Pretrained deep-learning ITS classifiers read the flanking regions, not the ITS2 barcode, and so fail on the amplicon that environmental fungal surveys sequence

Aaron O'Brien*[1], César Marín[2], and Pilar Parada[1]

[1]Centro de Biotecnología de Sistemas, Universidad Andrés Bello, Santiago, Chile
[2]Centro de Investigación e Innovación para el Cambio Climático (CiiCC), Facultad de Ciencias, Universidad Santo Tomás, Valdivia, Chile

Abstract

Deep-learning classifiers for the fungal internal transcribed spacer (ITS) report accuracies above 90% and are increasingly proposed for environmental metabarcoding. We benchmark two pretrained models, a convolutional network and a transformer sharing a training corpus of 5.23M sequences, against two established k-mer methods on 5,222 identical queries, one per genus, evaluated on both full-length ITS and the ITS2 subregion that dominates environmental sequencing. The design favours the classifiers: queries are drawn from the same public dataset they were trained on and stratified by whether a query's genus lies in their own label space, recovered from the distributed models, while the reference the k-mer methods consult mirrors that label space and excludes the queries themselves. Even so, on full-length ITS both classifiers are outperformed by both classical methods at every rank and in both strata: SINTAX recovers the correct family for 92.0% of seen-genus and 70.3% of novel-genus queries and best-hit alignment against a 56,327-sequence reference for 92.3% and 67.2%, against 77.9% and 57.7% for the transformer and 76.5% and 54.3% for the convolutional network. A hierarchical logistic regression on k-mer counts, fitted in ten minutes to 1.07% of the MycoAI training corpus, also exceeds both and places novel genera better than either search method, and refitted on ITS2 it recovers 89.2% of seen-genus families on that amplicon against 88.6% for best-hit alignment, so neither learned classification nor the amplicon is what fails. Restricting the identical records to ITS2 costs the k-mer methods 3.5 and 3.7 percentage points of seen-genus family accuracy but costs the classifiers 49.6 and 58.0, reducing them to 28.3% and 18.5%. An ablation identifies the cause. Grafting each query's unaltered ITS2 between the flanking regions of a donor record from a different phylum returns the *donor's* family for 34.4% of queries against the query's own for 4.2% in the convolutional model, and 63.7% against 0.8% in the transformer, from a baseline of 0.1% where no donor sequence is present. The models therefore read taxonomy principally from the flanking regions rather than from the ITS2 barcode, which explains the collapse and predicts the same failure for any subregion amplicon. Compounding this, on ITS2 the classifiers' output probability all but ceases to separate novel from known genera (AUROC 0.541 and 0.503, the latter at chance, against 0.866 for alignment identity and 0.785 for the SINTAX bootstrap), so the failure is not detectable from the models' own output. We recommend that reported accuracies for such models specify the amplicon region of the evaluation, state the length distribution of the training corpus, and include a same-query classical baseline.

*Corresponding author: aaron.obrien@unab.cl

1 Introduction

Metabarcoding of the fungal ITS region routinely recovers sequences with no close match in curated references such as UNITE [13], environmental "dark matter" novel at the genus level or below, and in many surveys such sequences are the majority of what is recovered [19]. Assigning taxonomy to this material is the central computational problem of environmental mycology, and a growing body of work applies deep learning to it [18, 12, 4, 7], reporting genus- and species-level accuracies above 90%, with independent benchmarking on full-length long-read ITS finding these models competitive with established methods for recovering community composition [16].

Those figures are difficult to translate into an expectation for environmental data, for two reasons. First, they are typically obtained under random cross-validation in which test sequences are drawn from the same labelled pool, frequently the same genera, as the training data. The extent of that overlap has been quantified: for the largest of the three MycoAI test sets, comprising 363,420 barcodes and the one on which overall performance is usually reported, the overlap analysis reported by Gao and colleagues finds complete overlap with the training set on both measures they report: 363,420 of 363,420 test barcodes are identical to a training barcode and 14,742 of 14,742 test species are present in training, each 100.00% ([7], Table 6). Their two smaller test sets overlap at 86.73% and 6.48% of barcodes, and they describe the two high-overlap sets as measuring performance on in-distribution sequences, naming the 6.48% set as the rigorous test of generalization. Where generalization is tested at all it is tested at the level of unseen species, with the query's genus and family still represented in training. The regime that characterizes dark matter, a sequence whose entire genus is absent, is not evaluated. Second, and less often remarked, these accuracies are obtained on full-length reference sequences, whereas environmental surveys sequence a primer-bounded subregion, most commonly ITS2. Whether performance on the former predicts performance on the latter has not been established.

We address both by benchmarking two pretrained classifiers against two long-standing k-mer methods, best-hit alignment and SINTAX, under conditions designed to be symmetric. Queries are drawn one per genus from the public release the classifiers were trained on, and stratified by whether their genus lies in the classifiers' own label space, recovered directly from the distributed models. The reference the k-mer methods consult is built to mirror that label space, so a seen-genus query has same-genus references available exactly as the classifiers had same-genus training data, and a novel-genus query has neither. The entire comparison is then run twice, on full-length ITS and on the ITS2 subregion of the identical records, and an ablation on the flanking regions establishes why the two differ.

The question is not untouched. Fujisawa and Imai [10] benchmark conventional and deep-learning methods for insect DNA barcoding under incomplete reference databases and report that detecting species absent from the reference is substantially harder than identifying those present, while finding both classes of method accurate for assignment itself. Our study differs in marker, kingdom and, most consequentially, in comparing amplicon regions, and we reach a stronger conclusion about the classifiers: on our data they lose to k-mer methods on assignment as well. Whether that difference reflects the marker, the reference, or the region mismatch we identify below is not settled by either study. A separate strand builds novelty handling into the model, either through Bayesian nonparametric priors that admit unobserved taxa at each rank [21] or through deep hierarchical Bayesian formulations [5]; a related line uses DNA as side information for zero-shot recognition of unseen taxa [6]. We do not evaluate these, and our conclusions concern the pretrained fixed-label classifiers currently offered for fungal ITS. Finally, two lines of work bound what any method can achieve on this marker and so bound our own baselines. dnabarcoder [20] showed that similarity cutoffs for fungal identification vary substantially between clades and predicted clade-specific cutoffs accordingly, which is why we treat a single global cutoff as a baseline to be interrogated rather than a standard to be defended; and the accuracy and precision of ITS-based identification have been assessed directly [11], which

is part of why we score recovery at family rank throughout rather than treating genus or species as the target.

Our results are unfavourable to the classifiers in a specific and, we think, consequential way. They are already outperformed by both classical methods on full-length ITS. On ITS2 they lose between 49.6 and 58.0 percentage points of family accuracy where the classical methods lose between 3.5 and 3.7, and the ablation locates the reason in a dependence on the flanking regions so strong that supplying the flanks of an unrelated phylum causes the transformer to return that phylum's family for most queries. And on ITS2 their confidence scores all but cease to carry novelty signal, so a practitioner receives confident names, mostly incorrect, with no internal indication of the problem. A companion study [14] addresses the complementary question of how abstention can be calibrated once a suitable score exists.

2 Methods

2.1 Classifiers and recovery of their label space

We evaluated the two architectures distributed with MycoAI [18]: a multi-head convolutional network (MycoAI-CNN, 246 MB) and a transformer (MycoAI-BERT, 1.1 GB), both pretrained on the UNITE+INSD dataset as released in 2023 [1], approximately 5.23M sequences after preprocessing [3], spanning 791 families and 3,695 genera, and both obtained from Zenodo record 10904344. A fixed-label classifier can be leakage-controlled only with respect to its own training set, since a genus cannot be withheld from an already-trained softmax head. We therefore recovered each model's label space directly from its per-rank label encoders rather than reconstructing a training set. The two label spaces were identical at every rank, so a single query set, stratification and reference serve both and the architectures are compared pairwise. Recovering the label space from the distributed model makes the stratification exact and independent of reference release, and the resulting taxon lists are deposited with the harness.

Two properties of the distributed artefacts bear on the ITS2 arm below and are worth stating before the results rather than after them. First, the training corpus consists almost entirely of full-length sequences, which carry the ITS1 spacer and the conserved 5.8S gene that a primer-bounded ITS2 amplicon does not. We describe these throughout as the flanking regions rather than as conserved context, since ITS1 is hypervariable and is itself a standard fungal barcode; what distinguishes them from ITS2 here is position in the training record, not conservation. The corpus is also markedly longer than an ITS2 amplicon: MycoAI preprocessing excludes sequences more than four standard deviations from a mean length of 558.0 bp (SD 126.2), as reported for that dataset by Gao and colleagues [7], against a median of 169 bp for the ITS2 queries. We note the relevant difference is compositional rather than metric, since Section 3.3 shows that restoring length alone does not restore accuracy. Second, neither the MycoAI publication [18] nor the distributed package scopes the models to full-length input, states a minimum query length, or warns on short queries; the `mycoai-classify` interface accepts primer-bounded ITS2 without error and returns a family probability for every sequence submitted. The ITS2 arm is therefore a use the distributed artefact permits and does not flag, which is the sense in which its behaviour there is a property of the released tool and not only of the architecture. Two further properties of the package belong in that account, since a paper about what practitioners would download should say what downloading it entails. The package contacts a telemetry service on import and raises rather than proceeding when no credentials are present, so it cannot be run in a non-interactive shell without disabling that call. And the distributed checkpoints cannot be loaded by PyTorch 2.6 or later, in which `torch.load` defaults to refusing to unpickle arbitrary classes; loading them requires overriding that default, which the package does not do. Neither affects anything reported below, all of which was produced under the pinned environment of §5, but both are met before a first prediction.

2.2 Query construction

Queries were drawn from the UNITE+INSD 2024 ITS reference distributed with `dnabarcoder` [20] (Zenodo 13336328), built from the 21.04.2024 release [2]. That is one release later than the one the classifiers were trained on, which bears on how the seen-genus stratum should be read: a seen-genus query is a sequence whose genus the models learned, but not necessarily a sequence they saw, so seen-genus recovery is a same-genus generalization test rather than a test of recall. The stratification itself is unaffected, being taken from the models' own label encoders rather than from any release. A query whose genus lies in the classifiers' label space is *seen*; one whose genus is absent is *novel-to-model*. Exactly one sequence was taken per genus, eliminating genus-level pseudoreplication so that each genus contributes a single independent test of placement and no genus is over-represented by sequence abundance. This yields 3,330 seen-genus and 1,892 novel-to-model queries. A further 252 genera fell outside the models' family space and were excluded, correct family placement being impossible by construction. Because one-per-genus sampling removes the abundance weighting a random draw from UNITE would carry, absolute recovery under this design is expected to fall below published accuracies obtained on abundance-weighted test sets; the comparisons of interest are between strata, between methods and between loci, all of which share the design.

2.3 Paired loci

Because ITS2 is shorter than the full ITS region on which the classifiers were trained, the comparison was run over two paired arms spanning the identical 5,222 records: the full ITS sequences (median query length 562 bp), and the ITSx-extracted [8] ITS2 subregion of those same records (median 169 bp), both distributed with the same release. Restricting the second arm to the record identifiers of the first makes region the only variable, so that any difference is attributable to the amplicon window rather than to taxon sampling.

2.4 Reference construction for the k-mer methods

To equalize the taxonomic knowledge available to all methods, the reference was built so that its coverage mirrors the classifiers' label space: a record was included if and only if its genus lies in that label space and it is not itself one of the queries, capped at twenty sequences per genus. This gives 56,327 sequences across 3,346 genera for the full-ITS arm and 54,108 for the ITS2 arm. Leakage control is thereby symmetric and automatic. A seen-genus query has same-genus references available, mirroring the classifiers having trained on that genus, while a novel-genus query has none, its genus lying outside the label space and therefore outside the reference by construction. To establish that the comparison does not depend on the cap, the full-ITS reference was rebuilt and rescored at five and fifty sequences per genus, giving references of 16,352 and 113,455 sequences.

2.5 Methods compared

Best-hit alignment. Queries were searched with `vsearch` [17] (`usearch_global`, best hit, identity floor 0.5, both strands) and assigned the taxonomy of the best hit; best-hit percentage identity served as its score. Of 5,222 queries, 5,220 obtained a hit on full ITS and 5,177 on ITS2; those without were scored as family-incorrect and maximally novel, no reference match meaning no placement.

SINTAX. We scored SINTAX [9] as implemented in `vsearch`, a k-mer bootstrap consensus classifier standard in metabarcoding pipelines, which abstains by withholding assignments below a confidence threshold. The identical reference was rewritten with full lineage headers, retaining 55,924 of 56,327 records on full ITS and 53,742 of 54,108 on ITS2, the 0.7% difference being records lacking a complete kingdom-to-genus lineage. Queries were classified with

`-sintax_cutoff 0.8`, the conventional setting, and per-query family-rank bootstrap confidence served as its score.

HiTaC. To separate the effect of learned classification from that of the sequence representation, we additionally trained HiTaC [12], a hierarchical classifier that fits a local logistic regression per parent node of the taxonomic tree on 6-mer frequency features. Unlike the MycoAI models, HiTaC is fitted by the user rather than distributed pretrained, so we trained it on the same label-space-mirrored reference the k-mer methods consult. Its label space is therefore that reference's taxa, novel-genus queries are novel to it by construction, and leakage control is symmetric with the other methods without any need to recover a training taxon list. We fitted it at two reference depths, 16,223 and 55,924 sequences after lineage completion, corresponding to 0.31% and 1.07% of the MycoAI training corpus, in approximately ten minutes each on eight CPU threads. The reference was rewritten with a terminal semicolon for HiTaC, whose header parser expects one and without it silently truncates the final rank label by one character. Every family value reported here is unchanged by that correction, family not being the terminal field.

Its confidence score requires a separately fitted hierarchical filter, which we did not train, so HiTaC contributes no novelty AUROC.

Classifiers. Both models were run through the distributed `mycoai-classify` interface, and the family-rank output probability P(family) served as each model's score. Inference on the 5,222 queries took approximately 2 seconds for the convolutional network and 295 seconds for the transformer on the same CPU.

2.6 Flank ablation

To test directly whether the classifiers depend on the regions flanking ITS2 rather than on ITS2 itself, we constructed seven arms over the same records, all sharing the query identifiers of the full-ITS arm so that a single truth table scores every arm. Each record's ITS2 was located within its own full ITS by substring match, succeeding for 5,221 of 5,222 records and yielding a median $5'$ flank of 360 bp, comprising ITS1 and 5.8S, and a median $3'$ flank of 26 bp, an LSU remnant. Region boundaries for the component arms were taken from ITSx [8] (v1.1.3, fungal profiles), which annotated ITS2 for 5,214 records and agreed with the release's own ITS2 boundary exactly for 5,188 of 5,212 comparable records; median ITS1 is 181 bp and median 5.8S 158 bp, so arm (iv) supplies roughly twice as much donor ITS1 as donor 5.8S and cannot on its own say which the models follow.

The arms are: (i) ITS2 alone; (ii) ITS2 extended on its $5'$ side with random `ACGT` sequence to the training-corpus mean of 558 bp, which adds length without adding information; (iii) ITS2 with its own flanks, which is the full-ITS arm; (iv) ITS2 grafted between the flanks of a donor record drawn from a different phylum, which adds genuine flanking context belonging to the wrong taxon; and three component arms that add back one native region at a time, (v) the record's own 5.8S followed by its own ITS2, (vi) its own ITS1 followed by its own ITS2, and (vii) its own ITS1 alone with no ITS2 at all. Sixteen phyla were available as donors and no donor shared its query's phylum. Arms (v) to (vii) use the query's own sequence throughout rather than a donor's, so they measure how much of full-length performance each region restores rather than which lineage a prediction follows; the two questions are separate and arm (iv) answers the second.

For arm (iv) each prediction was scored against the donor's lineage as well as the query's. Because the query's ITS2 is present and unaltered while only the flanks belong to the donor, agreement with the donor measures which part of the input the prediction follows, rather than inferring it from an accuracy profile. Arm (i) supplies the chance baseline for donor agreement, no donor sequence being present there.

Arms (ii) and (iv) are constructed inputs that no sequencing workflow would generate. Their purpose is diagnostic, and neither is an estimate of field performance. Arm (ii) in particular

confounds two things, since random sequence is both uninformative and potentially misleading, so its result should not be read as a pure length control.

2.7 Robustness to the ITS2 delimitation

The ITS2 arm depends on where the ITS2 boundary was drawn, which was done upstream by the release's authors. To test whether the collapse is an artefact of that choice, ITS2 was independently re-extracted from the same full-ITS queries with `ITSxRust` [15] (v0.2.2, `-region its2`, bundled fungal HMM profile set, default inclusion E-value 10^{-5}, no platform preset), and both extractions were classified on the identical subset of records. Because the queries are ITS-only sequences with a median 3′ LSU remnant of 26 bp, the HMM anchoring required to bound the 3′ end of ITS2 succeeded for 2,637 of 5,222 records, of which 1,754 derived from partial anchor chains; the comparison is therefore restricted to those 2,637 records and its absolute values are not directly comparable with those from the full query set. The corresponding competing interest is declared below.

2.8 Scoring and diagnostics

Family-level recovery was scored within each stratum by exact match against the reference taxonomy. Because all five methods were scored on the identical queries, differences between methods within a stratum were tested with McNemar's test on the discordant pairs, using the exact binomial form since some discordant counts are small, with Holm adjustment across the forty comparisons, ten pairs of methods in each of the two strata at each of the two loci, all at family rank. Novelty separability of each method's own score was quantified as the area under the ROC curve separating novel-genus from seen-genus queries, with percentile confidence intervals from 2,000 bootstrap resamples stratified on the two classes. As diagnostics distinguishing out-of-domain input from taxonomic failure we additionally report a per-rank accuracy profile from phylum to genus, and the internal coherence of each prediction, meaning whether the predicted family belongs to the predicted order in the reference taxonomy. The seen-genus stratum functions as a positive control throughout: its sequences come from the database the classifiers were trained on and their genera lie in the label space, so low recovery there indicates that the input, not the taxonomy, is out of distribution.

3 Results

3.1 Classical methods outperform the classifiers on full-length ITS

On full-length ITS, SINTAX recovered the correct family for 92.0% of seen-genus and 70.3% of novel-genus queries and best-hit alignment for 92.3% and 67.2%, against 77.9% and 57.7% for the transformer and 76.5% and 54.3% for the convolutional network (Table 1, Figure 1a,b). Genus-level recovery on seen genera placed the three leading methods well ahead of both distributed classifiers, HiTaC first at 84.8% against 84.4% for SINTAX and 84.3% for alignment, and 67.3% and 71.7% for the transformer and the convolutional network, although the two classifiers exchange places relative to their family ordering. A trained hierarchical classifier fitted to the same reference matched alignment exactly on seen genera at family rank (92.3%), led at genus rank and shared the top of the novel-genus ranking with SINTAX (71.4% against 70.3%, not distinguishable), exceeding best-hit alignment, having been fitted in about ten minutes to 1.07% of the MycoAI training corpus.

The comparison is therefore not between learned and classical methods, and framing it that way would misdescribe the result. Learned classification works well here: HiTaC is a trained, fixed-label classifier and it placed novel genera better than either search method. What underperforms is the pair of distributed MycoAI models, which are exceeded on full-length ITS

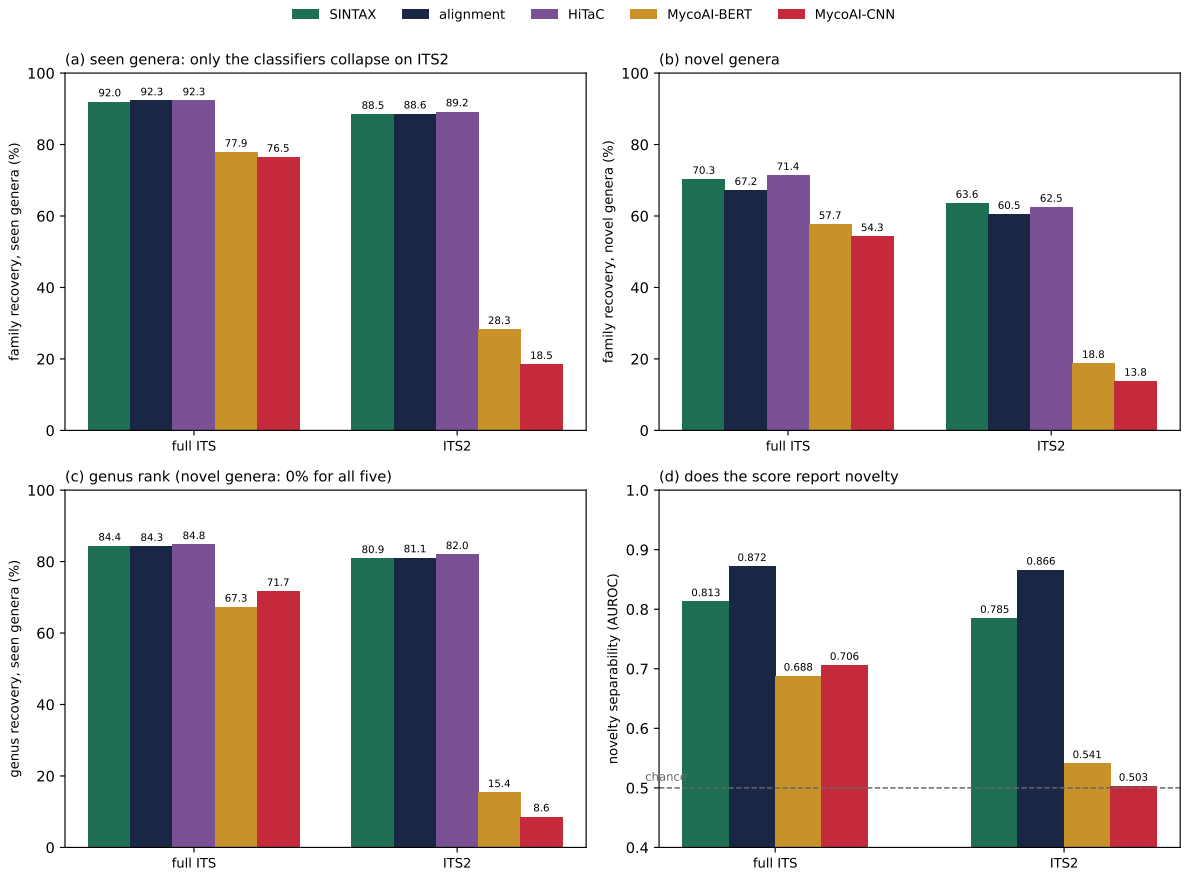

Figure 1: Five methods on 5,222 identical UNITE records, one per genus, stratified by whether a query's genus lies in the classifiers' training label space (3,695 genera, recovered from the distributed models) and correspondingly in the reference the k-mer methods consult. Each panel compares full-length ITS with the ITS2 subregion of the same records. (a) Family recovery on seen genera. (b) Family recovery on novel genera. (c) Genus recovery on seen genera; on novel genera it is exactly zero for all five methods at both loci, the correct label being absent from vocabulary and reference alike. (d) Novelty separability of each method's own score, with chance marked, for the four methods that report one; HiTaC's requires a filter we did not fit (§2.5).

by best-hit alignment, by SINTAX, and by a hierarchical logistic regression on k-mer counts fitted to one percent of their training data. That is a narrower claim than the one a learned-versus-classical framing would support, and a harder one to attribute to the choice of baseline.

Thirty of the forty pairwise contrasts survive McNemar's test after Holm adjustment. Nine of the ten that do not involve only the three leading methods, which on seen genera are mutually indistinguishable at both loci: 92.0%, 92.3% and 92.3% on full ITS and 88.5%, 88.6% and 89.2% on ITS2, every pairwise adjusted p equal to 1.00, with alignment against HiTaC splitting the discordant pairs exactly 75 to 75 at full length. On novel genera the result is a partial ordering rather than a tie. At full length SINTAX and HiTaC are indistinguishable (70.3% against 71.4%, adjusted $p = 1.00$) and both exceed alignment (1.9×10^{-6} and 1.8×10^{-5}); on ITS2 SINTAX exceeds alignment (9.7×10^{-7}) while HiTaC is distinguishable from neither (1.00 and 0.34). Every contrast involving a MycoAI model is significant in every cell at both loci. The tenth undistinguished contrast is the transformer against the convolutional network on seen genera at full length, 77.9% against 76.5%, adjusted $p = 0.071$.

Two things follow. Where a same-genus reference is available the three leading methods are interchangeable, at either locus, so the small gaps between them are not interpreted. Where

Table 1: Five methods on identical queries across both loci. Highest value in each column and locus in bold, which is not the same as a demonstrated lead: on seen genera the three leading methods are not distinguishable at either locus, and genus-rank contrasts were not tested (§3.1). Every method is given the resources available at the locus tested: the search methods a reference extracted at that locus, HiTaC a fit at it (§3.7), while the two MycoAI models are single distributed artefacts and are the only methods here that cannot be given it.

locus	method	family seen genus	family novel genus	genus seen genus	novelty AUROC
full ITS	SINTAX	92.0%	70.3%	84.4%	0.813
	alignment	**92.3%**	67.2%	84.3%	**0.872**
	HiTaC	**92.3%**	**71.4%**	**84.8%**	n/a
	MycoAI-BERT	77.9%	57.7%	67.3%	0.688
	MycoAI-CNN	76.5%	54.3%	71.7%	0.706
ITS2	SINTAX	88.5%	**63.6%**	80.9%	0.785
	alignment	88.6%	60.5%	81.1%	**0.866**
	HiTaC	**89.2%**	62.5%	**82.0%**	n/a
	MycoAI-BERT	28.3%	18.8%	15.4%	0.541
	MycoAI-CNN	18.5%	13.8%	8.6%	0.503

Genus recovery on novel genera is exactly 0.0% for every method at both loci and is omitted. HiTaC's novelty score requires a separately fitted filter which we did not train, so no AUROC is reported. Restricting the identical records to ITS2 costs the k-mer search methods 3.5 to 6.7 percentage points of family recovery and HiTaC, refitted at the locus, 3.1 to 8.9, against 38.9 to 58.0 for the two MycoAI models. HiTaC's off-locus values on ITS2, 47.4% and 30.8%, are in Table 5. Method contrasts were tested with McNemar's test on the discordant pairs with Holm adjustment across all forty comparisons, on the 5,221-query intersection of the five per-query tables, which is why percentages here may differ by up to 0.1 from those computed on the full query set. Contrasts were tested at family rank only. Thirty reach significance; the ten that do not are given in §3.1. The two ITS2 classifier values are the only ones near chance, and are the only ones for which an interval is quoted (§3.5).

it is not, the two that aggregate evidence across several relatives, SINTAX by bootstrapping k-mer subsamples to a consensus and HiTaC by fitting a classifier over k-mer frequencies, both exceed best-hit alignment, which commits to a single neighbour. That distinction matters most precisely when no same-genus relative exists, and it is the one place in the study where the search-based and classifier-based approaches separate in favour of the latter.

Degradation from seen to novel genera was of similar magnitude for every method, 21.7, 25.1, 20.9, 20.2 and 22.1 percentage points for SINTAX, alignment, HiTaC, the transformer and the convolutional model, so the MycoAI models' disadvantage is close to a constant offset rather than something specific to novel taxa. Capacity helped, but not usefully: the transformer exceeded the convolutional model in all four combinations of locus and stratum, by 1.4, 3.4, 9.8 and 5.0 percentage points, significant after adjustment in three of them, the exception being the 1.4-point gap on seen genera at full length (adjusted $p = 0.071$). It did so while being roughly four and a half times larger and about 140 times slower at inference, while remaining worse at genus rank on seen genera, and while closing none of the gap to the k-mer methods, whose advantage over it remains 14 points on full ITS and 60 on ITS2.

3.2 On the amplicon surveys actually sequence, only the classifiers collapse

Environmental fungal metabarcoding is dominated by primer-bounded ITS2 amplicons, so we repeated the comparison on the ITS2 subregion of the identical records with a reference built

by the same membership rule. The methods separate decisively (Table 1, Figure 1). Seen-genus family recovery fell by 3.5 percentage points for SINTAX, 3.7 for alignment and 3.1 for HiTaC refitted at the locus, but by 49.6 for the transformer and 58.0 for the convolutional network, reaching 28.3% and 18.5%. Novel-genus recovery behaved the same way, falling 6.7 points for both k-mer methods against 38.9 and 40.5 for the classifiers. Genus recovery on seen genera fell from 84.3% to 81.1% for alignment but from 71.7% to 8.6% for the convolutional network.

Because the records, the stratification and the reference construction rule are common to every method, this is not a property of the shorter amplicon. ITS2 carries sufficient information for all three methods that can be given it to retain 96.2%, 96.0% and 96.6% of their full-length family accuracy on seen genera, and between 88% and 91% on novel genera; what fails is the two distributed models' ability to use it. The per-rank profile locates the failure at the input. On ITS2, convolutional accuracy decayed monotonically from phylum (77.3%, barely above the accuracy of always naming the most abundant fungal phylum) through class (52.9%) and order (30.4%) to family (18.5%), with only 49.6% of predictions internally coherent in the sense that the predicted family belongs to the predicted order. The transformer behaved likewise (76.5%, 60.7%, 41.4%, 28.3%, coherence 55.0%). The corresponding full-ITS figures are 98.8%, 95.3%, 87.5%, 76.5% with coherence 90.5%, and 98.2%, 95.8%, 89.3%, 77.9% with coherence 92.2%. A model receiving an input outside its training domain does not degrade gracefully towards the correct higher ranks; it produces lineages whose own ranks contradict one another.

Query length within the ITS2 arm does not explain the effect and the two models do not even agree on its direction: family recovery across length quartiles rose then plateaued for the convolutional model (11.2, 21.7, 18.3, 22.6 per cent) and declined for the transformer (33.0, 34.1, 27.6, 18.9). No coherent account of the collapse follows from length in either direction, and the ablation below shows why: length is not the operative variable.

3.3 The classifiers read the flanking regions rather than the barcode

Section 3.2 locates the failure in the input without establishing which property of the input is responsible, and the training-corpus composition noted in Section 2.1 suggests the missing flanking regions. The ablation tests that directly and supports a stronger conclusion (Table 2).

Extending ITS2 with uninformative sequence to the training-corpus mean length made matters worse rather than better, reducing seen-genus family recovery from 18.5% to 6.0% in the convolutional model and from 28.3% to 7.5% in the transformer. Length is therefore not the operative variable, which also accounts for the inconsistent length gradients reported above.

Supplying genuine flanking context from the wrong taxon was worse still, reducing family recovery to 4.8% and 1.0% and phylum recovery to 36.4% and 11.7%, in both cases below the values obtained from ITS2 with no flanks at all. A model merely lacking context would be expected to perform no worse when context is restored, whatever its provenance. These models perform much worse, which indicates that the flanks are not supplementary but decisive. Prediction coherence moves in the opposite direction to accuracy, rising to 61.9% and 82.4% against 49.6% and 55.0% on ITS2 alone: the outputs become internally more consistent as they become less correct, which is the signature of a model confidently reporting a lineage, though not the query's.

Scoring arm (iv) against the donor's lineage identifies whose (Table 3). Agreement with the donor rises from between 0.1 and 0.7 per cent, where no donor sequence is present, to roughly a third at every rank for the convolutional model and to between 37 and 72 per cent for the transformer. The effect is markedly stronger in the transformer, and the two models differ in kind at phylum rank, where the convolutional model divides evenly between query and donor (37.8% each) while the transformer follows the donor almost seven times more often than the query. At family rank, the rank this benchmark concerns, both follow the donor predominantly. At family rank the transformer returns the donor's family for 63.7% of queries and the query's own for 0.8%, a ratio of roughly eighty to one, on sequences whose ITS2 is present and correct.

Table 2: Flank ablation, seen-genus stratum. Every arm covers the same records, 5,222 for arms (i) to (iii) and 5,221 for arm (iv), one record having no locatable flanks, and except where padding is added contains the same unaltered ITS2 sequences. Coherence is the proportion of predictions whose predicted family belongs to the predicted order.

model	arm	family recovery	phylum recovery	prediction coherence
MycoAI-CNN	(i) ITS2 alone	18.5%	77.3%	49.6%
	(ii) ITS2 padded to 558 bp	6.0%	62.7%	46.1%
	(iii) ITS2 with its own flanks	**76.5%**	**98.8%**	**90.5%**
	(iv) ITS2 with donor flanks	4.8%	36.4%	61.9%
	(v) own 5.8S + own ITS2	3.6%	78.8%	53.2%
	(vi) own ITS1 + own ITS2	8.6%	66.6%	48.0%
	(vii) own ITS1 alone	15.7%	68.7%	49.2%
MycoAI-BERT	(i) ITS2 alone	28.3%	76.5%	55.0%
	(ii) ITS2 padded to 558 bp	7.5%	50.6%	47.3%
	(iii) ITS2 with its own flanks	**77.9%**	**98.2%**	**92.2%**
	(iv) ITS2 with donor flanks	1.0%	11.7%	82.4%
	(v) own 5.8S + own ITS2	16.4%	93.4%	58.1%
	(vi) own ITS1 + own ITS2	11.2%	64.6%	49.7%
	(vii) own ITS1 alone	22.5%	74.1%	52.7%

Novel-genus family recovery follows the same ordering: 13.8, 3.9, 54.3, 3.2, 2.1, 6.3 and 9.5 per cent for the convolutional model and 18.8, 5.2, 57.7, 0.5, 11.6, 7.0 and 13.6 for the transformer. Arms (i) to (iii) cover 5,222 records and arm (iv) 5,221. The component arms cover the records ITSx annotated for the region in question: 5,210 for (v) and (vii), 5,213 for (vi). Percentages are therefore computed on seen-genus strata of 3,330, 3,321 and 3,323 respectively, a difference of under 0.3% of the stratum.

Arms (v) and (vi) add the query's own region, not a donor's. Neither is a sequence any protocol produces: ITS1 and ITS2 are never adjacent in nature, 5.8S lying between them, so arm (vi) alters layout as well as content and its value cannot be read as ITS1 plus ITS2 in their native arrangement.

Table 3: Whose taxonomy does the prediction follow? Agreement of arm (iv) predictions with the query's own lineage and with the donor's, pooled across strata (n = 5,221). The query's ITS2 is present and unaltered; only the flanks belong to the donor, which is always from a different phylum.

	MycoAI-CNN		MycoAI-BERT		baseline, arm (i)	
rank	query	donor	query	donor	query	donor
phylum	37.8%	37.8%	10.8%	**71.7%**	78.8%	0.7%
class	20.8%	36.3%	5.4%	**71.3%**	51.2%	0.4%
order	10.3%	38.2%	2.4%	**68.5%**	28.8%	0.3%
family	4.2%	**34.4%**	0.8%	**63.7%**	16.8%	0.1%
genus	0.9%	26.8%	0.3%	36.9%	5.5%	0.1%

Baseline phylum recovery (78.8%) is pooled across strata and so differs slightly from the seen-genus value of 77.3% in Table 2. The baseline column is the convolutional model on arm (i), where no donor sequence is present, so donor agreement there reflects chance coincidence of lineages. Agreement with both query and donor is impossible by construction, donors never sharing the query's phylum, so the lineages are disjoint at every rank.

Set against its 77.9% family recovery on native full-length input, it recovers the donor's family at 82% of the rate at which it recovers the correct one when unperturbed: arm (iv) does not present to it as a damaged sequence but as an intact sequence of the donor organism.

Comparing the two arms rules out the obvious alternative reading. In arm (ii) the added sequence is likewise the majority of the input but carries no taxonomic signal, and the result is degradation without misdirection. In arm (iv) the added sequence carries genuine but incorrect signal, and the prediction follows it. The effect is therefore not dilution by input proportion but the models weighting information in the flanking regions above information in ITS2. Which flanking component carries that weight is not resolved here, since arm (iv) transfers ITS1 and 5.8S together (§2.6); ITS1 is the larger of the two and is itself a barcode, so the donor's ITS1 is the more likely vehicle, but we have not separated them.

The two models differ in how strongly they follow the donor: measured as donor-phylum agreement relative to each model's own phylum recovery on native input, the convolutional model stands at 38% and the transformer at 73%. We do not read this as a capacity effect, since the two differ in architecture as well as in size and a transformer attending globally over 558 bp can integrate distant 5′ context in a way a convolutional stack with a bounded receptive field structurally cannot. What the comparison supports is narrower and is reinforced by §3.7: the more of the input a model is able to exploit, the more it loses when part of that input is withheld.

Restoring one native region at a time asks a different question, and its answer constrains the mechanism further (Table 2, arms (v) to (vii)). Neither component recovers full-length performance and both are worse than ITS2 alone. Adding the record's own 5.8S to its own ITS2 takes seen-genus family recovery from 18.5% to 3.6% in the convolutional model and from 28.3% to 16.4% in the transformer; adding its own ITS1 gives 8.6% and 11.2%. ITS1 presented alone, with no ITS2 at all, recovers 15.7% and 22.5%, which is no better than ITS2 alone. So these are not ITS1 classifiers that tolerate ITS2 input, and no single flanking region carries the signal the full record supplies.

What the models require is the complete assembly rather than any particular region within it. That is consistent with arm (iv) rather than in tension with it: given something with the shape of a whole record the prediction follows the flanking positions, even to another phylum's family, while given a fragment it does poorly whatever the fragment contains. Following the flanks and classifying from the flanks are different properties, and only the first is demonstrated here. It also explains why arm (ii) behaves as it does, since random padding and a genuine but partial reconstruction both leave an input that resembles no training record.

One dissociation within this is worth reporting separately, because it is the only place where adding context helps anything. In the transformer, appending 5.8S raises phylum recovery from 76.5% to 93.4%, close to the 98.2% it reaches on native full-length input, while family recovery falls from 28.3% to 16.4%. The conserved region restores deep placement and costs shallow discrimination at the same time. The convolutional model shows no such effect, moving from 77.3% to 78.8% at phylum, and we read the difference as the transformer being able to integrate a conserved block that a bounded receptive field cannot (§3.3). For a practitioner the consequence is narrow but real: a partial reconstruction can make phylum-level output look healthier while making the family call worse, so agreement at deep ranks is not evidence that the input is in distribution.

This is a property of how these models exploit the barcode rather than of ITS2 specifically, and it predicts the same failure for any protocol that sequences a subregion without its flanking regions.

3.4 The collapse does not depend on how ITS2 was delimited

Because the ITS2 arm inherits a boundary drawn by the release's authors, ITS2 was independently re-extracted from the same queries and both versions classified on the identical 2,637

records for which re-extraction succeeded. The two delimitations proved close: 85.1% of records were identical and 99.7% identical or nested, with a median length difference of zero and medians of 167 and 165 bp. Both models were run on both extractions. Convolutional seen-genus family recovery was 18.7% and 18.3%, novel-genus 15.1% and 14.7%, coherence 51.5% and 51.4%; the transformer gave 28.6% and 28.4%, 18.4% and 18.5%, with coherence 56.3% and 56.5%. Neither model moves by more than 0.4 percentage points between delimitations, and the transformer moves less than the convolutional network, 0.2 and 0.1 points against 0.4 and 0.4. Flank dependence and boundary sensitivity are therefore not the same property: the model that follows the flanking regions more strongly (§3.3) is the less affected by where the ITS2 boundary is drawn, which is what one would expect if what matters is whether the flanks are present at all rather than their exact extent. The subset is also representative of the whole, recovery on it being close to the values obtained on all 5,222 queries for both models, 18.7% against 18.5% and 28.6% against 28.3%, so the records for which HMM anchoring succeeded are not appreciably easier. Because the two extractions are near-identical in sequence, this establishes independence of implementation rather than of methodology, and does not exclude the possibility that a substantially different ITS2 window would behave differently.

3.5 On ITS2 the classifiers' confidence loses its novelty signal

The more consequential half of the result concerns what the models report about themselves. Treating each method's own score as a novelty signal, alignment identity separated novel from known genera with AUROC 0.872 on full ITS and 0.866 on ITS2, and the SINTAX bootstrap with 0.813 and 0.785: both degraded slightly and remained strongly informative. The classifiers' P(family) fell from 0.688 to 0.541 for the transformer and from 0.706 to 0.503 for the convolutional network (Figure 1d). The convolutional model's ITS2 value is indistinguishable from chance (95% CI 0.487–0.518) and the transformer's exceeds it only marginally (0.525–0.557), a distinction that does not survive contact with use: a conformal novelty flag calibrated on these same two scores at a 5% false-novelty rate detects 6.1% of genuinely novel genera for the convolutional model and 5.2% for the transformer while firing on 4.6% and 4.2% of known ones, against 42.7% detection at the same error rate for alignment identity on the same amplicon [14].

A practitioner applying either model to ITS2 amplicons therefore receives family predictions that are wrong most of the time, accompanied by confidence values that do not distinguish the sequences it can place from those it cannot. The failure is undetectable from the output alone, and would be invisible to any evaluation conducted on full-length sequences. This is the basis of our recommendation that reported accuracies specify the amplicon region of the evaluation and be accompanied by a same-query classical baseline.

3.6 No method can name a novel genus; they differ in whether they say so

At genus rank all five methods scored exactly 0.0% across all 1,892 novel-genus queries, at both loci. This is a structural rather than an empirical result, and it is shared: the correct genus is absent from the classifiers' output spaces and from the reference the k-mer methods consult, so none can return it. No increase in training data or model capacity alters this for a fixed-label model, and no reference supplies a name it does not contain.

What separates the methods is disclosure. Alignment reports a best-hit identity that falls when no close relative exists; SINTAX reports a bootstrap confidence that does the same and, by convention, withholds assignments below a threshold. Both therefore emit a quantity from which abstention can be constructed, and on ITS2 both retained that property. The classifiers report a probability over a fixed vocabulary which, on ITS2, was uninformative about novelty. Forced prediction from a closed set is thus not by itself the problem, since it applies to every method here. The problem is forced prediction without a score that degrades informatively, and the classifiers on ITS2 are the case where that combination bites.

3.7 Training at the amplicon locus removes the deficit

Table 4: HiTaC family recovery at two reference depths, on the same queries. Retention is ITS2 recovery as a percentage of full-ITS recovery.

	seen genus		novel genus		retention	
reference	full ITS	ITS2	full ITS	ITS2	seen	novel
16,223 sequences	90.4%	69.5%	70.5%	48.5%	77%	69%
55,924 sequences	92.3%	47.4%	71.4%	30.8%	51%	43%

For comparison, the k-mer search methods retain 96% of their seen-genus accuracy on ITS2 and the MycoAI models 24 to 36%. Both rows here are trained on full ITS; refitting on ITS2 raises retention to 96.6% (Table 5).

Table 5: HiTaC family recovery under every combination of training locus and test locus, on the same queries and at fixed reference membership. Diagonal cells match training and test locus; off-diagonal cells are the mismatch.

		tested on full ITS		tested on ITS2	
trained on	reference	seen	novel	seen	novel
full ITS	55,924	**92.3%**	**71.4%**	47.4%	30.8%
ITS2	53,742	67.7%	40.1%	**89.2%**	**62.5%**

Genus recovery on seen genera follows the same pattern, 84.8% and 42.4% for the full-ITS fit and 60.2% and 82.0% for the ITS2 fit; on novel genera it is exactly zero throughout, as for every other method (§3.6). The two references comprise the same records at the same per-genus depth, one extracted at each locus, so the fits differ in locus alone. The full-ITS row is the one reported in Table 4; the diagonal cells are the ones entered in Table 1.

Table 1 gives every method the resources available at the locus being tested, which for HiTaC means a separate fit at each. Comparing those two fits against both query sets isolates what the locus contributes (Table 5). Fitted and tested on ITS2, HiTaC recovered 89.2% of seen-genus and 62.5% of novel-genus families, against 47.4% and 30.8% for the same classifier fitted on full-length ITS and given the identical ITS2 queries. On-locus training returns 41.8 of the 44.9 percentage points lost on seen genera, and the recovered value sits with the search methods rather than beneath them: 89.2% against 88.6% for best-hit alignment and 88.5% for SINTAX, retaining 96.6% of its own full-length accuracy against their 96.2% and 96.0%, with 82.0% genus recovery against their 81.1% and 80.9%.

Two things follow, and the second is what bears on the MycoAI models. First, ITS2 is not the limitation. A hierarchical logistic regression over 6-mer frequencies, fitted in ten minutes to 1.07% of the MycoAI training corpus, places primer-bounded ITS2 amplicons about as well as alignment does, on the amplicon where the distributed models place 28.3% and 18.5%. Neither the shorter window, nor k-mer features, nor learned classification is what fails there. Second, the deficit is a mismatch between training and test locus rather than a property of the shorter sequence, which the reverse arm establishes: fitted on ITS2 and given full-length queries, the same classifier fell to 67.7% and 40.1%. Both diagonals are high and both off-diagonals low, so no account in terms of ITS2 carrying less usable information survives the comparison.

The two off-diagonals are not symmetric. Withholding the flanking regions from a classifier fitted with them cost 44.9 points; supplying them to one fitted without them cost 21.5. We report the asymmetry without resting anything on it, and note that it is not the comparison arms (ii) and (iv) of §3.3 make: both off-diagonal inputs here are the query's own sequence, whereas arm

(iv) substitutes another phylum's flanks. What it says is only that a locus-mismatched k-mer profile is damaged more by truncation than by extension.

Fitting the full-ITS model at two reference depths adds a second, smaller finding about what governs off-locus robustness (Table 4). Increasing the reference from 16,223 to 55,924 sequences raised full-ITS recovery, from 90.4% to 92.3% on seen genera, and halved ITS2 recovery, from 69.5% to 47.4%; retention fell from 77% to 51%. More training data made the classifier better at the locus it was trained on and substantially worse away from it. Read alongside the locus grid, the reading is straightforward: depth buys features specific to the training locus, so more is lost when the locus changes, and nothing is lost when it does not. The same direction appears in the alignment depth sweep below, where deeper references improved placement of known taxa and slightly impaired placement of novel ones. In both cases a larger reference is not uniformly better.

Together these results qualify the mechanism proposed in §3.3 without displacing it. The flank ablation shows directly that the MycoAI models weight ITS1 and 5.8S above ITS2, and that finding rests on interventions on the input rather than on any comparison across methods. What the present section adds is that the resulting deficit is specific to the mismatch and not intrinsic to the amplicon, since it disappears when a classifier of the same family is fitted at the locus, and that susceptibility to it grows with how much of the training record a model exploits rather than being fixed by architecture.

3.8 The comparison is not an artefact of reference depth

Because the reference available to the k-mer methods is a design choice, the full-ITS alignment arm was repeated at three per-genus depths (Table 6). Seen-genus recovery rose monotonically with depth, from 90.1% to 93.8%, as did genus recovery on that stratum (77.9% to 87.3%) and novelty separability (AUROC 0.836 to 0.887). Novel-genus recovery did not: it peaked at twenty sequences per genus (67.2%) and fell at fifty (64.3%).

Table 6: Sensitivity of the full-ITS alignment arm to reference depth. The classifiers do not consult the reference and their results are unchanged throughout.

cap per genus	reference size	family, seen	family, novel	AUROC
5	16,352	90.1%	66.1%	0.836
20	56,327	92.3%	67.2%	0.872
50	113,455	93.8%	64.3%	0.887

Alignment exceeds both classifiers on both strata at every depth tested.

The asymmetry follows from what depth can supply. A seen-genus query gains conspecifics and congeners as the reference deepens, so its best hit improves; a novel-genus query cannot gain a same-genus relative at any depth, while the growing reference offers more opportunities for a spurious best hit from the wrong family. Deeper references therefore assist known taxa and mildly impede novel ones. For the present purpose the relevant observation is that alignment exceeded both classifiers on both strata at every depth tested, so the ranking does not depend on this choice, and that the twenty-per-genus reference used throughout is close to optimal for the novel-genus stratum rather than generous to alignment.

4 Scope and limitations

Several limitations bound these results. First, two architectures were evaluated, but both are distributed by the same authors, share a training corpus, and have identical label spaces; the replication therefore controls for architecture and not for training data, label-space construction,

or development group, and a model from an independent lineage might behave differently. Second, both models were used as distributed, without fine-tuning on ITS2. A different classifier fitted on primer-bounded ITS2 does not show the collapse (§3.7), but that is evidence about what the deficiency is rather than a demonstration that fine-tuning these two models would remove it; our claim is about the pretrained artefacts practitioners would actually download rather than about the architectures in principle. Third, the classifiers were run through their own distributed interface with default settings; we did not tune inference-time options, and cannot exclude that some configuration would improve ITS2 performance. We note, however, that the interface exposes no option relating to query length or amplicon region and issues no warning on short input, so the configuration we used is the one a practitioner following the documentation would arrive at. Fourth, novelty is defined by absence from the models' label space, which reflects an earlier reference release, so a genus later split or renamed may be counted novel although the model encountered its sequences under a prior name; this biases the novel stratum towards better apparent classifier performance, making the observed degradation conservative. Fifth, one-per-genus sampling deliberately removes abundance weighting, so absolute values are not comparable with published abundance-weighted accuracies, only with each other. Sixth, the ablation establishing that the models weight the flanking regions above ITS2 uses constructed inputs, a padded fragment and an interphylum chimera, which no workflow would produce; they identify what the models attend to and are not estimates of performance on real data, and the padding arm additionally confounds absence of information with presence of misleading sequence. Finally, the robustness check on ITS2 delimitation compares two extractions agreeing on 99.7% of records, so it demonstrates independence of implementation rather than of methodology; a materially different ITS2 window might behave otherwise. Finally, the depth observation in §3.7 rests on two reference depths at one locus for a single classifier, enough to show that off-locus robustness is not fixed by architecture but not enough to characterise how it scales.

5 Reproducibility

Analyses used `vsearch` v2.31.0, `ITSx` v1.1.3, `ITSxRust` v0.2.2, and `mycoai-its` v0.0.5 with `torch` v2.2.2, `scikit-learn` v1.3.2, `numpy` v1.26.4 and `biopython` v1.87 on Python 3.10.20. Several of these are pinned by MycoAI's dependency constraints rather than being current releases, `scikit-learn` in particular being held at the 1.3 series, so they are reported as installed rather than as recommended. The harness is organised in three stages:

```
construct   extract_labels.py, build_pilot_queries.py,
            build_align_reference.py, build_sintax_reference.py,
            build_flank_arms.py
score       score_matched.py, score_sintax.py, score_hitac.py,
            score_donor_attribution.py, compare_its2_extractions.py,
            mcnemar_tests.py, auroc_ci.py, diagnose_pilot.py
drive       its2_arm.sh, sweep_depth.sh, hitac_refit.sh
figure      make_figA.py
```

`make_figA.py` draws Figure 1 from the same summary table the manuscript quotes, so figure and text cannot diverge. The per-rank label sets recovered from the distributed models are deposited as plain-text taxon lists, 3,695 genera and 791 families, so that the stratification is reproducible without redistributing the models. Representative commands:

```
extract_labels.py --model MycoAI-CNN.pt --outdir taxa
build_pilot_queries.py --classification unite.classification \
  --fasta unite.fasta --genus-file taxa/genus.txt \
  --family-file taxa/family.txt --max-per-genus 1
vsearch --usearch_global queries.fasta --db ref.fasta \
```

```
  --id 0.5 --top_hits_only
vsearch --sintax queries.fasta --db ref_sintax.fasta \
  --sintax_cutoff 0.8
itsxrust extract -i queries.fasta -o its2.fasta --region its2 --hmm F.hmm
mycoai-classify queries.fasta --model MycoAI-CNN.pt \
  --confidence --out pred.csv
```

All analysis scripts, the per-query scored tables every value here is computed from, and the recovered label sets are available under the MIT license at `https://github.com/ayobi/its-classifier-audit`. Query sets, ablation arms and references are not deposited, being rebuilt from UNITE by the scripts above; the pretrained classifiers and the UNITE release are cited below rather than redistributed.

6 Discussion

MycoAI is a substantial piece of work and the problem it addresses is real: assigning taxonomy to millions of ITS reads against a reference of comparable size is computationally awkward, and a model that classifies 10^5 sequences in minutes is a genuine contribution to how such surveys are run. Nor is our finding that learned classification is the wrong approach: the best novel-genus recovery we measured came from a trained classifier, and it was trained in ten minutes on one percent of MycoAI's data. Our finding is that the released MycoAI models are not ready for the data environmental mycology produces, and that the bar they fall below is lower than the comparison with alignment alone suggests. On full-length reference sequences they are already outperformed by two long-standing k-mer methods at every rank and in both strata. On the ITS2 amplicon that environmental surveys overwhelmingly generate they lose 49.6 and 58.0 percentage points of family accuracy where those same methods lose 3.5 and 3.7. Because records, stratification and reference rule were common to all five methods, the collapse cannot be attributed to the locus: ITS2 retains enough information for the classical methods to keep about 96% of their full-length seen-genus accuracy, and for a trained classifier refitted at that locus to keep 96.6%.

The ablation locates the failure precisely enough to say what would and would not fix it. These models weight taxonomic information in the regions flanking ITS2 above information in ITS2 itself, to the point that supplying the flanks of an unrelated phylum causes the transformer to return that phylum's family for most queries whose ITS2 is intact and correct. That is not a shortfall of context to be made up with more training data at the same locus. It is a dependence that guarantees failure whenever the flanking regions are absent, which is the normal condition in amplicon sequencing. The larger model is the more dependent, following the donor roughly twice as strongly as the convolutional one, and the same direction appears within a single method in §3.7, where the deeper-trained HiTaC also lost more when moved off-locus. We read these together as evidence that susceptibility grows with how much of the training record a model exploits, rather than as a claim about scaling, which two non-independent architectures cannot support. The constructive implication is that the deficiency is specific and addressable, and §3.7 demonstrates that rather than conjecturing it: the same hierarchical classifier, refitted on the ITS2 reference and given ITS2 queries, recovered 89.2% of seen-genus families where the full-ITS fit recovered 47.4%, placing it alongside best-hit alignment and SINTAX rather than beneath them. Ten minutes of compute on one percent of the MycoAI corpus, at the locus the application generates, was therefore sufficient to remove the deficit in a classifier of this kind. We did not fine-tune the MycoAI models themselves, so this establishes what the deficiency is rather than that it has been repaired in them, and our claim continues to concern the released artefacts rather than the architectures.

The compounding problem is that this failure is invisible from the models' own output. On ITS2 their confidence barely separated novel from known genera at all (AUROC 0.541 and 0.503,

the latter at chance), and a novelty flag calibrated on either detects genuine novelties at almost exactly the rate at which it fires on knowns, so a practitioner receives confident genus names, mostly wrong, with no internal signal that anything is amiss. Three recommendations follow for how such models are reported. Accuracies should be given for the amplicon region the intended application generates, since here the two differ by a factor of four and the full-length figure is actively misleading about the ITS2 case. The length distribution of the training corpus should be stated, since it determines the range over which a reported accuracy can be expected to hold and is not currently recoverable from the model card. And a same-query classical baseline should be standard, because on this benchmark SINTAX and best-hit alignment were not merely competitive but superior everywhere we looked, including on the novel genera that motivate learned approaches in the first place. That conclusion is not ours alone. In the benchmark table of a paper introducing a newer fungal foundation model, BLAST exceeds both MycoAI models at family and genus rank on the MycoAI test set, 94.7% and 93.1% against 93.9% and 87.8% for the convolutional network [7], in the same work whose introduction attributes "poor generalization to novel taxa" to traditional methods. The baseline is often already present in these papers and simply not read as the comparison it constitutes.

Two claims we might have made are not supported by our data, and we state them as such. Fixed label spaces are not the operative limitation: the k-mer methods consult a reference lacking the novel genus and are equally unable to name it, all four methods scoring exactly zero. And the classifiers' weakness is not specific to novel taxa, since all five methods degraded by a similar 20 to 25 percentage points from seen to novel genera, making the learned representations uniformly weaker rather than differentially weak on dark matter.

What does distinguish the methods is whether their scores degrade informatively. Alignment identity and the SINTAX bootstrap remained strong novelty signals on ITS2; the classifiers' probabilities fell to chance or within a whisker of it. A score that declines when the input is unfamiliar can be thresholded, calibrated, or used to abstain, and the machinery for turning such a score into a decision with a guaranteed error rate is the subject of a companion study [14]. A score that does not decline offers no purchase for any such mechanism, which is why we regard the loss of novelty signal on ITS2 as the more serious of the two failures reported here.

Author contributions

Aaron O'Brien conceived the study, built the harness, ran all analyses and wrote the original draft. César Marín critically reviewed and revised the manuscript, contributing edits throughout and additions to the literature cited. Pilar Parada acquired the funding (CORFO 23PTECCC-247149) and supervised the project. All authors read and approved the submitted version.

Competing interests

Two of the authors of this manuscript, Aaron O'Brien and Pilar Parada, are authors of `ITSxRust` [15], which Section 3.4 uses solely as an independent re-implementation for a robustness check. The reported result there is that the two extractions agree to within 0.4 percentage points for either model, which is neutral with respect to both tools, and the conclusion of that section would be unchanged had the re-extraction been performed with any other implementation.

Data and code availability

Harness scripts, the recovered label-space taxon lists, the per-query scored tables and the results summary are available at `https://github.com/ayobi/its-classifier-audit` under the

MIT license. Reference data: UNITE+INSD 2024 via Zenodo 13336328. Pretrained classifiers: Zenodo record 10904344.

Acknowledgements

This work was supported by CORFO grant 23PTECCC-247149 to Pilar Parada. César Marín thanks Fondecyt Regular Project No. 1240186 (ANID, Chile). We thank the Centro de Biotecnología de Sistemas for compute resources.

Use of generative AI

Anthropic's Claude and GitHub Copilot were used in preparing this work: to draft and revise manuscript text, to write and debug the analysis and figure scripts deposited with it, and to check bibliographic details. All analyses were run by the authors, all reported values were verified against the primary output, and the authors are responsible for the content of the manuscript.

References

[1] Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R.H., Kõljalg, U. (2023). *Full UNITE+INSD dataset for Fungi.* Version 18.07.2023. UNITE Community. doi:10.15156/BIO/2938065

[2] Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R.H., Kõljalg, U. (2024). *Full UNITE+INSD dataset for Fungi.* Version 21.04.2024. UNITE Community. doi:10.15156/BIO/2959330

[3] Romeijn, L. (2024). *MycoAI data.* Zenodo. doi:10.5281/zenodo.10946477

[4] Millan Arias, P., Sadjadi, N., Safari, M., Gong, Z., Wang, A.T., Haurum, J.B., Zarubiieva, I., Steinke, D., Kari, L., Chang, A.X., Lowe, S.C., Taylor, G.W. (2023). BarcodeBERT: transformers for biodiversity analysis. *arXiv* 2311.02401. doi:10.48550/arXiv.2311.02401

[5] Badirli, S., Picard, C.J., Mohler, G., Richert, F., Akata, Z., Dundar, M. (2023). Classifying the unknown: insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution* 14(6), 1515–1530. doi:10.1111/2041-210X.14104

[6] Badirli, S., Akata, Z., Mohler, G., Picard, C., Dundar, M.M. (2021). Fine-grained zero-shot learning with DNA as side information. In *Advances in Neural Information Processing Systems* 34, 19352–19362.

[7] Gao, T., Lowe, S.C., Furneaux, B., Chang, A.X., Taylor, G.W. (2025). BarcodeMamba+: advancing state-space models for fungal biodiversity research. *arXiv* 2512.15931 [cs.LG]. 11 pp. Accepted at the 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI, NeurIPS 2025. doi:10.48550/arXiv.2512.15931. Preprint as of writing; Table 6 there is the source of the training-test overlap figures quoted in §1.

[8] Bengtsson-Palme, J., Ryberg, M., Hartmann, M., et al. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution* 4(10), 914–919. doi:10.1111/2041-210X.12073

[9] Edgar, R.C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv.* doi:10.1101/074161

[10] Fujisawa, T., Imai, T. (2026). Performance and limitations of out-of-distribution detection for insect DNA barcoding. *Ecology and Evolution*. doi:10.1002/ece3.73112

[11] Lücking, R., Aime, M.C., Robbertse, B., et al. (2020). Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus* 11, 14. doi:10.1186/s43008-020-00033-z

[12] Miranda, F.M., Azevedo, V.C., Ramos, R.J., Renard, B.Y., Piro, V.C. (2024). HiTaC: a hierarchical taxonomic classifier for fungal ITS sequences compatible with QIIME2. *BMC Bioinformatics* 25, 228. doi:10.1186/s12859-024-05839-x

[13] Nilsson, R.H., Larsson, K.-H., Taylor, A.F.S., et al. (2019). The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research* 47(D1), D259–D264. doi:10.1093/nar/gky1022

[14] O'Brien, A., Marín, C., Parada, P. (2026). A calibrated novelty flag for fungal ITS metabarcoding: choosing the error rate at which sequences are declared new. *bioRxiv* 2026.07.29.741524. doi:10.64898/2026.07.29.741524

[15] O'Brien, A., Lagos, C., Fernández, K., Ojeda, B., Parada, P. (2026). ITSxRust: ITS region extraction with partial-chain recovery and structured diagnostics for long-read amplicon sequencing. *bioRxiv* 2026.02.25.707950. doi:10.64898/2026.02.25.707950 (version 0.2.2 used here)

[16] Graetz, A., Feng, J., Ringeri, A., Bird, A., Vu, D., Truong, C., Schwessinger, B. (2025). Benchmarking fungal species classification using Oxford Nanopore Technologies long-read ITS metabarcodes. *bioRxiv*. doi:10.1101/2025.07.02.661940

[17] Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. doi:10.7717/peerj.2584

[18] Romeijn, L., Bernatavicius, A., Vu, D. (2024). MycoAI: fast and accurate taxonomic classification for fungal ITS sequences. *Molecular Ecology Resources* 24(8), e14006. doi:10.1111/1755-0998.14006

[19] van Galen, L.G., Corrales, A., Truong, C., et al. (2025). The biogeography and conservation of Earth's 'dark' ectomycorrhizal fungi. *Current Biology* 35(11), R563–R574. doi:10.1016/j.cub.2025.03.079

[20] Vu, D., Nilsson, R.H., Verkley, G.J.M. (2022). dnabarcoder: an open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. *Molecular Ecology Resources* 22(7), 2793–2809. doi:10.1111/1755-0998.13651

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