TAXOSET: OPEN-REFERENCE TAXONOMIC ASSIGNMENT OF DNA BARCODES

Anonymous authors
Paper under double-blind review

ABSTRACT

A classifier can only return a name from its label space, and in DNA barcoding that space is the sequence reference, not the whole taxonomy. Our four curated corpora carry a barcode for 11.3% of the species it names. For a query from any of the other 88.7%, a method that can only return reference names has an accuracy of exactly zero, whatever the model. We introduce open-reference taxonomic assignment. The reference supplies the evidence, a query's nearest sequences and their lineages. The taxonomic hierarchy, 2.9 million taxa at all ranks, supplies the candidate names, sequenced or not. When the evidence locates a branch but cannot separate its unsequenced members, choosing one would invent information, so the method returns a set of names, whose coverage we never report without its size. TAXOSET implements this with a 0.34M-parameter network that learns how to decide, not what the taxa are, so it applies to names it never saw in training. Removing whole clades from the reference before retrieval makes novelty a label rather than an estimate. On ITS queries whose genus is absent, the returned set holds the true genus 69% of the time at a median of 11 names. At species rank it falls to 28%, where the limit is detecting novelty rather than the set itself. On a third-party benchmark whose novelty is real rather than simulated, TAXOSET reaches 82% on ITS, where no published method returns the absent genus at all.

1 INTRODUCTION

When a task admits labels a model was never trained on, any single prediction it makes for a query carrying one of them is wrong by construction. Open-set recognition answers by rejecting the query (Scheirer et al., 2013); a rejection does not say which labels remain. DNA barcoding is a case where the remaining labels can be listed: a named hierarchy enumerates every admissible name, so what a rejection withholds can be returned and measured. Barcoding identifies an organism from a short genomic region (Hebert et al., 2003); metabarcoding extends it to mixed environmental samples, surveying whole communities at once (Taberlet et al., 2012; Deiner et al., 2017). Taxonomic assignment gives a query sequence a name at a chosen rank, such as species, genus, or family.

Bioinformatic approaches for taxonomic annotation of DNA barcodes are based on sequence similarity against annotated sequences in reference molecular databases. The search returns the most similar reference sequences, called *hits*, and their lineages are what the method reads. But the sequence reference and the taxonomy are different resources. For one marker, let R be the reference sequences with known lineages, and let T be the named taxonomic hierarchy, built from the Catalogue of Life and, for Bacteria, the Genome Taxonomy Database (Bánki et al., 2026; Parks et al., 2022). Unlike R, T also holds taxa for which no sequence of that marker is available; we call such a taxon *unsequenced* for that marker. T lists about 2.4 million species, and our four curated corpora for ITS, 12S, 16S and COI carry barcodes for 11.3% of them.

This creates a structural limit. If the true taxon is absent from R, a method that can only return reference names cannot predict it, whatever the model. We call such a method *reference-bound*, and every family in use is one. Alignment methods copy the name of a close reference sequence (Altschul et al., 1990; Rognes et al., 2016); supervised classifiers choose among the reference labels (Wang et al., 2007; Miranda et al., 2024); recent deep models improve the sequence representation without changing the output label (Millan Arias et al., 2026; Gao & Taylor, 2024). Methods that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

model novelty place probability on taxa unobserved in training, but report them as new branches rather than as names drawn from a taxonomy (Zito et al., 2023).

Consider a family F with three genera, G_1, G_2, and G_3. The reference contains barcodes for G_1 but none for G_2 or G_3. A query from G_2 retrieves hits from G_1. Those hits identify the correct family F, but not the true genus. The hierarchy makes G_2 and G_3 available as candidates, yet the sequence evidence cannot distinguish them. Choosing one would add information the data does not support. The appropriate output is therefore the set $\{G_2, G_3\}$.

We call this task *open-reference taxonomic assignment*. A query is novel at rank r when its true rank-r taxon is named in T but absent from the marker-specific reference R. Novel therefore means absent from the reference, not unknown to science. PROTAX places a query on a taxonomy that reference sequences only partly populate, giving every branch a probability, unsequenced branches included (Somervuo et al., 2016). We ask a complementary question: what should a system return when several of those unsequenced taxa remain consistent with the evidence? Our answer is a prediction set, ordered only as far as the evidence orders it, and evaluated by both its coverage and its size.

We introduce TAXOSET, which takes retrieved sequences as evidence and the hierarchy as its name space. The candidates are assembled for each query rather than fixed at training time, so its size does not depend on $|T|$ and it can score names never seen during training. What it learns is how to pool the evidence over those candidates; the decision to leave the reference is calibrated afterwards, at a held false-positive rate.

Our contributions are:

1. **Open-reference formulation.** A classifier's label space is fixed at training time. Ours is built for each query, from retrieval and the hierarchy, and assignment becomes set-valued prediction (Sadinle et al., 2019) over it. Indistinguishable taxa are scored through their parent, and their number corrects the ranking of candidates.
2. **Clade exclusion moved into training.** Clades are removed to test a classifier on absent taxa; we remove them from the reference before retrieval, across four markers, and train on the result, so that the rank at which a query becomes novel is a label the network learns from rather than an estimate.
3. **A compact, transferable method.** TAXOSET places a query with a 0.34M-parameter DeepSets network (Zaheer et al., 2017). We test it against a rolled-back reference and a third-party benchmark, and show that the same weights transfer across markers and retrieval engines. We also report the settings in which it does not improve on unlearned rules given the same hits and hierarchy.

2 PROBLEM SETTING: ASSIGNMENT BEYOND THE REFERENCE

We write ℓ for the map from a sequence to its lineage, so that $\ell(R)$ is the set of reference-backed taxa and $\ell(R) \subsetneq T$. T has seven ranks: kingdom, phylum, class, order, family, genus and species. We evaluate assignment at the three deepest (family, genus and species), where reference coverage is sparse enough for novelty to arise; the upper ranks enter only through lineages, which supply the candidate parents. We write E for the evidence a query retrieves, its hits in the reference sequence database and their lineages, and call it the query's episode.

The system decides at each evaluated rank, independently of the others. It returns either a reference-backed name or a set of hierarchy-backed names, and a novelty decision chooses between the two. The evidence ranks that set down to groups of names it cannot separate, and no further (Equation 1). The set is empty only when retrieval and the hierarchy expose no candidate at all; the system then returns nothing, which we call an abstention.

A reference-bound method cannot return a taxon outside $\ell(R)$, whatever the retrieval evidence or the model. We call this a *structural zero*: it follows from the label space alone. Four held-out query groups separate the cases: known species, whose species are present in R, and queries whose true taxon is first unsequenced at species, genus, or family rank. For a query novel at rank r, the ranks above r remain reference-supported, whereas rank r does not.

Why the open-world output is a set. At a given rank, let U_g be the unsequenced children of a candidate parent g. Permuting the members of U_g leaves E unchanged, so the division of $\pi(g \mid E)$, the mass the model gives to the unsequenced children of g taken together, is underdetermined by the evidence. Among all divisions of that mass, the uniform one is the unique maximum-entropy solution, and we adopt it:

$$\Pr[s^* = k \mid E] = \frac{\pi(g_k \mid E)}{|U_{g_k}|}, \tag{1}$$

where s^* is the true taxon of the query and g_k is the parent of candidate k.

This concerns the information reaching the model, not the organisms. Other variables, such as geography, ecology or another marker, could separate two siblings. The placement network receives none of them, only similarity metrics at the DNA sequence level, so the uniform allocation is what its architecture produces. Two sequenced siblings can be just as indistinguishable, through hybridization or shared ancestry (Section 6). We still return one name for them, because their sequences are in the reference and the evidence is allowed to choose between them; the set is for the case where no sequence exists to choose from.

Equation 1 has three consequences. Even with the correct parent, a single-name prediction succeeds with probability at most $1/|U_g|$. If only b names can be returned, the best list holds the b candidates with the largest $\pi(g_k \mid E)/|U_{g_k}|$, which in log space means subtracting $\log |U_{g_k}|$ from the parent score. And returning every unsequenced child of the correct parent contains the true taxon whenever parent placement is correct.

Appendix B.3 tests whether the evidence supports a stable sibling ordering after all. It finds no consistent signal in eight of the nine marker–rank settings, so we keep this conservative rule everywhere.

3 METHOD: SCORING AND SET-VALUED OUTPUT

Figure 1 presents the method at training and inference time. Both stages use the same retrieval and scoring procedure, but training uses a reduced reference to simulate missing taxa. The method then combines the retrieved sequences with the taxonomic hierarchy to propose names absent from the reference. We call the whole approach, retrieval and placement network together, TAXOSET.

3.1 FROM HITS TO TAXONOMIC CANDIDATES

For each query, we use VSEARCH (Rognes et al., 2016) to retrieve the 20 most similar reference sequences $r_1, \ldots, r_{20}$. Each hit r_j provides its lineage $\ell(r_j)$ and an evidence vector e_j containing its alignment identity and within-episode z-score. A 97% match is more informative when the other hits are close to 80% than when they are all close to 96%. The z-score captures this difference and also makes the evidence more comparable across markers. Additional features, such as alignment coverage or embedding similarity, can also be included, as shown in Section 4.4.

Scoring all 2,907,182 names in T for each query would be both inefficient and uninformative: the episode contains no evidence for names not exposed by a hit's parent. We instead build a candidate set on the fly. At each rank, it contains the taxa represented by the retrieved sequences, at most twenty, and their parents' unsequenced children.

3.2 SCORING, ROUTING AND PREDICTION SETS

Equation 1 determines how unsequenced taxa are scored: the parent carries the score, divided among its unsequenced children. Scoring them separately would only let the network read the order in which the hierarchy lists them.

The network uses only the retrieved hits and should not depend on their order. We therefore use a DeepSets encoder (Zaheer et al., 2017). Each hit is encoded separately, mean pooling over these representations summarises the episode, and a second encoder maps each hit together with that summary. At each taxonomic rank, the network then produces scores for the hits, the candidate

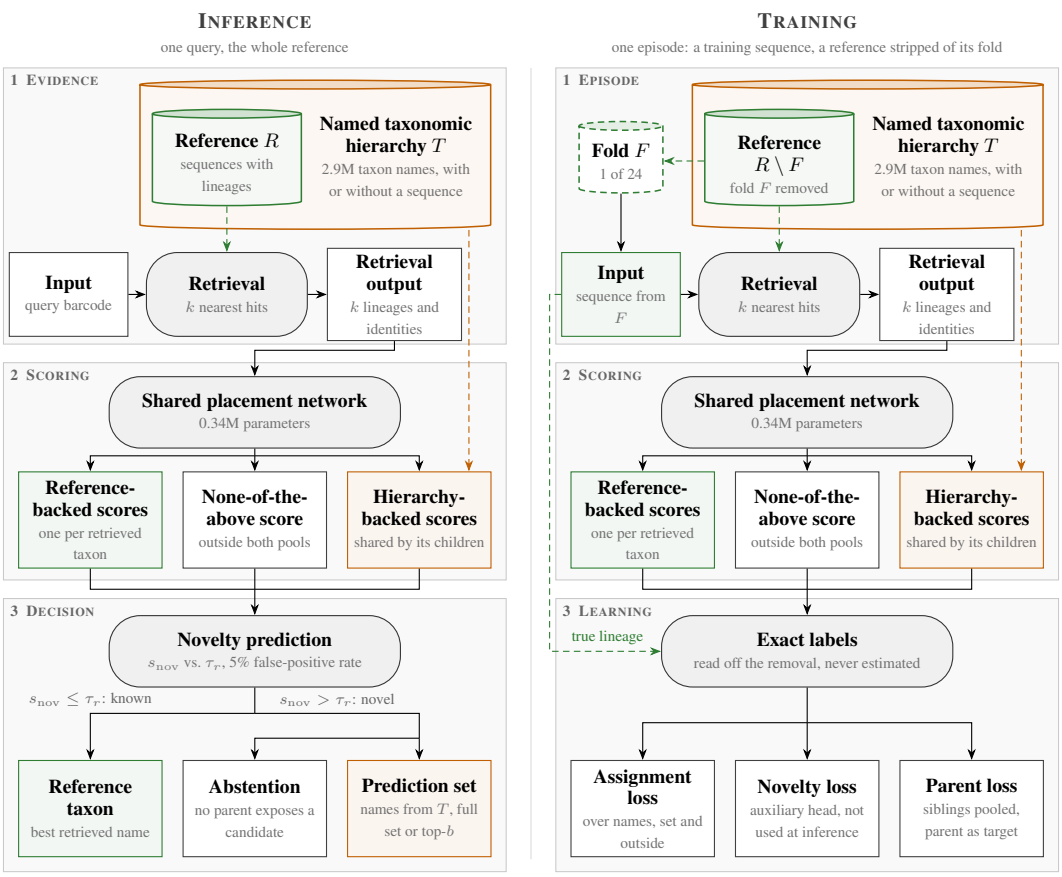


Figure 1: **The same pipeline at inference and in training.** Ovals are computations, rectangles are data, cylinders are databases. Green marks what comes from the sequenced reference R, orange what comes from the named hierarchy T; R is drawn inside T because it covers a small part of it. **Left**, inference: at each evaluated rank the two candidate pools are scored together with a none-of-the-above option, and a calibrated novelty score chooses between a reference-backed name and a set of hierarchy-backed names, which collapses to an abstention when no retrieved parent exposes a candidate. **Right**, training: the first two bands are unchanged, but retrieval searches a reference stripped of one of 24 clade folds, so the query's taxon is named in T and absent from $R \setminus F$, which is what makes every open-world label exact rather than estimated.

parents, and a none-of-the-above option. It scores the candidates an episode supplies rather than a fixed set of classes, so its size is independent of $|T|$ and it can score taxa that were never seen during training. An attention encoder with eleven times more parameters performs better in only one of nine settings (Appendix B, Table B.1). The full architecture is shown in Figure B.1.

For each reference-backed taxon, the network combines the score of its best hit with the agreement among its hits, in a proportion it sets for each episode. A softmax then compares the reference-backed taxa, the hierarchy-backed set taken as a single option, and none-of-the-above. The probability assigned to the last two forms the novelty score. The same softmax therefore ranks the candidate taxa and determines whether the query should be considered novel.

The novelty threshold is calibrated on held-out known queries so that only 5% of them are incorrectly classified as novel. Below this threshold, the method returns the highest-scoring reference-backed taxon. Above it, the method returns the hierarchy-backed taxa, ranked according to Equation 1. It abstains when no such candidate is available. We evaluate both the complete set and a list limited to $b = 10$ names, always reporting their size.

The example introduced earlier also occurs in the data. An ITS query from the unsequenced genus *Rhynchosporium* retrieves hits from two families. The hierarchy then exposes six unsequenced genera, including *Rhynchosporium*.

3.3 Training without taxonomic leakage

To recognize taxa that are absent from the reference, the network must encounter this situation during training. We reproduce it by temporarily removing clades from the reference R while keeping their names in the hierarchy T. At each evaluated rank, the clades are divided into 24 folds. The sequences from a fold F are used as queries and searched against $R \setminus F$, ensuring that their taxa are absent from the reference. To also learn the known case, one quarter of the episodes use the reference without removing a fold.

The loss contains three terms at each rank. The first is a cross-entropy over the same three options the softmax compares. The second learns to identify the correct parent when the true taxon is absent from the reference, following Equation 1. The third trains an auxiliary head to detect novelty. This head supports training but is not used at inference time.

The clades used for evaluation are removed before constructing the reference, training data, and calibration data. The model therefore never sees a sequence of the evaluation taxa during training; only their names remain, in the hierarchy. We verify for every marker that no evaluation query occurs in the reference or in training, and that its taxon at the evaluated rank is absent from the reference lineages. We use no sequence augmentation and apply the same hyperparameters to all four markers. Optimization details are provided in Appendix B.

4 Experiments

4.1 Data and evaluation design

We evaluate four markers spanning organisms and rates of evolution: ITS for fungi and plants, from CurateMake (Gardette et al., 2026); 12S for metazoans, from MIDORI2 (Leray et al., 2022); 16S for bacteria, from GTDB (Parks et al., 2022); and COI for insects, from BIOSCAN-1M (Gharaee et al., 2023) (Table 1, Appendix A). For each marker, queries fall into the four groups of Section 2: known species, and taxa first unsequenced at species, genus or family. Withholding a clade from the reference to measure performance on absent taxa is clade exclusion; Wood & Salzberg (2014) named it. Benchmarks since have measured how classifiers handle organisms absent from the reference (Ye et al., 2019); defining novelty by the rank at which a taxon is first absent follows Orsholm et al. (2026). We add that four-group split and the use of the same removals to build training episodes.

Table 1: Reference data after curation, and the evaluation groups. Query groups hold known species, or taxa first unsequenced at species, genus or family rank. Counts are fixed across seeds; a dash means the corpus provides no usable known-species split.

		Reference used for retrieval				Evaluation queries			
Marker	Source	Sequences	Species	Genera	Families	Known sp.	Novel sp.	Novel gen.	Novel fam.
ITS	CurateMake	970,750	89,645	12,589	1,880	14,999	14,999	14,974	14,985
12S	MIDORI2	94,085	43,389	17,934	3,811	14,998	14,998	14,999	14,991
16S	GTDB SSU	32,658	32,658	11,113	3,253	–	6,496	8,827	10,120
COI	BIOSCAN-1M	44,658	5,427	1,827	280	8,099	7,232	7,883	15,000

At reference-supported ranks we report top-1 accuracy with 95% confidence intervals over three seeds. At novel ranks we report how often a novel query is routed at the threshold of Section 3.2, the coverage of the returned set among routed queries, and *end-to-end coverage*, the fraction of all queries that are both routed and covered. Every coverage figure comes with a set size, since returning more names raises coverage on its own. Intervals at novel ranks are bootstrapped over parent clades rather than queries, because barcodes from one clade are not independent.

Best-hit, the vote rule and a top-20 adaptation of PROTAX's decision rule receive exactly the same hits as TaxoSet. PROTAX is our main learned baseline, and we also retrain its released PROTAX-GPU engine on our reference and splits (Appendix C.2). Broader controls come from SINTAX

(Edgar, 2016), HiTaC (Miranda et al., 2024) and BayesANT (Zito et al., 2023) (Appendix C.1), and from a 500M-parameter nucleotide transformer (Appendix C.3). Four ablations then remove one ingredient at a time and leave the rest untouched: the learning (Appendix D.7), the set encoder and the sibling-size correction (Appendix B), and the retriever (Appendix F).

4.2 Can learned placement improve reference-backed assignment?

We start where the correct taxon is present in the reference, which measures how the network reads its hits without the open-reference output. Table 2 covers known species and the reference-supported ranks of novel-species and novel-genus queries.

Table 2: Reference-backed assignment using identical retrieved hits (top-1 mean $\pm$ CI_{95} over three seeds). "Known species" is the standard closed-reference task; the other rows ask for the closest higher rank that is still represented in the reference. Scores are shown without the novelty gate to isolate taxonomic placement. The PROTAX column is the top-20-restricted adaptation of Appendix C.2, on the same hits; its seed spread is below 0.001 everywhere. Bold marks the best score in each row.

Marker	Query group $\rightarrow$ rank	TaxoSet	best-hit	vote	PROTAX adapt.
ITS	known sp. $\rightarrow$ species	**0.752±0.001**	0.731	0.571	0.659
ITS	novel sp. $\rightarrow$ genus	**0.859±0.000**	0.837	0.823	0.803
ITS	novel gen. $\rightarrow$ family	**0.611±0.002**	0.527	0.605	0.507
16S	novel sp. $\rightarrow$ genus	0.798±0.000	**0.809**	0.479	0.748
16S	novel gen. $\rightarrow$ family	**0.784±0.005**	0.740	0.758	0.690
COI	known sp. $\rightarrow$ species	0.988±0.000	**0.990**	0.786	0.968
COI	novel sp. $\rightarrow$ genus	**0.620±0.002**	0.615	0.410	0.566
COI	novel gen. $\rightarrow$ family	**0.603±0.004**	0.597	0.512	0.524
12S	known sp. $\rightarrow$ species	0.778±0.001	**0.788**	0.415	0.701
12S	novel sp. $\rightarrow$ genus	0.803±0.001	**0.811**	0.542	0.741
12S	novel gen. $\rightarrow$ family	**0.730±0.001**	0.715	0.629	0.660

On average, TaxoSet is above both best-hit and vote in seven of the eleven rows. For known species it improves ITS from 0.731 to 0.752, while best-hit stays slightly ahead on COI (0.990 against 0.988) and 12S (0.788 against 0.778). The gains are larger at higher ranks, for example ITS family (0.611 against 0.527) and 16S family (0.784 against 0.740). The network therefore benefits when several hits agree, while best-hit stays strong when its first match is already reliable. The PROTAX adaptation, on the same hits, remains below both in all eleven rows.

4.3 Can the method return names absent from the reference R?

Figure 2 shows that, for ITS novel genera, the full set reaches 0.694 ± 0.006 coverage with a median of 11 names. Limiting the set to ten names gives 0.568. Full-set coverage reaches 0.639 on 12S and 0.709 on COI novel genera, but the median COI set contains 1,493 names and is therefore impractical. Table D.1 reports routing, conditional coverage and set-size quantiles, and Appendix D the budgets of five and twenty names.

To isolate the benefit of learning, we compare the network with best-hit expansion, round-robin expansion and sibling-adjusted vote on the same hits and hierarchy (Appendix D.7). Across paired clade-blocked comparisons the network has 18 wins, 3 losses and 6 undecided cases. 16 wins survive Benjamini–Hochberg correction and 12 survive Bonferroni. All three losses occur for 12S novel species (Table D.7; the resampling scheme and per-taxon accuracies are in Appendix D.8). Learning therefore improves routing and ranking in most settings, not all. These baselines isolate learning as the only changed component: they retain the same retrieval, hierarchy, and sibling correction, while replacing the learned routing and ranking with fixed rules.

The margin does not come from the encoder either. Swapping it for the attention model of Appendix B, eleven times larger, changes each of these margins by at most 0.013, and turns a win into a loss only where the margin is already indistinguishable from zero (last column of Table D.7). On the three losses, the larger model loses by more. What learning contributes to here is the decision, not the capacity of the model that makes it.

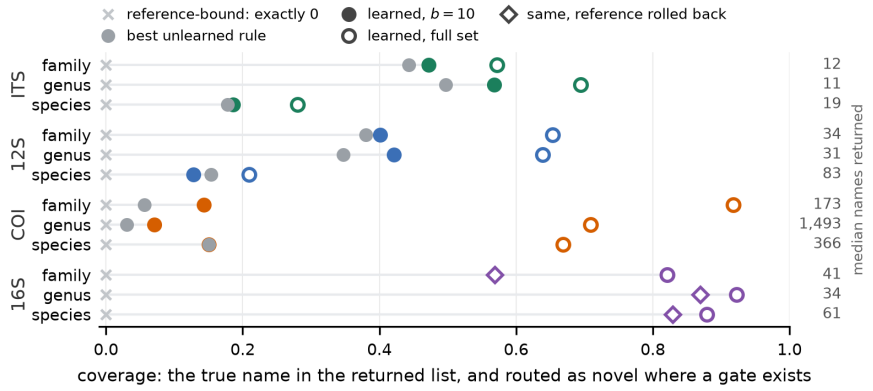


Figure 2: **Assigning taxa absent from the sequence reference.** Crosses mark the structural zero of reference-bound methods. Filled dots show TaxoSet at a budget of ten names and open circles its full prediction set; the number at right is the median full-set size. Gray dots are the best unlearned rule on the same retrieval and hierarchy. Coverage is the end-to-end event that a query is routed as novel and its true name occurs in the returned set. 16S carries open markers only: with one sequence per species it has no known query, so no threshold can be held at a false-positive target and its rows do not pay the routing cost the other three do. Its diamonds repeat the measurement against GTDB r207, where the missing taxa are historical rather than removed (Section 4.5).

Retrained on our references and splits, PROTAX-GPU is within the network's seed variation only on COI. On 12S it routes four times fewer novel-species queries; 16S has no known queries for calibration. With matched top-20 inputs, the network has higher end-to-end coverage in all nine calibrable settings; eight differences survive Bonferroni correction (Appendix C.2).

An external benchmark. To test generalization beyond our own benchmark design, we evaluated TAXOSET on the independent `taxclass` benchmark of Orsholm et al. (2026). We used its reference, queries, hierarchy, novelty definition, and published predictions, and trained TAXOSET from scratch on its training data alone. Novelty arises from a geographic shift for COI and a methodological shift for ITS, rather than from simulated taxon removal. For genera represented in the reference, TAXOSET ranks first on COI (0.792 versus 0.769) and fourth of ten on ITS (0.799 versus 0.817 for SINTAX). For genera absent from the reference, published methods recover the correct name for only four of 3,780 COI queries and none of 4,500 ITS queries, whereas TAXOSET reaches end-to-end coverage of 0.695 and 0.822, with median set sizes of 122 and 37, respectively. Most of this coverage comes from the open-reference formulation rather than the learned gate: without the gate, retrieval and the hierarchy alone reach 0.721 and 0.851 (Appendix G).

4.4 CAN THE LEARNED STRATEGY BE REUSED?

Across markers. The network reads retrieval scores and taxonomic structure, not the DNA sequence itself, so it should not be tied to one marker. We train it on three markers and test it on the fourth. Placement barely notices: with no weight changed, the transferred network stays within 0.034 of the native one in ten of the eleven cells of Table 2 and is ahead in four, gaining 0.032 on 16S novel genera. With the novelty threshold recalibrated on the target marker, end-to-end coverage stays within five points of native training in eight of nine settings (Appendix E).

Across retrieval engines. We also replace VSEARCH with top-20 retrieval on cosine similarity between embeddings from BarcodeBERT (Millan Arias et al., 2026) or the nucleotide transformer (Dalla-Torre et al., 2025), keeping the same weights and again recalibrating only the threshold. This transfers less well than the marker swap, and placement suffers where routing does not. Placement is tied to the scale it was trained on: repeated under the new engines, Table 2 loses 25 to 35 points and the network's margin over best-hit on identical hits goes with it, from seven cells of eleven to four. The routing decision is not tied to it, and coverage then follows retrieval quality. For ITS novel genera, BarcodeBERT reduces family-level Recall@20 from 0.813 to 0.484 and coverage

from 0.694 to 0.086 (Appendix F). The network can be reused with another retriever, but it cannot recover taxa that retrieval does not expose.

4.5 Temporal validation: rolling the reference back one release

Leave-clade-out simulates missing taxa. To test genuine historical absence we hold the r220 taxonomy fixed and roll the reference back to GTDB r207: the reference keeps the $33{,}982$ sequences whose species r207 already names, and the $24{,}120$ species that gained their first sequence afterwards become the queries. We train and calibrate on r207 alone. Their names are in the r220 taxonomy; their sequences are not in the historical reference.

The diamonds of Figure 2 carry this experiment, beside the circles of the controlled split, both read without a gate for the reason the caption gives. New species and genera lose about five points each. New families fall from 0.821 to 0.569 over $1{,}690$ queries, only $1{,}153$ of which have a genuinely absent family; r207 lists the remaining 537 under different names. Parent placement reaches 0.747 ± 0.001 for new species and 0.777 ± 0.001 for new genera (Appendix A.2). The mechanism survives a real historical absence, while the family result marks where controlled removal is optimistic.

5 Related work

Reference-bound assignment. Among the alignment- and composition-based methods compared in Section 4, SINTAX and IDTAXA can stop at a supported higher rank rather than commit to an unsupported fine-rank prediction (Edgar, 2016; Murali et al., 2018). Phylogenetic placement comes closest to overcoming this restriction: EPA-ng places a query on the most likely branch of a reference phylogenetic tree, locating an absent lineage near its closest represented relatives without naming it (Barbera et al., 2019). Naming requires a separate step and remains limited to names already present on the tree. On benchmarks where novelty is real rather than simulated, EPA-ng was strongest for COI and composition-based methods for ITS (Orsholm et al., 2026).

Taxonomies beyond the sequence reference. PROTAX is the closest predecessor of TaxoSet. It assigns probability to named taxa without reference sequences, and to an unknown branch for units missing from the taxonomy, from user-supplied evidence such as BLAST similarity (Somervuo et al., 2016); instantiations exist for arthropod COI (Roslin et al., 2022) and fungal ITS (Abarenkov et al., 2018), and PROTAX-GPU scales it to millions of references (Li et al., 2024). What separates it from our setting is informational, not computational. When several sibling taxa lack reference sequences, PROTAX assigns them identical probabilities and evaluates the pooled group as correct whenever it contains the true taxon. TaxoSet instead returns that group explicitly as a prediction set and reports both its coverage and its size. Each released instantiation is trained for one marker and one clade on its own taxonomy, so none predicts over the hierarchy our splits are scored against, which is why we retrain it (Appendix C.2). Nevertheless, PROTAX models two things we do not: units absent from the hierarchy, and mislabeled references. Stalder et al. (2025) combine sequences, a fish phylogeny and geographic co-occurrence to predict species absent from their sequence reference; we ask instead what retrieval and a named hierarchy alone can return.

Learned barcode representations and evaluation boundaries. BarcodeBERT, BarcodeMamba and BarcodeMAE learn representations from large barcode collections (Millan Arias et al., 2026; Gao & Taylor, 2024; Safari et al., 2025); others build on general genomic models (Dalla-Torre et al., 2025), several markers (Ye et al., 2026), or multimodal supervision (Gong et al., 2025). These supply retrieval evidence, but their reported "unseen" setting must be read carefully. Their hold-out applies to the labels, not to the self-supervised pretraining corpus, so a species absent from fine-tuning may still have been read there as unlabelled sequence; it may also be supplied as an inference-time key. Neither is the same as removing the taxon from the sequence reference, which is what our splits and our temporal test do (Appendix C.3).

Predicting classes unseen in training. When those classes are enumerated at inference, the setting is close to generalized zero-shot learning (Xian et al., 2019), except that a taxon name alone cannot distinguish unsequenced siblings, which is what motivates set-valued prediction (Sadinle et al., 2019). Open-set recognition and out-of-distribution detection formalize the rejection decision (Scheirer et al., 2013; Fang et al., 2021); the hierarchy is what turns it into a choice among named alternatives. In barcoding that decision is measurably harder than the identification it guards (Fuji-

sawa & Imai, 2026). Viewed as amortized inference, the network learns one procedure reused across episodes rather than a fixed mapping to training labels, which relates it to prior-data fitted networks (Müller et al., 2022; Hollmann et al., 2025).

6 Limitations and discussion

What learning adds. Learning decides when to use the candidate names that retrieval and the hierarchy produce, and how to rank them. Higher coverage also requires returning more names, and some prediction sets are too large to be useful.

Limits of the sequences and retrieval. Two things limit performance: what the barcode carries, and what retrieval finds. ITS copies can vary within one organism (Lücking et al., 2020), and hybridization, introgression, and incomplete lineage sorting can give different species the same barcode (Yang et al., 2017), which no model reading one barcode can separate. Informativeness is not enough: the correct taxon must also appear among the retrieved hits, so top-20 recall bounds what any method reading them can reach. For known ITS species that bound is 93.3% and the network reaches 75.2%; the 18.1 points between them are headroom, not accuracy a stronger model is guaranteed to recover, since the true taxon can be in the list without being separable. For COI the bound is 99.8% and best-hit already reaches 99.0%, leaving little room for learning. A longer search raises the bound on ITS novel genera from 0.813 to 0.863, and retraining at fifty hits recovers none of it (Appendix D.5). Methods that read the whole reference recover part of what the shortlist loses (Appendix D.6). For ITS novel genera and families, the full set contains the true taxon in 99% of episodes where retrieval exposes the correct parent (Appendix D.4).

Transfer and generalization. The marker-transfer and retriever-swap experiments keep the network weights fixed, so they mainly test whether the learned decision rule transfers. The rollback experiment is more realistic because its missing taxa were genuinely absent. More markers, clades, and historical releases are needed to establish how broadly the method generalizes.

Data and deployment. We assume that each query is a valid sequence from the target marker. In practice, metabarcoding data may contain off-target sequences, chimeras, or low-quality reads, which TaxoSet does not detect. The none-of-the-above score describes support outside the two candidate pools; it is not a contamination detector. An open-set method would reject such a read; TaxoSet gives it a name instead, so quality-control and decontamination steps of query sequences are needed before TaxoSet analyses. Reference labels may also be incorrect or outdated, and the hierarchy may contain conflicting classifications (Appendix A.3). A query mislabeling is therefore not always a model error and may require taxonomic expertise or independent data. Finally, calibration controls the overall false-positive rate but not the rate within each biological group, and stricter subgroup control rejects more useful predictions (Appendix D.3).

7 Conclusion

We introduced open-reference taxonomic assignment, where reference sequences provide the evidence and a named hierarchy provides the candidate names, for any hierarchy T with $\ell(R) \subseteq T$. We also introduced TaxoSet, a 0.34M-parameter network that scores each query's candidates through the parents its hits expose, and so reaches taxa never seen during training.

By removing entire clades before retrieval, TaxoSet knows exactly which queries belong to taxa that are absent from the reference. Across four markers, the returned set contains the true genus for 69% of ITS queries whose genus is absent, with a median of 11 names. On an independent benchmark, published methods almost never recover these absent genera.

These sets could later be refined using information that is not contained in the DNA sequence, such as geography, season, ecology, or the presence of other species, either as a filter on the returned set or as a prior on the parent before the set is formed. Separating the hierarchy of names from the sequence reference makes it possible to propose taxa without reference sequences, and a small network is enough to use them. The less of the hierarchy a reference covers, the more there is to gain, and every marker we tested is in that regime.

REPRODUCIBILITY STATEMENT

All named hierarchies are deterministic given pinned data releases (CoL COL26.3, doi:10.48580/dgwxb; GTDB r220; per-marker sources with versions and DOIs listed in the appendix); leave-clade-out episodes are regenerated by a single script from fold partitions; all results are means over three seeds with Student-t CIs; training the placement network takes $\sim$75 minutes on one A100. At inference the network processes $\sim 10^5$ queries per second on one A100 (one second per 100k reads, all three ranks and hierarchy-backed candidates included); end-to-end latency is dominated by retrieval (33 queries per second on 16 CPU threads against the ITS reference of $970{,}750$ sequences), which parallelizes trivially and shrinks further under the standard dereplication of environmental reads. The ITS reference database is itself a published, auditable artifact (Gardette et al., 2026). The supplementary material is a self-contained repository: hierarchy construction, the placement network, episode generation, the analysis scripts that rebuild every results table and both results figures from the frozen per-episode dumps (the data-inventory and temporal-design tables transcribe the pinned releases and the temporal protocol), and the raw evaluation logs and per-episode outcome vectors behind every cell of Table 2, from which that table regenerates with one command, the run logs of the classical and Bayesian baselines, `results/classical/`, whose `hitac.log` carries the fitting times quoted in Appendix C.1, and the scripts and logs of the released-PROTAX-GPU campaign (Appendix C.2), and the conversion and scoring scripts of the external benchmark beside the per-episode outcome vectors its table is built from (Appendix G). The heavy artifacts (hierarchies, dumps, evaluation hit lists, profile alignments and the placement-network weights for the three reported seeds) are deposited separately with a checksum manifest.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

We used large language models at most stages of this work, and set out the division of labor plainly. They did not choose the research question, design the method, or decide any scientific claim. They did not generate or alter experimental data or reported measurements: every value in this paper was computed by the released code on the pinned data releases.

We used them to interrogate intermediate ideas, to write and debug analysis scripts, to design checks and verify the manuscript's claims against that code, to draft and revise prose and LaTeX, and to produce reviewer-style feedback on early versions. That feedback informed our own review; it did not replace it. We read, verified and revised everything they produced, and we are responsible for the entire submission, including any text or artifact produced with their help.

REFERENCES

Kessy Abarenkov, Panu Somervuo, R. Henrik Nilsson, Paul M. Kirk, Tea Huotari, Nerea Abrego, and Otso Ovaskainen. Protax-fungi: a web-based tool for probabilistic taxonomic placement of fungal internal transcribed spacer sequences. *New Phytologist*, 220(2):517–525, 2018. doi: 10.1111/nph.15301.

Kessy Abarenkov, R. Henrik Nilsson, Karl-Henrik Larsson, et al. The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Research*, 52(D1):D791–D797, 2024. doi: 10.1093/nar/gkad1039.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/S0022-2836(05)80360-2.

Olaf Bánki, Yury Roskov, Markus Döring, et al. Catalogue of Life Checklist (COL26.3, base release). ChecklistBank, 2026. Version 2026-03-12, dataset 314455. doi: 10.48580/dgwxb.

Pierre Barbera, Alexey M. Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, 68(2):365–369, 2019. doi: 10.1093/sysbio/syy054.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025. doi: 10.1038/s41592-024-02523-z.

Kristy Deiner, Holly M. Bik, Elvira Mächler, Mathew Seymour, Anaïs Lacoursière-Roussel, Florian Altermatt, Simon Creer, Iliana Bista, David M. Lodge, Natasha de Vere, Michael E. Pfrender, and Louis Bernatchez. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21):5872–5895, 2017. doi: 10.1111/mec.14350.

Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Adam Gudyś. FAMSA: fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, 6:33964, 2016. doi: 10.1038/srep33964.

Robert C. Edgar. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 2016. doi: 10.1101/074161. Preprint; never peer-reviewed.

Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *International Conference on Machine Learning*, volume 139, pp. 3122–3132, 2021.

Tomochika Fujisawa and Takashi Imai. Performance and limitations of out-of-distribution detection for insect DNA barcoding. *Ecology and Evolution*, 16(2):e73112, 2026. doi: 10.1002/ece3.73112.

Tiancheng Gao and Graham W. Taylor. BarcodeMamba: state space models for biodiversity analysis. In *NeurIPS Workshop on Foundation Models for Science*, 2024. arXiv:2412.11084.

Auguste Gardette, Eugeni Belda, Edi Prifti, and Jean-Daniel Zucker. CurateMake: an auditable workflow for multi-source ITS reference database harmonisation and phylogenetic validation. *bioRxiv*, 2026. doi: 10.64898/2026.07.10.737676. Preprint.

Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, et al. A step towards worldwide biodiversity assessment: the BIOSCAN-1M insect dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43593–43619, 2023. doi: 10.52202/075280-1889.

ZeMing Gong, Austin T. Wang, Xiaoliang Huo, Joakim Bruslund Haurum, Scott C. Lowe, Graham W. Taylor, and Angel X. Chang. CLIBD: bridging vision and genomics for biodiversity monitoring at scale. In *International Conference on Learning Representations*, 2025. arXiv:2405.17537.

Paul D. N. Hebert, Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270 (1512):313–321, 2003. doi: 10.1098/rspb.2002.2218.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Anders Lanzén, Steffen L. Jørgensen, Daniel H. Huson, Markus Gorfer, Svenn Helge Grindhaug, Inge Jonassen, Lise Øvreås, and Tim Urich. CREST: classification resources for environmental sequence tags. *PLoS ONE*, 7(11):e49334, 2012. doi: 10.1371/journal.pone.0049334.

Matthieu Leray, Nancy Knowlton, and Ryuji J. Machida. MIDORI2: a collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environmental DNA*, 4(4):894–907, 2022. doi: 10.1002/edn3.303.

Roy Li, Sujeevan Ratnasingham, Iuliia Zarubiieva, Panu Somervuo, and Graham W. Taylor. PROTAX-GPU: a scalable probabilistic taxonomic classification system for DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1904):20230124, 2024. doi: 10.1098/rstb.2023.0124.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Robert Lücking, M. Catherine Aime, Barbara Robbertse, Andrew N. Miller, et al. Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus*, 11(1):14, 2020. doi: 10.1186/s43008-020-00033-z.

Pablo Millan Arias, Niousha Sadjadi, Monireh Safari, ZeMing Gong, Austin T. Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Dirk Steinke, Lila Kari, Angel X. Chang, Scott C. Lowe, and Graham W. Taylor. BarcodeBERT: transformers for biodiversity analyses. *Bioinformatics Advances*, 6(1):vbag054, 2026. doi: 10.1093/bioadv/vbag054. arXiv:2311.02401.

Fábio M. Miranda, Vasco C. Azevedo, Rommel J. Ramos, Bernhard Y. Renard, and Vitor C. Piro. HiTaC: a hierarchical taxonomic classifier for fungal ITS sequences compatible with QIIME2. *BMC Bioinformatics*, 25(1):228, 2024. doi: 10.1186/s12859-024-05839-x.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022.

Adithya Murali, Aniruddha Bhargava, and Erik S. Wright. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, 6(1):140, 2018. doi: 10.1186/s40168-018-0521-5.

Johanna Orsholm, Alessandro Zito, Panu Somervuo, Jesse P. Harrison, Markus Koskela, Otso Ovaskainen, Mariana P. Braga, Nicolas Chazot, Tomas Roslin, and Brendan Furneaux. Discovering the unseen: a performance comparison of taxonomic classification methods for unknown DNA barcodes. *Methods in Ecology and Evolution*, 17(9):2574–2593, 2026. doi: 10.1111/2041-210x.70358.

Donovan H. Parks, Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, 50(D1):D785–D794, 2022. doi: 10.1093/nar/gkab776.

Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016. doi: 10.7717/peerj.2584.

Luuk Romeijn, Andrius Bernatavicius, and Duong Vu. MycoAI: fast and accurate taxonomic classification for fungal ITS sequences. *Molecular Ecology Resources*, 24(8):e14006, 2024. doi: 10.1111/1755-0998.14006.

Tomas Roslin, Panu Somervuo, Mikko Pentinsaari, Paul D. N. Hebert, Jireh Agda, Petri Ahlroth, et al. A molecular-based identification resource for the arthropods of Finland. *Molecular Ecology Resources*, 22(2):803–822, 2022. doi: 10.1111/1755-0998.13510.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341.

Monireh Safari, Pablo Millan Arias, Scott C. Lowe, Lila Kari, Angel X. Chang, and Graham W. Taylor. Enhancing DNA foundation models to address masking inefficiencies. In *ICLR Workshop on AI for Nucleic Acids (AI4NA)*, 2025. BarcodeMAE; arXiv:2502.18405.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Panu Somervuo, Sonja Koskela, Juho Pennanen, R. Henrik Nilsson, and Otso Ovaskainen. Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, 32(19):2920–2927, 2016. doi: 10.1093/bioinformatics/btw346.

Steven Stalder, Théophile Sanchez, Michele Volpi, Stéphanie Manel, David Mouillot, Arnaud Auber, Morgane Bruno, Virginie Marques, Camille Albouy, and Loïc Pellissier. Zero-shot deep learning for the annotation of unknown eDNA sequences using co-occurrences and phylogenetic embeddings. *PLOS Computational Biology*, 21(12):e1013776, 2025. doi: 10.1371/journal.pcbi.1013776.

Pierre Taberlet, Eric Coissac, François Pompanon, Christian Brochmann, and Eske Willerslev. Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8):2045–2050, 2012. doi: 10.1111/j.1365-294X.2012.05470.x.

Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007. doi: 10.1128/AEM.00062-07.

Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014. doi: 10.1186/gb-2014-15-3-r46.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768.

Jia Yang, Lucía Vázquez, Xiaodan Chen, Huimin Li, Hao Zhang, Zhanlin Liu, and Guifang Zhao. Development of chloroplast and nuclear DNA markers for Chinese oaks (Quercus subgenus Quercus) and assessment of their utility as DNA barcodes. *Frontiers in Plant Science*, 8:816, 2017. doi: 10.3389/fpls.2017.00816.

Bin Ye, Junfeng Xia, Xia Wan, Min Wu, and Satoshi Chiba. Multi-barcoding-based species identification using a hierarchical attention network with staged curriculum learning. *Methods in Ecology and Evolution*, 17(4):1143–1160, 2026. doi: 10.1111/2041-210x.70264.

Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, 2019. doi: 10.1016/j.cell.2019.07.010.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

Alessandro Zito, Tommaso Rigon, and David B. Dunson. Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. *Methods in Ecology and Evolution*, 14(2):529–542, 2023. doi: 10.1111/2041-210X.14009.

A Data, hierarchy, and evaluation splits

A.1 Data sources and curation

Table 1 reports the curated FASTA files used for retrieval, after the evaluation clades have been removed. Before that split, our four corpora hold a barcode for 11.3% of the species in the hierarchy; the retrieval references cover 6.5%. Both figures are printed by `taxoset.analysis.species_coverage`.

ITS comes from CurateMake (Gardette et al., 2026), 12S from MIDORI2 GB272 (Leray et al., 2022), 16S from GTDB r220 (Parks et al., 2022), and COI from BIOSCAN-1M (Gharaee et al., 2023). For 12S, we use the UNIQ version, which keeps every distinct haplotype of a species rather than one sequence per species; this is what makes a known-species evaluation possible. The 16S file, `bac120_ssu_reps_r220.fna`, holds the longest 16S sequence of each representative bacterial genome. Archaea are therefore present in the hierarchy but absent from this reference. COI metadata come from `BIOSCAN_Insect_Dataset_metadata.tsv`.

We apply the same curation to all four markers. We keep only sequences with a full lineage from kingdom to species, remove alignment gaps, match names to the Catalogue of Life through the cached correction map, and drop lineages that are not in the frozen inventory, since those taxa cannot be named. We then remove every clade kept for evaluation from the reference.

Curation reduces ITS from 5,189,645 to 2,058,106 sequences and COI from 1,128,313 to 84,441 specimens. After removing the evaluation clades, the references hold 970,750 ITS and 44,658 COI sequences. Another 1,027,356 ITS sequences are used neither as references nor as queries, which keeps every query group at 15,000 or fewer. For 16S, all 58,102 sequences are used: 32,658 as references and 25,444 as queries. Table 1 counts only queries with at least one hit, which explains the small differences from these totals.

The hierarchy combines Catalogue of Life COL26.3 (Bánki et al., 2026) with GTDB r220 for Bacteria. It holds 2,907,182 names at all ranks with their parent–child relationships, but no sequence evidence. The cited GTDB paper describes release R06-RS202; our data come from r220. Where a rank is missing, we reuse the name above it, sometimes with a suffix. These filler names are 2.8% of genus names but only 0.0 to 0.9% of returned names, by marker and novelty rank.

A.2 Temporal split

For the temporal test of Section 4.5, we keep the r220 taxonomy fixed and roll the sequence reference back to GTDB r207. Training and calibration also use r207 alone. A query counts as temporal when its taxon is absent from that release at the evaluated rank, even if a higher rank is present. Table A.1 compares this historical absence with leave-clade-out, before the novelty gate of Section 3.2 and with the full prediction set, no size cap. The right parent, the genus of a new species or the family of a new genus, is found in 0.747 ± 0.001 and 0.777 ± 0.001 of cases.

16S has no known-species group, so it cannot set its usual novelty threshold, which the other markers set on known queries. To test the full pipeline anyway, we use the conditional threshold instead: it is set on training episodes in which a sibling clade was removed one rank below, so it needs no known query, and the trainer already computes it for 16S and routes with it. Table A.2 applies that gate to both 16S splits and caps the returned set at 64 names.

The gate costs a lot in both splits. For novel genera, end-to-end coverage is 0.246 under leave-clade-out and 0.292 under the rollback; for novel families, 0.119 and 0.101. The wide ungated gap at family rank in Table A.1 mostly disappears after the gate: end to end, the two splits finish within 0.02 of each other, and the rollback is ahead at genus rank. Leave-clade-out makes retrieval and the hierarchy look better than a historical absence does. The full pipeline does not keep that advantage.

At species rank the conditional threshold fails in both splits, through distinct failure modes. Under leave-clade-out it drops to zero and routes every query, a gate in name only; under the rollback it goes to its maximum and routes none, so the pipeline abstains. A threshold that jumps from one extreme to the other on two similar splits is not one we would use at species rank.

Table A.1: Coverage of the full prediction set for 16S, without the gate and without a size cap. Temporal queries have a name and a sequence in GTDB r220 but no sequence in the r207 reference at the evaluated rank. Each query group holds at most 15,000 sequences.

Taxon absent at	Temporal queries	Leave-clade-out	Temporal
Species	15,000	0.879	0.829
Genus	6,690	0.923	0.869
Family	1,690	0.821	0.569

Table A.2: 16S through the conditional gate. Each group is evaluated at its novel rank, and the returned set is capped at 64 names, so coverage is at most the uncapped value of Table A.1. *Routed* is the share of queries the gate calls novel; *end to end* counts a query as covered only if it is routed and its true name is in the set. n is given as leave-clade-out / rollback. Means over three seeds.

			leave-clade-out		rollback to r207	
group	rank	n	routed	end to end	routed	end to end
novel species	species	6,496 / 15,000	1.000	0.752	0.000	0.000
novel genus	genus	8,827 / 6,690	0.304	0.246	0.420	0.292
novel family	family	10,120 / 1,690	0.195	0.119	0.274	0.101

A.3 TAXONOMIC HARMONIZATION AND LINEAGE CONFLICTS

The merged hierarchy sometimes places the same organism under two different lineages above the genus. For example, an organism may appear under both `Animalia/Chordata/Teleostei` and `Obazoa/Opisthokonta/Craniata`. If we remove one lineage, the taxon can stay in the reference under the other, which would bias the evaluation.

We count names that appear under more than one lineage. This gives an upper bound on the affected cases. The bound is worst at novel-genus rank, where it is 0.0% for COI, 0.2% for 16S, 3.6% for 12S, and 12.5% for ITS; for ITS it is 1.8% at novel-species rank and 3.4% at novel-family rank. These values probably overstate the problem, because the same name can legitimately appear under different nomenclatural codes.

Removing these cases does not remove the gain for novel ITS genera; it grows a little, the interval going from $[+0.022, +0.030]$ to $[+0.022, +0.033]$. At a fixed query, hits that share no normalized rank with it and hits that share three ranks differ by only 0.01 identity points. So most conflicts are close relatives placed under different higher ranks, not unrelated sequences.

B PLACEMENT NETWORK ARCHITECTURE AND TRAINING

B.1 SCORE POOLING AND NETWORK DIMENSIONS

Let J_c be the hits assigned to reference-backed candidate c. Its score at rank r combines its strongest hit, the evidence from all its hits, and its share of the identity-weighted vote:

$$S_r(c) = \alpha_r(E) \max_{j \in J_c} u_{r,j} + \big(1 - \alpha_r(E)\big)\, \mathrm{LSE}_{j \in J_c}\, u_{r,j} + \lambda_r(E)\, V_c, \tag{2}$$

where LSE is the log-sum-exp, $V_c = \log\big(\sum_{j \in J_c} w_j / \sum_j w_j\big)$ is the candidate's share of the vote, and $w_j = \max(s_j - 50,\, 1)^2$ weights hit j by its identity margin over the 50% search floor, with s_j on a 0 to 100 scale. Setting $\alpha_r = 1$ and $\lambda_r = 0$ in Equation 2 gives the best-hit rule; $\alpha_r = 0$ with a large λ_r gives the vote rule. The two unlearned rules are the two ends of the same parameterization.

The network reads 20 hits, each described by its identity and its z-score within the episode. It does not use their retrieval order. At each taxonomic rank r it produces three kinds of score, drawn as the three heads of Figure B.1: a hit score $u_{r,j}$ for each hit, a parent score $a_{r,g}$ for each retrieved parent g, and a *none-of-the-above* score ω_r. The parent token pools the parent's hits and adds its best identity, its vote share and a rank embedding; a layer normalization and a rank-specific linear readout turn it into $a_{r,g}$, which all the parent's unsequenced children share. Each parent is counted once: counting every child would multiply a parent's share by its number of unsequenced children, about 2 to 16 in practice. The episode summary token, [CLS] in the figure, gives ω_r and, through a six-output head, (α_r, λ_r) for the three ranks.

The probabilities of the hierarchy-backed option and of *none-of-the-above* are added to form the novelty score. A query routed as novel receives the hierarchy-backed set, or abstains if the set is empty; otherwise it receives the highest-scoring reference-backed name. The set is returned even when *none-of-the-above* dominates the novelty score. That is deliberate: the set is the best available answer under the evidence, and the alternative would be a single reference-backed name the network has just judged unlikely. The set is empty for at most 3.9% of ITS novel-family queries and under 1% elsewhere.

B.2 LOSS AND OPTIMIZATION

At each rank, training combines three terms: a cross-entropy over the three options above, whose target is the true reference-backed candidate when retrieval exposes it, the hierarchy-backed option when it does not but the true parent is still reachable, and *none-of-the-above* when neither holds; a cross-entropy for placing an absent taxon under the correct parent; and a binary cross-entropy for an auxiliary novelty head used only during training, whose label the clade removal gives exactly. All episodes contribute, with equal weights across terms and ranks.

For 15% of episodes we remove the query's family from the reference; for 30% its genus; for 30% its species. The remaining 25% use the intact reference with the query's own hit removed, so that

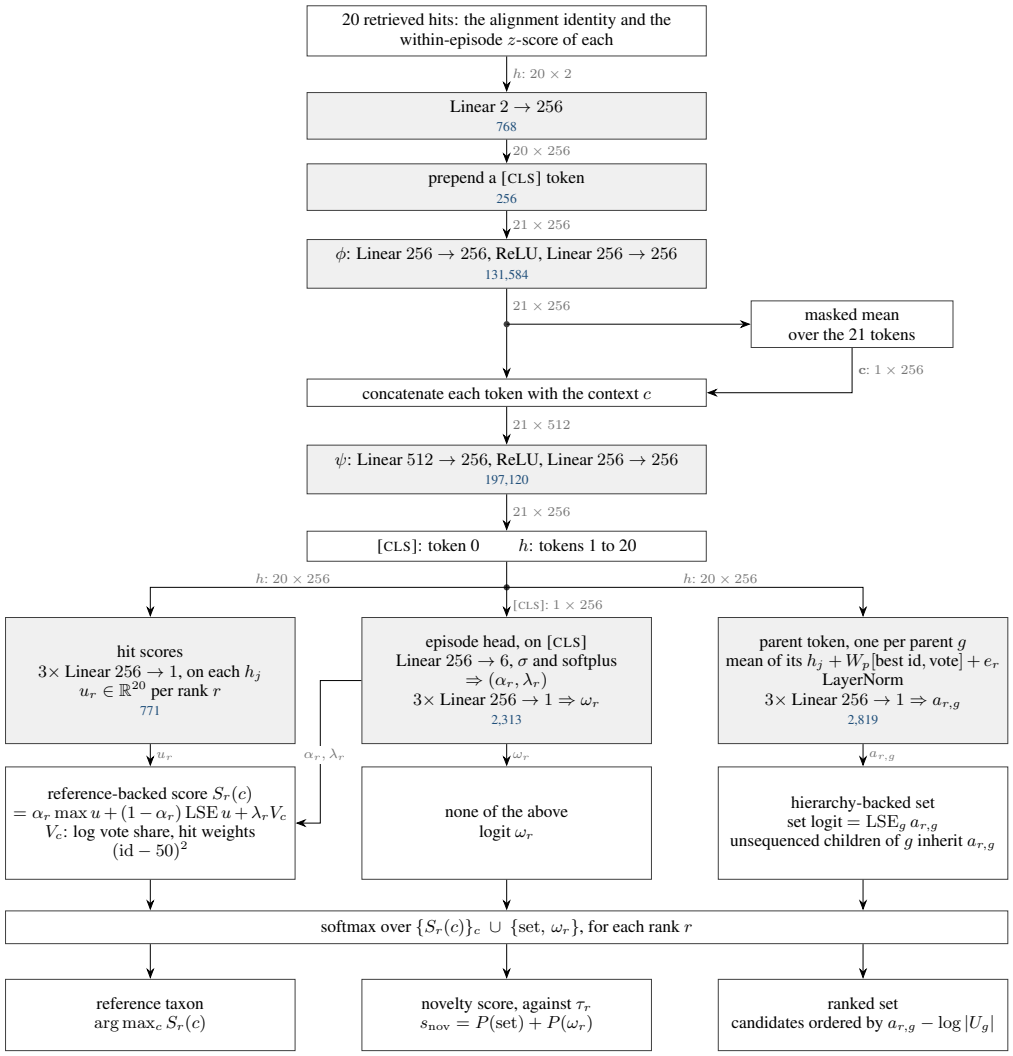


Figure B.1: **The placement network, layer by layer.** Shaded boxes hold parameters, with the count in blue; white boxes hold none. The drawn layers hold 335,631 of the network's 336,405 parameters, 97.7% of them in ϕ and ψ; the remaining 774 belong to the auxiliary novelty head, never read at inference, and to three unused weights.

no episode is solved by retrieving the query itself. We weight each episode's loss by $(n_g n_f)^{-1/4}$, where n_g and n_f count the training episodes of its genus and family, so that common clades do not dominate, and scale the weights to average one.

For every marker, we use AdamW (Loshchilov & Hutter, 2019), a learning rate of 3×10^{-4} with cosine decay, batches of 1,024, gradient clipping at norm 1, 40 epochs, and three seeds.

B.3 Ablations

The set encoder. Replacing the set encoder with four layers of 8-head attention raises the parameter count from 0.34M to 3.7M and makes training two to five times slower. On the same episodes, losses, calibration and queries, attention improves only one of nine settings, 12S novel families; the other eight intervals include zero (Table B.1). We keep the smaller encoder.

The sibling-size correction. Equation 1 shares each parent's probability equally among its $|U_g|$ unsequenced children. Compared with giving every child the full parent score, this improves expected top-1 within the returned list in all eight novel-species and novel-genus settings, by up to 1.9

Table B.1: Attention minus mean-pooling coverage on identical queries. This control predates the final gate calibration of Table D.1, so only the paired difference is reported. The bootstrap resamples clades, separately for each seed; the table gives the widest interval and the largest p-value across seeds.

marker	novel rank	attention − mean pooling	CI (clade-blocked)	p
ITS	family	-0.0005	[-0.0056,+0.0088]	0.9570
ITS	genus	+0.0058	[-0.0035,+0.0176]	0.4805
ITS	species	+0.0020	[-0.0082,+0.0111]	0.0955
12S	family	+0.0170	[+0.0032,+0.0398]	0.0055
12S	genus	-0.0030	[-0.0153,+0.0106]	0.9622
12S	species	+0.0040	[-0.0055,+0.0122]	0.4737
COI	family	+0.0000	[+0.0000,+0.0000]	1.0000
COI	genus	-0.0129	[-0.0721,+0.0007]	0.3934
COI	species	+0.0088	[-0.0095,+0.0345]	0.5008

points for 16S novel genera. A tunable penalty $\gamma \log |U_g|$ generalizes it, equal sharing being $\gamma = 1$. Coverage keeps rising past 1 in six of the eight cells and peaks between 2.5 and 4. Choosing γ on half the parent clades and scoring it on the other half, one cell survives Bonferroni over the eight: ITS novel genera, $+0.022$ with interval $[+0.006, +0.044]$, at $\gamma = 3$. We keep $\gamma = 1$, which follows from equal sharing and needs no marker-specific tuning.

Ranking siblings. Finding the correct parent does not by itself tell its unsequenced children apart. To test whether the hits still carry a signal for individual children, we compare a diagnostic readout that scores each child separately with 20 random shuffles of the child labels, for each of three seeds.

Eight of the nine marker–rank settings show no stable ordering: the learned result falls inside the shuffling interval, or changes direction across seeds. COI novel species is the one exception, where expected top-1 within the returned list rises from about 0.03 under shuffling to between 0.064 and 0.072. We cannot tell what produces that weak signal, so we keep the same conservative rule for every marker and give siblings equal scores.

For the $14{,}999$ ITS novel-species queries, choosing one tied name at random would be correct 21.8% of the time. Returning the set better reflects what the evidence supports.

C Additional baseline protocols and results

C.1 Classical and Bayesian baselines

Table C.1 compares several methods on ITS. SINTAX uses the full reference with its standard 0.8 threshold. BayesANT and HiTaC do not scale to it, so they run on a subset drawn with a fixed seed: at most two sequences per species and at most $40{,}000$ sequences. We rerun best-hit and TaxoSet on that same subset, so the table separates the method from the amount of reference it was given. Every row is scored on the full lineage down to the deepest rank the reference still supports for that group. At the group's own novel rank every reference-bound method is at a structural zero, so those cells are not shown.

At the reduced scale TaxoSet leads both novel cells, 0.790 against 0.769 for best-hit at genus rank and 0.594 against 0.561 for HiTaC at family rank, and matches best-hit on known species. The reduced reference costs every method a great deal, as the two best-hit rows show, but it does not reorder them.

BayesANT rarely detects novelty: its novelty branch fires on 3.0% of known-species queries, 2.7% of novel-species queries, 0.1% of novel-genus queries, and never at family rank. HiTaC's cost grows faster than its data: fitting $4{,}000$ sequences takes 9 minutes and $40{,}000$ takes 2 hours 52 minutes, whatever the thread count. The full reference would need about a week. Both timings come from the run log named in the reproducibility statement.

Table C.1: Classical and Bayesian baselines on ITS, scored on full lineages, abstentions counted as errors. These tools cannot read our retrieved hits, so the table is a control, not a matched comparison. BayesANT and HiTaC only run at a reduced reference scale; the best-hit and TAXOSET rows at that same scale separate the method from the amount of reference. Bold marks the best of the four reduced-scale rows.

method	protocol	known → sp.	novel sp. → gen.	novel gen. → fam.
TaxoSet	full reference (Table 2)	0.752	0.859	0.611
best-hit	full reference (Table 2)	0.731	0.837	0.527
vote	full reference (Table 2)	0.571	0.823	0.605
SINTAX	full reference, cutoff 0.8	0.592	0.786	0.413
TAXOSET	40k subsample	0.345	**0.790**	**0.594**
best-hit	40k subsample	**0.348**	0.769	0.539
HiTaC	40k subsample	0.306	0.764	0.561
BayesANT	40k subsample	0.247	0.712	0.477

C.2 COMPARISON WITH PROTAX

PROTAX spreads probability over the visible branches of a taxonomy and an *unknown* branch, from similarity features (Somervuo et al., 2016); PROTAX-GPU scales it to millions of sequences (Li et al., 2024). The released models are built for one marker and one clade, FinPROTAX for Finnish arthropod COI (Roslin et al., 2022) and Protax-fungi for fungal ITS (Abarenkov et al., 2018). Using them as released would change the reference and the label space at the same time, so their predictions cannot be scored against our hierarchy. We therefore use two protocols. We retrain the released PROTAX-GPU engine on our data, and we adapt its decision rule to our top-20 hits, which gives a paired test on identical inputs and covers ITS, where the engine cannot run.

PROTAX-GPU on our splits. We keep the released engine unchanged, forward pass and gradient included, and train it on the same references, hierarchy and splits as TAXOSET, with an *unknown* branch under every internal node. PROTAX needs aligned sequences. For each marker we build one profile HMM from a FAMSA alignment (Deorowicz et al., 2016) of 2,000 references and take p-distance over match columns only, as FinPROTAX does. Sequences covering under 60% of the core columns are dropped, since two sequences with no column in common have no defined distance; this keeps all COI queries, 76% of 16S queries and 65% of 12S queries, the rest being fragments.

We exclude ITS. Built the same way, its profile has 34 core columns, against $1{,}504$ for 16S, 652 for COI and 593 for 12S, and even then only half the corpus covers enough of them to give a defined distance: $503{,}281$ of $970{,}750$ references and $534{,}691$ of $1{,}030{,}750$ queries. PROTAX scores a pair by the share of matching positions among the valid ones, and with at most 34 valid positions that score takes at most 361 distinct values, which cannot separate hundreds of thousands of fungal lineages. The hypervariable spacers do not align to one profile across kingdoms. The engine computes the distance itself from aligned columns and takes no similarity from elsewhere; feeding it pairwise identities would mean replacing the one part we kept untouched. The top-20 adaptation below covers ITS instead.

We fit the coefficients on the released objective for 30 epochs of $20{,}000$ references, with three seeds, using Adam instead of the released gradient descent, which is what made the 2.9-million-node 12S tree converge; the estimator does not depend on the optimizer. Placement at a rank is the argmax over the named nodes of that depth. The novelty score is the total probability on *unknown* branches, thresholded at the paper's 5% false-positive rate on $2{,}000$ known queries, or by leave-one-out over $1{,}500$ references on 16S, which has no known query.

Table C.2 gives the outcome. On COI, PROTAX-GPU and TAXOSET stay within seed spread of each other, both below the identity threshold; the released gate moves by nearly a factor of two across seeds, ours by a few thousandths. On 12S it routes 0.07 of novel-species queries against 0.26 for TAXOSET, and its coverage stays far below ours. On 16S its leave-one-out threshold never fires and coverage is zero on all three seeds; our calibration uses conditional negatives (Appendix D.3) and needs no known query. Without the gate, taking the argmax over all named candidates, sequenced or not, the released engine comes within half a point of best-hit on COI known species and is below

both best-hit and TAXOSET everywhere else. Full taxonomy enumeration is therefore not enough: PROTAX-GPU is competitive only on COI, the marker its alignment model was designed for.

Table C.2: The released PROTAX-GPU at the deployed gate, joint coverage, mean ± CI_{95} over three seeds. Every gate targets the same 5% false-positive rate on the marker's known queries. The released pipeline keeps near-full-length sequences only, so its query count n is given next to ours. 16S has no known query and no gate. Bold marks the best mean.

Marker	Group	n released/harness	PROTAX-GPU released	TaxoSet	identity threshold
COI	novel sp.	7,231/7,232	0.300±0.218	0.340±0.009	**0.592±0.000**
COI	novel gen.	7,883/7,883	0.367±0.215	0.557±0.006	**0.580±0.000**
12S	novel sp.	9,855/14,998	0.036±0.008	0.074±0.004	**0.188±0.000**
12S	novel gen.	9,717/14,999	0.047±0.039	**0.520±0.004**	0.454±0.000

Adaptation to the top-20 hits. For a comparison on identical inputs, we adapt PROTAX to the same top-20 hits, episodes and splits as TAXOSET. From the order down, each node splits its mass over its visible children, one class for its named unsequenced children, and *unknown*, through a multinomial on four shared coefficients per level: an intercept, a has-reference flag, the best identity under the branch, and the second best. The unsequenced siblings of one parent enter as one branch of multiplicity k whose mass is shared as $1/k$, following Equation 1. *Unknown* scores zero, as in FinPROTAX, and the mislabeling probability is zero at inference. Each level is trained on the true path, the removed clade labeling the unsequenced class exactly. The novelty score is one minus the mass of the visible sequenced candidates, calibrated like every other method's. The returned set holds one entry per parent at the per-child share of its unsequenced class. Unlike full PROTAX, a sequenced taxon missing from the top 20 counts as unknown.

These results are not comparable with published PROTAX accuracies, which count any prediction within a group of unsequenced taxa as correct; here a prediction is correct only if the returned set contains the true name. Table C.3 gives coverage and Table C.4 the paired, clade-blocked comparison. TAXOSET has higher coverage in all nine settings where a gate can be calibrated, from $+0.095$ on 12S novel species to $+0.549$ on COI novel species, and eight differences survive Bonferroni correction. Most of the gap is routing: on novel COI species the adaptation routes 0.28 of queries against 0.95 for TAXOSET. The released engine shows that part of the large COI gap comes from the top-20 restriction rather than from PROTAX's decision rule. The 12S margin agrees with the released engine, and ITS has no released counterpart. The all-hit-parents rule in the table is the full-set counterpart of the round-robin ordering of Table D.6.

Table C.3: The top-20 PROTAX adaptation at the deployed gate, joint coverage, mean ± CI_{95} over three seeds. Its novelty mass is held to the same 5% false-positive target as every other method, realized within $[0.023, 0.055]$, the threshold saturating below target at species rank; hits, hierarchy and splits are identical. 16S admits no gate. Bold marks the best mean.

Marker	Group	PROTAX adapt.	TaxoSet	best-hit	all-hit parents
ITS	novel sp.	0.056±0.001	**0.281±0.009**	0.229	0.261
ITS	novel gen.	0.148±0.006	**0.694±0.006**	0.425	0.610
ITS	novel fam.	0.365±0.000	**0.572±0.002**	0.243	0.522
12S	novel sp.	0.119±0.005	0.209±0.006	0.188	**0.229**
12S	novel gen.	0.138±0.009	**0.639±0.007**	0.455	0.550
12S	novel fam.	0.313±0.011	0.654±0.003	0.456	**0.665**
COI	novel sp.	0.122±0.000	0.669±0.007	0.592	**0.685**
COI	novel gen.	0.351±0.010	**0.709±0.000**	0.580	0.707
COI	novel fam.	0.707±0.010	**0.918±0.000**	0.799	0.918

C.3 CONTROL WITH LEARNED SEQUENCE REPRESENTATIONS

A classifier limited to the taxa in its reference cannot predict an absent taxon, whatever its encoder. So we ask a narrower question: when whole clades are held out, can a large sequence model, in place of alignment retrieval, better predict a rank that is still supported?

Table C.4: TAXOSET minus the top-20 PROTAX adaptation in joint coverage on identical queries, each routed by its own gate at the same false-positive target. The bootstrap resamples clades, separately for each seed; the table gives the widest interval and largest p-value.

marker	novel rank	TaxoSet − adaptation	CI (clade-blocked)	p
ITS	family	+0.2068	[+0.1014,+0.3276]	<0.0001
ITS	genus	+0.5470	[+0.4405,+0.6300]	<0.0001
ITS	species	+0.2353	[+0.1878,+0.2805]	<0.0001
12S	family	+0.3425	[+0.2036,+0.5034]	<0.0001
12S	genus	+0.4955	[+0.4311,+0.5585]	<0.0001
12S	species	+0.0954	[+0.0790,+0.1126]	<0.0001
COI	family	+0.2104	[+0.0012,+0.5508]	0.0149
COI	genus	+0.3534	[+0.1965,+0.5069]	<0.0001
COI	species	+0.5493	[+0.4132,+0.6785]	<0.0001

We fully fine-tune `nucleotide-transformer-v2-500m-multi-species`: last hidden states averaged over the attention mask, then dropout and a linear layer over the training families, under cross-entropy. Sequences are 6-mer tokenized and cut at 512 tokens; learning rates are 3×10^{-5} for the backbone and 5×10^{-4} for the head. Training the whole backbone is the setting most favorable to the model, and the one most exposed to overfitting. We compare it with a k-mer nearest-neighbor baseline trained on the same data.

This control uses its own exports and split, not those of the main evaluation: $58{,}102$ sequences for 16S, $84{,}441$ for COI and a $200{,}000$-sequence subsample for ITS, larger than the matching retrieval references because taxonomic normalization merged and dropped lineages after export. Each run holds out 20% of the genera that can be held, in a single partition rather than the 24 leave-clade-out folds of the main evaluation; a family with one genus is never held, and one genus per family always stays in training, so the family label is always reachable. This split is easier than the main one. Its results compare only with baselines on the same split, never with Table 2. The partition depends on the export and the seed alone, so every arm below draws the same one.

On 16S the 500M model reaches 0.751 family accuracy after eight epochs, with training loss still falling, against 0.762 for the k-mer baseline; a 250M version reaches 0.713. Neither clears a nearest neighbor over 6-mer counts, which points to overfitting on the taxa seen in training. BarcodeBERT, fine-tuned as a classifier here and not used as the retriever of Appendix F, reaches 0.854 on COI against 0.609 for the baseline. But it was pretrained on about a million COI barcodes from the collection our COI queries come from, so this measures transfer from sequences already seen. On 16S and ITS, where it has no such exposure, it falls below the baseline: 0.659 against 0.762, and 0.548 against 0.589. The nucleotide transformer, whose pretraining holds at most a few dozen invertebrate mitochondrial genes, clears the baseline on ITS, 0.705 against 0.589, and is the one arm of the control that does.

Learned representations can therefore improve prediction when pretraining matches the target domain. They do not by themselves expand the label space.

The method on the control's own split. To read our own method on the control's terms, we built a split whose reference is the training side of that same partition and whose queries are its held-out genera, and trained the placement network on it, nothing else changed. The novelty gate is off for every column: on this split a novel-genus query has its family in the reference, so nothing the gate routes can be right at that rank. Table C.5 gives the family argmax on the same held-out queries.

On 16S the network leads every column, by eight points over the best the control reaches. On ITS it is far ahead of both models but four points behind the vote rule on its own hits, a loss Table D.7 already records. On COI it beats best-hit by three thousandths and loses to both foundation models.

The COI cell is the one to distrust. The same BarcodeBERT is 25 points above the k-mer baseline on COI and 10 and 4 points below it on 16S and ITS, where it has no pretraining exposure; part of what the COI cell measures is recall of sequences already seen. NT-500M has no such exposure and still reaches 0.809 there, because COI is the marker where the first hit is almost always right.

Table C.5: TAXOSET under the control's protocol: leave-genus-out, novel genus scored at family rank, same partition and queries for every column. A query that retrieves nothing counts as wrong, one on 16S and 106 on ITS. TAXOSET is mean ± CI_{95} over three seeds; the two unlearned rules run on its hits; each foundation model is beside the k-mer baseline of its run. †BarcodeBERT is pretrained on COI barcodes from the collection our COI queries come from, so this cell is not prediction on unseen sequences.

		on our retrieval			on the raw sequence		
marker	n	TAXOSET	best-hit	vote	BarcodeBERT	NT-500M	k-mer
16S	9,996	**0.846**±0.002	0.787	0.815	0.659	0.751	0.762
COI	11,653	0.697±0.003	0.694	0.587	**0.854**†	0.809	0.609
ITS	36,717	0.795±0.005	0.614	**0.834**	0.548	0.705	0.589

That leaves little room for anyone. Reading 20 hits can only beat reading one where the true family is in the list and the first hit misses it, and that share is 0.16 on 16S, 0.30 on ITS and 0.07 on COI. Every gain on our side of the table sits inside it: on COI the network is 0.3 points above best-hit within a 7-point zone; on ITS it is 18 points above within a 30-point zone, and the vote rule 22.

Is it the model or the way we tuned it? Training a 500M backbone end to end on tens of thousands of sequences invites overfitting, so a poor result could be the tuning rather than the representation. We reran 16S with LoRA adapters (Hu et al., 2022) on a frozen backbone, and with no backbone training at all, a linear probe on the pooled last layer. Table C.6 gives the six arms on one partition, so the k-mer baseline is one number and the last column a paired difference.

Table C.6: 16S family accuracy under three tuning regimes, beside the k-mer nearest-neighbor baseline. All six arms share one partition, so the baseline is one number and the last column is a paired difference. Eight epochs, weight decay 0.01, seed 0.

model	tuning	model	k-mer	difference
NT-500M	full	0.751	0.762	−0.011
NT-500M	LoRA	0.727	0.762	−0.035
NT-500M	frozen	0.391	0.762	−0.371
BarcodeBERT	full	0.659	0.762	−0.103
BarcodeBERT	LoRA	0.638	0.762	−0.124
BarcodeBERT	frozen	0.234	0.762	−0.528

LoRA lands within a few points of full fine-tuning on both models, and behind it; the frozen probe is far behind. The gap to a k-mer nearest neighbor is therefore not an artifact of training the whole backbone: a parameter-efficient recipe reproduces it, and the representation has to be adapted at all to be useful here.

D COMPLETE OPEN-WORLD RESULTS

D.1 COVERAGE AND SET SIZE

Table D.1 reports end-to-end coverage, the share of queries that are routed as novel and whose returned set holds the true name, together with its two parts, the routed share and the coverage conditional on routing, and the median and 90th percentile of the set size. Set size changes what a coverage means. For novel genera, ITS reaches 0.694 with a median of 11 names; COI reaches 0.709 with a median of 1,493.

D.2 LIST SIZE AND NOVELTY THRESHOLD

Table D.6 reports coverage for lists capped at 5, 10 or 20 names, for every marker. Varying the novelty threshold on ITS shows that at genus and family rank a higher threshold routes far fewer queries without improving coverage, while at species rank routing is the limit: without a gate the complete set covers 0.932 of queries, but the calibrated threshold routes only 0.323 (seed 0). We

Table D.1: End-to-end coverage at the deployed gate, averaged over three seeds, with the routed share, the coverage conditional on routing, clade-blocked intervals, and returned-set sizes. The gate is calibrated on a separate taxon split. 16S is omitted: it has no usable known-species split.

marker	novel rank	realized FPR	routed	routed-set cov.	**end-to-end cov.**	per taxon	CI (clade)	med $\|U\|$	q90 $\|U\|$
ITS	species	0.046	0.317	0.885	**0.281**	0.398	[0.24,0.31]	19	196
ITS	genus	0.046	0.841	0.825	**0.694**	0.718	[0.63,0.75]	11	1,068
ITS	family	0.048	0.928	0.617	**0.572**	0.549	[0.42,0.69]	12	32
12S	species	0.051	0.262	0.800	**0.209**	0.244	[0.20,0.23]	82	878
12S	genus	0.048	0.761	0.839	**0.639**	0.654	[0.58,0.69]	31	915
12S	family	0.041	0.862	0.759	**0.654**	0.635	[0.53,0.76]	34	298
COI	species	0.045	0.950	0.704	**0.669**	0.746	[0.53,0.78]	366	1,953
COI	genus	0.049	0.985	0.720	**0.709**	0.720	[0.57,0.82]	1,493	4,427
COI	family	0.050	1.000	0.918	**0.918**	0.785	[0.34,1.00]	173	272

therefore report the calibrated threshold with the complete set and its size, and never tune it on evaluation data.

D.3 Novelty calibration

Our main threshold is set on held-out known taxa so that it marks 5% of them as novel. Figure D.1 shows end-to-end coverage as that rate is allowed to grow. Changing it has little effect on COI, but many novel ITS and 12S species are missed because they are not routed as novel.

A stricter, conditional threshold can instead be set on hard negatives: training episodes in which the query's sibling clade was removed one rank below, so that the gated rank is still known. For ITS novel genera it raises the threshold from 0.40 to 0.89. The routed share then falls from 0.84 to 0.38 and end-to-end coverage from 0.69 to 0.28, for a false-positive rate of 0.5%. This option is useful when a false novelty call costs more than a missed novel taxon; it is not the threshold behind our main results.

Without choosing any threshold, the novelty-score AUROC ranges from 0.76 for ITS novel species to 1.00 for COI novel families. It is not measured on 16S, which has no known query.

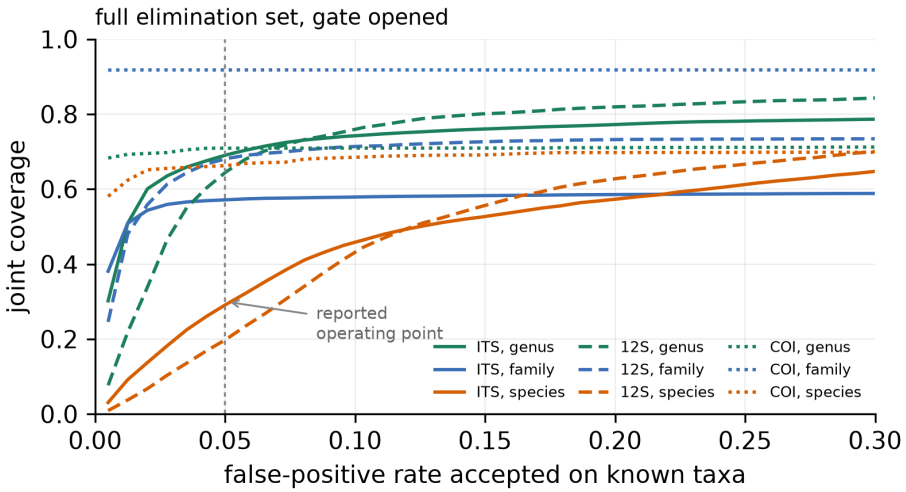


Figure D.1: End-to-end coverage of the complete prediction set as the accepted false-positive rate on known taxa grows. Color gives the novelty rank, line style the marker. The dotted line marks the threshold used in the main results.

D.4 Retrieval bound

The hierarchy can expose the true name only if retrieval finds its correct parent. Table D.2 measures how often it does within the first 20 hits, at the deepest rank the reference supports: for a query novel

at rank r, the parent rank $r - 1$. This is an upper bound set before the network runs. For ITS novel genera and families, the complete set holds the true name in 99% of the cases where the correct parent is retrieved. The gap is wider elsewhere, so the table is a decomposition, not a claim that the network always reaches the bound.

Table D.2: Recall@20 at the deepest rank supported by the reference. The dash marks 16S, which has no usable known-species split.

	known	novel species	novel genus	novel family
ITS	0.933	0.934	0.821	0.605
12S	0.909	0.918	0.866	0.741
16S	–	0.880	0.924	0.822
COI	0.998	0.719	0.724	0.922

D.5 Search depth and list length

A deeper search can mean two things. The first is how hard the search looks: VSEARCH stops after a budget of rejected candidates, and raising that budget while still reporting 20 hits already moves the bound, at no cost to the network, whose input stays 20 wide. The second is the length of the list, which the network would have to be retrained to read. Table D.3 separates them. Column *paper* is the search behind the reported results and reproduces Table D.2 within 0.01 on every cell; column 20 is the same list length under a search allowed to look four times longer; the other columns lengthen the list.

The two parts fall unevenly across markers. The largest single gain comes from search effort alone: 16S novel families move from 0.821 to 0.890 at the same 20 hits, and the three longer lists add only two points on top. On COI, effort alone is worth 4.4 points at novel species and 2.8 at novel genus, more than doubling the list adds. On ITS, effort is worth almost nothing and length a great deal: novel genera move from 0.813 to 0.863 at 50 hits and 0.897 at 200. One cell moves the wrong way, ITS novel families, where a longer search finds better-scoring wrong hits that push the true parent out of the first 20.

Table D.3: Recall of the true parent at the deepest reference-supported rank, as the search changes. *paper* is the accept and reject budget behind the reported results, reporting 20 hits; the numbered columns raise both budgets with the depth and report that many hits. Seed 0.

		hits reported				
marker	group	paper	20	50	100	200
ITS	novel sp.	0.932	0.934	0.948	0.957	0.965
ITS	novel gen.	0.813	0.817	0.863	0.882	0.897
ITS	novel fam.	0.595	0.585	0.672	0.714	0.759
12S	novel sp.	0.916	0.930	0.948	0.959	0.963
12S	novel gen.	0.864	0.871	0.889	0.901	0.910
12S	novel fam.	0.736	0.750	0.804	0.829	0.858
16S	novel sp.	0.879	0.890	0.900	0.905	0.909
16S	novel gen.	0.923	0.927	0.936	0.940	0.945
16S	novel fam.	0.821	0.890	0.907	0.916	0.923
COI	novel sp.	0.718	0.762	0.771	0.793	0.808
COI	novel gen.	0.724	0.752	0.784	0.798	0.814
COI	novel fam.	0.918	0.934	0.965	0.978	0.983

Reading more hits. Table D.3 shows that a longer search puts the true parent in the list more often. The network cannot use it. We rebuilt every episode at 50 hits and retrained all four markers, three seeds each, nothing else changed. Table D.4 gives the result. On ITS, where the bound had the most to give, five points at 50 hits, the three cells move by less than three thousandths. On COI the network gains 1.5 points at novel species, but best-hit gains 1.3 on the same lists, so the margin between them does not move. On 12S and 16S the family rank loses two points.

A third arm locates the difficulty: the network trained at 20, reading 50-hit episodes without retraining. On COI this costs almost nothing and reaches the same 0.636 as the retrained network. On 12S it loses three points that retraining only half recovers. On 16S at family rank it does better than the retrained network, 0.793 against 0.766. So a longer list is not simply more evidence: it shifts what the network reads, retraining does not reliably compensate, and on one cell it makes things worse.

A list of 20 hits is therefore a trade-off the data support, not a setting left untuned. A longer list widens what is reachable and makes the first hit less trustworthy at the same time, and the two effects cancel. Compute is not what settles it: the network is linear in the list length, and end-to-end latency is dominated by the search either way.

Table D.4: Reference-backed accuracy when every episode is rebuilt at 50 hits. The last column loads the network trained at 20 and reads 50-hit episodes with it, so it separates the change of evidence from the retraining. Mean over three seeds; one seed in the last column.

marker	cell	twenty	fifty	twenty's weights at fifty
ITS	known sp.	0.752	0.751	–
ITS	novel sp. → gen.	0.859	0.860	–
ITS	novel gen. → fam.	0.611	0.613	–
12S	known sp.	0.778	0.773	–
12S	novel sp. → gen.	0.803	0.798	0.773
12S	novel gen. → fam.	0.730	0.708	0.694
16S	novel sp. → gen.	0.798	0.798	0.781
16S	novel gen. → fam.	0.784	0.766	0.793
COI	known sp.	0.988	0.988	–
COI	novel sp. → gen.	0.620	0.634	0.636
COI	novel gen. → fam.	0.603	0.611	0.606

D.6 When retrieval misses

A shortlist of 20 hits is a hard constraint, and several methods on the external benchmark have none: they score the whole reference, or a k-mer profile of it. If the shortlist is what limits us, those methods should do best exactly where it fails. We split the benchmark's reference-backed group in two: a query is reachable when the true genus is among the lineages of our 20 hits, and blind when it is not. On COI, $1{,}280$ of $9{,}435$ queries are blind; on ITS, 431 of $7{,}056$.

On the blind COI queries the best published method is phylogenetic placement (Barbera et al., 2019) at 0.276, against 0.838 on the reachable ones; IDTAXA reaches 0.145 and SINTAX 0.019 there, although both are above 0.845 on the reachable half. On the blind ITS queries the best is 0.086. Building a tree over the entire reference therefore recovers about a quarter of what our shortlist loses on COI, and the other three quarters are lost to everyone.

D.7 What learning adds

Retrieval and the hierarchy already produce a list without any training. Table D.6 therefore compares TaxoSet with three unlearned rules on the same hits, hierarchy, false-positive target and list sizes, and Table D.7 holds the paired, clade-blocked intervals behind the count of Section 4.3: 18 wins, 3 losses and 6 undecided cases out of 27 cells. Sixteen wins survive Benjamini–Hochberg correction and 12 survive Bonferroni; the three losses, the 12S novel-species cells, one per list size, survive Bonferroni as well.

The losses have a measurable cause. Learning can only help where the evidence leaves room to overrule the nearest reference sequence. Take the queries whose best hit names the wrong genus and ask how far below it the right genus sits; Table D.5 gives the median. On ITS and COI the right genus is half an identity point behind, close enough for a rule that weighs the whole list to overturn the first hit. On 12S and 16S it is more than a full point behind: the evidence points the other way, and there is nothing for pooling to gain. These are the two markers where the network sits below best-hit at this rank in Table 2.

Table D.5: Queries whose species is absent and whose best hit names the wrong genus. The last column is the identity the right genus gives up to the wrong best hit, in percentage points, at the median. Seed 0, evaluation hit lists.

marker	best hit wrong	true genus in the list	its rank	identity behind
ITS	2,444	0.584	3	0.50
12S	2,834	0.557	3	1.20
16S	1,243	0.367	3	1.10
COI	2,784	0.267	5	0.50

D.8 Per-taxon accuracy and uncertainty

Taxonomic data are imbalanced, so Table D.8 reports accuracy both per query and averaged across true taxa. The second measure gives each taxon equal weight. It is usually lower because common taxa tend to be easier, but it does not change the overall conclusions.

Table D.8: Accuracy per query and per true taxon for reference-backed placement, reported in that order. Seed 0.

marker	known → species	novel species → genus	novel genus → family
ITS	0.751/0.709	0.859/0.753	0.610/0.509
12S	0.777/0.745	0.803/0.737	0.730/0.624
16S	–	0.798/0.734	0.783/0.673
COI	0.988/0.974	0.619/0.584	0.605/0.451

For comparisons between prediction sets, we estimate uncertainty by a paired bootstrap over parent clades, not over queries. Each resample draws parent clades with replacement and keeps every query of the selected clades; this is the clade-blocked interval the tables report. We compute an interval for each seed, then report the lowest lower bound and the highest upper bound across the three seeds. Table D.7 also applies Bonferroni correction over its 27 comparisons. The claims in the main text rest on these intervals, which are wider than intervals over individual queries would be.

E Transfer across markers

For each target marker, we train the network on the other three and apply it without changing any weight. Only the novelty threshold is calibrated, on known queries of the target marker.

Table E.1 is Table 2 with one column added, so the two compare cell by cell: top-1 at the deepest reference-supported rank, no gate, which isolates placement from routing. Placement transfers. The transferred network stays within 0.034 of the native one in ten of the eleven cells, is ahead in four, and the median difference is -0.003. Against best-hit it wins six cells of eleven, the native network seven: the whole cost of never having seen the marker is one cell. Two rows stand out. On 16S novel genera the transferred network is ahead by 0.032; 16S has one sequence per species, which makes its episodes the poorest of the four, so training on the other three regularizes better than training on this one. The single real loss is COI novel species, -0.060, the cell where Appendix D.6 finds the least room above a fixed rule: where there is almost nothing to win, a shift in the input costs the whole margin.

Table E.2 reports the deployed setting instead: joint coverage at the gate, against the network trained on the target marker and two unlearned rules on identical hits. Its numbers are not comparable with Table 2: the gate is on, the rank scored is the novel one rather than the one above it, and its best-hit column is the best-hit expansion rule, not a single name. The realized false-positive rate stays within $[0.045, 0.051]$ of the 5% target. Table E.3 adds the paired, clade-blocked test against the native network: the gap stays below five points in eight of nine settings, is indistinguishable from zero in four, and reverses for COI novel species. The main exception is ITS novel genera, 0.588 against 0.694.

Against the strongest unlearned rule, all-hit parents, the two networks and the rule split the full-set cells. This restates Section 4.3: at the complete set the learned contribution is routing, and it travels

Table D.6: End-to-end coverage against the unlearned rules, averaged over three seeds. Every method uses the same retrieval, hierarchy, false-positive target and list cap. Bold marks the best result at each cap.

method	realized FPR	routed	e2e@b=5	e2e@b=10	e2e@b=20
ITS, novel species					
identity gate + hierarchy, best-hit order	0.043	0.297	0.143	0.175	0.205
identity gate + hierarchy, round-robin	0.043	0.297	0.137	0.170	0.203
identity gate + hierarchy, vote/$\lvert U\rvert$	0.043	0.297	0.145	0.178	0.208
TaxoSet, learned gate	0.046	0.317	**0.151**	**0.186**	**0.219**
ITS, novel genus					
identity gate + hierarchy, best-hit order	0.043	0.752	0.376	0.440	0.491
identity gate + hierarchy, round-robin	0.043	0.752	0.413	0.482	0.525
identity gate + hierarchy, vote/$\lvert U\rvert$	0.043	0.752	0.434	0.497	0.530
TaxoSet, learned gate	0.046	0.841	**0.497**	**0.568**	**0.614**
ITS, novel family					
identity gate + hierarchy, best-hit order	0.043	0.883	0.312	0.391	0.494
identity gate + hierarchy, round-robin	0.043	0.883	0.322	0.427	0.505
identity gate + hierarchy, vote/$\lvert U\rvert$	0.043	0.883	0.361	0.443	0.508
TaxoSet, learned gate	0.048	0.928	**0.387**	**0.472**	**0.551**
12S, novel species					
identity gate + hierarchy, best-hit order	0.050	0.276	**0.129**	**0.154**	**0.176**
identity gate + hierarchy, round-robin	0.050	0.276	0.100	0.122	0.149
identity gate + hierarchy, vote/$\lvert U\rvert$	0.050	0.276	0.110	0.137	0.161
TaxoSet, learned gate	0.051	0.262	0.105	0.128	0.147
12S, novel genus					
identity gate + hierarchy, best-hit order	0.050	0.666	0.281	0.347	0.398
identity gate + hierarchy, round-robin	0.050	0.666	0.254	0.332	0.387
identity gate + hierarchy, vote/$\lvert U\rvert$	0.050	0.666	0.277	0.347	0.404
TaxoSet, learned gate	0.048	0.761	**0.333**	**0.421**	**0.491**
12S, novel family					
identity gate + hierarchy, best-hit order	0.050	0.890	0.274	0.370	0.472
identity gate + hierarchy, round-robin	0.050	0.890	0.271	0.371	0.470
identity gate + hierarchy, vote/$\lvert U\rvert$	0.050	0.890	0.288	0.380	0.483
TaxoSet, learned gate	0.041	0.862	**0.298**	**0.402**	**0.517**
COI, novel species					
identity gate + hierarchy, best-hit order	0.045	0.966	**0.094**	**0.151**	0.223
identity gate + hierarchy, round-robin	0.045	0.966	0.071	0.112	0.164
identity gate + hierarchy, vote/$\lvert U\rvert$	0.045	0.966	0.089	0.140	0.187
TaxoSet, learned gate	0.045	0.950	0.092	0.150	**0.226**
COI, novel genus					
identity gate + hierarchy, best-hit order	0.045	0.983	0.015	0.030	0.057
identity gate + hierarchy, round-robin	0.045	0.983	0.012	0.022	0.041
identity gate + hierarchy, vote/$\lvert U\rvert$	0.045	0.983	0.014	0.026	0.047
TaxoSet, learned gate	0.049	0.985	**0.036**	**0.071**	**0.134**
COI, novel family					
identity gate + hierarchy, best-hit order	0.045	1.000	0.027	0.056	0.114
identity gate + hierarchy, round-robin	0.045	1.000	0.026	0.053	0.110
identity gate + hierarchy, vote/$\lvert U\rvert$	0.045	1.000	0.025	0.054	0.110
TaxoSet, learned gate	0.050	1.000	**0.066**	**0.143**	**0.297**

Table D.7: Margin in end-to-end coverage between TAXOSET and the best unlearned rule. Intervals are paired, clade-blocked 95% bootstrap envelopes over three seeds. Bold marks an interval excluding zero; * and † mark results that survive Benjamini–Hochberg and Bonferroni correction over 27 cells. The last column gives the attention control's mean margin ± Student-t CI over three seeds; it is not part of the primary count.

cell	budget	TaxoSet	clade-blocked interval	attention control
ITS species	b=5	**+0.0065***	[+0.001,+0.013]	+0.0004 ± 0.0077
ITS species	b=10	**+0.0081***	[+0.001,+0.016]	+0.0010 ± 0.0091
ITS species	b=20	**+0.0112***	[+0.003,+0.020]	+0.0030 ± 0.0109
ITS genus	b=5	**+0.0630**†	[+0.007,+0.136]	+0.0553 ± 0.0027
ITS genus	b=10	**+0.0714**†	[+0.016,+0.142]	+0.0624 ± 0.0029
ITS genus	b=20	**+0.0839**†	[+0.025,+0.153]	+0.0745 ± 0.0029
ITS family	b=5	+0.0263	[-0.044,+0.084]	+0.0134 ± 0.0056
ITS family	b=10	+0.0290	[-0.037,+0.081]	+0.0196 ± 0.0034
ITS family	b=20	**+0.0436**	[+0.000,+0.082]	+0.0331 ± 0.0078
12S species	b=5	**-0.0238**†	[-0.034,-0.014]	-0.0330 ± 0.0053
12S species	b=10	**-0.0256**†	[-0.036,-0.015]	-0.0353 ± 0.0064
12S species	b=20	**-0.0290**†	[-0.041,-0.017]	-0.0373 ± 0.0066
12S genus	b=5	**+0.0527**†	[+0.032,+0.074]	+0.0499 ± 0.0044
12S genus	b=10	**+0.0738**†	[+0.049,+0.098]	+0.0725 ± 0.0064
12S genus	b=20	**+0.0878**†	[+0.060,+0.118]	+0.0856 ± 0.0063
12S family	b=5	+0.0095	[-0.007,+0.029]	+0.0140 ± 0.0094
12S family	b=10	**+0.0212**	[+0.001,+0.045]	+0.0290 ± 0.0116
12S family	b=20	**+0.0338***	[+0.005,+0.069]	+0.0424 ± 0.0112
COI species	b=5	-0.0026	[-0.026,+0.014]	-0.0016 ± 0.0110
COI species	b=10	-0.0002	[-0.037,+0.028]	-0.0004 ± 0.0125
COI species	b=20	+0.0036	[-0.066,+0.069]	-0.0090 ± 0.0330
COI genus	b=5	**+0.0204**†	[+0.008,+0.034]	+0.0210 ± 0.0014
COI genus	b=10	**+0.0409**†	[+0.017,+0.068]	+0.0422 ± 0.0024
COI genus	b=20	**+0.0768**†	[+0.032,+0.129]	+0.0783 ± 0.0044
COI family	b=5	**+0.0393**†	[+0.013,+0.044]	+0.0391 ± 0.0005
COI family	b=10	**+0.0870**†	[+0.028,+0.098]	+0.0866 ± 0.0017
COI family	b=20	**+0.1824**†	[+0.058,+0.206]	+0.1814 ± 0.0040

to a marker the network never saw. Gated transfer cannot be tested on 16S, with no known query; ungated, the transferred network scores 0.795 and 0.816 there, against 0.809 and 0.740 for best-hit.

Table E.1: Table 2 with one column added. *Unseen marker* is trained on the other three markers and evaluated here without updating any weight; every other column is Table 2's, on the same splits, queries and hits, which is why best-hit reproduces it exactly. Top-1, no novelty gate, mean ± CI_{95} over three seeds. Bold marks the better of the two networks in each row.

Marker	Query group → rank	TaxoSet	TaxoSet (unseen marker)	best-hit	vote
ITS	known sp. → species	**0.752±0.001**	0.735±0.002	0.731	0.571
ITS	novel sp. → genus	0.859±0.000	**0.861±0.000**	0.837	0.823
ITS	novel gen. → family	**0.611±0.002**	0.577±0.001	0.527	0.605
12S	known sp. → species	0.778±0.001	**0.785±0.003**	0.788	0.415
12S	novel sp. → genus	**0.803±0.001**	0.802±0.001	0.811	0.542
12S	novel gen. → family	**0.730±0.001**	0.725±0.003	0.715	0.629
16S	novel sp. → genus	**0.798±0.000**	0.795±0.000	0.809	0.479
16S	novel gen. → family	0.784±0.005	**0.816±0.025**	0.740	0.758
COI	known sp. → species	0.988±0.000	**0.989±0.000**	0.990	0.786
COI	novel sp. → genus	**0.620±0.002**	0.560±0.006	0.615	0.410
COI	novel gen. → family	**0.603±0.004**	0.599±0.019	0.597	0.512

Table E.2: Leave-one-marker-out transfer at the deployed gate: joint coverage, mean ± CI_{95} over three seeds. TAXOSET (unseen marker) is trained on the other three markers; its novelty threshold is calibrated on the target marker's known queries, and the unlearned rules run on identical hits. Scored at the rank where the query becomes novel; *best-hit exp.* is the best-hit expansion rule, not a single name, so the columns are not those of Table 2. Intervals measure variation across seeds, not clade-blocked uncertainty. Bold marks the highest mean in each row. 16S is omitted, having no known query to calibrate a gate on.

Marker	Query group	TaxoSet (unseen marker)	TaxoSet (native)	best-hit exp.	all-hit parents
ITS	novel sp.	0.240±0.006	**0.281±0.009**	0.236	0.268
ITS	novel gen.	0.588±0.010	**0.694±0.006**	0.427	0.615
ITS	novel fam.	0.548±0.010	**0.572±0.002**	0.243	0.523
12S	novel sp.	0.205±0.010	0.209±0.006	0.188	**0.229**
12S	novel gen.	0.597±0.030	**0.639±0.007**	0.455	0.550
12S	novel fam.	0.648±0.011	0.654±0.003	0.456	**0.665**
COI	novel sp.	0.683±0.004	0.669±0.007	0.592	**0.685**
COI	novel gen.	0.707±0.000	**0.709±0.000**	0.580	0.707
COI	novel fam.	**0.918±0.000**	0.918±0.000	0.799	0.918

Table E.3: Native TAXOSET minus unseen-marker TAXOSET, joint coverage on identical queries, each routed by its own gate at the same false-positive target. The bootstrap resamples clades, separately for each seed; the table gives the widest interval and largest p-value.

marker	novel rank	native TaxoSet − unseen-marker TaxoSet	CI (clade-blocked)	p
ITS	family	+0.0246	[+0.0118,+0.0403]	<0.0001
ITS	genus	+0.1055	[+0.0324,+0.1988]	<0.0001
ITS	species	+0.0418	[+0.0234,+0.0593]	<0.0001
12S	family	+0.0069	[-0.0142,+0.0279]	0.6474
12S	genus	+0.0430	[+0.0248,+0.0616]	<0.0001
12S	species	-0.0018	[-0.0114,+0.0068]	0.9484
COI	family	+0.0000	[+0.0000,+0.0000]	1.0000
COI	genus	+0.0024	[-0.0004,+0.0076]	0.7342
COI	species	-0.0122	[-0.0251,-0.0026]	0.0030

F TRANSFER ACROSS RETRIEVAL ENGINES

At inference, we replace VSEARCH with cosine top-20 retrieval on frozen embeddings from BarcodeBERT (Millan Arias et al., 2026) or `nucleotide-transformer-v2-500m` (Dalla-Torre et al., 2025). The placement network keeps its VSEARCH weights; only the novelty threshold is recalibrated.

The retriever writes $100 \cos\theta$ in place of the alignment identity. The within-episode z-score is invariant to that change, so the network's second input is unchanged by construction; the first input and the vote weights are not. Embedding similarities crowd near the top: on ITS novel genera, alignment identity runs from 65.6 at the tenth percentile to 98.5 at the ninetieth, BarcodeBERT from 88.8 to 97.7. What matters for the vote is how much the weights differ inside one episode, and that collapses: their coefficient of variation falls from 0.115 under alignment to 0.015 under BarcodeBERT and 0.003 under the transformer. Under either encoder the vote term is nearly constant and carries almost no signal.

Table F.3 reports each engine's Recall@20, which bounds every cell of Table F.2, and coverage tracks it. On ITS novel genera, BarcodeBERT lowers family-rank recall from 0.813 to 0.484; coverage then falls from 0.694 to 0.086, the recalibrated gate routing 0.22 of queries instead of 0.84 at the same false-positive rate. The collapse is a calibrated refusal, not a silent error. COI is the one setting where an embedding engine retrieves better than alignment, 0.809 against 0.724; it is also BarcodeBERT's pretraining domain, whose corpus overlaps the evaluated clades, so that bound is optimistic.

Table F.1 repeats Table 2 under each engine: same rows, same queries, top-1 at the deepest reference-supported rank, no gate. Placement does not survive the swap. It falls by 25 to 35 points on ITS, 12S and 16S, and the margin over best-hit on identical hits goes with it: the network is ahead in seven of eleven cells under VSEARCH, six under BarcodeBERT and four under the transformer. What survives is the decision: Table F.2 scores the deployed system, and there TAXOSET beats the best-hit expansion rule in 25 of 27 cells. Together the two tables show that the placement weights are tied to the scale they were trained on, while the routing decision is not. COI novel genera under BarcodeBERT is the one cell that goes the other way, 0.747 against 0.709 at the gate.

These results add together a change in retrieval and a change in input scale. Separating the two would require retraining the placement network on the new scores, which we did not do.

Table F.1: Table 2 repeated under two other retrieval engines: same rows, same splits, top-1 at the deepest reference-supported rank, no novelty gate. Only the 20 hits change; the placement weights are the VSEARCH-trained ones. Cosine always returns 20 neighbors while VSEARCH can return none, so the embedding columns score up to 26 more queries of the $15{,}000$ in a group. *Best-hit* is a single name here, the lineage of the first hit, not the expansion rule of Table F.2. Mean $\pm$ CI_{95} over three seeds; best-hit depends only on the hit list. Bold marks the better of the two within each engine.

		VSEARCH		BarcodeBERT		NT-500M	
Marker	Query group → rank	TaxoSet	best-hit	TaxoSet	best-hit	TaxoSet	best-hit
ITS	known sp. → species	**0.752**±0.001	0.731	**0.500**±0.001	0.483	**0.500**±0.003	0.487
ITS	novel sp. → genus	**0.859**±0.000	0.837	**0.643**±0.001	0.635	**0.710**±0.000	0.709
ITS	novel gen. → family	**0.611**±0.002	0.527	**0.266**±0.001	0.264	**0.405**±0.006	0.393
16S	novel sp. → genus	0.798±0.000	**0.809**	0.486±0.003	**0.496**	0.467±0.004	**0.475**
16S	novel gen. → family	**0.784**±0.005	0.740	**0.425**±0.007	0.401	**0.460**±0.009	0.450
COI	known sp. → species	0.988±0.000	**0.990**	0.942±0.003	**0.950**	0.830±0.034	**0.886**
COI	novel sp. → genus	**0.620**±0.002	0.615	0.506±0.003	**0.514**	0.251±0.024	**0.359**
COI	novel gen. → family	**0.603**±0.004	0.597	**0.675**±0.002	0.669	0.362±0.028	**0.380**
12S	known sp. → species	0.778±0.001	**0.788**	0.427±0.002	**0.441**	0.429±0.007	**0.454**
12S	novel sp. → genus	0.803±0.001	**0.811**	0.401±0.002	**0.402**	0.449±0.002	**0.466**
12S	novel gen. → family	**0.730**±0.001	0.715	**0.308**±0.000	0.299	0.386±0.001	**0.391**

Table F.2: The same placement weights with three retrieval engines: joint coverage at the deployed gate, mean $\pm$ CI_{95} over three seeds. TAXOSET is the VSEARCH-trained network of Table 2; the hits come from the named engine, and only the gate quantiles are re-estimated on that engine's known queries. Best-hit is the best-hit expansion rule on identical hits. Bold marks the larger mean within each engine. Table F.3 gives the bound every cell lives under.

		VSEARCH		BarcodeBERT		NT-500M	
Marker	Group	TaxoSet	best-hit exp.	TaxoSet	best-hit exp.	TaxoSet	best-hit exp.
ITS	novel sp.	**0.281±0.009**	0.229	**0.044±0.003**	0.026	**0.053±0.006**	0.024
ITS	novel gen.	**0.694±0.006**	0.425	**0.086±0.007**	0.026	**0.108±0.022**	0.049
ITS	novel fam.	**0.572±0.002**	0.243	**0.074±0.006**	0.020	**0.122±0.026**	0.040
12S	novel sp.	**0.209±0.006**	0.188	**0.015±0.003**	0.009	**0.027±0.002**	0.015
12S	novel gen.	**0.639±0.007**	0.455	**0.045±0.005**	0.015	**0.069±0.007**	0.032
12S	novel fam.	**0.654±0.003**	0.456	**0.078±0.002**	0.015	**0.129±0.010**	0.038
COI	novel sp.	**0.669±0.007**	0.592	**0.451±0.019**	0.325	**0.120±0.014**	0.101
COI	novel gen.	**0.709±0.000**	0.580	**0.747±0.030**	0.555	0.163±0.081	**0.176**
COI	novel fam.	**0.918±0.000**	0.799	**0.832±0.001**	0.557	0.186±0.097	**0.331**

G AN EXTERNAL BENCHMARK

Every other comparison in this paper scores baselines we retrained, on splits we built. Here we test TAXOSET on the independent `taxclass` benchmark of Orsholm et al. (2026), keeping its references, queries, taxonomy, definition of novelty and published predictions: twelve competing methods on COI and nine on ITS.

Table F.3: Recall@20 of each engine at the deepest reference-supported rank: the share of queries whose true taxon at that rank is among the 20 hits. Retrieval is deterministic, so there is no seed spread.

Marker	Group	VSEARCH	BarcodeBERT	NT-500M
ITS	known sp.	0.931	0.686	0.723
ITS	novel sp. → genus	0.932	0.776	0.844
ITS	novel gen. → family	0.813	0.484	0.611
12S	known sp.	0.907	0.541	0.582
12S	novel sp. → genus	0.916	0.553	0.632
12S	novel gen. → family	0.864	0.533	0.618
16S	novel sp. → genus	0.879	0.589	0.585
16S	novel gen. → family	0.923	0.629	0.670
COI	known sp.	0.998	0.981	0.895
COI	novel sp. → genus	0.718	0.669	0.512
COI	novel gen. → family	0.724	0.809	0.602

Benchmark. Novelty comes from a real change of data source. For COI, training barcodes are Finnish and test barcodes German; for ITS, training sequences come from culture collections and test sequences from UNITE (Abarenkov et al., 2024). The references hold 35,556 COI and 10,776 ITS sequences, the test sets 26,901 and 17,226. The taxonomy names 193,730 and 38,959 complete lineages, most of which the reference does not contain. Grouping queries by the deepest rank absent from the reference, the rule of Section 2, gives both markers a known group, so the threshold calibrates on both. Its seven ranks are not ours: for COI they run from class to species, so what we call the family rank is its tribe, while genus and species coincide.

Protocol. We convert the benchmark to the layout of Section 3.3, retrieve its test sequences against its training set with the same VSEARCH settings, and train TAXOSET from scratch, hyperparameters unchanged, three seeds. We reuse no data from our own corpora. Published predictions are read as distributed, from the column each file names `genus`; a method whose file cannot be parsed is dropped rather than scored, so a parse failure never reads as a low accuracy. Every method is scored over the whole query group, so an abstention, or a query that retrieved nothing, counts as wrong.

Table G.1: TAXOSET on the `taxclass` benchmark of Orsholm et al. (2026), whose reference, queries, hierarchy and novelty are all external to this paper; the network is trained on that benchmark's training set alone. Reference-backed rows give top-1 genus accuracy, mean ± CI_{95} over three seeds, beside the best published method scored on the same queries with the same rule. Open-reference rows give end-to-end coverage, routed as novel and covered by the complete set. The bound column is that coverage before the gate; it needs no training, so the gap between the two is what the network is responsible for. Every published method is at a structural zero on these rows. Abstentions count as errors throughout.

marker	cell	n	TaxoSet	best published	ungated bound	median $\lvert U \rvert$
COI	novel sp. → genus	9,435	**0.792**±0.001	0.769 (IDTAXA)	–	–
COI	novel gen. → genus	3,780	**0.695**±0.003	0.001 (PROTAX)	0.721	122
ITS	novel sp. → genus	7,056	**0.799**±0.004	0.817 (SINTAX)	–	–
ITS	novel gen. → genus	4,500	**0.822**±0.002	0.000	0.851	37

Results. Table G.1 gathers the four cells. When the genus is in the reference, TAXOSET ranks first on COI with 0.792, against 0.769 for IDTAXA (Murali et al., 2018). On ITS it scores 0.799 and ranks fourth of ten, 1.8 points behind SINTAX (Edgar, 2016) and within half a point of RDP-NBC and IDTAXA. Table G.2 gives the whole field. Four methods sit within two points of each other at the top of the ITS panel: at this rank the task is close to solved, and the room to tell methods apart is thin.

When the genus is absent, published methods identify four of 3,780 COI queries and none of 4,500 ITS queries: the structural zero of Section 2, measured on somebody else's benchmark. TAXOSET reaches end-to-end coverage of 0.695 on COI with a median set of 122 names, and 0.822 on ITS with a median set of 37. Before novelty routing, retrieval and the hierarchy alone place the true

Table G.2: Every method of the `taxclass` benchmark on its reference-backed cell: top-1 genus accuracy on queries whose species is absent from the reference but whose genus is present. Scored over the whole group, so an abstention counts as wrong. Published predictions are read as distributed; TAXOSET is the mean over three seeds. BLAST names no genus on 88% of the COI queries, which is why it sits so low: it declines rather than errs. Methods, in the order they first appear: (Murali et al., 2018; Barbera et al., 2019; Edgar, 2016; Wang et al., 2007; Romeijn et al., 2024; Somervuo et al., 2016; Zito et al., 2023; Altschul et al., 1990; Lanzén et al., 2012).

COI, 9,435 queries		ITS, 7,056 queries	
method	genus acc.	method	genus acc.
TaxoSet	0.792	SINTAX	0.817
IDTAXA	0.769	RDP-NBC	0.804
EPA-ng (taxtree)	0.762	IDTAXA	0.803
EPA-ng (phyltree)	0.760	TaxoSet	0.799
SINTAX	0.734	MycoAI-BERT	0.655
EPA-ng (freetree)	0.693	BayesANT	0.611
RDP-NBC	0.683	BLAST	0.560
MycoAI-BERT	0.658	PROTAX	0.512
PROTAX	0.627	MycoAI-CNN	0.397
MycoAI-CNN	0.505	CREST4	0.206
BayesANT	0.419		
BLAST	0.112		
CREST4	0.111		

genus in the set for 0.721 and 0.851 of queries, a quantity no training touches; routing costs the remaining 2.6 and 2.9 points. As in Section 4.3, most of the gain comes from making absent names available as candidates, and the network is what decides when to use them, at a held false-positive rate.