for species classification is that it can avoid the sequence alignment process of different DNA barcoding by extracting embedding with consistent dimensions and greatly simplifying the application of DNA barcodes in species classification. We used the extract_imbedding.py script (https://github.com/Bruce-ywj/ERNIE-RNA) to extract the RNA secondary structure embedding of each DNA barcode with dimension 12×768.

2.2 | Model architecture

We proposed a staged curriculum learning architecture for hierarchical taxonomic classification of multimodal DNA barcoding sequences. The training strategy contains three progressive stages, integrating six molecular markers—COI, 16S rRNA, H3, 18S rRNA, ITS1 and ITS2—through dynamic feature fusion and hierarchical attention mechanisms. This strategy was designed based on two main considerations. First, species classification inherently reflects a hierarchical structure of different taxonomy levels, which can be accurately captured by a hierarchical attention network to model relationships among features across different hierarchical taxonomy levels. Second, the availability of multi-type DNA barcoding data varies significantly; for example, COI, as one of the most typical DNA barcoding markers, has a large number of available samples, while other barcoding types have relatively fewer available samples, resulting in a lower coverage of records and a higher proportion of missing values. To mitigate the interference of low-quality modality data on model accuracy, we employed a multi-stage, progressive curriculum learning strategy for model training (Figure 1).

2.2.1 | COI features engineering

Given that foundational models based on 5 million COI barcodes have been developed (Gharaee et al., 2024; Gong et al., 2024), in

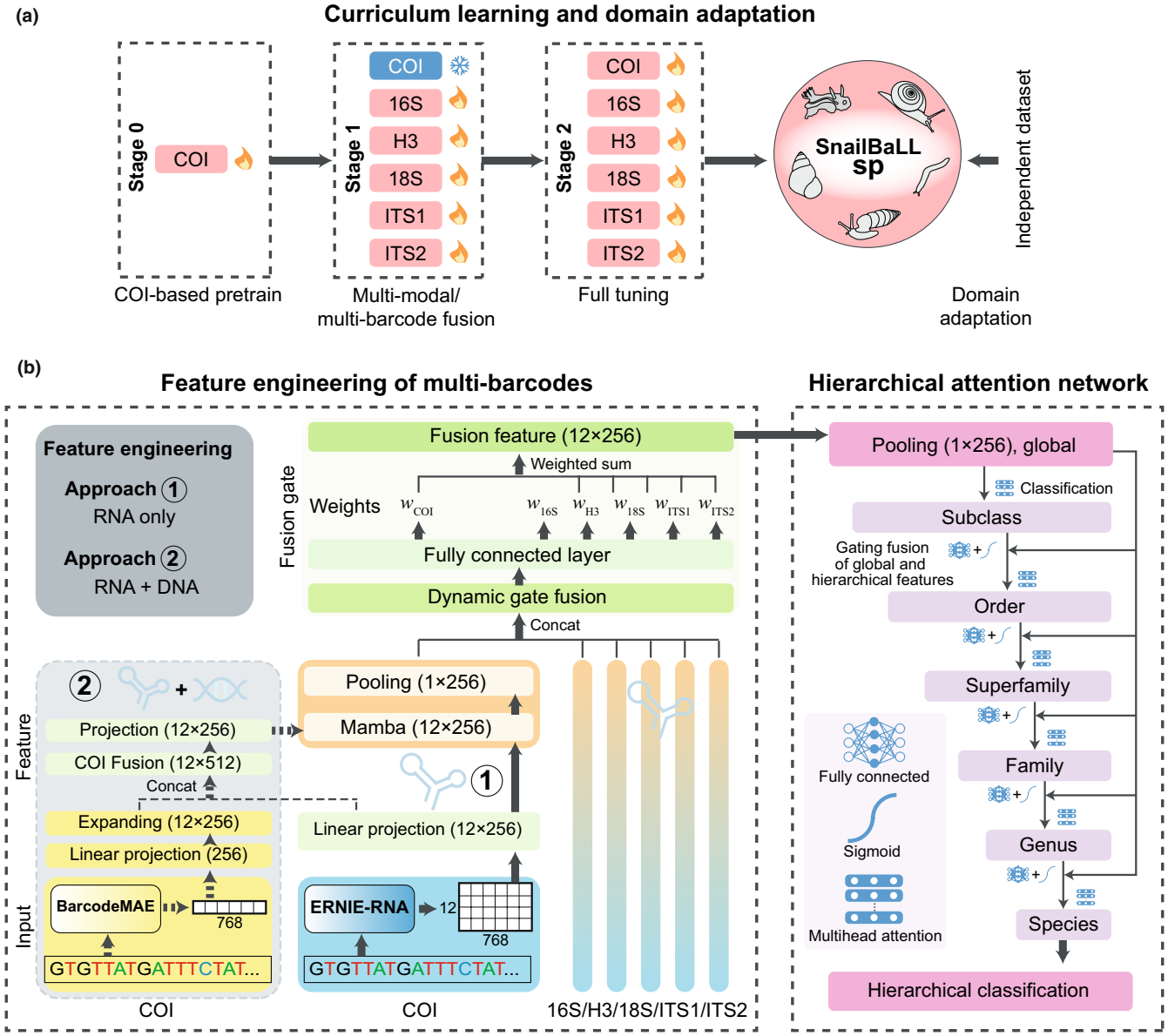

FIGURE 1 Architecture of SnailBaLLsp, a model for Gastropoda species classification based on DNA barcodes represented by the RNA large language model using a three-stage curriculum learning strategy and hierarchical attention network. (a) The staged curriculum learning and domain adaptation. The flame and snow coin represent the data/modality that is active or frozen in the corresponding stage. (b) Blocks of feature engineering of multi-barcodes and the hierarchical attention network.

addition to the RNA secondary structure emphasized in this study, the DNA barcoding embedding might also enhance our model's predictive performance. Specifically, embedding features output by COI barcoding foundational models such as BarcodeMAE (Safari et al., 2025) could be beneficial. To investigate the roles of RNA secondary structure embedding and DNA barcoding embedding in the model, we adopted two distinct feature engineering approaches for the COI sequences. The first approach used only the RNA secondary structure features of COI sequences extracted by the foundation model ERNIE-RNA (Yin et al., 2025). The second approach fused the RNA secondary structure features with the COI barcoding embedding features output by BarcodeMAE (Figure 1b). Following the two approaches, the processed features were separately fed into the Mamba module. This setup enabled the subsequent fusion with RNA secondary structure embedding from other barcode types, before being input into the hierarchical attention neural network. By comparing model performance using the two different approaches of COI feature engineering, we validated both the effectiveness of the features themselves and the efficacy of the model architecture.

2.2.2 | Staged curriculum learning

We firstly applied the approach 1 of COI feature engineering—the RNA secondary structure embedding only—to construct our model. We used the pre-trained foundation model ERNIE-RNA (Yin et al., 2025) to encode each DNA barcode sequence into a 12×768 embedding of RNA secondary structure for input representation. For modality $m \in \{\text{COI}, 16S, H3, 18S, ITS1, ITS2\}$, the embedding $E_m \in \mathbb{R}^{12\times768}$ is projected into a shared latent space using a linear layer $H_m = W_m E_m + b_m$, where W_m and b_m are learnable parameters. The projected embedding $H_m \in \mathbb{R}^{12\times256}$ then fed into the subsequent modules. Then, we performed a curriculum learning with three stages.

Firstly, we only used the COI modality to train the model. The model minimizes the hierarchical loss $\mathcal{L}_{\text{hier}}$ while other modalities are disabled:

$$\mathcal{L}_0 = \mathcal{L}_{\text{hier}}\left(f_{\theta_{\text{COI}}}\left(H_{\text{COI}}\right), y\right) \tag{1}$$

where f denotes the classification network, H_{COI} is trainable parameters, and y denotes taxonomic labels.

Secondly, we performed multimodal fusion training by freezing the parameters of the COI projection layer and the corresponding trained Mamba and hierarchical classification layers while activating the fusion parameters of the other five modalities. We designed a gated fusion mechanism to dynamically fuse the multimodal features. We computed a pooled representation for each modality and then generated the fusion weights of the modalities:

$$\alpha = \text{softmax}\left(\text{MLP}\left(\text{concat}\left(p_{\text{COI}}, p_{16S}, \cdots, p_{\text{ITS2}}\right)\right)\right) \tag{2}$$

where $\alpha \in \mathbb{R}^6$ and $\sum \alpha_i = 1$. The fused representation was calculated as $F = \sum_m \alpha_m \odot H_m$. MLP represents multilayer perceptrons.

Thirdly, we unfroze all parameters in the model and performed full fine-tuning for joint optimization.

We then applied the approach 2 of COI feature engineering—the fusion of RNA secondary structure embedding and COI barcode feature embedding—to construct a parallel model. In brief, we firstly calculated the fused features of the two embeddings:

$$F_{\text{COI-fused}} = \text{RELU}\left(\text{LayerNorm}\left(W\left[H_{\text{COI}}; H_{\text{COI-MAE}}\right] + b\right)\right) \tag{3}$$

where $H_{\text{COI}} \in \mathbb{R}^{12\times256}$ is the projected feature of RNA secondary structure of COI, $H_{\text{COI-MAE}} \in \mathbb{R}^{12\times256}$ is the expanded features of COI barcode embedding extracted by BarcodeMAE (Safari et al., 2025), and W is the weight of the fusion layer. Among them, H_{COI} is output by linear projection of the original RNA secondary structure of COI, and $H_{\text{COI-MAE}}$ is firstly output by linear projection of the original COI embedding from BarcodeMAE and then obtained by transforming it into a 3D tensor and replicating 12 times in the sequence dimension of the three dimensions for ensuring the global information can be obtained from any position. This processing ensures the embeddings of two modalities can be fused in the same dimension (Figure 1b). Consistently, the number of modalities became seven and the fusion weights of the seven modalities were $\alpha = \text{softmax}\left(\text{MLP}\left(\text{concat}\left(p_{\text{COI}}, p_{\text{COI-MAE}}, p_{16S}, \cdots, p_{\text{ITS2}}\right)\right)\right)$, where $\alpha \in \mathbb{R}^7$ and $\sum \alpha_i = 1$.

2.2.3 | Hierarchical attention network

The fused feature F is processed by a Mamba block which is a state-space model (SSM) with selective scanning (Gao & Taylor, 2024). The Mamba block outputs a sequence of feature vectors $S \in \mathbb{R}^{12\times256}$. By averaging the sequence, we obtain a global representation:

$$g = \frac{1}{12}\sum_{t=1}^{12} S^{(t)} \tag{4}$$

Then, we generated hierarchical predictions through gated attention across the six taxonomic levels $l \in \{\text{subclass}, \text{order}, \text{superfamily}, \text{family}, \text{genus}, \text{species}\}$. For the first level subclass, the self-attention output of the global representation is calculated:

$$f^{(0)} = \text{MultiHeadAttention}(g, g, g) \tag{5}$$

For the other taxomomy levels when $l \geq 1$:

$$z^{(l)} = \sigma\left(\text{Linear}^{(l)}\left(\text{concat}\left(f^{(l-1)}, g\right)\right)\right)$$

$$\tilde{f}^{(l)} = z^{(l)} \odot f^{(l-1)} + \left(1 - z^{(l)}\right) \odot g$$

$$f^{(l)} = \text{MultiHeadAttention}\left(\tilde{f}^{(l)}, \tilde{f}^{(l)}, \tilde{f}^{(l)}\right)$$

$$\hat{y}^{(l)} = \text{softmax}\left(\text{Linear}^{(l)}\left(f^{(l)}\right)\right) \tag{6}$$

Then, we optimize the hierarchical loss combined weighted cross-entropy (CE):

$$\mathcal{L}_{\text{hier}} = \sum_{l} \omega_l \cdot \text{CE}\left(\hat{y}^{(l)}, y^{(l)}\right) \tag{7}$$

Through hyperparameter tuning, we choose the best hyperparameter based on the performance. We use a group of hierarchical weights (e.g. $\omega = [1.0,0.8,0.6,0.4,0.2,0.1]$) corresponding to the weights of six taxonomy levels, subclass, order, superfamily, family, genus and species, in the hierarchical loss function. This setting is based on the assumption that accurate high-level classification is the foundation of accurate low-level classification, which allows the model to focus more on higher-level classification accuracy, thereby ensuring the correctness of hierarchical classification. We also use learning rate 10^{-4}, batch size 64 and early stopping with patience 30 epochs for training with 300 epochs.

2.2.4 | Domain adaptive

In order to further improve the model's generalization and prediction performance for barcode data of unknown species, we implemented a domain adaptation approach built upon our curriculum-trained model to align feature distributions between the source (the training and validation dataset) and target (the independent testing dataset) domain.

In brief, we firstly transformed sequence features into a domain-agnostic representation to extract the domain-invariant features:

$$h_{\text{align}} = \text{RELU}\left(W_2 \cdot \text{RELU}\left(W_1 \cdot h + b_1\right) + b_2\right) \tag{8}$$

where h is the pooled sequence feature, $W_1 \in \mathbb{R}^{256\times512}$, $W_2 = \mathbb{R}^{512\times256}$, and b is learnable parameters.

Then, we constructed an adversarial domain classifier to distinguish source and target domains:

$$d = \sigma\left(W_d^{(2)} \cdot \text{RELU}\left(W_d^{(1)} \cdot h_{\text{align}} + b_d^{(1)}\right) + b_d^{(2)}\right) \tag{9}$$

Thirdly, we used a gradient reversal layer (GRL) and reversed the gradient of domain classification loss during backpropagation to facilitate the feature extractor in generating domain-invariant features:

$$\nabla_\theta^{\text{GRL}} = -\alpha \cdot \nabla_\theta \mathcal{L}_{\text{domain}} \tag{10}$$

where $\alpha = 0.1$ controls adaptation strength.

Finally, we performed adversarial optimization to calculate the joint loss combined hierarchical classification and domain confusion based on binary cross-entropy (BCE):

$$\mathcal{L} = \mathcal{L}_{\text{hier}}\left(f\left(X_{\text{source}}\right), y_{\text{hier}}\right) + \lambda \cdot \text{BCE}\left(d\left(X_{\text{batch}}\right), y_{\text{domain}}\right) \tag{11}$$

where $X_{\text{batch}} = \text{concat}\left(X_{\text{source}}, X_{\text{traget}}\right)$, y_{hier} is the hierarchical taxonomic labels, $y_{\text{domain}} \in \{0, 1\}$ is the domain labels, and λ is the balancing parameter.

2.3 | Multimodal barcode analysis

To validate that multiple types of DNA barcoding provide richer and more accurate features for species classification than using a single COI barcode, we evaluated the impact of multimodal barcoding integration on model performance. We performed pairwise integration of COI with each of the other five barcoding types and trained models under the staged curriculum learning architecture (Figure 1a), which produced a fused model in Stage 1 and a fully fine-tuned model in Stage 2 for each pair of barcodes. We then comparatively analysed these models against COI-only models and models integrating all six barcodes. Since RNA secondary structure embedding was applied to all barcoding types, we represented COI features using only Approach 1 with only RNA secondary structure embedding of COI rather than by Approach 2 (Figure 1b) to ensure fair comparison in this analysis.

2.4 | Comparison analysis

We compared SnailBaLLsp's predictions with the recent state-of-the-art (SOTA) foundational models specifically designed for COI barcode sequences, namely BarcodeMamba (Gao & Taylor, 2024) and BarcodeMAE (Safari et al., 2025). Both models were enhanced upon BarcodeBERT and have been reported to achieve significantly higher performance than BarcodeBERT (Arias et al., 2025; Gao & Taylor, 2024; Safari et al., 2025); therefore, here comparison analysis to BarcodeBERT was not involved. Particularly, BarcodeMAE leveraged a training set comprising 5 million insect COI barcoding samples and potentially had high generalization capabilities. Consequently, we used BarcodeMamba and BarcodeMAE to evaluate the predictive performance on our gastropod COI barcoding data and compared their results with those of SnialBaLLsp on the testing and the independent testing dataset. Other studies, such as CLIBD, used the same 5 million insect samples of COI barcoding dataset as that used in BarcodeMAE to integrate insect images and achieved notable performance (Gharaee et al., 2024; Gong et al., 2024). However, our model SnialBaLLsp is designed only for DNA barcodes and does not currently integrate image data of Gastropoda. Therefore, CLIBD was excluded from our comparative analysis.

We processed the training and testing datasets of Gastropoda into input datasets suitable for BarcodeMamba and BarcodeMAE, respectively, according to the formats of the CandianInverterbrates1.5 M (Arias et al., 2025; Gao & Taylor, 2024) and Bioscan-5 M datasets (Gharaee et al., 2024). In detail, the datasets for BarcodeMamba consist of supervisied_train, supervisied_val and supervisied_test (https://github.com/bioscan-ml/BarcodeMamba), which correspond to the training, validation and testing datasets used in SnailBaLLsp. The datasets for BarcodeMAE consist of supervisied_train and unseen (https://github.com/bioscan-ml/BarcodeMAE). Because the formatted data for BarcodeMamba and BarcodeMAE do not include subclass and superfamily labels, comparison analysis was performed

only at the order, family, genus and species levels. Here, the supervisied_train set corresponded to the training set used by SnailBaLLsp. The unseen set corresponded to SnailBaLLsp's testing set and unseen independent testing set (Table 2), respectively. Using this setup, we fine-tuned the pre-trained BarcodeMAE model on supervisied_train and then evaluated its classification performance separately on SnailBaLLsp's testing set and the unseen independent testing set.

2.5 | Case study and model transfer

To further evaluate the predictive capability of SnailBaLLsp on new case data and analyse its transferability to barcode classification tasks beyond gastropods, we constructed two new datasets: a Gastropoda case dataset and a multi-barcode dataset for Bivalvia that belongs to the same phylum Mollusca as Gastropoda. The Gastropoda case dataset was accessed from GenBank (from 2025/06/01 to 2025/10/31), ensuring no overlap with training, testing or independent testing datasets (Table 1). This dataset comprises 1056 COI and other barcode sequences (Table 2). They were treated as label-unknown data during prediction, and their known taxonomic labels (Table 1) were used as ground-truth to evaluate the performance. The Bivalvia dataset includes multi-barcode sequences accessed from GenBank searched by order keywords (Table S4) and COI sequences filtered from BOLDistilled_COI_Mar2025 (Prosser et al., 2025) processed identically to gastropod data. After data cleaning and preprocessing, we assigned hierarchical taxonomic labels of the Bivalvia dataset based on the NCBI taxonomy system using NCBItaxa function in ete3 module (Huerta-Cepas et al., 2016), which is different from the MolluscaBase taxonomy system. This mainly considers excluding the impact of classification systems on model transfer. The final dataset contained 35,934 COI sequences across 3060 species (Table 1), along with other barcode types with sample coverage rates as low as 0.1%–5.4% (Table 2).

We applied SnailBaLLsp (RNA + DNA) models, including three staged and domain-adapted models—which demonstrated high classification accuracy—to the Gastropoda case dataset for hierarchical prediction. To investigate the reasons for decreased accuracy at the species level, we integrated prediction results from both the independent unseen dataset and the case dataset, focusing specifically on COI sequences in terms of sequence length and genetic divergence. We firstly compared sequence lengths between success and failure prediction sequences. Then, to assess whether sequence divergence contributes to species-level misclassification, we performed comparison analysis between the genetic distance between failure and success prediction sequences within the same species (Failure–Success distance) and the mean intraspecific distance, which was represented by the genetic distance among success prediction sequences of the same species (Success–Success distance).

Furthermore, we fully fine-tuned each SnailBaLLsp (RNA + DNA) model on the Bivalvia dataset and evaluated the hierarchical classification accuracy of the fine-tuned models to assess the transferability of SnailBaLLsp across taxa beyond Gastropoda.

3 | RESULTS

Sequence augmentation contributed to improving prediction accuracy. It increased the number of sequences per category at each taxonomy level, which alleviated the distribution imbalance of classifications at each taxonomy level to a certain extent (Figure S1A). Compared to the model trained on the original sequences, the model trained on augmented data had higher classification accuracy at every taxonomy level across the validation, testing and independent testing datasets (Figure S1B).

The three-stage curriculum learning architecture based on a hierarchical attention network was effective and achieved a stable training process on the multimodal barcode data of Gastropoda and high prediction accuracy at every taxonomy level (Figure S2). During stage 0 using only COI with the RNA secondary structure features (Approach 1, Figure 1), the validation loss decreased while prediction accuracy for the validation dataset increased steadily at each taxonomy level. The COI-based model in stage 0 achieved prediction accuracies exceeding 80% and 90% for species level and genus level, respectively, with accuracies ranging from 91%–99% for levels of family, superfamily, order and subclass. The multimodal fusion model in stage 1 achieved accuracies above 84% for all taxonomy levels, representing a 4% improvement for species level over stage 0. Similarly, the multimodal full-tuning model in stage 2 also achieved an exceeding 4% improvement for species compared to stage 0 (Figure S2).

The multi-type DNA barcoding provided benefits for species classification than using only the COI barcode at every taxonomy level (Figure 2). Fusing COI with any other barcode improved the classification accuracy at each taxonomy level, with the COI + 16S fusion model showing the most notable improvement (Figure 2). This might be due to the higher sample coverage of 16S sequences (16.4%) in the dataset compared to other barcoding types. At the genus level, the COI + 18S fusion model also demonstrated an improvement, while at the species level, the COI + H3 fusion model exhibited a marked improvement. Among these, the model fusing all barcode types performed best, though its improvement over the COI + 16S model was not significant.

The staged learning strategy and the domain adaptive method consistently improved prediction performance of our model (Figure 3). Compared to the COI-based model in Stage 0, the multimodal full tuning model in Stage 2 enhanced the species-level classification accuracy from 81.85% to 84.22% on the testing dataset and from 42.88% to 43.64% on the unseen dataset. Although there were no obvious changes in accuracy and F1 score between Stage 2 (multimodal full tuning) and Stage 1 (multimodal fusion) on the validation and testing sets, a notable improvement was observed on the unseen dataset. After performing the domain adaptation, the multimodal full tuning model (Stage 2) further enhanced the species-level classification accuracy to 84.99% on the testing dataset and to 46.47% on the unseen dataset (Figure 3). The application of domain adaptive method not only preserved high performance on the validation and testing sets, but also significantly enhanced performance on the

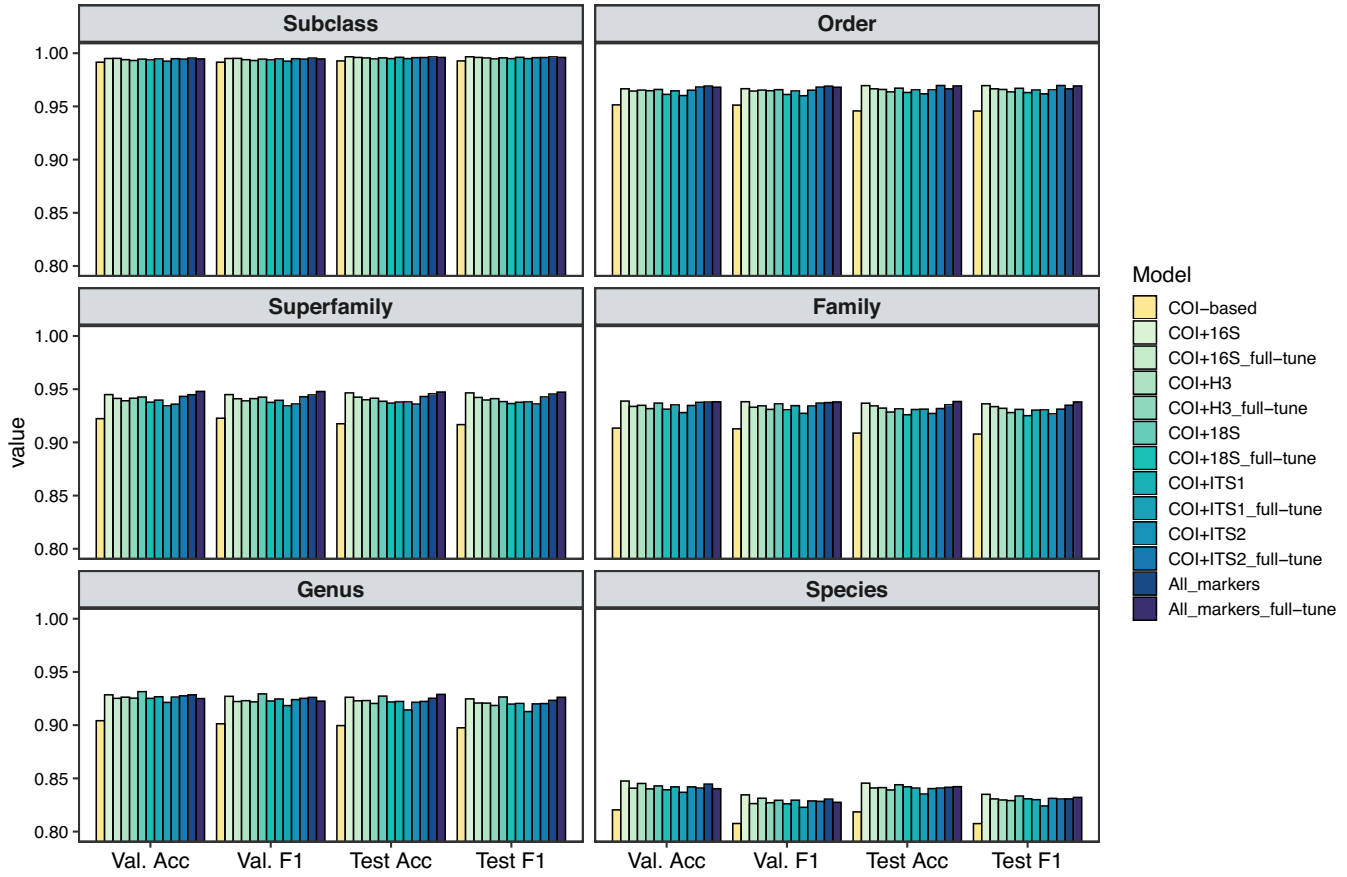

FIGURE 2 Comparison of the performance of COI-based model by using COI feature engineering of Approach 1, fusing COI with one of the other modalities, fusing all modalities of all barcode markers and corresponding full parameter fine-tuning models at each taxonomy level. The *x* axis label 'Val.' represents the validation dataset, 'Acc' represents the global accuracy, and 'F1' represents the weighted F1 score of the multi-classification prediction task.

unseen datasets, which was beneficial for improving generalization ability. Overall, through staged learning for multimodal barcodes and the application of domain adaptive method, the species-level classification accuracy was totally enhanced by 3.14% on the testing dataset and by 3.79% on the unseen dataset. Similarly, the weighted F1 score at the species level was enhanced by 0.0318 on the testing dataset and by 0.0337 on the unseen dataset (Figure 3). For levels from order to genus, the accuracy and weighted F1 scores on the testing dataset were also enhanced by 2.35%–2.98% and 0.0191–0.0367, respectively; those on the unseen dataset were enhanced by 1.81%–3.08% and 0.0085–0.0181, respectively. At the subclass level, the enhancements were not obvious (Figure 3).

The fusion of RNA secondary structure embedding features and DNA barcode features in COI barcodes further improved the classification accuracy at every taxonomy level, particularly for the unseen dataset (Figure S3; Figure 4). In Approach 2 of COI feature engineering (Figure 1), the RNA secondary structure feature of COI was replaced by the fusion of both the RNA secondary structure embedding and the DNA barcoding embedding. After three-stage learning and domain adaptive application, the trained model showed slight performance improvements on the testing dataset (Figure S3), while achieving a higher accuracy of 55.02% and a weighted F1 score of 0.5596 on the unseen dataset at the species level, representing increases of 8.55% and 0.0794 compared to the previous model using only RNA secondary structure embedding of the six barcode types (Figure 4). In addition, incorporating DNA barcode embedding of COI with the RNA secondary structure embedding of the six barcoding types also improved performance at the genus, family and order levels (Figure 4). However, under Approach 2 of COI engineering, model performance in Stage 1 and Stage 2 appeared to differ little from that in Stage 0, which was trained only on COI features of RNA fusing DNA (Figure 4).

Our model SnailBaLLsp outperformed the SOTA barcode-based foundational models for species classification at every taxonomy level. Both BarcodeMamba and BarcodeMAE utilized only COI data of a multi-barcode Gastropoda dataset constructed in this study, whereas SnailBaLLsp utilized multimodal features of multiple barcoding types. Notably, the pre-trained BarcodeMamba model exhibited competent prediction only at the order level but almost failed at all other taxonomy levels (Figure S3). The pre-trained BarcodeMAE model achieved good classification accuracy across all taxonomy levels when fine-tuned on the COI barcode data of our Gastropoda dataset. On the testing dataset, BarcodeMAE performed comparably to

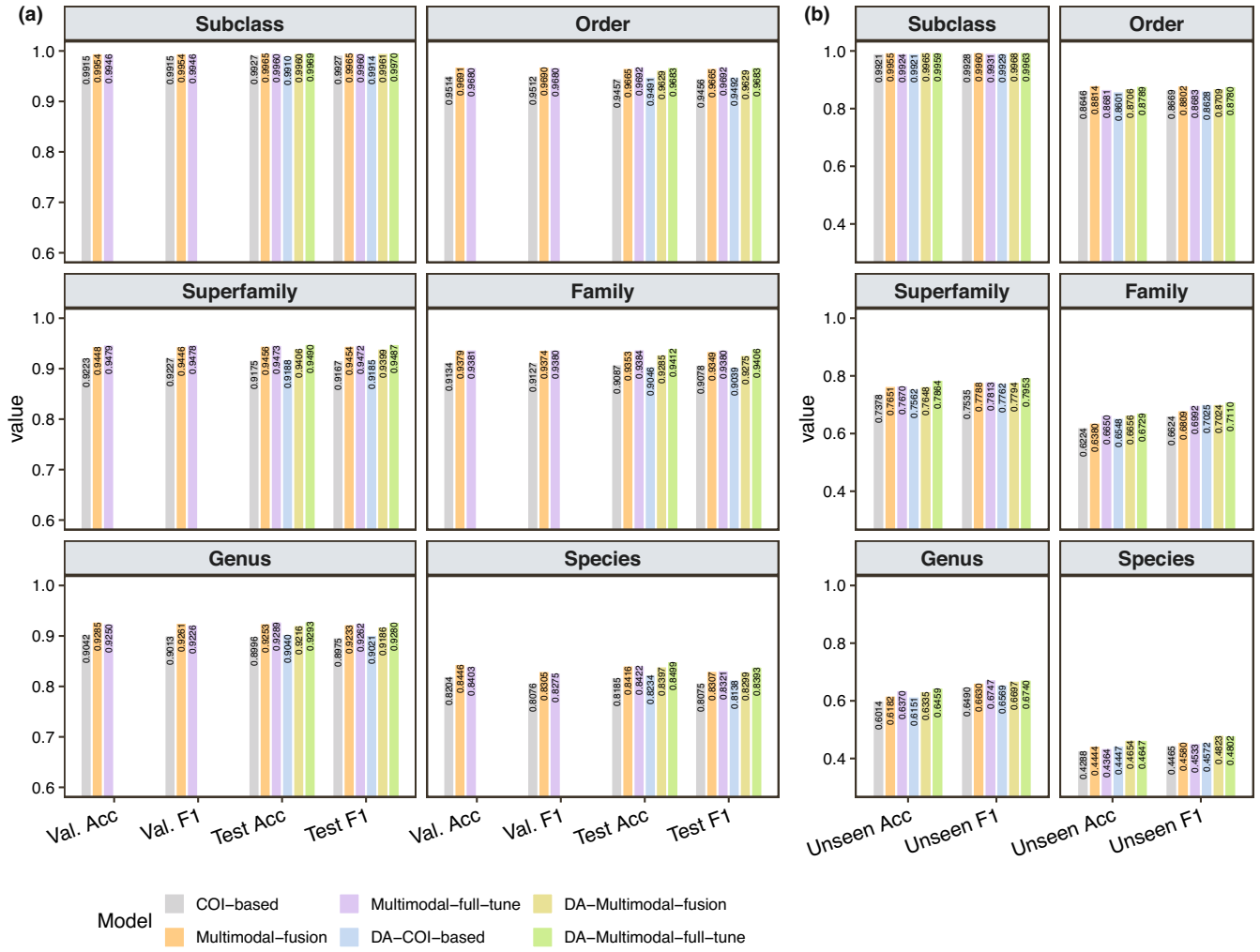

FIGURE 3 Comparison of the performance of the three-staged learning and the domain adaptive method on the validation dataset, testing dataset and unseen dataset by using COI feature engineering of Approach 1. (a) The global accuracy and the weighted F1 score on the validation dataset and the testing dataset (from GenBank). (b) The global accuracy and weighted F1 score of the unseen dataset. The model names containing 'DA' represent models applied domain adaptive method. The *x* axis label 'Val.' represents validation dataset, 'Indep.' represents independent test dataset, 'Acc' represents the accuracy, and 'F1' represents the weighted F1 score of the multi-classification prediction task.

SnailBaLLsp (RNA) which trained only on RNA secondary structure features of the six barcode types (Figure S3). On the unseen dataset, BarcodeMAE even outperformed SnailBaLLsp (RNA) (Figure 4). However, after fusing BarcodeMAE-derived COI barcoding embedding with RNA secondary structure embedding of all barcodes (Approach 2, Figure 1), the retrained SnailBaLLsp (RNA + DNA) surpassed BarcodeMAE on the unseen dataset (Figure 4). This indicates that, on the one hand, BarcodeMAE trained on 5 million COI barcode samples indeed provides accurate features for COI barcoding; on the other hand, SnailBaLLsp's staged hierarchical attention framework can use the comprehensive features of multi-type barcodes for accurate species classification and enhancing generalization ability.

SnailBaLLsp provides a pipeline for hierarchical classification of label-unknown multi-barcode sequences. Users can input newly sequenced gastropod barcode data for hierarchical prediction. Our evaluation on the Gastropoda case dataset demonstrated prediction accuracy comparable to that on the independent unseen dataset (Figure 5a). Analysis of misclassified cases from both the independent unseen and Gastropoda case datasets using the four models of SnailBaLLsp (RNA + DNA) revealed that 3 orders, 22 families and 104 genera were ineffectively predicted by at least three models (Table S5). Analysis of sequence and species counts at the order, family and genus levels for both success and failure prediction groups revealed that failed orders were associated with lower numbers of constituent families and species, indicating reduced sequence diversity. Failed families were correlated with lower average sequences per genus and per species, while failed genera were linked to insufficient average sequences per species and low total sequence count (Figure S4). These findings suggest that misclassified orders, families and genera may suffer from inadequate training data coverage, lower diversity or small sample sizes, leading to model performance degradation.

Additionally, we provide scripts for full fine-tuning on new taxonomic groups. Users can conveniently fine-tune SnailBaLLsp using multi-barcode training data from their taxon of interest modifying

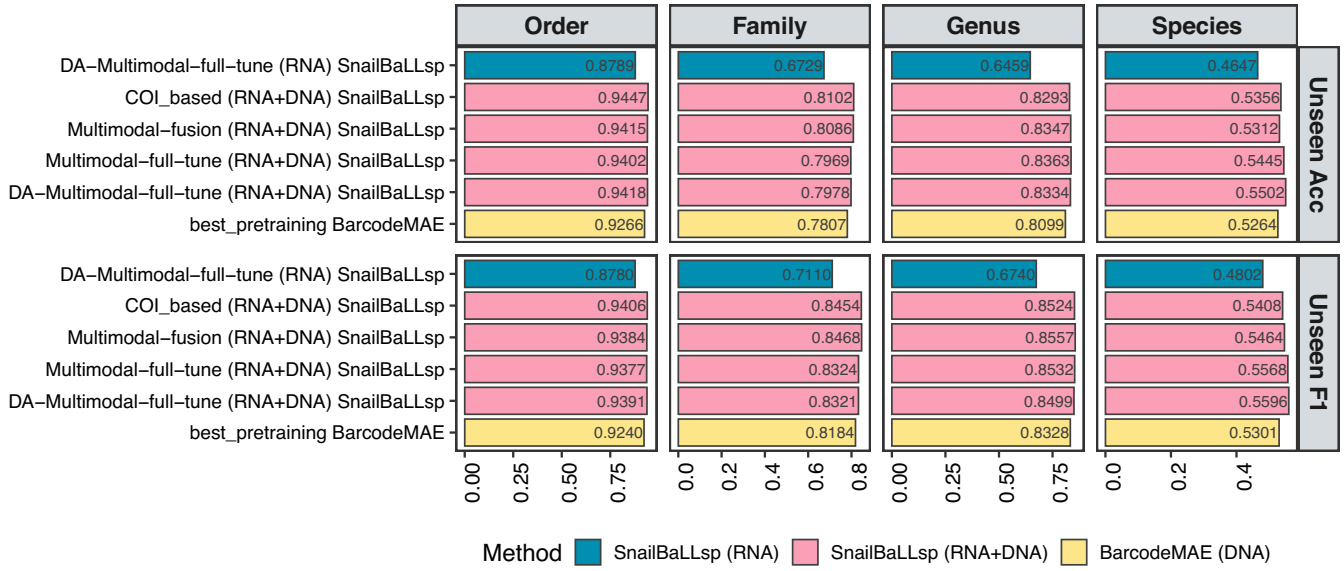

FIGURE 4 Comparison analysis of accuracy and F1 score predicted on the unseen dataset across different taxonomy levels among various models. SnialBaLLsp (RNA) is trained using RNA secondary structure embedding features from multiple barcoding types, following Approach 1 as illustrated in Figure 1b. Its performance in domain adaptation training is represented by blue bars. SnialBaLLsp (RNA + DNA) is trained using fusion features incorporated RNA secondary structure embedding of multiple barcoding types with DNA barcoding embedding of COI, following Approach 2 as illustrated in Figure 1b. Its performance in the three training stages and domain adaptation training is represented by red bars. The performance fine-tuned using pre-trained best model of BarcodeMAE is represented by yellow bars.

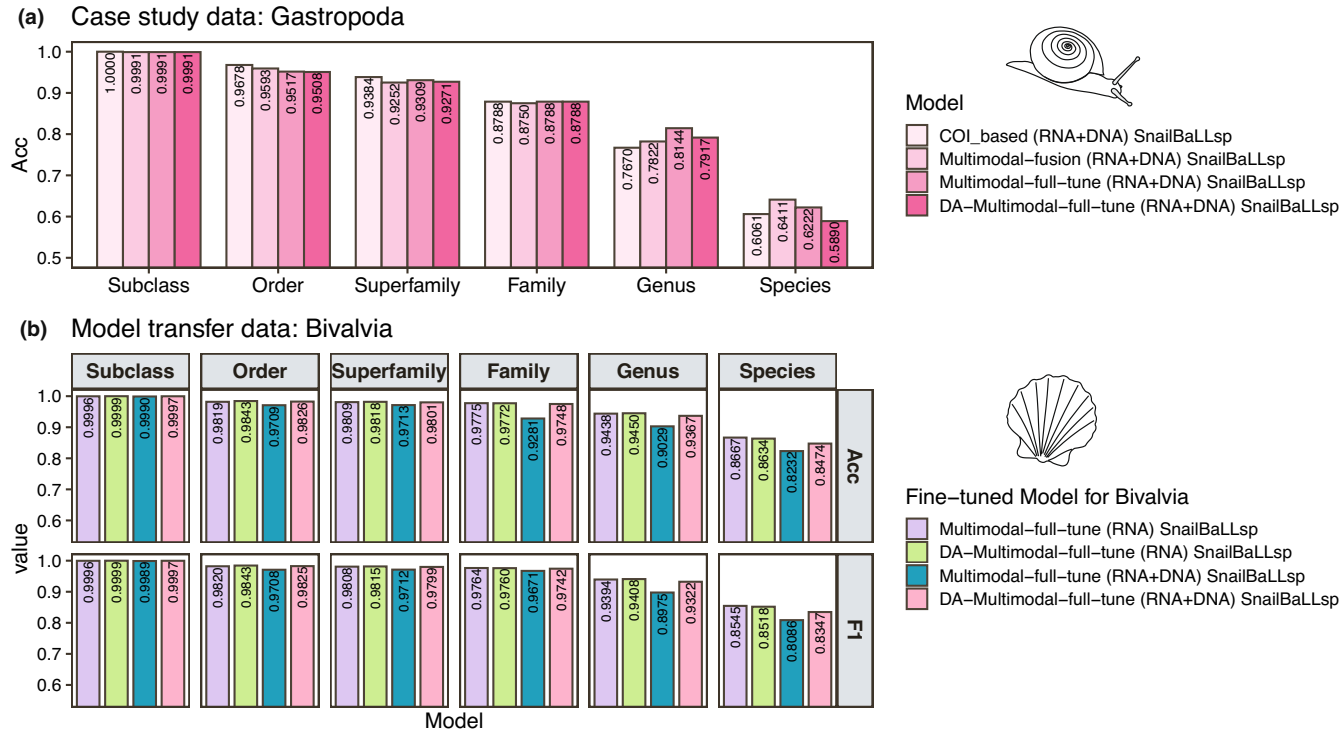

FIGURE 5 SnailBaLLsp enables predictions on label-unknown data and can be successfully transferred to hierarchical classification tasks for non-gastropod classes. (a) Prediction accuracy at each taxonomic level on the Gastropoda case dataset for the four SnailBaLLsp models integrating RNA secondary structure embedding and DNA barcode embedding. (b) Testing set accuracy and weighted F1-score of the four SnailBaLLsp models (after Stage 2 and domain adaptation) following full fine-tuning on multi-barcode data of Bivalvia.

the category numbers at each hierarchical taxonomy level and key parameters (e.g. learning rate, epochs). In this study, fine-tuning SnailBaLLsp on Bivalve data yielded excellent performance across multiple models (Figure 5b), demonstrating its potential as a universal framework for cross-taxon hierarchical classification.

Furthermore, we examined the sequence length and genetic distance of COI sequences using predictions from the DA-Multimodal-full-tune model. Most correctly predicted sequences were longer than misclassified ones, suggesting that sequence length influences prediction accuracy (Figure S5). Crucially, in most cases (271/470 = 57.7%), the genetic distance between failure and success prediction sequences (Failure–Success distance) exceeded the intraspecific distance among success prediction sequences (Distance ratio > 1; Figure 6a). Although some cases showed smaller Failure–Success distances (Distance ratio < 1), the difference between distances in this group ($p = 0.012$; Figure S6) was substantially less significant than in the Distance ratio > 1 group ($p < 2.22e\text{-}16$; Figure S6). This pattern was more pronounced at the species level: the maximum excess of Failure–Success over Success–Success distance reached 168-fold (16,801%), with most species showing >100% difference (Figure 6b). In contrast, the maximum deficit of Failure–Success distance was only 83.4% (Figure 6c). Notably, 80.9% (241/298) of Failure–Success distances in the Distance ratio > 1 group still significantly exceeded the 2.38% intraspecific divergence threshold, which was derived from the median Success–Success distance (Figure 6d, Figure S7). These findings indicate that

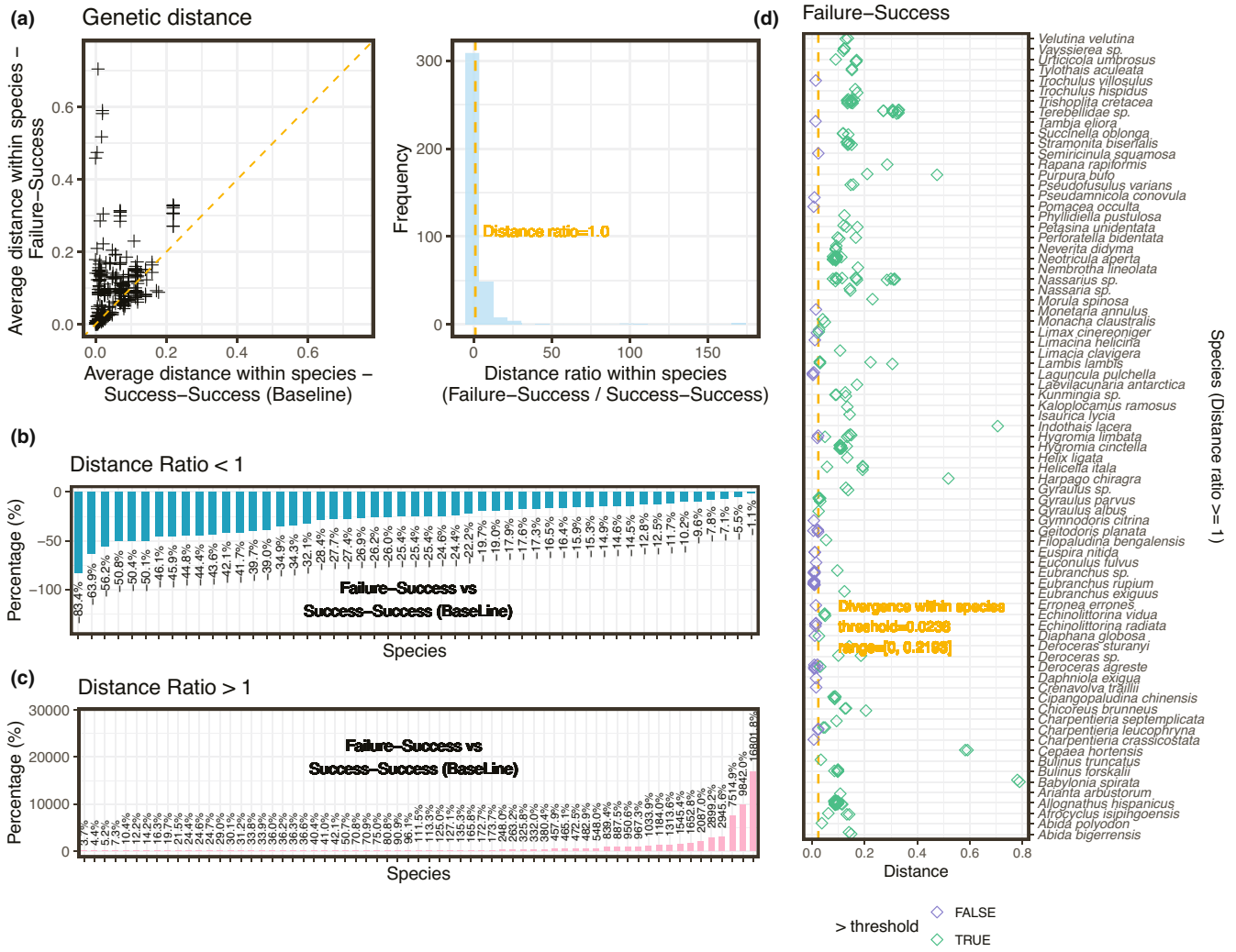

FIGURE 6 Comparison of genetic distances between failure–success prediction sequences (Failure–Success) and mean intraspecific distances among success prediction sequences (Success–Success) within the same species. Sequences are predicted using domain-adapted SnailBaLLsp (RNA + DNA) model. (a) Distribution of Failure–Success and Success–Success distances per species relative to the $y = x$ diagonal. Most points lie above the diagonal, indicating that the Failure–Success distance is generally greater. The histogram (right) shows the distribution of the distance ratio (Failure–Success/Success–Success) with a demarcation at ratio = 1. (b) For cases with a distance ratio < 1, the percentage difference between Failure–Success and Success–Success distances does not exceed 84%. (c) For cases with a distance ratio > 1, the percentage difference is extreme (up to 168-fold), with most exceeding 100%. (d) For species with a distance ratio > 1, most Failure–Success distances exceed the intraspecific divergence threshold. The range of intraspecific divergence is also shown. The orange line represents the threshold derived from the median Success–Success distance.

COI sequence divergence within species potentially contributes to species-level misclassification.

4 | DISCUSSION

We constructed a deep learning model, SnailBaLLsp, to conduct species classification prediction for major groups of Gastropoda. It is capable of handling multi-type barcoding data with incomplete sample coverage while simultaneously providing predictions across all taxonomy levels from subclass to species. The model addressed two critical challenges. First, due to the large number of COI samples in the current barcoding database and the scarcity of other types of DNA barcoding, and the inability to achieve sample matching, previous models relied only on the single COI barcode (Arias et al., 2025; Gao & Taylor, 2024; Safari et al., 2025), thereby ignoring the incomplete lineage sorting effects in species classification and systematic research (Ahmed et al., 2022; Antil et al., 2023; Liu et al., 2017; Moritz & Cicero, 2004; Steenwyk et al., 2023; Yang & Rannala, 2010). Second, although trained on data with clear labels at each taxonomy level, previous models did not effectively integrate prior taxonomic knowledge. Their classification predictions at each level were fragmented, resulting in high prediction accuracy at the final species level but low accuracy at higher taxonomy levels (Arias et al., 2025; Gao & Taylor, 2024). Although some studies have proposed taxonomic text information to map DNA barcoding and image modalities (Gharaee et al., 2024; Gong et al., 2024), their accuracy for unseen genus-level classification was only 47.03% (Gong et al., 2024), which remain relatively low. This might be due to the lack of hierarchical constraints from higher to lower taxonomy levels. To address the two challenges, SnailBaLLsp employs a hierarchical attention network and a staged curriculum learning method, first pre-training on the COI sequence dataset with full sample coverage and then incrementally training with other barcode sequences that have lower sample coverage. This architecture mitigates data distribution inconsistency caused by large differences in sample coverage across multimodal DNA barcoding data. Additionally, SnailBaLLsp integrates a hierarchical attention mechanism to fuse prior knowledge across taxonomy levels, enabling progressive constraints from higher to lower taxonomic features (Hong et al., 2021; Liu et al., 2025; Zhang et al., 2025; Zhou et al., 2025) and thereby emphasizing the multi-granularity of species classification priors.

SnailBaLLsp was trained on a multi-type barcoding dataset of the main taxa of Gastropoda with complete labels from subclass to species. This dataset has the potential to become one of the benchmark datasets for supervised learning in species classification based on DNA barcoding. Our fully labelled dataset of Gastropoda contains 76,577 distinct original COI sequences and has been expanded to 101,607 sequences through augmentation. It is comparable to the current largest insect COI barcoding dataset, Bioscan-5 M, which contains 118,051 COI sequences used for supervised training (Gharaee et al., 2024). In addition, our dataset involves a total of 11,295 Gastropoda species, which is also comparable to the 11,846 insect species in Bioscan-5 M's supervised training set (Gharaee et al., 2024). A novel feature of our Gastropoda dataset is that it provides not only the largest number and most comprehensive sample coverage of COI sequences of Gastropoda but also includes other common barcoding data of molecular markers such as 16S rRNA (with a sample coverage of 16.4%), H3, 18S rRNA, ITS1 and ITS2, which are not available in most existing barcode datasets. However, due to data availability constraints, the current Gastropoda dataset is limited to two major subclasses. Future work could incorporate data from the remaining Gastropoda subclasses. It should be noted that we employed supervised learning methods and have not yet used unlabelled Gastropoda sequence data. Future studies could expand to include unlabelled Gastropoda data for unsupervised or semi-supervised learning. One potential prospect is that the barcoding data we constructed for this important group, Gastropoda, could serve as an additional training data source for current barcoding-based foundation models (Arias et al., 2025; Gao & Taylor, 2024; Safari et al., 2025), potentially further enriching COI feature representations. Moreover, our dataset does not include species image data. The combination of images and DNA has been shown to improve species identification accuracy (Hofmann et al., 2024; Yang et al., 2022), which also represents a growing trend in integrative taxonomy studies (Hunt et al., 2025; Karbstein et al., 2024; Meier et al., 2025; Miralles et al., 2024). Further research on integrating DNA barcoding with image data for gastropod species identification is therefore warranted.

Our results confirm that the RNA secondary structure features of barcode sequences can serve as effective species classification features (Acharya et al., 2022; Mohanty et al., 2023; Pere et al., 2024), expanding the application of RNA secondary structure of COI barcode in species identification. When models were constructed using only RNA secondary structure features for each barcoding (Approach 1, Figure 1), performance significantly surpassed that of BarcodeMamba and achieved comparable performance to BarcodeMAE on the testing dataset. However, our performance on the unseen dataset still did not reach that of BarcodeMAE (Figure 4). A likely reason is that our gastropod dataset does not contain a large volume of unlabelled data, and the overall sample size is far smaller than the 5-million-sample dataset used by BarcodeMAE (Gharaee et al., 2024; Safari et al., 2025). This resulted in the model built solely on RNA secondary structure features exhibiting slightly lower generalization capability on unseen species compared to the SOTA foundation model. However, when we further fused the COI barcoding embedding derived from BarcodeMAE, the model's predictive power for the unseen dataset exceeded that of BarcodeMAE. This outcome demonstrates two key points: (1) BarcodeMAE currently represents the optimal and indispensable approach for representing COI barcoding features; and (2) fusing the RNA secondary structure features with specialized DNA barcoding features can enhance model performance. Regrettably, apart from COI, there are currently no specialized foundation models pre-trained on other DNA barcoding sequences (such as 16S and ITS). This is one reason why our study utilized only RNA secondary structure features for the other five

barcoding types. While numerous cross-species DNA foundation models have been currently developed to capture deep features of DNA sequence, such as DNABERT-2 (Zhou et al., 2024), GPN-MSA (Benegas et al., 2025), Caduceus (Schiff et al., 2024), HyenaDNA (Nguyen et al., 2023), Evo2 (Brixi et al., 2025). However, the main focus of these foundation models has been mostly on biomedical tasks such as variant effect prediction (e.g. Benegas et al., 2025) and gene functional element prediction (e.g. Nguyen et al., 2023). There is relatively little research on their application in the field of species classification, and their species classification performance was not as good as the foundation models trained specifically for barcoding sequences (Arias et al., 2025; Gao & Taylor, 2024; Safari et al., 2025). This presents a potential research direction for ecologists and bioinformaticians: developing specialized barcoding classification models tailored for species identification and systematic analysis. Such models would be easier to train, require less computational resources, be more lightweight than genome-based foundation models and could potentially deliver superior species classification accuracy, which would significantly advance species classification and biodiversity research.

SnailBaLLsp provides an end-to-end intelligent barcode-based species classification method for major Gastropoda taxa for the first time. It can predict classification at taxonomy levels from subclass to species—for input consisting of either a single COI barcoding or a combination of multiple barcoding types. In practical research scenarios, if sequencing all the six barcoding types is not feasible, ensuring at least the two typical barcoding types of COI and 16S could still improve species classification accuracy (Figure 2). Another advantage of SnailBaLLsp is that it can achieve prediction accuracy of over 79.78% at the family level and other higher levels even when the classification of a barcoding sequence is still difficult to determine in lower taxonomy levels (such as genus and species). This makes it a powerful tool for determining high-taxonomy-level classification of unknown or taxonomically challenging specimens. This feature holds promising potential for application in studies that use DNA metabarcoding techniques to detect multiple species within environmental or biological samples (Kartzinel et al., 2025; Meier et al., 2024). For instance, it can accurately assign specimens to the family level within mixed-species samples, preventing misclassification at higher taxonomic levels.

It is reported that Gastropods show wide variation in COI and other mitochondrial sequences between species and populations (David et al., 2022; Davison et al., 2024; Thomaz et al., 1997). For species with both success and failure predictions by SnailBaLLsp, most misclassified sequences exhibited higher intraspecific genetic divergence. This indicates that the decreased species-level performance does not reflect methodological limitations but is largely influenced by the extreme intraspecific COI divergence (Figure 6), which was commonly observed in gastropods (David et al., 2022; Davison et al., 2024; Thomaz et al., 1997). These findings suggest that future improvements should focus on enhancing model training to better recognize high divergent sequences and accurately extract their species-specific features. Furthermore, sequence diversity and quantity across taxonomic levels, as well as quality factors such as sequence length, significantly impact prediction accuracy (Figures S4 and S5). For instance, species of Enidae, which have disputed taxonomic status (Ye et al., 2024), exemplify this issue due to the limited number of available barcoding sequences for their respective genera and species, making them difficult to predict successfully (Table S5). We anticipate that model performance can be further improved as both the volume and quality of gastropod barcode data continue to grow. When applying the current SnailBaLLsp model, users are advised to ensure sequence length is as complete as possible and to exercise caution when interpreting lower-level predictions for taxon groups with lower documented accuracy. It is recommended to initially rely on higher-level taxonomic predictions for preliminary screening, followed by additional phylogenetic and taxonomic validation for finer-level classifications.

SnailBaLLsp still has limitations. It currently does not include other, less species-rich gastropod subclasses or unlabelled gastropod barcoding data. Consequently, its performance in predicting lower taxonomy levels such as species level for unknown species needs further improvement. In addition, due to the lack of specialized DNA foundation models for other barcode types, or the unclear efficacy of existing DNA foundation models in capturing the deep features of DNA barcoding, features in DNA level for barcodes other than COI have not yet been incorporated. Future work could integrate deep DNA features for other barcode types into our framework. Finally, at present, only supervised learning methods have been used in this study. In the future, unsupervised or semi-supervised learning could be combined to further improve the classification accuracy of Gastropoda species.

Overall, our deep learning model SnailBaLLsp constructed by multi-type DNA barcoding data will serve as an important exemplary model for species identification. It does not seek to replace taxonomists, but rather provides an important interface for integrating molecular (DNA package barcode) and morphological (image) data using artificial intelligence in future integrative taxonomy studies, thereby promoting future biodiversity research, ecological and environmental monitoring and invasive alien species management.

AUTHOR CONTRIBUTIONS

Conceptualization: Bin Ye. Methodology: Bin Ye, Junfeng Xia. Resources: Junfeng Xia, Min Wu, Xia Wan, Satoshi Chiba. Writing—original draft: Bin Ye. Writing—review and editing: Junfeng Xia, Min Wu, Xia Wan, Satoshi Chiba. Funding acquisition: Bin Ye, Min Wu, Satoshi Chiba.

ACKNOWLEDGEMENTS

This work was supported by Project funded by China Postdoctoral Science Foundation (2025M772813), National Natural Science Foundation of China (31872196) and the Japan Society for the Promotion of Science (KAKENHI-17H04611).

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70264.

DATA AVAILABILITY STATEMENT

Training, test and prediction code used in this study are available via https://github.com/BRUCEBINYE/SnailBaLLsp and https://doi.org/10.5281/zenodo.17780323 (Ye et al., 2025). Training and test datasets used in our model are available via https://doi.org/10.5061/dryad.ttdz08m9c (Ye et al., 2026a). The Supporting Information Figures and Tables are available via https://doi.org/10.5281/zenodo.18357507 (Ye et al., 2026b).

ORCID

Bin Ye https://orcid.org/0000-0002-5359-4389

REFERENCES

Acharya, G. C., Mohanty, S., Dasgupta, M., Sahu, S., Singh, S., Koundinya, A. V. V., Kumari, M., Naresh, P., & Sahoo, M. R. (2022). Molecular phylogeny, DNA barcoding, and ITS2 secondary structure predictions in the medicinally important eryngium genotypes of East Coast region of India. *Genes*, *13*, 1678. https://doi.org/10.3390/genes13091678

Ahmed, S., Ibrahim, M., Nantasenamat, C., Nisar, M. F., Malik, A. A., Waheed, R., Ahmed, M. Z., Ojha, S. C., & Alam, M. K. (2022). Pragmatic applications and universality of DNA barcoding for substantial organisms at species level: A review to explore a way forward. *BioMed Research International*, *2022*, 1846485. https://doi.org/10.1155/2022/1846485

Antil, S., Abraham, J. S., Sripoorna, S., Maurya, S., Dagar, J., Makhija, S., Bhagat, P., Gupta, R., Sood, U., Lal, R., & Toteja, R. (2023). DNA barcoding, an effective tool for species identification: A review. *Molecular Biology Reports*, *50*, 761–775. https://doi.org/10.1007/s11033-022-08015-7

Arias, P. M., Sadjadi, N., Safari, M., Gong, Z., Wang, A. T., Haurum, J. B., Zarubiieva, I., Steinke, D., Kari, L., Chang, A. X., Lowe, S. C., & Taylor, G. W. (2025). BarcodeBERT: Transformers for biodiversity analysis. *arXiv*, 2311.02401. https://doi.org/10.48550/arXiv.2311.02401

Benegas, G., Albors, C., Aw, A. J., Ye, C., & Song, Y. S. (2025). A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, *43*, 1960–1965. https://doi.org/10.1038/s41587-024-02511-w

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41–48). Association for Computing Machinery.

Bjerge, K., Geissmann, Q., Alison, J., Mann, H. M. R., Høye, T. T., Dyrmann, M., & Karstoft, H. (2023). Hierarchical classification of insects with multitask learning and anomaly detection. *Ecological Informatics*, *77*, 102278. https://doi.org/10.1016/j.ecoinf.2023.102278

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., … Hie, B. L. (2025). Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025.2002.2018.638918. https://doi.org/10.1101/2025.02.18.638918

Chen, Y., Zhu, Z., Zhu, S., Qiu, L., Zou, B., Jia, F., Zhu, Y., Zhang, C., Fang, Z., Qin, F., Fan, J., Wang, C., Yu, G., & Gao, Y. (2025). SCKansformer: Fine-grained classification of bone marrow cells via kansformer backbone and hierarchical attention mechanisms. *IEEE Journal of Biomedical and Health Informatics*, *29*, 558–571. https://doi.org/10.1109/JBHI.2024.3471928

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Coleman, A. W. (2015). Nuclear rRNA transcript processing versus internal transcribed spacer secondary structure. *Trends in Genetics*, *31*, 157–163. https://doi.org/10.1016/j.tig.2015.01.002

David, P., Degletagne, C., Saclier, N., Jennan, A., Jarne, P., Plénet, S., Konecny, L., François, C., Guéguen, L., Garcia, N., Lefébure, T., & Luquet, E. (2022). Extreme mitochondrial DNA divergence underlies genetic conflict over sex determination. *Current Biology*, *32*, 2325–2333.e2326. https://doi.org/10.1016/j.cub.2022.04.014

Davison, A., Chowdhury, M., Johansen, M., Uliano-Silva, M., Blaxter, M., & Wellcome Sanger Institute Tree of Life, p. (2024). High heteroplasmy is associated with low mitochondrial copy number and selection against non-synonymous mutations in the snail *Cepaea nemoralis*. *BMC Genomics*, *25*, 596. https://doi.org/10.1186/s12864-024-10505-w

Elhamod, M., Diamond, K. M., Maga, A. M., Bakis, Y., Jr, B., Henry, L., Mabee, P., Dahdul, W., Leipzig, J., Greenberg, J., Avants, B., & Karpatne, A. (2022). Hierarchy-guided neural network for species classification. *Methods in Ecology and Evolution*, *13*, 642–652. https://doi.org/10.1111/2041-210X.13768

Fan, J., Zhou, N., Peng, J., & Gao, L. (2015). Hierarchical learning of tree classifiers for large-scale plant species identification. *IEEE Transactions on Image Processing*, *24*, 4172–4184. https://doi.org/10.1109/TIP.2015.2457337

Gao, T., & Taylor, G. W. (2024). BarcodeMamba: State space models for biodiversity analysis. *arXiv*, 2412.11084v11081. https://doi.org/10.48550/arXiv.2412.11084

Gharaee, Z., Lowe, S. C., Gong, Z., Millan Arias, P., Pellegrino, N., Wang, A. T., Haurum, J. B., Eyriay, I., Kari, L., & Steinke, D. (2024). Bioscan-5m: A multimodal dataset for insect biodiversity. *Advances in Neural Information Processing Systems*, *37*, 36285–36313. https://doi.org/10.52202/079017-1144

Gong, Z., Wang, A. T., Huo, X., Haurum, J. B., Lowe, S. C., Taylor, G. W., & Chang, A. X. (2024). CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. *arXiv*, 2405.17537. https://doi.org/10.48550/arXiv.2405.17537

Han, A. Y., Yi, K. M., Kim, K. T., & Choi, J. Y. (2025). Hierarchical feature attention learning network for detecting object and discriminative parts in fine-grained visual classification. *IEEE Access*, *13*, 19533–19544. https://doi.org/10.1109/ACCESS.2025.3534444

Hofmann, M., Kiel, S., Kösters, L. M., Wäldchen, J., & Mäder, P. (2024). Inferring taxonomic affinities and genetic distances using morphological features extracted from specimen images: A case study with a bivalve data set. *Systematic Biology*, *73*, 920–940. https://doi.org/10.1093/sysbio/syae042

Hong, J., Gao, R., & Yang, Y. (2021). CrepHAN: Cross-species prediction of enhancers by using hierarchical attention networks. *Bioinformatics*, *37*, 3436–3443. https://doi.org/10.1093/bioinformatics/btab349

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, *33*, 1635–1638. https://doi.org/10.1093/molbev/msw046

Hunt, R., Reyes-Hernández, J. L., Shaw, J. J., Solodovnikov, A., & Pedersen, K. S. (2025). Integrating deep learning derived morphological traits and molecular data for total-evidence phylogenetics: Lessons from digitized collections. *Systematic Biology*, *74*, 453–468. https://doi.org/10.1093/sysbio/syae072

Ito, K., Miura, K., Aoki, T., Otake, Y., Makino, W., & Urabe, J. (2023). Zooplankton classification using hierarchical attention branch

network. In H. Lu, M. Blumenstein, S.-B. Cho, C.-L. Liu, Y. Yagi, & T. Kamiya (Eds.), *Pattern recognition* (pp. 409–419). Cham.

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., Wagner, N. D., Emerson, B. C., Albach, D. C., Scheu, S., Bradler, S., de Vries, J., Irisarri, I., Li, H., Soltis, P., Mäder, P., & Wäldchen, J. (2024). Species delimitation 4.0: Integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*, *39*, 771–784. https://doi.org/10.1016/j.tree.2023.11.002

Kartzinel, T. R., Hoff, H. K., Divoll, T. J., Littleford-Colquhoun, B. L., Anderson, H., Burak, M. K., Kuzmina, M. L., Musili, P. M., Rogers, H., Troncoso, A. J., & Kartzinel, R. Y. (2025). Global availability of plant DNA barcodes as genomic resources to support basic and policy-relevant biodiversity research. *Molecular Ecology*, *34*, e17712. https://doi.org/10.1111/mec.17712

Li, R., Ratnasingham, S., Zarubiieva, I., Somervuo, P., & Taylor, G. W. (2024). PROTAX-GPU: A scalable probabilistic taxonomic classification system for DNA barcodes. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *379*, 20230124. https://doi.org/10.1098/rstb.2023.0124

Liu, H., Jiang, P., Tao, Z., Wan, M., & Sun, Q. (2025). Tighnari: Multi-modal plant species prediction based on hierarchical cross-attention using graph-based and vision backbone-extracted features. *arXiv*, 2501.02649. https://doi.org/10.48550/arXiv.2501.02649

Liu, J., Jiang, J., Song, S., Tornabene, L., Chabarria, R., Naylor, G. J. P., & Li, C. (2017). Multilocus DNA barcoding – Species identification with multilocus data. *Scientific Reports*, *7*, 16601. https://doi.org/10.1038/s41598-017-16920-2

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research*, *35*, D26–D31. https://doi.org/10.1093/nar/gkl993

Meier, R., Hartop, E., Pylatiuk, C., & Srivathsan, A. (2024). Towards holistic insect monitoring: Species discovery, description, identification and traits for all insects. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *379*, 20230120. https://doi.org/10.1098/rstb.2023.0120

Meier, R., Lawniczak, M. K. N., & Srivathsan, A. (2025). Illuminating entomological dark matter with DNA barcodes in an era of insect decline, deep learning, and genomics. *Annual Review of Entomology*, *70*, 185–204. https://doi.org/10.1146/annurev-ento-040124-014001

Miralles, A., Puillandre, N., & Vences, M. (2024). DNA barcoding in species delimitation: From genetic distances to integrative taxonomy. In R. DeSalle (Ed.), *DNA barcoding: Methods and protocols* (pp. 77–104). Springer.

Mohanty, S., Mishra, B. K., Dasgupta, M., Acharya, G. C., Singh, S., Naresh, P., Bhue, S., Dixit, A., Sarkar, A., & Sahoo, M. R. (2023). Deciphering phenotyping, DNA barcoding, and RNA secondary structure predictions in eggplant wild relatives provide insights for their future breeding strategies. *Scientific Reports*, *13*, 13829. https://doi.org/10.1038/s41598-023-40797-z

Moritz, C., & Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLoS Biology*, *2*, e354. https://doi.org/10.1371/journal.pbio.0020354

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., & Ré, C. (2023). HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv*. https://doi.org/10.48550/arXiv.2306.15794

Noumida, A., & Rajan, R. (2022). Multi-label bird species classification using hierarchical attention framework. In *2022 IEEE 19th India Council International Conference (INDICON)* (pp. 1–6). IEEE.

Pere, K., Mburu, K., Muge, E. K., Wagacha, J. M., & Nyaboga, E. N. (2024). Molecular identification, genetic diversity, and secondary structure predictions of Physalis species using ITS2 DNA barcoding. *BMC Plant Biology*, *24*, 1178. https://doi.org/10.1186/s12870-024-05889-6

Pollock, L. J., Kitzes, J., Beery, S., Gaynor, K. M., Jarzyna, M. A., Mac Aodha, O., Meyer, B., Rolnick, D., Taylor, G. W., Tuia, D., & Berger-Wolf, T. (2025). Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity*, *1*, 166–182. https://doi.org/10.1038/s44358-025-00022-3

Prosser, S. W. J., Floyd, R. M., Thompson, K. A., Monckton, S. K., & Hebert, P. D. N. (2025). BOLDistilled: Automated construction of comprehensive but compact DNA barcode reference libraries. *Molecular Ecology Resources*, *25*, e70043. https://doi.org/10.1111/1755-0998.70043

Qin, J., Martínez, L., Pedrycz, W., Ma, X., & Liang, Y. (2023). An overview of granular computing in decision-making: Extensions, applications, and challenges. *Information Fusion*, *98*, 101833. https://doi.org/10.1016/j.inffus.2023.101833

Ratnasingham, S., & Hebert, P. D. N. (2007). Bold: The barcode of life data system (http://www.barcodinglife.org/). *Molecular Ecology Notes*, *7*, 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS One*, *8*, e66213. https://doi.org/10.1371/journal.pone.0066213

Romeijn, L., Bernatavicius, A., & Vu, D. (2024). MycoAI: Fast and accurate taxonomic classification for fungal ITS sequences. *Molecular Ecology Resources*, *24*, e14006. https://doi.org/10.1111/1755-0998.14006

Safari, M., Arias, P. M., Lowe, S. C., Kari, L., Chang, A. X., & Taylor, G. W. (2025). Enhancing DNA foundation models to address masking inefficiencies. *arXiv*, 2502.18405. https://doi.org/10.48550/arXiv.2502.18405

Sanchez, T., Stalder, S., Lamperti, L., Brosse, S., Frossard, A., Leugger, F., Rozanski, R., Zong, S., Manel, S., Medici, L., Kuhn, F., Han, X., Mestrot, A., Albouy, C., Volpi, M., & Pellissier, L. (2025). ORDNA: Deep-learning-based ordination for raw environmental DNA samples. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.70033

Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., & Kuleshov, V. (2024). Caduceus: Bi-directional equivariant long-range DNA sequence modeling. *arXiv*, 2403.03234. https://doi.org/10.48550/arXiv.2403.03234

Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X., & Rokas, A. (2023). Incongruence in the phylogenomics era. *Nature Reviews Genetics*, *24*, 834–850. https://doi.org/10.1038/s41576-023-00620-x

Thomaz, D., Guiller, A., & Clarke, B. C. (1997). Extreme divergence of mitochondrial DNA within species of pulmonate land snails. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *263*, 363–368. https://doi.org/10.1098/rspb.1996.0056

Wang, Q., Song, Y., Du, Y., Yang, Z., Cui, P., & Luo, B. (2024). Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species: A phylogenetic perspective. *Ecological Informatics*, *80*, 102538. https://doi.org/10.1016/j.ecoinf.2024.102538

Wang, X., Chen, Y., & Zhu, W. (2022). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 4555–4576. https://doi.org/10.1109/TPAMI.2021.3069908

Weinbach, B. C., Akerkar, R., Nilsen, M., & Arghandeh, R. (2025). Hierarchical deep learning framework for automated marine vegetation and fauna analysis using ROV video data. *Ecological Informatics*, *85*, 102966. https://doi.org/10.1016/j.ecoinf.2024.102966

Yang, B., Zhang, Z., Yang, C. Q., Wang, Y., Orr, M. C., Wang, H., & Zhang, A. B. (2022). Identification of species by combining molecular and morphological data using convolutional neural networks. *Systematic Biology*, *71*, 690–705. https://doi.org/10.1093/sysbio/syab076

Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, *107*, 9264–9269. https://doi.org/10.1073/pnas.0913022107

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). *Hierarchical attention networks for document classification* (pp. 1480–1489). Association for Computational Linguistics.

Ye, B., Hirano, T., Xia, J.-F., Morii, Y., Wu, M., & Chiba, S. (2024). Phylogenetic analysis of Chinese Enidae: Inference from genome-wide analyses and morphological studies on three selected genera. *Zoological Research: Diversity and Conservation*, *1*, 247–261. https://doi.org/10.24272/j.issn.2097-3772.2024.007

Ye, B., Xia, J., Wan, X., Wu, M., & Chiba, S. (2025). Data from: BRUCEBINYE/SnailBaLLsp: SnailBaLLsp v1.0.2 (v1.0.2). *Zenodo*. https://doi.org/10.5281/zenodo.17780323

Ye, B., Xia, J., Wan, X., Wu, M., & Chiba, S. (2026a). Data from: Multi-barcoding-based Gastropoda identification using hierarchical attention network with staged curriculum learning. *Dryad*. https://doi.org/10.5061/dryad.ttdz08m9c

Ye, B., Xia, J., Wan, X., Wu, M., & Chiba, S. (2026b). Data from: Supplementary information for multi-barcoding-based Gastropoda identification using hierarchical attention network with staged curriculum learning. *Zenodo*. https://doi.org/10.5281/zenodo.18357507

Yin, W., Zhang, Z., Zhang, S., He, L., Zhang, R., Jiang, R., Liu, G., Wang, J., Zhang, X., Qin, T., & Xie, Z. (2025). ERNIE-RNA: An RNA language model with structure-enhanced representations. *Nature Communications*, *16*, 10076. https://doi.org/10.1038/s41467-025-64972-0

Zhang, H., Kuang, Z., Peng, X., He, G., Peng, J., & Fan, J. (2020). Aggregating diverse deep attention networks for large-scale plant species identification. *Neurocomputing*, *378*, 283–294. https://doi.org/10.1016/j.neucom.2019.10.077

Zhang, T., Chen, L., Zhu, H., & Wong, G. (2025). Mammalian piRNA target prediction using a hierarchical attention model. *BMC Bioinformatics*, *26*, 50. https://doi.org/10.1186/s12859-025-06068-6

Zhou, W., Song, W., Zhang, Z., Zhang, F., Teng, Z., & Tian, Z. (2025). Hierarchical feature aggregation with mixed attention mechanism for single-cell RNA-seq analysis. *Expert Systems with Applications: An International Journal*, *260*, 125340. https://doi.org/10.1016/j.eswa.2024.125340

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., & Liu, H. (2024). DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv*, 2306.15006. https://doi.org/10.48550/arXiv.2306.15006

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Sequence data augmentation enhances diversity across various taxonomy levels and improves model prediction performance.

Figure S2. Training loss curves, Top-1 Accuracy, and F1 Scores on the validation dataset across taxonomy levels for the three-stage curriculum learning.

Figure S3. Comparison analysis of accuracy and F1 score predicted on the testing dataset across different taxonomy levels among various models.

Figure S4. Sequence characteristics of failure predictions at the order, family, and genus levels.

Figure S5. (**a**) Scatter plot of sequence lengths for success versus failure predictions within each taxon across order, family, and genus levels.

Figure S6. Differences between intraspecies Failure–Success and Success–Success distances.

Figure S7. Proportion of Failure–Success distances exceeding the intraspecific divergence threshold, for species with Distance Ratio (Failure-Success/Success-Success) >1.

Table S1. Classification information and GenBank search keywords of main Gastropoda species according to the taxonomy system of MolluscaBase in the Training-Validation-Testing dataset.

Table S2. Classification information and GenBank search keywords of main Gastropoda species according to the taxonomy system of MolluscaBase in the Independent Testing 1 dataset.

Table S3. Dynamically enhancement of sequence number according to original sequence number at the genus level.

Table S4. Classification information and GenBank search keywords of main Bivalvia orders in the model transfer dataset.

Table S5. Taxa that have not been successfully predicted.

How to cite this article: Ye, B., Xia, J., Wan, X., Wu, M., & Chiba, S. (2026). Multi-barcoding-based species identification using a hierarchical attention network with staged curriculum learning. *Methods in Ecology and Evolution*, *17*, 1143–1160. https://doi.org/10.1111/2041-210x.70264