Received: 1 March 2024 | Accepted: 16 July 2024

DOI: 10.1111/2041-210X.14440

Methods in Ecology and Evolution

RESEARCH ARTICLE

Environmental niche models improve species identification in DNA barcoding

Cai-qing Yang[1] | Ying Wang[1] | Xin-hai Li[2,3] | Jing Li[1] | Bing Yang[1] | Michael C. Orr[4,5] | Ai-bing Zhang[1]

[1]College of Life Sciences, Capital Normal University, Beijing, China

[2]Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[3]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

[4]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[5]Entomologie, Staatliches Museum für Naturkunde Stuttgart, Stuttgart, Germany

Correspondence
Ai-bing Zhang
Email: zhangab2008@cnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 32200343 and 32170421; Natural Science Foundation of Beijing Municipality, Grant/Award Number: 5232001; Chinese Academy of Sciences President's International Fellowship Initiative program, Grant/Award Number: 2024PVC0046; Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 14th Five-year Plan, Grant/Award Number: BPHR20220114; National Key Research and Development Program of China, Grant/Award Number: 2023YFC2606600; Academy for Multidisciplinary Studies, Capital Normal University

Handling Editor: Laura Graham

Abstract

1. Recent advances in DNA barcoding have immeasurably advanced global biodiversity research in the last two decades. However, inherent limitations in barcode sequences, such as hybridization, introgression or incomplete lineage sorting can lead to misidentifications when relying solely on barcode sequences.
2. Here, we propose a new Niche-model-Based Species Identification (NBSI) method based on the idea that species distribution information is a potential complement to DNA barcoding species identifications. NBSI performs species membership inference by incorporating niche modelling predictions and traditional DNA barcoding identifications.
3. Systematic tests across diverse scenarios show significant improvements in species identification success rates under the newly proposed NBSI framework, where the largest increase is from 4.7% (95% CI: 3.51%–6.25%) to 94.8% (95% CI: 93.19%–96.06%). Additionally, obvious improvements were observed when using NBSI on potentially ambiguous sequences whose genetic nearest neighbours belongs to another species or more than two species, which occurs commonly with species represented by single or short DNA barcodes.
4. These results support our assertion that environmental factors/variables are valuable complements to DNA sequence data for species identification by avoiding potential misidentifications inferred from genetic information alone. The NBSI framework is currently implemented as a new R package, 'NicheBarcoding', that is open source under GNU General Public Licence and freely available from https://CRAN.R-project.org/package=NicheBarcoding.

KEYWORDS
biodiversity, DNA barcoding, incomplete lineage sorting, integrated taxonomy, niche model, phylogenetic niche conservatism

1 | INTRODUCTION

DNA barcoding has evolved over two decades and is now widely recognized as the optimal solution for large-scale, species-level identification. It has greatly expanded in use from its origins in taxonomy to become a standard tool in ecological and evolutionary research (Fiser Pečnikar & Buzan, 2014; Joly et al., 2014). Following innovations in DNA sequencing (Taylor & Harris, 2012), nearly sixteen million (15,824,647) barcodes from over three-hundred fifty thousand (351,521) species have been accumulated in BOLD Systems (The Barcode of Life Data, http://www.boldsystems.org; Ratnasingham & Hebert, 2007) as of February 2024. Information gathered from DNA barcodes has proven invaluable for understanding interspecies interactions (García-Robledo et al., 2013; Kamo et al., 2018; Pfenninger et al., 2007; Santos et al., 2011; Zhang et al., 2020); assessing biodiversity in species-rich, difficult-to-access and poorly catalogued ecosystems (Ashfaq et al., 2018; Chen et al., 2016; Hao et al., 2020; Mora et al., 2011; Wu et al., 2023); monitoring illegal trade in animal byproducts (Asis et al., 2016; Domingo-Roura et al., 2006; Sultana et al., 2018); identifying exotic species reliably and quickly (Ficetola et al., 2008; Pejovic et al., 2016; Porco et al., 2012); verifying the identity of medicinal plants (Chen et al., 2014; Gong et al., 2018; Yang et al., 2020); and more.

However, challenges persist due to the inherent limitations of mitochondrial DNA (mtDNA) sequences (Chesters et al., 2015; Rubinoff & Holland, 2005). Studies have shown that identical COI sequences do not guarantee species concordance, with a 6% chance of misidentification (Meier et al., 2006). Non-monophyly among barcode genes, reflecting discordance between gene tree and species tree, further complicates identification (Chesters et al., 2015; Funk & Omland, 2003; Ross, 2014). These are often linked to mitochondrial hybrid introgression, incomplete lineage sorting, and male-biased gene flow (e.g. Despres, 2019; McGuire et al., 2007; Moritz & Cicero, 2004; Talavera et al., 2013; Wallis et al., 2017), but can also stem from human factors such as tree inference methods, inaccurate reference taxonomy, under-sampling, or other operational factors (Bergsten et al., 2012; Lim et al., 2012; Meier et al., 2008; Mutanen et al., 2016; Wiemers & Fiedler, 2007). The development of integrative taxonomy concepts and techniques serves to improve the issue of species misidentification caused by reliance on single data sources (Borges et al., 2016; Collins & Cruickshank, 2013; Orr et al., 2022; Spiers et al., 2022; Will et al., 2005; Wright et al., 2019). For example, Yang et al. (2022) has presented a convolutional neural network method (MMNet) that integrates morphological and molecular data for species identification with high accuracy across various taxa. However, Yang et al. (2022)'s method may also encounter big challenges when morphological data are limited due to the lack of photographs, the presence of tiny or damaged specimens.

In such cases, instead of morphology, geospatial data, with its generality, accessibility, quantifiability and high species specificity, may serve as an additional source of information to complement barcode sequence data for species identification. The increased volume of geotagged DNA barcoding samples (e.g. BOLD Systems) offer valuable reference data, rendering niche model analysis an ideal aspect for integrative species identification without morphological information (Ballesteros & Hormiga, 2018; Leaché et al., 2009; Raxworthy et al., 2007; Ruiz-Sanchez & Sosa, 2010; Scriven et al., 2016). However, niche modelling is rarely used for species identification at present, typically employed as supplementary evidence in the delineation of closely related species (Ballesteros & Hormiga, 2018; Duran et al., 2019; Leaché et al., 2009; Orr et al., 2014; Rissler & Apodaca, 2007; Wijayathilaka et al., 2018; Wilson et al., 2012). Although ecological information reflects the long-term adaption of species to certain environments, it may provide useful information for species identifications to avoid potential assignment errors encountered when solely basing identification on barcode sequences.

Here, we propose a Niche-model-Based Species Identification (NBSI) framework to reinforce species identifications in traditional DNA barcoding by incorporating both information from DNA barcodes and niche model-based species distribution. Systematic tests under different scenarios using both simulated and empirical datasets are performed to show that environmental factors/variables could be valuable complements to DNA sequence data for species identification. We also provide an R package 'NicheBarcoding', to enable free and easy use of this framework by researchers worldwide.

2 | A NEWLY PROPOSED FRAMEWORK: NBSI

Integrating data from different types is a complex and challenging task that requires careful consideration and progressive development (Orr et al., 2022). Our novel framework proceeds as follows: Initially, we employ DNA barcoding identification to determine a potential species k for an unknown sample, and derive the membership probabilities $P_{b,k}$ based on the barcode sequence. Subsequently, we construct an ecological niche model and make a prediction for that potential species to estimate the membership probabilities $P_{e,k}$ for the unknown sample based on ecological variables. Finally, considering the topological monophyly of the barcode sequences (i.e. whether intraspecific genetic distances are smaller than interspecific ones), we assign weights to the probabilities derived from both the sequence model and the ecological niche model according to their confidence levels (SR_b and $SR_{e,k}$), resulting in the final integrated outcome of NBSI. The notation table of the parameters involved in the NBSI framework can be found in Table S1.

2.1 | Probability based on barcode sequence

To determine the intrinsic discriminative capacity of barcode sequence data for different species, we calculate the topological monophyly of the reference sequence dataset (Ref_b). The *nj* function from the 'ape' package (Paradis et al., 2004; Popescu et al., 2012) in R-4.2.1 is used to estimate phylogenetic structure based on genetic distances, and the *monophyly* function from the 'spider'

package (Brown et al., 2012) is applied to identify a list of species with sequence monophyly $\mathrm{MonoList}_b$. In this context, species with only one sequence (i.e. singleton) in the dataset are treated as non-monophyletic.

Then, for the identification of DNA barcoding, we choose a non-tree-based Bayesian approach (Jin et al., 2013; Nielsen & Matz, 2006; Zhang et al., 2017) to establish a *naiveBayes* model B using the 'e1071' package (https://CRAN.R-project.org/package=e1071). The trained model B is then queried with Ref_b, and the identification results for each sample in Ref_b are tallied. When the identified species matches the actual species and the Bayesian posterior probability of species membership is ≥ 0.99, it is recorded as a true positive (TP); whereas the probability is < 0.99, it is a false negative (FN). Conversely, if the identified species does not match the actual species but the probability is ≥ 0.99, it is a false positive (FP); whereas if the probability is < 0.99, it is a true negative (TN). The confidence of model B (Success Rate, SR) is calculated as (Yang et al., 2022; Zhang et al., 2012):

$$\mathrm{SR}_b = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \times 100\%.$$

Using the query sequence Que_b in model B, we can obtain the potential target species k for Que_b and the membership probability of Que_b belonging to species k, denoted as $P_{b,k}$.

2.2 | Probability based on ecological niche

For the identification procedure of the ecological niche, the model Niche_k of species k ($k = 1, 2, \ldots, n$) is built by a nonlinear method with good performance, MAXENT (Phillips et al., 2006), using environmental variables dataset Ref_e. According to the ROC curve, we can obtain the model evaluation parameters, including specificity, sensitivity, threshold ($T_{e,k}$), and accuracy ($\mathrm{SR}_{e,k}$). The threshold value here represents the predicted probability of the modelled species $\mathrm{Ref}_{e,k}$ belongs to the corresponding species k when self-inquiring to the model Niche_k.

Using the query environmental variables Que_e, we infer the ecological niche model constructed for species k and then obtain $\hat{E}_{\mathrm{que},k}$. To ensure the comparability of results from different species' models, we employ the Fuzzy method (Shi et al., 2018; Zhang et al., 2012) for standardization by calculating the following three parameters:

$$x = 1 - \hat{E}_{\mathrm{que},k},$$

$$\theta_1 = 1 - T_{e,k},$$

$$\theta_2 = 1 - \max\left\{\hat{E}_{i,k} \mid 1 \leq i \leq n; i \neq k\right\}.$$

The probability $P_{e,k}$ that the query environment variable Que_e belongs to the potential target species k is obtained by the following fuzzy-set function (Shi et al., 2018; Zhang et al., 2012):

$$P_{e,k} = f(x,\theta_1,\theta_2) = \begin{cases} 1, & x \leq \theta_1 \\ 1-2\times\left(\dfrac{x-\theta_1}{\theta_2-\theta_1}\right)^2, & \theta_1 < x \leq \dfrac{\theta_1+\theta_2}{2} \\ 2\times\left(\dfrac{x-\theta_2}{\theta_2-\theta_1}\right)^2, & \dfrac{\theta_1+\theta_2}{2} < x \leq \theta_2 \\ 0, & x > \theta_2 \end{cases}.$$

2.3 | Calculation of the final integrated outcome

Following the setting of threshold value in Shi et al. (2018)'s study (these values could also be set by users), when $P_{b,k} \geq 0.95$ and $P_{e,k} = 1$, both information sources confirm the potential species k as correct, resulting in $\mathrm{NBSI} = 1$; while $P_{b,k} < 0.90$ and $P_{e,k} = 0$, both sources indicate the potential species k as incorrect, resulting in $\mathrm{NBSI} = 0$. In cases where the results do not match either of the above scenarios, it may suggest a discrepancy between the two sources or an ambiguous situation that requires further investigation. We should consider whether the species k is included in the list of species with sequence monophyly $\mathrm{MonoList}_b$. If yes, then based on the model averaging algorithm (Buckland et al., 1997; Burnham & Anderson, 2004; Liu et al., 2023), weights are assigned to the probability results from both sources according to their confidence levels:

$$\mathrm{NBSI} = \frac{\mathrm{SR}_b}{\mathrm{SR}_b+\mathrm{SR}_{e,k}} \times P_{b,k} + \frac{\mathrm{SR}_{e,k}}{\mathrm{SR}_b+\mathrm{SR}_{e,k}} \times P_{e,k} = \frac{\mathrm{SR}_b \times P_{b,k} + \mathrm{SR}_{e,k} \times P_{e,k}}{\mathrm{SR}_b+\mathrm{SR}_{e,k}}.$$

If not, the species k is considered non-monophyletic, indicating a low reliability of the barcode result $P_{b,k}$. In this case, the ecological niche model result is directly used, with $\mathrm{NBSI} = P_{e,k}$.

3 | MATERIALS AND METHODS

To explore the performance of this approach, we tested the NBSI framework in both simulated and empirical datasets across a range of geographic scales while varying genetic diversity, to comprehensively compare the integrated use of both distribution and barcode data to the use of traditional DNA barcoding methods in isolation. All datasets involved in this study are given in Appendices S1 and S2.

3.1 | Generation of simulated datasets

In the simulations, we considered two factors that may influence the success rate of species identification in NBSI: (1) the number of individuals of each species in the reference dataset; and (2) the monophyly of the gene tree compared with its true species tree. These are main factors affecting the reliability of reference libraries built using barcode information. The detailed simulation strategy is described as follows.

3.1.1 | Simulation of gene trees and barcode sequences

Species trees were generated with the coalescent process using the *rcoal* function in the package 'ape' (Paradis et al., 2004; Popescu et al., 2012) in R. We set three gradients of virtual species numbers (20, 50 and 100) to simulate reference datasets with differing species richness. Within each species tree, coalescent simulations were also performed to generate gene trees (Zhang et al., 2008) using the *simSeqfromSp* function in the package 'phybase' (Liu & Yu, 2010). To ensure dataset comparability, the sample size of all simulated datasets was set to 100 virtual individuals. Thus, the dataset with 20 species would have five individuals in each species, while the 50-species dataset would have two per species and the 100-species dataset would be an all-singleton dataset, which can adequately test the performance of our approach when handling commonplace datasets with limited sampling. The fundamental coalescent parameter, θ, was set to three gradients of 0.05, 0.1 and 0.2. As θ refers to the mutation rate per site in molecular sequence data (Yang, 2006), increases in this parameter cause corresponding increases in intraspecific genetic distance. When the intraspecific distance exceeds the interspecific distance between neighbours, a non-monophyletic gene tree would be recovered, thereby decreasing the success rate of barcoding species identifications (Meier et al., 2006).

The simulated sequences were evolved along these gene trees using the GTR model to generate a set of sequence matrices with the *simSeqfromSp* function implemented in the package 'phybase' (Liu & Yu, 2010). We set the sequence length to 658 base pairs, consistent with one of the most commonly used COI barcode fragments (LCO-HCO: Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, et al., 2003). Parameter values used in GTR model were derived from Roe and Sperling (2007) study with the following settings: base frequencies A = 0.3255, C = 0.1014, G = 0.1206, T = 0.4525; and rates matrix AC = 10.6213, AG = 16.7683, AT = 8.8273, CG = 1.5416, CT = 122.9118 and GT = 1.000.

3.1.2 | Simulation of distribution patterns

Nineteen bioclimatic variables were downloaded from WorldClim (version 1.4 with 2.5 arc minute resolution; http://www.worldclim.org/; Hijmans et al., 2005) to quantify the ecological space of each simulated individual. We used the *randomPoints* function in the 'dismo' package (https://CRAN.R-project.org/package=dismo) to randomly generate 5000 real coordinates as background points, and extracted their bioclimatic variables as an available matrix by applying the *extract* function in the 'raster' package (https://CRAN.R-project.org/package=raster) with the downloaded WorldClim data. Then, the environmental data were randomly sampled from the variable matrix and matched to the virtual individuals of each simulated dataset.

In order to test whether the barcoding and ecological information of a reference dataset are significantly correlated, we calculated the pairwise genetic distance and ecological distance of each dataset at the species level using the K80 model (Kimura, 1980) and the Euclidean method, respectively, and applied the Mantel statistic (Mantel, 1967) using the *mantel* function in the package 'vegan' with 999 random permutations (Ballesteros & Hormiga, 2018; Oksanen et al., 2020). A significantly correlated result from the Mantel test implies some type of association between the barcode sequence and ecological fitness, whereas an uncorrelated result means that the two properties are independent.

In total, $3 \times 3 = 9$ datasets were simulated with different settings (Table 1). Additionally, to evaluate the effectiveness of our method in handling extreme situations, we selected sequences that were potentially misidentified by DNA barcoding to constitute another set of tested datasets. The genetic nearest neighbours of these targeted sequences belonged to either another species or more than two species. Given the intrinsic limitations of mtDNA and the complexity of human factors (see Section 1), relying solely on sequence information in these scenarios could lead to ambiguous species assignments (Zhang et al., 2012).

3.2 | Collection of empirical datasets

We also tested our framework with six empirical datasets retrieved from online databases and empirical studies, comprising different real-world genetic diversity and various ecological conditions. These datasets covered different taxonomic scales and gene markers, and each was tested using the commonly used bioclimatic variables from WorldClim (version 1.4 with 2.5 arc minute resolution; http://www.worldclim.org/; Hijmans et al., 2005).

Four of the six datasets were downloaded from BOLD Systems (The Barcode of Life Data, http://www.boldsystems.org; Ratnasingham & Hebert, 2007), including both barcode sequences (COI fragments) and geographic coordinates of collection localities: a nematode dataset (order-level: Spirurida, Chromadorea), a reptile dataset (family level: Scincidae, Reptilia Squamata), two insect datasets (family level: Limacodidae, Insecta Lepidoptera; genus-level: *Theretra*, Insecta Lepidoptera Sphingidae). Specimens from the 'hawk-moths' dataset (family level: Sphingidae, Insecta Lepidoptera) was collected using light-traps in Zhejiang, China from 2017 to 2018; we recorded the geographic information for each sample and sequenced their COI fragments following procedures in Jin et al. (2013). The '*Dendrolimus*' dataset (genus-level: *Dendrolimus*, Insecta Lepidoptera Lasiocampidae) was taken from Dai et al. (2012). The researchers found that the phylogenetic relationships of distantly related species were clearly resolved by barcode sequence; whereas the three closely related species (*D. punctatus*, *D. tabulaeformis*, *D. spectabilis*) could not be resolved (Dai et al., 2012).

All datasets were further cleaned by eliminating records which did not fulfil these criteria: (1) complete taxonomic information (family, genus and species names), (2) valid coding gene fragments and (3) available geographic coordinates. Table 2 lists the basic information of each dataset as well as the Mantel test results between barcode sequences

TABLE 1 Settings and Mantel tests of simulated datasets.

Dataset no.	Number of species	Number of individuals of each species	θ	Barcoding gap[a]	MantelStat.[b] (r)	MantelSig.[c] (p-value)	Number of ambiguous sequences[d]	Proportion of ambiguous sequences (%)
1	20	5	0.05	−0.61	−0.0323	0.620	8	8.00
2	20	5	0.1	−0.59	−0.1029	0.825	15	15.00
3	20	5	0.2	−0.64	−0.0627	0.630	16	16.00
4	50	2	0.05	−0.62	−0.0365	0.926	51	51.00
5	50	2	0.1	−0.62	−0.0159	0.697	57	57.00
6	50	2	0.2	−0.62	−0.0335	0.922	64	64.00
7	100	1	0.05	−0.49	0.0121	0.323	100	100.00
8	100	1	0.1	−0.49	−0.0292	0.780	100	100.00
9	100	1	0.2	−0.55	−0.0378	0.895	100	100.00

[a]The difference between the maximum intra-specific genetic distance and the minimum inter-specific genetic distance.
[b]The statistic r of the Mantel test between genetic and environmental Euclidean distance matrix.
[c]Mantel test p-values. If $p \leq 0.05$, then the two-distance matrix are significantly correlated; if $p > 0.05$, then they are uncorrelated.
[d]The sequences whose genetic nearest neighbours belongs to either another species or more than two species.

TABLE 2 Basic information and Mantel tests of empirical datasets.

No.	Datasets	Size	Number of species	Percentage of singletons (%)	Barcoding gap[a]	MantelStat.[b] (r)	MantelSig.[c] (p-value)	Number of ambiguous sequences[d]	Proportion of ambiguous sequences (%)
1	Spirurida	63	24	37.50	−0.0533	0.4412	**0.004****	11	17.46
2	Scincidae	110	32	53.12	−0.1985	0.1883	**0.003****	19	17.27
3	Limacodidae	325	297	91.58	−0.1607	0.0257	0.207	286	88.00
4	hawk-moths[e]	666	51	17.65	−0.1738	0.0808	**0.050***	419	62.91
5	*Theretra*	455	46	23.91	−0.0709	0.0676	0.229	43	9.45
6	*Dendrolimus*[f]	145	7	0	−0.0537	0.1000	0.276	19	13.57

[a]The difference between the maximum intra-specific genetic distance and the minimum inter-specific genetic distance.
[b]The statistic r of the Mantel test between genetic and environmental Euclidean distance matrix.
[c]Mantel test p-values. If $p \leq 0.05$, then the two distance matrices are significantly correlated; if $p > 0.05$, they are uncorrelated. *$p \leq 0.05$; **$p \leq 0.01$.
[d]The sequences whose genetic nearest neighbours belongs to either another species or more than two species.
[e]Zhejiang hawk-moths dataset that collected using light-traps.
[f]Chinese *Dendrolimus* dataset from Dai et al. (2012).

and ecological fitness. We also selected the sequences that were potentially misidentified by DNA barcoding as we did for the simulated datasets, and found that those so-called 'ambiguous sequences' were prevalent in 9.45% to 88.00% of the empirical datasets we selected (Table 2). The global distribution of specimens is shown in Figure 1.

3.3 | Ecological variables used in niche modelling

For testing the empirical datasets, 19 bioclimatic variables from WorldClim (version 1.4 with 2.5 arc minute resolution; http://www.worldclim.org/; Hijmans et al., 2005) were applied to construct niche models. To avoid the influence of potential correlations among variables on ecological niche modelling, we implemented the variance inflation factor (VIF; Fox & Monette, 1992; Lai et al., 2022) to assess multicollinearity. The *vif* function from the 'car' package (https://CRAN.R-project.org/package=car) was employed to estimate VIF values for each environmental variable. If there were variables with VIF values ≥10, the variable with the highest VIF value would be removed. Then, the remaining variables were reassembled into a new set, and the VIF values were recalculated collectively. This iterative process continued, with the highest VIF variable being eliminated at each step, until all variables in the revised set exhibited VIF values below the threshold of 10. In the testing procedures of this study, the retained variables for each niche modelling have been documented in the Appendices S3 and S4.

3.4 | Test and success rate calculation

For each dataset and the corresponding dataset composed of ambiguous sequences (the sequences whose genetic nearest

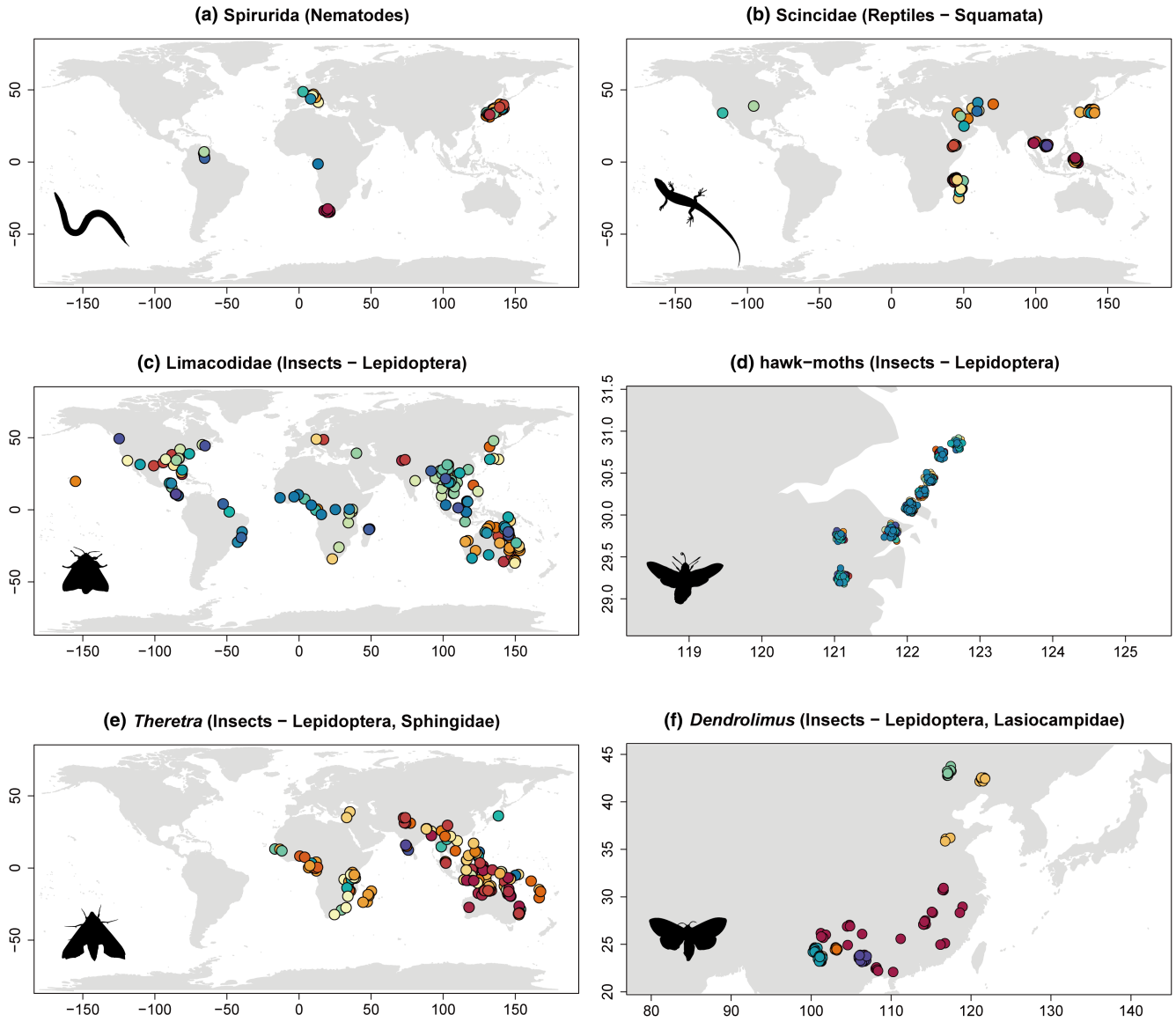

FIGURE 1 Empirical datasets and their global distributions, including four BOLD-downloaded datasets: (a) Spirurida (nematodes), (b) Scincidae (reptiles, Squamata), (c) Limacodidae (insects, Lepidoptera) and (e) *Theretra* (insects, Lepidoptera Sphingidae); and two field-sampled or literature-published datasets: (d) Zhejiang hawk-moths (insects, Lepidoptera), (f) *Dendrolimus* (insects, Lepidoptera Lasiocampidae; Dai et al., 2012). Points in different colours represent different species.

neighbours belong to another species or more than two species; see Tables 1 and 2), we implemented 1000 replicates of the leave-one-out cross-validation test (Yang et al., 2022; Zhang et al., 2012) to obtain the species identification results and its credibility, $P_{b,k}$ and NBSI, for each inquiry within the simulated and empirical datasets. A series of thresholds were set at 0.90, 0.95 and 0.99 to accept the identification results, calculating the success rate for both the traditional barcoding identification SR_b and the niche-model-based identification SR_{NBSI}. True positive (TP) and true negative (TN) query results were both considered successful inquiries ($SR = (TP + TN) / 1000 \times 100\%$; Yang et al., 2022; Zhang et al., 2012).

4 | RESULTS

4.1 | Performance in simulated datasets

The success rates of NBSI versus traditional DNA barcoding for the completely simulated datasets and their ambiguous sequence datasets are given in Figure 2A,B, respectively, including the various permutations detailed above. Detailed statistics including sensitivity and specificity results are listed in Table S2.

Given that all complete simulated datasets consisted of 100 virtual individuals, the number of virtual individuals per species in the 20-species, 50-species and 100-species datasets were 5, 2 and 1,

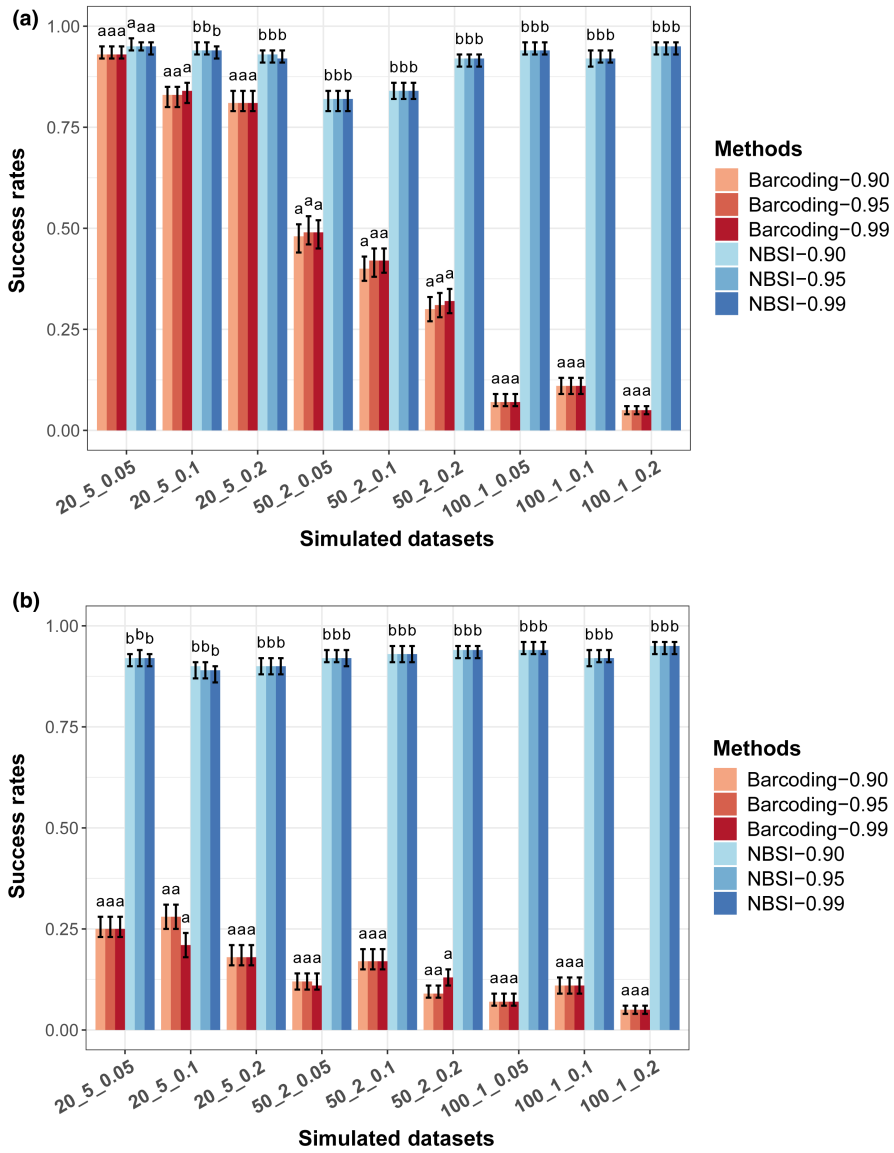

FIGURE 2 Success rates of NBSI (blue bars) compared with traditional DNA barcoding (red bars) when testing (A) completely simulated datasets and (B) their corresponding ambiguous sequence datasets. The colour gradient of the bars, from light to dark, represents the thresholds of acceptance for identification results, set at 0.90, 0.95 and 0.99. The *X*-axis lists the nine simulated datasets, each configured with a virtual species count of 20 (5 individuals per species), 50 (2 individuals per species) or 100 (1 individual per species), and further specified the molecular mutation rate by coalescent parameters (θ) set at 0.05, 0.1 and 0.2; the *Y*-axis indicates the percentage of successful identifications (true positive and true negative assignments) among 1000 leave-one-out cross-validation test. The error bars indicate 95% confidence intervals, and different lowercase letters suggest significant differences at the level of 0.05 using Tukey's 'Honest Significant Differences' test (Miller, 1981; Yandell, 1997).

respectively. The success rate of traditional DNA barcoding identification was highest in the three 20-species datasets, where each species had a relatively greater number of reference sequences; while the success rate was lower in the three 50-species datasets. For the three 100-species datasets, when tested using the leave-one-out method, singletons were removed as pseudo-query sequences, resulting in the absence of matching species members in the reference database for the target species, leading to a high proportion of false positives (FP) and an exceptionally low success rate (Figure 2A). For datasets with equal numbers of species, the success rate of traditional DNA barcoding identification decreases as the sequence variability increases (θ=0.05, 0.1, 0.2; Figure 2A).

Working better overall, the niche-model-based identification approach significantly improved the identification success rate in eight of the nine scenarios (Figure 2A). The largest improvement occurred in the '100_1_0.2' dataset at all acceptance thresholds, where the success rate improved from 4.7% (95% CI: 3.51%–6.25%) to 94.8% (95% CI: 93.19%–96.06%); while the smallest was the '20_5_0.05' dataset in threshold 0.99, where the success rate improved from 93.3% (95% CI: 91.52%–94.73%) to 95.0% (95% CI: 93.41%–96.23%). The latter represents an ideal scenario where the reference database contains sufficient replicates for each species, and exhibits low intraspecific genetic variability, enabling a relatively high level of identification accuracy based solely on barcode sequence information. Even in this context, our proposed method is capable of further enhancing the results, which implies that the integration of ecological information confers an intrinsic advantage.

For the tests on the nine ambiguous sequence datasets, all the pseudo-query samples extracted by leave-one-out simulation could certainly be misidentified by traditional DNA barcoding. The success rates of these datasets are exceptionally low regardless of the number of virtual individuals per species in the reference library (Figure 2B). As expected, the niche-model-based identification approach significantly improving the success rates in all scenarios. The largest improvement was still from the '100_1_0.2' dataset, because it is composed of singletons; while the smallest

was for the '20_5_0.1' dataset in threshold 0.95, where the success rate improved from 27.7% (95% CI: 24.97%–30.61%) to 89.2% (95% CI: 87.07%–91.02%). Meanwhile, these improvements are consistent across datasets with varying settings of θ, supporting our assertion that the integration of ecological information can enhance species identifications by avoiding potential misidentifications that arise solely from sequence data.

4.2 | Performance in empirical datasets

Similar findings were observed in the test results of empirical datasets (Figure 3). Our new framework performed well across the various empirical scenarios overall, with significant improvements in five of the six cases and, minimally, comparable success to traditional barcoding; and with significant improvements in all tests of the ambiguous sequence datasets (Figure 3A,B). The number of testing replications, the threshold and calculation of the success rate for empirical datasets were consistent with those of the simulated datasets. Detailed statistics including sensitivity and specificity results are given in Table S3.

In the complete empirical datasets, the largest improvement of NBSI over traditional DNA barcoding was in the 'Limacodidae' dataset at threshold 0.90, where the success rate improved from 17.3% (95% CI: 15.03%–19.82%) to 67.4% (95% CI: 64.38%–70.28%); while the smallest was for the '*Dendrolimus*' dataset, where the success rate improved from 94.1% (95% CI: 92.41%–95.44%) to 96.4% (95% CI: 95.00%–97.43%). In the former case, the proportion of singletons reached as high as 91.58%, with 88% of the samples classified as ambiguous sequences (Table 2), resulting in an extremely low species identification accuracy relying solely on sequence information. Our method is adept at handling the situation under these dire circumstances. In contrast, the latter lacks singletons in its reference library and contains only 13.57% ambiguous sequences, where traditional barcoding methods already achieve a relatively high success rate. Yet datasets

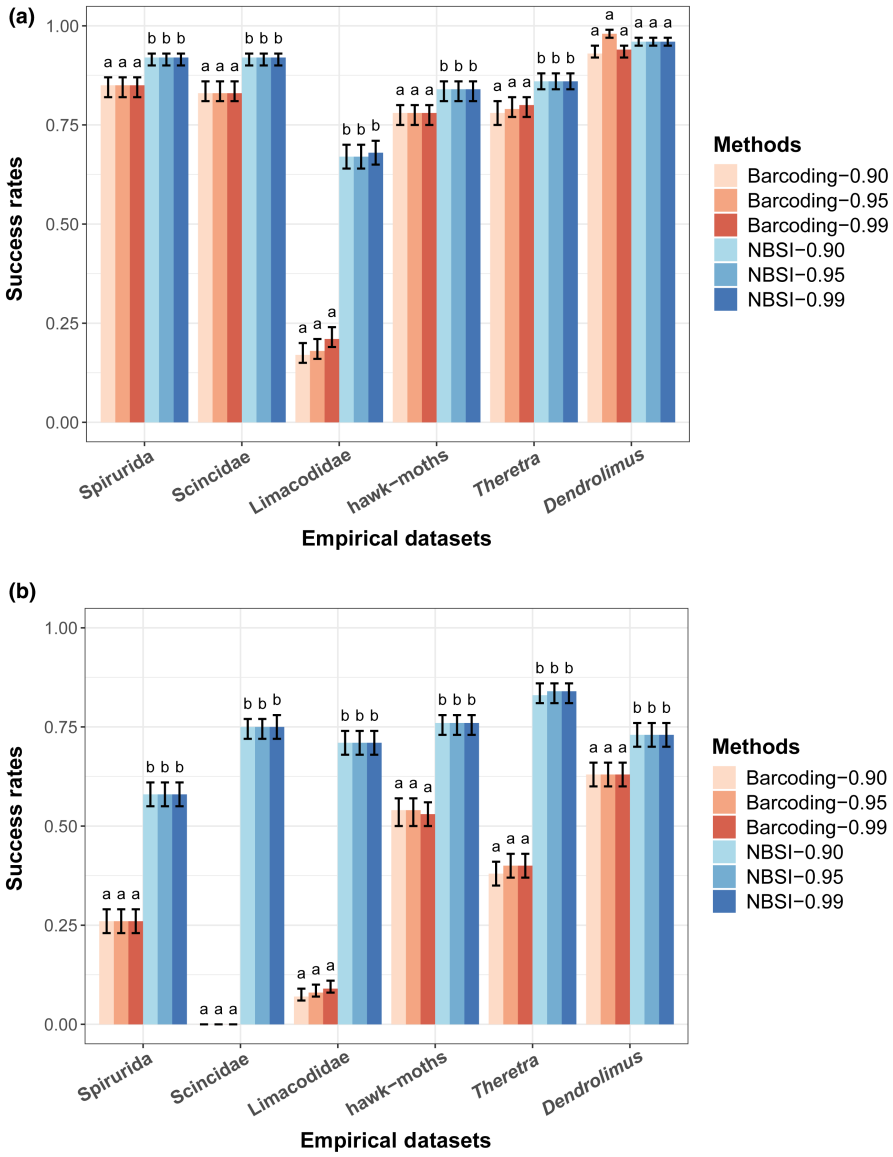

FIGURE 3 Success rates of NBSI (blue bars) compared with traditional DNA barcoding (coral bars) when testing (A) completely empirical datasets and (B) their corresponding ambiguous sequence datasets. The colour gradient of the bars, from light to dark, represents the thresholds of acceptance for identification results, set at 0.90, 0.95 and 0.99. The *X*-axis lists the six empirical datasets, including four BOLD-downloaded datasets, a field-sampled and a literature-published datasets; the *Y*-axis indicates the percentage of successful identifications (true positive and true negative assignments) among 1000 leave-one-out cross-validation test. The error bars indicate 95% confidence intervals, and different lowercase letters suggest significant differences at the level of 0.05 using Tukey's 'Honest Significant Differences' test (Miller, 1981; Yandell, 1997).

that meet such conditions are rare in practice. In the testing of the other four datasets, the magnitude of improvement is relatively similar (Figure 3A). They represent datasets composed of different taxonomic ranks (including orders, families, families in a specific region, and genera), which are more reflective of the general conditions encountered in empirical studies.

For the tests on the corresponding ambiguous sequence datasets, traditional DNA barcoding generated more misidentifications than with the completely empirical datasets. Figure 3B shows that even in the ideal dataset, '*Dendrolimus*', the success rates were 63%, while the 'Scincidae' dataset had zero correct identification among 1000 leave-one-out tests. As expected, the NBSI approach significantly improved success rates in all scenarios (Figure 3B). The largest improvement presents to the 'Scincidae' dataset in threshold 0.99, where the success rate improved from 0 to 75.1% (95% CI: 72.28%–77.73%); while the smallest was for the '*Dendrolimus*' dataset in all acceptance thresholds, where the success rate improved from 63.2% (95% CI: 60.12%–66.18%) to 73.1% (95% CI: 70.22%–75.80%). In these scenarios where misidentifications are almost certain, our method can still achieve a success rate of up to 84% (see the results of '*Theretra*' dataset).

5 | DISCUSSION

Using both simulated and empirical datasets, we demonstrate that the newly proposed niche-model-based species identification (NBSI) framework can successfully integrate both DNA sequence and environmental information of species to greatly improve species identification, when compared to traditional DNA barcoding. Under the NBSI framework, the use of these two relatively independent data sources enables each to act as a check on the other in case one source is biased in some manner. Phenomena such as non-monophyly and the other discussed issues with traditional DNA barcoding represent non-trivial challenges that are difficult or sometimes impossible to overcome with barcoding alone. If only ecological information were used, this could also lead to issues, as unrelated species may commonly converge upon similar niches and this could cause the assignment of entirely unrelated species as the same thing. In the present study, we construct our approach by applying a non-tree-based Bayesian approach (Jin et al., 2013; Nielsen & Matz, 2006), a nonlinear MAXENT model (Phillips et al., 2006), and the commonly used WorldClim variables (version 1.4 with 2.5 arc minute resolution; http://www.worldclim.org/; Hijmans et al., 2005).

Although the NBSI framework sounds theoretically perfect, there are factors that may impact its performance under real-world applications. For instance, due to the inherent limitations, the application of ecological niche models (ENMs) requires caution when dealing with species with rare distributions, as their accuracy is highly dependent on the quality of species occurrence and environmental data (Deb et al., 2017; Murphy & Smith, 2021). Precisely for this reason, niche modelling with presence-only data may have limited predictive accuracy due to spatial sampling bias (SSB), and this is challenging to fully correct for without independent test data (Baker et al., 2024). In fact, our NBSI approach also provides a solution to some extent, which is to use the niche model for a secondary check only after the barcode has identified a potential target species *k* (see the Sections 2.1 and 2.2 for details). Therefore, the outcomes of the niche model do not directly dictate the species affiliation of unknown samples, but they can validate or refute the preliminary results of barcoding identifications. Nonetheless, there remains a need for heightened vigilance regarding this issue, particularly when the models assume that sampling data are random or representative (Yackulic et al., 2013). Some studies have proposed solutions to such problems through innovative statistical methods, such as the joint classification-occupancy model (Spiers et al., 2022; Wright et al., 2019), the two-species false-positive N-mixture model (Clement et al., 2022), and so on. These will have a positive effect on improving misclassification and should be further considered when implementing the NBSI method. Concurrently, these limitations suggest that an optimized NBSI method should also provide users with flexible choices in niche modelling approaches, and a variety of available SSB correction methods (Baker et al., 2024) could be integrated into the process in the future.

The unavoidable, inherent limitations of sequence information determine the upper limits of the success rate of species identification using traditional DNA barcoding. In complex evolutionary processes, barcode sequences often exhibit low interspecific divergence and high intraspecific variation, leading to indistinct barcoding gaps (e.g. Table 2), which render them ineffective for distinguishing all species (Xu et al., 2017). Regardless of whether algorithms are distance-based, tree-based or character-based, they cannot thoroughly solve the prevalent problem of shared haplotypes among different species for barcoding markers, let alone the artificial inter-specific similarities that result from sampling bias (Chesters et al., 2015; Rubinoff & Holland, 2005). Here, we extracted these potentially ambiguous sequences whose genetic nearest neighbours belong to another species or more than two species, and found that NBSI significantly improved the success rate of species identifications in these challenging situations. We also demonstrated that the NBSI framework does not introduce more errors than traditional DNA barcoding when faced with completely simulated and empirical datasets, and in most cases, it outperformed traditional DNA barcoding method that only uses DNA sequence information for species identification. Figure 4 shows the relationship between improvements of success rates and proportions of ambiguous sequences in each dataset. With increasing proportion of ambiguous sequences in these complete datasets, the improvements by NBSI over traditional barcoding become larger. This trend is significant across both simulated ($r = 0.9898$; p-value = 1.165e-22) and empirical datasets ($r = 0.7919$; p-value = 8.979e-05), suggesting that the NBSI framework performs significantly better across tests when ambiguous sequences are prevalent.

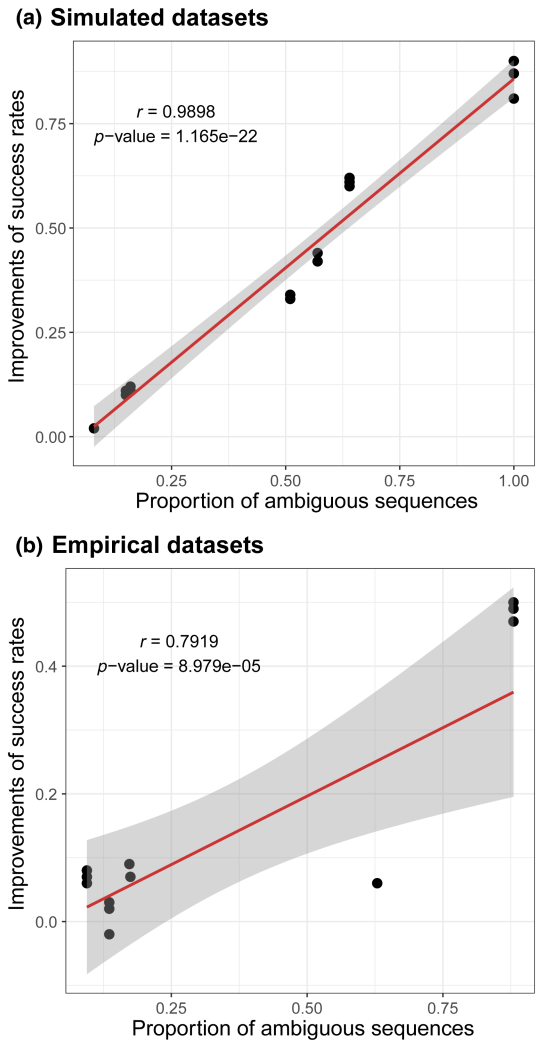

FIGURE 4 Improvements of success rates from traditional DNA barcoding to the NBSI framework in the (a) simulated and (b) empirical datasets. The *X*-axis indicates proportions of ambiguous sequences and the *Y*-axis indicates improvements of success rates. Correlations were measured by Spearman's rank correlation coefficient (Hollander & Wolfe, 1973).

Global online datasets of bioclimatic variables such as WorldClim (version 1.4/2.1; http://www.worldclim.org/; Hijmans et al., 2005) and ENVIREM (version 1.0; https://envirem.github.io/; Title & Bemmels, 2018) are widely used in niche-related research because of their worldwide coverage. However, such macro-data are not useful when species differ primarily in their microhabitat requirements. The selection of macro- or micro-environmental variables depends on the distribution of species in the reference library, the ecological requirements of the species being studied and the availability of such data. Coarse-scale datasets are more commonly available for the species with large distributions, or the species distributed across many latitudes and ecoregions, while high-resolution data are more appropriate for species with narrow or sympatric distributions, or those with very strict microhabitat requirements. As such, the type of data used is often a balancing act between what is optimal and what is feasible, with considering the impact of the calculation method of bioclimatic variables on the model (Bede-Fazekas & Somodi, 2020). In this study, test datasets were selected from widely distributed species at large geographic scales, and we used WorldClim variables to demonstrate the power of our NBSI framework, given the importance and common usage of these variable sets. For those species with strict microhabitat requirements or for studies at smaller spatial scales, users of the 'NicheBarcoding' package could collect or collate their own micro-environmental information and use that instead of broad-scale datasets.

As described in Hutchinson (1957), the fundamental niche of a species is an 'n-dimensional hypervolume', in which every point that corresponds to a state of the environment would permit a species to exist indefinitely. Species which overlap in geo-space likely also overlap in eco-space (at coarser grain sizes), but overlapping of species in eco-space does not indicate that they are sympatric in geo-space (Pulliam, 2000; Qiao et al., 2016; Soberón & Nakamura, 2009). The two-dimensional ecological spaces of part of the genetically close related species in the empirical datasets are mapped through principal component analysis (Figure 5), where the NBSI method improves the success rate by taking advantage of the non-overlapping characteristics of their ecological niche. Other novel quantitative methods (e.g. Lu et al., 2021; Osorio-Olvera et al., 2020; van der Veen et al., 2021) are also helpful in displaying the effect of ecological niche similarity on NBSI. Undersampling of species points in geo-space may hinder the generation of accurate niche models, and, consequently, a query sampled from a different geographic area may not be assigned to its correct species based on niche models (false-negative identifications). To avoid this issue, we sampled species points in eco-space instead of geo-space, but we suggest that large sample sizes with true records should be used whenever possible regardless, to fully represent species' true eco-spaces (Arlé et al., 2021) and better understand their limits.

In addition to the geographical scale, the dynamics of ecological niche conservatism within different eco-evolutionary processes on a larger time scale should also be taken into account. For example, closely related species pairs that have recently diverged may have very similar ecological niches (Peterson, 2011); such ecological similarity may be promoted by the presence of predators, or conversely, lead to the differentiation of ecological niches due to long-term competitive interactions (Haraldsson & Thébault, 2023). Therefore, before using the NBSI framework, users can gather prior knowledge about the relationship between genetic information and ecological niche information in the reference database, as shown in Table 2. Yet whether the improvement effect of our approach is still predictable in these complex ecological contexts, remains to be further verified.

6 | CONCLUSIONS

Improving species identifications from DNA barcodes using multiple types of information is a critical pursuit. DNA barcoding will only become more important in the future as we finally begin to document the enormous undescribed diversity of insects and other poorly

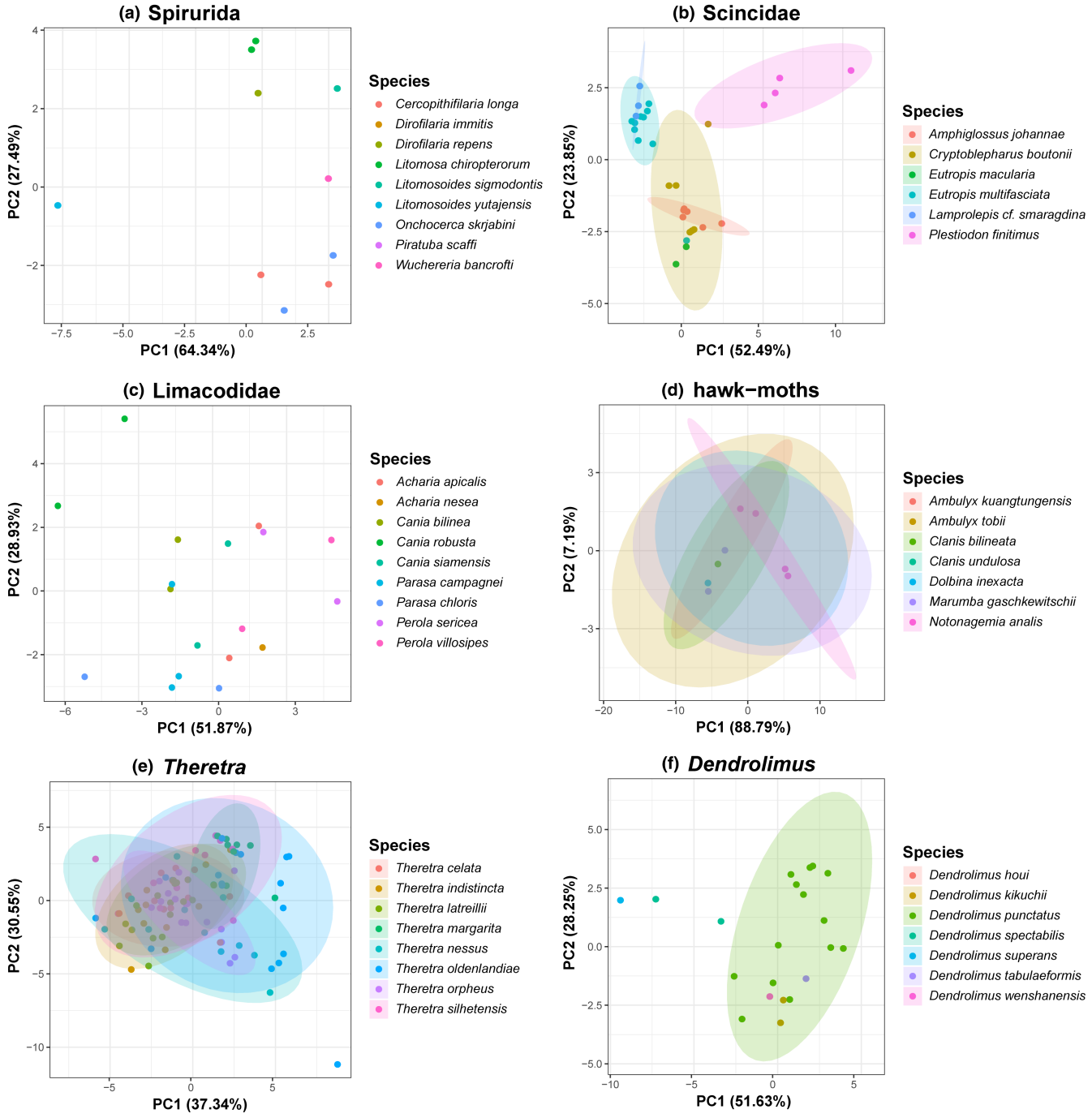

FIGURE 5 Ecological space of part of the genetically close related species in the empirical datasets: (a) Spirurida (nematodes), (b) Scincidae (reptiles, Squamata), (c) Limacodidae (insects, Lepidoptera), and (e) *Theretra* (insects, Lepidoptera Sphingidae); and two field-sampled or literature-published datasets: (d) Zhejiang hawk-moths (insects, Lepidoptera), (f) *Dendrolimus* (insects, Lepidoptera Lasiocampidae; Dai et al., 2012).

known taxa throughout the Global South (Orr et al., 2020). Our niche-model-based species identification framework enables joint decision-making based on ecological and molecular data, representing a significant advancement for DNA barcoding. Modern barcoding approaches and many historical collectors both prioritize locality information collection, making this a widely available and powerful complement for species-level identifications for both museum- and field-based efforts.

Our thorough tests show that ecological information can successfully and significantly improve DNA barcoding under the NBSI framework, including in difficult scenarios where traditional DNA barcoding methods typically fail. We extended the use of niche modelling further to fully take advantage of the vast quantity of distributional data available, as few other data types can compete with the prevalence of barcode data. This is especially true in the Global South where locality information is far easier to generate than molecular

data due to technical or funding limitations (Orr et al., 2020), and many of these areas are tropical and hyper-diverse (containing many undescribed species; Giam et al., 2012), further benefiting from such more powerful, integrated methods. Even where local languages can make georeferencing of historical specimens difficult for foreign researchers working with specimens from such developing countries, this would simply encourage collaboration with local scientists, yet another positive outcome of the increasing interdisciplinarity of DNA barcoding.

The NBSI method will also become better as the accuracy of niche estimation continues to improve (Kass et al., 2021; Pichler & Hartig, 2021). As DNA barcoding technology matures for taxonomy and biodiversity studies (Cristescu, 2014; Hebert, Cywinska, et al., 2003; Hebert & Gregory, 2005; Yang et al., 2022), it will become increasingly vital to integrate ecological and other information to improve these methods.

AUTHOR CONTRIBUTIONS

Ai-bing Zhang designed the study; Cai-qing Yang performed the research; Cai-qing Yang and Ai-bing Zhang wrote the codes, and tested most of the examples. Ying Wang, Xin-hai Li and Michael C. Orr analysed parts of the data; Jing Li and Bing Yang collected some of the datasets; Cai-qing Yang and Ai-bing Zhang wrote the first draft of the manuscript, and all authors contributed substantially to revisions and gave final approval for publication.

ACKNOWLEDGEMENTS

We are deeply grateful for the guidance and insights provided by Jie Zhou and Xiao Liang regarding the issue of model averaging. We also grateful to the editors and the anonymous reviewers for the suggestions provided on our manuscript which have enabled our approach to better serve the relevant researchers. This research was supported by the Natural Science Foundation of China (32200343, 32170421), Beijing Municipal Natural Science Foundation (5232001), Chinese Academy of Sciences President's International Fellowship Initiative program (2024PVC0046), Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 14th Five-year Plan (BPHR20220114), National Key Research and Development Program of China (2023YFC2606600), and Academy for Multidisciplinary Studies, Capital Normal University.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

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.14440.

DATA AVAILABILITY STATEMENT

NicheBarcoding is now freely available from https://CRAN.R-project.org/package=NicheBarcoding (Yang et al., 2024a). Source code can also be found on Github https://github.com/Yangcq-Ivy/NicheBarcoding and Zenodo https://doi.org/10.5281/zenodo.13270695 (Yang et al., 2024b). All releases will be distributed on Microsoft Windows, Mac OSX and Linux platforms. The datasets and the Supporting Information tests in this study can be found at: https://doi.org/10.5061/dryad.rbnzs7h96 (Yang et al., 2024c).

ORCID

Cai-qing Yang https://orcid.org/0009-0003-9997-8374
Ying Wang https://orcid.org/0000-0002-5789-2850
Xin-hai Li https://orcid.org/0000-0003-4514-0149
Jing Li https://orcid.org/0000-0003-3234-8055
Bing Yang https://orcid.org/0000-0002-9610-1935
Michael C. Orr https://orcid.org/0000-0002-9096-3008
Ai-bing Zhang https://orcid.org/0000-0003-3450-5421

REFERENCES

Arlé, E., Zizka, A., Keil, P., Winter, M., Essl, F., Knight, T., Weigelt, P., Jiménez-Muñoz, M., & Meyer, C. (2021). BRACATUS: A method to estimate the accuracy and biogeographical status of georeferenced biological data. *Methods in Ecology and Evolution*, *12*, 1609–1619. https://doi.org/10.1111/2041-210x.13629

Ashfaq, M., Sabir, J. S. M., El-Ansary, H. O., Perez, K., Levesque-Beaudin, V., Khan, A. M., Rasool, A., Gallant, C., Addesi, J., & Hebert, P. D. N. (2018). Insect diversity in the Saharo-Arabian region: Revealing a little-studied fauna by DNA barcoding. *PLoS One*, *13*(7), e0199965. https://doi.org/10.1371/journal.pone.0199965

Asis, A. M., Lacsamana, J. K., & Santos, M. D. (2016). Illegal trade of regulated and protected aquatic species in The Philippines detected by DNA barcoding. *Mitochondrial DNA Part A DNA Mapping, Sequencing, and Analysis*, *27*, 659–666. https://doi.org/10.3109/19401736.2014.913138

Baker, D. J., Maclean, I. M. D., & Gaston, K. J. (2024). Effective strategies for correcting spatial sampling bias in species distribution models without independent test data. *Diversity and Distributions*, *30*, e13802. https://doi.org/10.1111/ddi.13802

Ballesteros, J. A., & Hormiga, G. (2018). Species delimitation of the North American orchard-spider *Leucauge venusta* (Walckenaer, 1841) (Araneae, Tetragnathidae). *Molecular Phylogenetics and Evolution*, *121*, 183–197. https://doi.org/10.1016/j.ympev.2018.01.002

Bede-Fazekas, Á., & Somodi, I. (2020). The way bioclimatic variables are calculated has impact on potential distribution models. *Methods in Ecology and Evolution*, *11*, 1559–1570. https://doi.org/10.1111/2041-210x.13488

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G. N., Ribera, I., Nilsson, A. N., Barraclough, T. G., & Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, *61*, 851–869. https://doi.org/10.1093/sysbio/sys037

Borges, L. M. S., Hollatz, C., Lobo, J., Cunha, A. M., Vilela, A. P., Calado, G., Coelho, R., Costa, A. C., Ferreira, M. S. G., Costa, M. H., & Costa, F. O. (2016). With a little help from DNA barcoding: Investigating the diversity of Gastropoda from the Portuguese coast. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep20226

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M. C., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, *12*, 562–565. https://doi.org/10.1111/j.1755-0998.2011.03108.x

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*, 603. https://doi.org/10.2307/2533961

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference—Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. https://doi.org/10.1177/0049124104268644

Chen, R., Jiang, L. Y., Chen, J., & Qiao, G. X. (2016). DNA barcoding reveals a mysterious high species diversity of conifer-feeding aphids in the mountains of southwest China. *Scientific Reports*, *6*, 20123. https://doi.org/10.1038/srep20123

Chen, S., Pang, X., Song, J., Shi, L., Yao, H., Han, J., & Leon, C. (2014). A renaissance in herbal medicine identification: From morphology to DNA. *Biotechnology Advances*, *32*, 1237–1244. https://doi.org/10.1016/j.biotechadv.2014.07.004

Chesters, D., Zheng, W. M., Zhu, C. D., & Yu, D. (2015). A DNA barcoding system integrating multigene sequence data. *Methods in Ecology and Evolution*, *6*, 930–937. https://doi.org/10.1111/2041-210x.12366

Clement, M. J., Royle, J. A., & Mixan, R. J. (2022). Estimating occupancy from autonomous recording unit data in the presence of misclassifications and detection heterogeneity. *Methods in Ecology and Evolution*, *13*, 1719–1729. https://doi.org/10.1111/2041-210x.13895

Collins, R. A., & Cruickshank, R. H. (2013). The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, *13*(6), 969–975. https://doi.org/10.1111/1755-0998.12046

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*, 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Dai, Q.-Y., Gao, Q., Wu, C.-S., Chesters, D., Zhu, C.-D., & Zhang, A.-B. (2012). Phylogenetic reconstruction and DNA barcoding for closely related pine moth species (*Dendrolimus*) in China with multiple gene markers. *PLoS One*, *7*, e32544. https://doi.org/10.1371/journal.pone.0032544

Deb, C. R., Jamir, N. S., & Kikon, Z. P. (2017). Distribution prediction model of a rare orchid species (*Vanda bicolor* Griff.) using small sample size. *American Journal of Plant Sciences*, *8*, 1388–1398. https://doi.org/10.4236/ajps.2017.86094

Despres, L. (2019). One, two or more species? Mitonuclear discordance and species delimitation. *Molecular Ecology*, *28*, 3845–3847. https://doi.org/10.1111/mec.15211

Domingo-Roura, X., Marmi, J., Ferrando, A., López-Giráldez, F., Macdonald, D. W., & Jansman, H. A. H. (2006). Badger hair in shaving brushes comes from protected Eurasian badgers. *Biological Conservation*, *128*, 425–430. https://doi.org/10.1016/j.biocon.2005.08.013

Duran, D. P., Herrmann, D. P., Roman, S. J., Gwiazdowski, R. A., Drummond, J. A., Hood, G. R., & Egan, S. P. (2019). Cryptic diversity in the North American *Dromochorus* tiger beetles (Coleoptera: Carabidae: Cicindelinae): A congruence-based method for species discovery. *Zoological Journal of the Linnean Society*, *186*, 250–285. https://doi.org/10.1093/zoolinnean/zly035

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*, 423–425. https://doi.org/10.1098/rsbl.2008.0118

Fiser Pečnikar, Z., & Buzan, E. V. (2014). 20 years since the introduction of DNA barcoding: From theory to application. *Journal of Applied Genetics*, *55*, 43–52. https://doi.org/10.1007/s13353-013-0180-y

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*, 178–183. https://doi.org/10.1080/01621459.1992.10475190

Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, *34*, 397–423. https://doi.org/10.1146/annurev.ecolsys.34.011802.132421

García-Robledo, C., Erickson, D. L., Staines, C. L., Erwin, T. L., & Kress, W. J. (2013). Tropical plant-herbivore networks: Reconstructing species interactions using DNA barcodes. *PLoS One*, *8*(1), e52967. https://doi.org/10.1371/journal.pone.0052967

Giam, X., Scheffers, B. R., Sodhi, N. S., Wilcove, D. S., Ceballos, G., & Ehrlich, P. R. (2012). Reservoirs of richness: Least disturbed tropical forests are centres of undescribed species diversity. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 67–76. https://doi.org/10.1098/rspb.2011.0433

Gong, L., Qiu, X. H., Huang, J., Xu, W., Bai, J. Q., Zhang, J., Su, H., Xu, C. M., & Huang, Z. H. (2018). Constructing a DNA barcode reference library for southern herbs in China: A resource for authentication of southern Chinese medicine. *PLoS One*, *13*, e0201240. https://doi.org/10.1371/journal.pone.0201240

Hao, M. D., Jin, Q., Meng, G. L., Yang, C. Q., Yang, S. Z., Shi, Z. Y., Tang, M., Liu, S. L., Li, Y. N., Zhang, D., Su, X., Shih, C., Sun, Y. R., Zhou, X., & Zhang, A. B. (2020). Regional assemblages shaped by diverse historical and contemporary factors: Evidence from a species-rich insect group. *Molecular Ecology*, *29*, 2492–2510. https://doi.org/10.1111/mec.15412

Haraldsson, M., & Thébault, E. (2023). Emerging niche clustering results from both competition and predation. *Ecology Letters*, *26*, 1200–1211. https://doi.org/10.1111/ele.14230

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*, 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D., & Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology*, *54*, 852–859. https://doi.org/10.1080/10635150500354886

Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding animal life: Cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, *270*, 96–99. https://doi.org/10.1098/rsbl.2003.0025

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978. https://doi.org/10.1002/joc.1276

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. John Wiley & Sons.

Hutchinson, G. E. (1957). Concluding remark. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*, 415–427. https://doi.org/10.1007/978-3-642-68836-2_20

Jin, Q., Han, H., Hu, X., Li, X., Zhu, C., Ho, S. Y., Ward, R. D., & Zhang, A. B. (2013). Quantifying species diversity with a DNA barcoding-based method: Tibetan moth species (Noctuidae) on the Qinghai-Tibetan plateau. *PLoS One*, *8*, e64428. https://doi.org/10.1371/journal.pone.0064428

Joly, S., Davies, T. J., Archambault, A., Bruneau, A., Derry, A., Kembel, S. W., Peres-Neto, P., Vamosi, J., & Wheeler, T. A. (2014). Ecology in the age of DNA barcoding: The resource, the promise and the challenges ahead. *Molecular Ecology Resources*, *14*, 221–232. https://doi.org/10.1111/1755-0998.12173

Kamo, T., Kusumoto, Y., Tokuoka, Y., Okubo, S., Hayakawa, H., Yoshiyama, M., Kimura, K., & Konuma, A. (2018). A DNA barcoding method for identifying and quantifying the composition of pollen species collected by European honeybees, *Apis mellifera* (Hymenoptera: Apidae). *Applied Entomology and Zoology*, *53*, 353–361. https://doi.org/10.1007/s13355-018-0565-9

Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., Soley-Guardia, M., & Anderson, R. P. (2021). ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution*, *12*, 1602–1608. https://doi.org/10.1111/2041-210x.13628

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*, 111–120. https://doi.org/10.1007/BF01731581

Lai, J., Zou, Y., Zhang, J., & Peres-Neto, P. R. (2022). Generalizing hierarchical and variation partitioning in multiple regression and canonical analyses using the rdacca.hp R package. *Methods in Ecology and Evolution*, *13*, 782–788. https://doi.org/10.1111/2041-210x.13800

Leaché, A. D., Koo, M. S., Spencer, C. L., Papenfuss, T. J., Fisher, R. N., & McGuire, J. A. (2009). Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 12418–19623. https://doi.org/10.1073/pnas.0906380106

Lim, G. S., Balke, M., & Meier, R. (2012). Determining species boundaries in a world full of rarity: Singletons, species delimitation methods. *Systematic Biology*, *61*, 165–169. https://doi.org/10.1093/sysbio/syr030

Liu, L., & Yu, L. L. (2010). Phybase: An R package for species tree analysis. *Bioinformatics*, *26*, 962–963. https://doi.org/10.1093/bioinformatics/btq062

Liu, Q., Charleston, M. A., Richards, S. A., Holland, B. R., & Jermiin, L. (2023). Performance of Akaike information criterion and Bayesian information criterion in selecting partition models and mixture models. *Systematic Biology*, *72*, 92–105. https://doi.org/10.1093/sysbio/syac081

Lu, M., Winner, K., & Jetz, W. (2021). A unifying framework for quantifying and comparing n-dimensional hypervolumes. *Methods in Ecology and Evolution*, *12*, 1953–1968. https://doi.org/10.1111/2041-210x.13665

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*, 209–220. https://doi.org/10.1007/s00253-002-1013-9

McGuire, J. A., Linkem, C. W., Koo, M. S., Hutchison, D. W., Lappin, A. K., Orange, D. I., Lemos-Espinal, J., Riddle, B. R., & Jaeger, J. R. (2007). Mitochondrial introgression and incomplete lineage sorting through space and time: Phylogenetics of crotaphytid lizards. *Evolution*, *61*, 2879–2897. https://doi.org/10.1111/j.1558-5646.2007.00239.x

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, *55*, 715–728. https://doi.org/10.1080/10635150600969864

Meier, R., Zhang, G., & Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the 'barcoding gap' and leads to misidentification. *Systematic Biology*, *57*, 809–813. https://doi.org/10.1080/10635150802406343

Miller, R. G. (1981). *Simultaneous statistical inference*. Springer.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., & Worm, B. (2011). How many species are there on Earth and in the ocean? *PLoS Biology*, *9*, e1001127. https://doi.org/10.1371/journal.pbio.1001127

Moritz, C., & Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLoS Biology*, *2*, 1529–1531. https://doi.org/10.1371/journal.pbio.0020354

Murphy, S. J., & Smith, A. B. (2021). What can community ecologists learn from species distribution models? *Ecosphere*, *12*, e03864. https://doi.org/10.1002/ecs2.3864

Mutanen, M., Kivela, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., Huemer, P., Dinca, V., van Nieukerken, E. J., Lopez-Vaamonde, C., Vila, R., Aarvik, L., Decaens, T., Efetov, K. A., Hebert, P. D., Johnsen, A., Karsholt, O., Pentinsaari, M., Rougerie, R., … Godfray, H. C. J. (2016). Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology*, *65*, 1024–1040. https://doi.org/10.1093/sysbio/syw044

Nielsen, R., & Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology*, *55*, 162–169. https://doi.org/10.1080/10635150500431239

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *vegan: Community ecology package*. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

Orr, M. C., Ascher, J. S., Bai, M., Chesters, D., & Zhu, C. D. (2020). Three questions: How can taxonomists survive and thrive worldwide? *Megataxa*, *1*, 19–27. https://doi.org/10.11646/megataxa.1.1.4

Orr, M. C., Feijó, A., Chesters, D., Vogler, A. P., Bossert, S., Ferrari, R. R., Costello, M. J., Hughes, A. C., Krogmann, L., Ascher, J. S., Zhou, X., Li, D.-Z., Bai, M., Chen, J., Ge, D., Luo, A., Qiao, G., Williams, P. H., Zhang, A.-B., … Zhu, C.-D. (2022). Six steps for building a technological knowledge base for future taxonomic work. *National Science Review*, *9*, nwac284. https://doi.org/10.1093/nsr/nwac284

Orr, M. C., Koch, J. B., Griswold, T. L., & Pitts, J. P. (2014). Taxonomic utility of niche models in validating species concepts: A case study in *Anthophora* (*Heliophila*) (Hymenoptera: Apidae). *Zootaxa*, *3846*, 411–429. https://doi.org/10.11646/zootaxa.3846.3.5

Osorio-Olvera, L., Lira-Noriega, A., Soberón, J., Peterson, A. T., Falconi, M., Contreras-Díaz, R. G., Martínez-Meyer, E., Barve, V., Barve, N., & Qiao, H. (2020). NTBOX: An R package with graphical user interface for modelling and evaluating multidimensional ecological niches. *Methods in Ecology and Evolution*, *11*, 1199–1206. https://doi.org/10.1111/2041-210x.13452

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Pejovic, I., Ardura, A., Miralles, L., Arias, A., Borrell, Y. J., & Garcia-Vazquez, E. (2016). DNA barcoding for assessment of exotic molluscs associated with maritime ports in northern Iberia. *Marine Biology Research*, *12*, 168–176. https://doi.org/10.1080/17451000.2015.1112016

Peterson, A. T. (2011). Ecological niche conservatism: A time-structured review of evidence. *Journal of Biogeography*, *38*, 817–827. https://doi.org/10.1111/j.1365-2699.2010.02456.x

Pfenninger, M., Nowak, C., Kley, C., Steinke, D., & Streit, B. (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic *Chironomus* (Diptera) species. *Molecular Ecology*, *16*, 1957–1968. https://doi.org/10.1111/j.1365-294X.2006.03136.x

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Pichler, M., & Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, *12*, 2159–2173. https://doi.org/10.1111/2041-210x.13687

Popescu, A. A., Huber, K. T., & Paradis, E. (2012). ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics*, *28*, 1536–1537. https://doi.org/10.1093/bioinformatics/bts184

Porco, D., Decaëns, T., Deharveng, L., James, S. W., Skarżyński, D., Erséus, C., Butt, K. R., Richard, B., & Hebert, P. D. N. (2012). Biological invasions in soil: DNA barcoding as a monitoring tool in a multiple taxa survey targeting European earthworms and springtails in North America. *Biological Invasions*, *15*, 899–910. https://doi.org/10.1007/s10530-012-0338-2

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters*, *3*, 349–361. https://doi.org/10.1046/j.1461-0248.2000.00143.x

Qiao, H., Peterson, A. T., Campbell, L. P., Soberón, J., Ji, L., & Escobar, L. E. (2016). NicheA: Creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, *39*, 805–813. https://doi.org/10.1111/ecog.01961

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

Raxworthy, C. J., Ingram, C. M., Rabibisoa, N., & Pearson, R. G. (2007). Applications of ecological niche modeling for species delimitation: A review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. *Systematic Biology*, *56*, 907–923. https://doi.org/10.1080/10635150701775111

Rissler, L. J., & Apodaca, J. J. (2007). Adding more ecology into species delimitation: Ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*). *Systematic Biology*, *56*, 924–942. https://doi.org/10.1080/10635150701703063

Roe, A. D., & Sperling, F. A. H. (2007). Patterns of evolution of mitochondrial cytochrome coxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution*, *44*, 325–345. https://doi.org/10.1016/j.ympev.2006.12.005

Ross, H. A. (2014). The incidence of species-level paraphyly in animals: A re-assessment. *Molecular Phylogenetics and Evolution*, *76*, 10–17. https://doi.org/10.1016/j.ympev.2014.02.021

Rubinoff, D., & Holland, B. S. (2005). Between two extremes: Mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology*, *54*, 952–961. https://doi.org/10.1080/10635150500234674

Ruiz-Sanchez, E., & Sosa, V. (2010). Delimiting species boundaries within the neotropical bamboo *Otatea* (Poaceae: Bambusoideae) using molecular, morphological and ecological data. *Molecular Phylogenetics and Evolution*, *54*, 344–356. https://doi.org/10.1016/j.ympev.2009.10.035

Santos, A. M., Besnard, G., & Quicke, D. L. (2011). Applying DNA barcoding for the study of geographical variation in host-parasitoid interactions. *Molecular Ecology Resources*, *11*, 46–59. https://doi.org/10.1111/j.1755-0998.2010.02889.x

Scriven, J. J., Whitehorn, P. R., Goulson, D., & Tinsley, M. C. (2016). Niche partitioning in a sympatric cryptic species complex. *Ecology and Evolution*, *6*, 132812–132839. https://doi.org/10.1002/ece3.1965

Shi, Z. Y., Yang, C. Q., Hao, M. D., Wang, X. Y., Ward, R. D., & Zhang, A. B. (2018). FuzzyID2: A software package for large data set species identification via barcoding and metabarcoding using hidden Markov models and fuzzy set methods. *Molecular Ecology Resources*, *18*, 666–675. https://doi.org/10.1111/1755-0998.12738

Soberón, J., & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(Suppl 2), 19644–19650. https://doi.org/10.1073/pnas.0901637106

Spiers, A. I., Royle, J. A., Torrens, C. L., & Joseph, M. B. (2022). Estimating species misclassification with occupancy dynamics and encounter rates: A semi-supervised, individual-level approach. *Methods in Ecology and Evolution*, *13*, 1528–1539. https://doi.org/10.1111/2041-210x.13858

Sultana, S., Ali, M. E., Hossain, M. A. M., Asing, A., Naquiah, N., & Zaidul, I. S. M. (2018). Universal mini COI barcode for the identification of fish species in processed products. *Food Research International*, *105*, 19–28. https://doi.org/10.1016/j.foodres.2017.10.065

Talavera, G., Dinca, V., & Vila, R. (2013). Factors affecting species delimitations with the GMYC model: Insights from a butterfly survey. *Methods in Ecology and Evolution*, *4*, 1101–1110. https://doi.org/10.1111/2041-210X.12107

Taylor, H. R., & Harris, W. E. (2012). An emergent science on the brink of irrelevance: A review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, *12*, 377–388. https://doi.org/10.1111/j.1755-0998.2012.03119.x

Title, P. O., & Bemmels, J. B. (2018). ENVIREM: An expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling. *Ecography*, *41*, 291–307. https://doi.org/10.1111/ecog.02880

van der Veen, B., Hui, F. K. C., Hovstad, K. A., Solbu, E. B., & O'Hara, R. B. (2021). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, *12*, 1288–1300. https://doi.org/10.1111/2041-210x.13595

Wallis, G. P., Cameron-Christie, S. R., Kennedy, H. L., Palmer, G., Sanders, T. R., & Winter, D. J. (2017). Interspecific hybridization causes long-term phylogenetic discordance between nuclear and mitochondrial genomes in freshwater fishes. *Molecular Ecology*, *26*, 3116–3127. https://doi.org/10.1111/mec.14096

Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist? - A case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, *4*, 8. https://doi.org/10.1186/1742-9994-4-8

Wijayathilaka, N., Senevirathne, G., Bandara, C., Rajapakse, S., Pethiyagoda, R., & Meegaskumbura, M. (2018). Integrating bioacoustics, DNA barcoding and niche modeling for frog conservation—The threatened balloon frogs of Sri Lanka. *Global Ecology and Conservation*, *16*, e00496. https://doi.org/10.1016/j.gecco.2018.e00496

Will, K. W., Mishler, B. D., & Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, *54*(5), 844–851. https://doi.org/10.1080/10635150500354878

Wilson, J. S., Clark, S. L., Williams, K. A., & Pitts, J. P. (2012). Historical biogeography of the arid-adapted velvet ant *Sphaeropthalma arota* (Hymenoptera: Mutillidae) reveals cryptic species. *Journal of Biogeography*, *39*, 336–352. https://doi.org/10.1111/j.1365-2699.2011.02580.x

Wright, W. J., Irvine, K. M., Almberg, E. S., Litt, A. R., & Yoccoz, N. (2019). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, *11*, 71–81. https://doi.org/10.1111/2041-210x.13315

Wu, Y. H., Hou, S. B., Yuan, Z. Y., Jiang, K., Huang, R. Y., Wang, K., Liu, Q., Yu, Z. B., Zhao, H. P., Zhang, B. L., Chen, J. M., Wang, L. J., Stuart, B. L., Chambers, E. A., Wang, Y. F., Gao, W., Zou, D. H., Yan, F., Zhao, G. G., … Che, J. (2023). DNA barcoding of Chinese snakes reveals hidden diversity and conservation needs. *Molecular Ecology Resources*, *23*, 1124–1141. https://doi.org/10.1111/1755-0998.13784

Xu, S. Z., Li, Z. Y., & Jin, X. H. (2017). DNA barcoding of invasive plants in China: A resource for identifying invasive plants. *Molecular Ecology Resources*, *18*, 128–136. https://doi.org/10.1111/1755-0998.12715

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., Veran, S., & O'Hara, R. B. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*, 236–243. https://doi.org/10.1111/2041-210x.12004

Yandell, B. S. (1997). *Practical data analysis for designed experiments*. Chapman & Hall.

Yang, B., Zhang, Z. X., Yang, C. Q., Wang, Y., Orr, M. C., Wang, H. B., & 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, C. Q., Lv, Q., & Zhang, A. B. (2020). Sixteen years of DNA barcoding in China: What has been done? What can be done? *Frontiers in Ecology and Evolution*, *8*, 57. https://doi.org/10.3389/fevo.2020.00057

Yang, C. Q., Wang, Y., Li, X. H., Li, J., Yang, B., Orr, M. C., & Zhang, A. B. (2024a). *NicheBarcoding: Niche-model-based species identification*. R package version 1.8. https://CRAN.R-project.org/package=NicheBarcoding

Yang, C. Q., Wang, Y., Li, X. H., Li, J., Yang, B., Orr, M. C., & Zhang, A. B. (2024b). NicheBarcoding (Version v0.0.1.8000). *Zenodo*, https://doi.org/10.5281/zenodo.13270695

Yang, C. Q., Wang, Y., Li, X. H., Li, J., Yang, B., Orr, M. C., & Zhang, A. B. (2024c). Data from: Environmental niche models improve species identification in DNA barcoding. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.rbnzs7h96

Yang, Z. H. (2006). *Computational molecular evolution*. Oxford University Press.

Zhang, A. B., Hao, M. D., Yang, C. Q., & Shi, Z. Y. (2017). BarcodingR: An integrated R package for species identification using DNA barcodes. *Methods in Ecology and Evolution*, *8*, 627–634. https://doi.org/10.1111/2041-210X.12682

Zhang, A. B., Muster, C., Liang, H. B., Zhu, C. D., Crozier, R., Wan, P., Feng, J., & Ward, R. D. (2012). A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Molecular Ecology*, *21*, 1848–1863. https://doi.org/10.1111/j.1365-294X.2011.05235.x

Zhang, A. B., Sikes, D. S., Muster, C., & Li, S. Q. (2008). Inferring species membership using DNA sequences with back-propagation neural networks. *Systematic Biology*, *57*, 202–215. https://doi.org/10.1080/10635150802032982

Zhang, X. M., Shi, Z. Y., Zhang, S. Q., Zhang, P., Wilson, J. J., Shih, C., Li, J., Li, X. D., Yu, G. Y., & Zhang, A. B. (2020). Plant-herbivorous insect networks: Who is eating what revealed by long barcodes using high-throughput sequencing and Trinity assembly. *Insect Science*, *28*, 127–143. https://doi.org/10.1111/1744-7917.12749

SUPPORTING INFORMATION

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

Table S1: The notation table of the parameters involved in the NBSI framework.

Table S2: The detailed statistics of tests of simulated datasets.

Table S3: The detailed statistics of tests of empirical datasets.

Appendix S1: All generated simulated datasets.

Appendix S2: All collected empirical datasets.

Appendix S3: The variables retained for niche modelling in every single test of simulated datasets.

Appendix S4: The variables retained for niche modelling in every single test of empirical datasets.

How to cite this article: Yang, C.-q., Wang, Y., Li, X.-h., Li, J., Yang, B., Orr, M. C., & Zhang, A.-b. (2024). Environmental niche models improve species identification in DNA barcoding. *Methods in Ecology and Evolution*, *15*, 2343–2358. https://doi.org/10.1111/2041-210X.14440