

Does Tropical Adaptation Help Plant Language Models Generalize? A Split-Sensitivity Evaluation

Gabriela Morales Soto

Institute of Computing - UNICAMP
Campinas-SP, Brazil, 13083-852
g299213@ic.unicamp.br

Helio Pedrini

Institute of Computing - UNICAMP
Campinas-SP, Brazil, 13083-852
helio@ic.unicamp.br

Abstract

Plant species assemblages can contain climate-related information, but it is unclear whether plant language models remain useful when observations are incomplete. We compare a sparse LightGBM baseline with Pl@ntBERT and a tropically adapted variant combining vocabulary expansion and domain-adaptive pre-training. The models predict four WorldClim variables from species composition in São Paulo state under progressively shortened species lists. The clearest pattern occurs for precipitation of the driest month (BIO14). On the original split, A2-MT shows positive differences over LightGBM under severe incompleteness after hierarchical bootstrap and multiplicity correction. Because BIO14 was selected post hoc, we retrain the models on five alternative assignments of a nominal 40 km spatial-block design. LightGBM is higher on complete lists in four of five partitions, while positive BIO14 differences under incomplete lists are directionally more stable but retain corrected support in only two partitions. Additional checks show that the pattern persists with a longer context window, weakens under frequency-weighted removal, and narrows to Max-10 and Max-5 when both model families receive truncation-augmented training. These results show that within-split robustness does not substitute for retraining across alternative spatial partitions.

1 Introduction

A species assemblage, the set of species observed in a locality, can carry information about climate because species differ in their environmental tolerances. This principle underlies methods such as the Modern Analogue Technique (Imbrie and Kipp, 1971; Overpeck et al., 1985), which relates assemblage composition to environmental conditions observed in reference data.

Biodiversity assemblage data resemble other incomplete environmental records used elsewhere in

this workshop: both require inferring a climate-related signal from partial, unevenly sampled observations. We frame our contribution as methodological transfer of NLP-style sequence models to non-textual structured records under spatial dependence, with climate inference as the motivating application rather than a demonstrated operational tool.

We study whether a plant language model can use this signal in a tropical setting. Pl@ntBERT (Leblanc et al., 2024, 2025) applies a BERT-style sequence model, an architecture originally built to learn from sentences by predicting words from their context, to species lists instead of text. Its pretraining data, however, are predominantly European vegetation plots, so many tropical taxa and co-occurrence patterns may be underrepresented in what it originally learned.

We evaluate four WorldClim bioclimatic targets: annual mean temperature (BIO1), annual precipitation (BIO12), precipitation of the driest month (BIO14), and precipitation seasonality (BIO15). BIO14 showed the strongest pattern on the original split and was therefore selected post hoc for the retraining-based split-sensitivity analysis. We treat this follow-up as an internal sensitivity analysis rather than a confirmatory test.

The practical motivation is incomplete biodiversity observation. Species occurrence databases such as the Global Biodiversity Information Facility (GBIF) can have uneven sampling across space and taxa (Hughes et al., 2021). In our corpus, observed richness is strongly right-skewed: the median 5 km grid cell contains 24 species, compared with a mean of 58.9. We therefore ask whether a language-model pipeline retains useful climate signal when species lists are shortened, and whether it outperforms a sparse LightGBM baseline built from binary species-presence features under those conditions.

A broader methodological question is whether

robustness within one fixed split implies robustness to the choice of split itself. We provide three main contributions: (i) we evaluate how a combined tropical-adaptation pipeline—train-only vocabulary expansion followed by domain-adaptive pretraining—changes PI@ntBERT transfer across continuous climate regression and an exploratory phytophysiology-classification task; (ii) we compare the adapted model with a sparse LightGBM baseline under simulated species-list incompleteness; and (iii) for the strongest post-hoc target, BIO14, we distinguish within-split resampling from full retraining across five alternative assignments of one nominal 40 km spatial-block design.

2 Related Work

Assemblage composition has long been used to infer environmental conditions. The Modern Analogue Technique compares target assemblages with modern reference assemblages observed under known climates (Imbrie and Kipp, 1971; Overpeck et al., 1985), while the coexistence approach reconstructs climatic intervals from overlapping taxon tolerances (Mosbrugger and Utescher, 1997). Our study examines a related inference problem using large occurrence datasets and a domain-adapted plant language model.

PI@ntBERT (Leblanc et al., 2024, 2025) adapts a BERT-style architecture (Devlin et al., 2019) to abundance-ordered species lists and evaluates tasks including masked-species prediction and habitat classification. Our study examines transfer to tropical assemblages and continuous bioclimatic regression. Domain-adaptive pretraining (DAPT), continued pretraining on in-domain unlabeled data before task fine-tuning, has improved downstream performance in other domains (Gururangan et al., 2020). We test whether vocabulary expansion combined with DAPT improves performance in this setting and whether any observed gain persists across data splits.

Spatially structured evaluation is important when nearby observations share environmental information. Roberts et al. (2017) show that conventional cross-validation can underestimate predictive error under spatial dependence and recommend structure-aware splitting. Ploton et al. (2020) similarly report weaker ecological prediction under spatial validation. These results motivate our distinction between resampling an existing test set and retraining on a newly constructed spatial partition.

3 Data and Experimental Design

This section describes the data, model variants, and experimental protocols used in our evaluation.

3.1 Data Sources and Analysis Subsets

We use three data sources with distinct roles. GBIF occurrence records provide the plant observations used to construct cell-level species assemblages and binary species-presence features. IF-SP 2020 vegetation polygons (Instituto Florestal, 2020) define the analysis subset and provide labels for the exploratory classification task reported in Appendix E. WorldClim 2.1 (Fick and Hijmans, 2017) provides the four regression targets and is never used as model input.

For GBIF (GBIF.org, 2026), we retain plant records with valid coordinates, species-level identification, and `occurrenceStatus == PRESENT`, after applying the geolocation filters used in preprocessing. Records are assigned to a 5 km grid, and cells with fewer than 10 records or 5 species are removed. This yields an ecological corpus of 4,014 cells, 503,859 occurrence records, and 13,355 unique species.

To define the analysis subset, we intersect grid cells with IF-SP 2020 vegetation polygons and retain cells whose dominant mapped class covers at least 30% of the intersected IF-SP area. The resulting IFSP30 subset contains 1,077 cells. A post hoc geometric check found that 10.5% of these cells (113 of 1,077) have centroids falling just outside São Paulo’s official political boundary, at a maximum distance of 3.5 km (median 950 m); this is consistent with a boundary mismatch between the IF-SP 2020 polygons and the administrative boundary. A border-cell exclusion and matched placebo analysis suggest that these cells do not uniquely drive the reported BIO14 pattern (Appendix F). The original fixed partition contains 751 training and 326 test cells; it was inherited from an earlier phase of the project, and its exact generation procedure (random, spatial, or otherwise) could not be reconstructed from the retained artifacts. This uncertainty is one motivation for the retraining-based split-sensitivity analysis below. After excluding cells with missing climate targets, the regression analysis uses 749 training and 321 test cells. Each cell is represented as a deduplicated sequence of species names in case-insensitive lexicographic order. This representation differs from the abundance-oriented ordering used in PI@ntBERT pretraining,

a mismatch considered in the Limitations.

3.2 Bioclimatic Targets

The four targets defined in Section 1 are extracted from WorldClim 2.1 (Fick and Hijmans, 2017) at 2.5 arc-min resolution by cell centroid. At São Paulo’s latitude, this corresponds to approximately 4.25 to 4.64 km, comparable to the 5 km analysis grid. Their means (SDs) in the analyzed data are BIO1 19.6°C (2.3°C), BIO12 1628 mm (407 mm), BIO14 49 mm (26 mm), and BIO15 55.7 (12.9). Climate variables are used only as regression targets and are never provided as model inputs.

3.3 Model Variants

Throughout the paper, A0 and A2 denote the two adaptation states evaluated in this study. A0 refers to direct transfer from the original Pl@ntBERT checkpoint without tropical adaptation. A2 refers to the tropically adapted model obtained through the expansion of the train-only vocabulary followed by domain-adaptive pretraining in the training cells. The suffixes follow the experimental configurations used here: FT identifies the direct full-fine-tuning configuration used with A0, whereas MT and ST distinguish multi-target and single-target fine-tuning of A2. The resulting configurations are A0-FT, A2-MT, and A2-ST.

Our baseline consists of four independent LightGBM regressors, one per climate target, trained on binary species-presence features. The feature space is restricted to species observed in the training set. Exact hyperparameters are reported in Appendix D.

A0-FT starts from the original Pl@ntBERT checkpoint (Leblanc et al., 2024, 2025) and fine-tunes the full model end to end with a linear regression head that jointly predicts the four climate targets. No tropical adaptation is applied.

A2-MT starts from the A2 model defined above and is fine-tuned end to end to jointly predict the same four climate targets. We repeat this procedure across five training seeds, 42 through 46. Exact training settings are reported in Appendix D.

A2-ST uses the same A2 model but is fine-tuned only for precipitation of the driest month. It serves as a secondary control for whether the pattern observed for this variable depends on joint supervision from the four climate targets. A2-ST and the frozen-probe analyses are evaluated only on the original fixed split and are not retrained on the new spatial-block split.

We also freeze the A0 and A2 representations and fit ridge regressors under three pooling strategies to test whether tropical adaptation changes linearly accessible climate signal before task-specific fine-tuning.

3.4 Incomplete-List Evaluation

We evaluate complete species lists and six simulated incompleteness settings: proportional retention of 75%, 50%, or 25% of the original species, and hard caps of 20, 10, or 5 species per cell. Retained species are sampled without replacement, and their relative order is preserved. Five truncation seeds, 101 through 105, define the same shortened inputs for all compared models. The hard-cap settings do not use abundance information because the stored assemblages contain species identities only. A0-FT is evaluated on complete lists only.

3.5 Retraining-Based Spatial Partition Sensitivity

To test sensitivity to the original partition, we construct a new spatial-block split without using the previous train/test assignment. We take the same 1,077 labeled cells, set aside their original train/test assignment, and regroup the underlying 5 km grid into disjoint blocks of 8×8 cells, corresponding to a nominal 40 km block width (Figure 1). Using a fixed random seed, whole blocks rather than individual cells are assigned to the test set until roughly 30% of the data falls there, which yields 748 training and 329 test cells across 104 and 53 blocks respectively, with no block or cell shared between the two sides; after removing cells with missing climate values, 743 training and 327 test cells remain.

We retrain LightGBM, A0-FT, and all five seeds of A2-MT from scratch on this new split, with vocabulary expansion and domain-adaptive pretraining once again restricted to the new training cells only; we do not retrain A2-ST or the frozen probes on this split. For the incomplete-list evaluation on this split we use five of the seven original settings (complete list, 25%, and the three hard caps), and we freeze the exact list of which cells, sparsity levels, and truncation seeds go together into a single manifest of 6,867 rows, reused unchanged for both LightGBM and A2-MT so that the two models are always tested on exactly the same inputs.

The split-generation procedure is deterministic and versioned: it groups the 5 km lattice into nominal 8×8 -cell blocks, randomly permutes whole

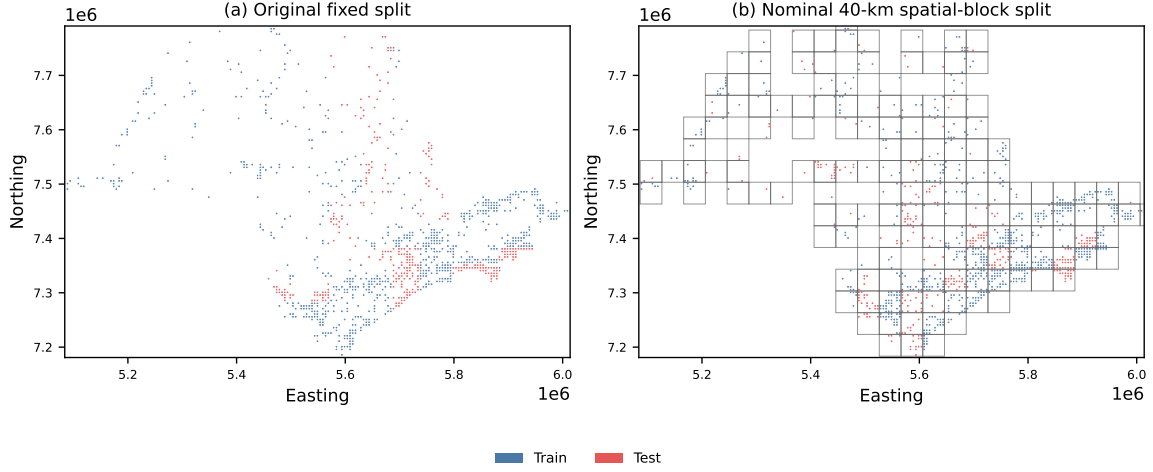


Figure 1: Spatial structure of the original fixed partition and the initial blocked retraining partition. (a) Original fixed train/test split. (b) Partition 1 of the nominal 40 km spatial-block design, shown as a representative example (Partitions 2-5 use the same blocking scheme; Section 5.3). Training and test cells are shown separately; panel (b) also shows the spatial blocks used to construct the partition. No cell or block is shared between training and test in panel (b).

blocks under the partition seed, and assigns blocks to the test set until the cumulative test-cell count first reaches or exceeds the 30% target. Split seeds, manifest hashes, and additional provenance are reported in Appendix D.

We refer to this split as Partition 1. To assess sensitivity to the specific block assignment, we generated four additional partitions (Partitions 2-5) under the identical nominal 40 km blocking design, using seeds drawn before inspecting outcomes on those four newly generated partitions (Appendix D). All five partitions reuse the same 1,077 cells; their test sets partially overlap, so we treat them as sensitivity analyses rather than independent replications (Section 5.3). Across Partitions 1-5, the raw train/test counts are 748/329, 736/341, 746/331, 752/325, and 743/334 cells, respectively, with test-block counts of 53, 55, 48, 37, and 40. All five partitions have zero train-test block overlap.

Exploratory vegetation-type classification and full training configurations are reported in Appendices E and D, respectively.

3.6 Evaluation and Uncertainty

Our main contrast is $\Delta R^2 = \overline{R^2_{A2-MT}} - \overline{R^2_{LightGBM}}$, where A2 values are averaged over training seeds and, for truncated conditions, over truncation realizations. The original-split analysis also reports a seed-ensemble estimand in the Appendix A. Uncertainty is assessed with 10,000-replicate hierarchical bootstrap procedures. For

the retrained Partitions 1-5 analyses, each replicate resamples test cells, A2 training seeds, and truncation seeds with replacement; the sampled truncation realization is kept paired between A2-MT and LightGBM. We report 95% percentile intervals and two-sided add-one bootstrap tail areas. Benjamini-Hochberg adjustment is applied across the five completeness levels within each partition. Because the five partitions reuse cells and have overlapping test sets, we do not treat their 25 split-by-level contrasts as independent tests. Accordingly, cross-partition sign counts are treated only as descriptive summaries and are not assigned a combined significance test. Full definitions are provided in Appendix A.

4 Full-List Performance and Transfer

We first evaluate all models on complete species lists using the original fixed train/test split.

4.1 Full-List Regression

Species identity alone yields predictive performance for several climate targets. On complete test lists, the sparse LightGBM baseline reaches $R^2=0.51$ for BIO14 and $R^2=0.48$ for BIO12 (Table 1). However, most predictable variation is geographic: a LightGBM model using only centroid coordinates and elevation reaches $R^2=0.88-0.98$ across the four targets (Appendix G), and adding species features does not improve on geography alone. A partial Mantel analysis controlling for

geographic distance is nevertheless consistent with a small residual association between composition and climate for BIO1 and BIO14 ($r=0.05-0.06$, $p = 0.001$ for both BIO1 and BIO14, robust to two distance metrics), but not for BIO12. We therefore interpret species composition as carrying modest climate-related information beyond location, rather than as a substitute for geographic predictors; BIO14 is the target for which this residual signal also aligns with the subsequent incompleteness analysis.

The adapted A2-MT pipeline has higher five-seed mean R^2 than direct A0-FT by approximately 0.25 for BIO14 and 0.21 for BIO12. This A0-A2 contrast is descriptive; we did not run the hierarchical bootstrap or false-discovery-rate procedure for it. A2-MT also does not uniformly exceed LightGBM. BIO1 is highly seed-sensitive, BIO12 remains slightly below the baseline mean, and BIO15 is similar but variable. For BIO14, the A2-MT mean is 0.537 versus 0.510 for LightGBM, but the full-list hierarchical interval includes zero ($CI_{95}=[-0.058, +0.117]$). A2-ST reaches $R^2=0.498$ on the full BIO14 test set.

As an exploratory cross-task check, we also compare A0 and A2 on phytophysiology classification (Appendix E). A2 is descriptively higher than A0 on macro-F1, balanced accuracy, and accuracy, including a 0.10 absolute increase in accuracy. Relative to LightGBM, however, the comparison is mixed: A2 has higher accuracy but lower macro-F1 and balanced accuracy. Because this analysis uses only three seeds and no hierarchical bootstrap or multiplicity correction, we treat it as secondary evidence about A0-to-A2 transfer rather than evidence that the adapted pipeline is generally superior.

5 Performance under Incomplete Species Lists

We now assess performance under simulated species-list incompleteness, starting with the original split.

5.1 BIO14 under Severe Incompleteness (Original Split)

Figure 2 shows $\Delta R^2 = R^2_{A2-MT} - R^2_{LightGBM}$ for BIO14 across the seven list-retention settings on the original split. For the typical-pipeline summary, intervals include zero when lists retain about 30 or more species on average. Larger positive differences appear in the more severe settings that retain

roughly 20 or fewer species (Table 5, Appendix B).

The result should not be interpreted as a universal 20-species threshold. The 25% and max-20 settings retain similar mean richness (15.5 and 17.3 species) through different mechanisms (proportional retention versus a hard cap), and their performance differs. The missingness mechanism therefore matters in addition to the number of retained species.

The typical-pipeline comparison survives within-family FDR correction at 25%, max-20, max-10, and max-5. The 25%, max-20, and max-5 settings also survive the global correction across all 84 tests. At the full, 75%, and 50% settings, the observed mean difference is positive but the hierarchical interval includes zero.

5.2 Spatial Sensitivity: Resampling vs. Retraining

We examined this pattern with respect to spatial variation using two complementary approaches, which yielded contrasting results.

The first check resamples the existing test set into spatial blocks at three scales (10, 20, 40 km, four offsets each; twelve partitions total), reusing the already-fitted models. The BIO14 advantage stayed positive across all twelve partitions at 25%, max-20, and max-5; max-10 was less stable (positive in eight of twelve). This reuses the original fitted models and test set, so it tests sensitivity to how that set is grouped, not whether the effect recurs under retraining on different data.

The retraining-based analysis addresses that question by fitting LightGBM, A0-FT, and all five A2-MT seeds on Partition 1, the initial nominal 40 km blocked assignment described in Section 3.5, and it tells a noticeably different story (Partition 1 row of Table 6, Appendix C). For BIO14, both models score lower on this split, and the full-list ranking seen originally reverses: where A2-MT's full-list mean had exceeded LightGBM's on the original split (0.537 against 0.510, Table 1, though that gap was not itself statistically supported), on the retrained split LightGBM reaches an R^2 of 0.399 for the same variable while A2-MT reaches only 0.329 on average across five seeds, and direct transfer without adaptation falls to -0.018 . This reversal is informative on its own, since it shows that which model looks better on complete lists can flip between the original fixed split and Partition 1, independently of anything to do with sparsity.

On Partition 1, A2-MT remains descriptively

Table 1: Climate regression on full test lists ($n = 321$), original split. Language-model values are mean $\pm$ SD over five seeds; LightGBM is a single run. Bold marks the highest mean per target.

Model	BIO1	BIO12	BIO14	BIO15
LightGBM	0.219	0.482	0.510	0.283
A0-FT	-0.067 ± 0.077	0.255 ± 0.064	0.284 ± 0.047	0.101 ± 0.039
A2-MT	0.082 ± 0.169	0.461 ± 0.024	0.537 ± 0.028	0.260 ± 0.085
A2-ST _{BIO14}	—	—	0.498 ± 0.062	—

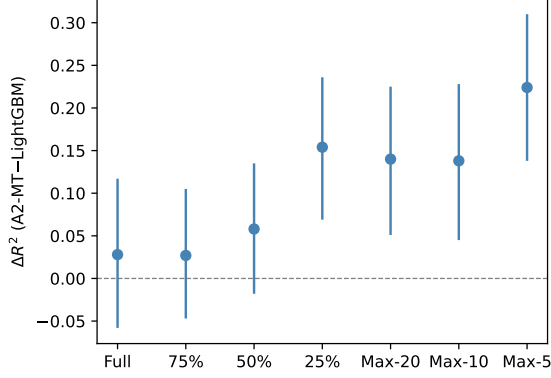


Figure 2: BIO14 performance difference (ΔR^2 , A2-MT minus LightGBM) on the original fixed split, typical-pipeline summary. Points show observed ΔR^2 ; bars show 95% hierarchical bootstrap intervals. † marks significance after within-family and global correction; * marks within-family correction only (Table 5, Appendix B).

above LightGBM at 25% retention and max-20, but all five hierarchical confidence intervals include zero (Table 6), and no level survives Benjamini-Hochberg correction within the five-level family. The original BIO14 advantage did not retain inferential support on this alternative blocked partition: part of the descriptive pattern persists, but its statistical support does not.

5.3 Retraining Sensitivity Across Five Partitions

Figure 3 and Table 2 summarize BIO14 ΔR^2 across all five 40-km partitions (Partitions 1-5; per-partition detail in Appendix C).

Across complete lists, A2-MT exceeds LightGBM in only one of five partitions. Under incomplete lists, positive differences recur in all five partitions at 25% and Max-5, and in four of five at Max-20 and Max-10 (Table 2). These cross-partition summaries are descriptive because the partitions overlap.

Inferential support is less stable. No contrast survives within-partition BH adjustment in

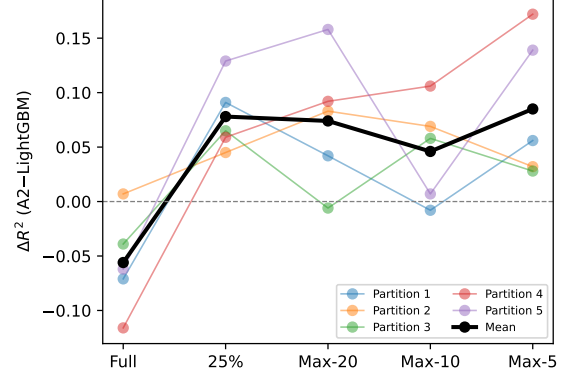


Figure 3: BIO14 ΔR^2 (A2-MT minus LightGBM) across the five 40-km spatial-block partitions. Thin lines are individual partitions; the thick black line is the cross-partition mean (Table 2). No level survives Benjamini-Hochberg correction in Partitions 1-3.

Partitions 1-3. In Partition 4, only Max-5 remains supported ($\Delta R^2 = +0.172$, 95% CI $[+0.061, +0.291]$, $q = 0.007$). In Partition 5, 25%, Max-20, and Max-5 remain supported (all $q = 0.016$). Thus, the direction of the incomplete-list effect is more stable than its inferential strength. Partitions 4 and 5, which retain adjusted contrasts, have the fewest test blocks (37 and 40, versus 48–55) and higher BIO14 test-set variance (SD 28.0–28.3 versus 21.5–23.1). For a fixed absolute prediction error, greater target variance can increase R^2 because the total sum of squares is larger. With only five overlapping partitions, however, test-block count and target variance are confounded, and we cannot attribute the difference to either factor. We did not identify a partition characteristic that cleanly explains the difference.

5.4 Other Climate Targets

No comparable typical-pipeline pattern is observed consistently for BIO1, BIO12, or BIO15. BIO12 shows isolated seed-ensemble advantages under some incomplete-list conditions on the original split, but this result depends on ensembling and is treated as secondary. BIO14 is therefore the

Table 2: Descriptive BIO14 ΔR^2 (A2-MT–LightGBM) across five 40-km partitions. Partitions overlap; not independent replications.

Level	Mean	Median	Range	A2 higher
Full	−0.056	−0.062	[−0.116, +0.007]	1/5
25%	+0.078	+0.065	[+0.045, +0.129]	5/5
Max-20	+0.074	+0.083	[−0.006, +0.158]	4/5
Max-10	+0.046	+0.058	[−0.008, +0.106]	4/5
Max-5	+0.085	+0.056	[+0.028, +0.172]	5/5

only target showing a repeated typical-pipeline difference across multiple severe-incompleteness settings. In particular, BIO1 performance is weak for both model families, so the present setup should not be interpreted as a successful temperature-reconstruction model; BIO15 likewise shows no consistent A2-MT advantage over LightGBM.

6 Additional Robustness Checks

The BIO14 analysis motivated six additional robustness and diagnostic checks targeting potential confounds and mechanisms not addressed above. Table 3 summarizes verdicts; full detail is in Appendices H–M.

Table 3: Summary of additional robustness and diagnostic checks related to the BIO14 results (original split unless noted).

Check	Outcome
Sequence length (64→128 tokens)	Same qualitative verdict at all 7 levels (Appendix H)
Vocabulary vs. DAPT ablation	Both components improve over A0 descriptively; combined gain is roughly additive (Appendix I)
Frequency-weighted removal	25% and max-20 lose significance; max-10 and max-5 survive (Appendix J)
Symmetric truncation exposure	LightGBM is higher at 5/7 levels; A2-MT is higher only at Max-10 and Max-5 (Appendix K)
Block-level bootstrap (Partitions 4–5)	All 4 previously supported contrasts retain positive intervals, with wider CIs (Appendix L)
Spatial buffer (10 km exclusion)	A2-MT’s advantage over A0-FT widens, not narrows (Appendix M)

The most consequential of these is the symmetric-exposure check. LightGBM’s original training used complete species-presence vectors, while A2’s domain-adaptive pretraining used five random species-order reorderings per cell as augmentation. These reorderings change sequence order but do not expose A2 to truncated species lists. Thus, neither original pipeline was explicitly trained on the incomplete-list conditions used at evaluation. The symmetric-exposure experi-

ment addresses this issue directly by retraining both model families with truncated training examples. Retraining LightGBM on a 23,219-row truncation-augmented training set (Appendix K) raises its BIO14 R^2 across all seven levels, exceeding A2-MT at full, 75%, 50%, 25%, and max-20; A2-MT remains descriptively higher only at Max-10 and Max-5. A size-matched control trained on bootstrap-resampled complete rows is lower than truncation-augmented LightGBM at every level for which the control was run, supporting a truncation-exposure effect rather than a sample-size explanation. This narrows the severe-sparsity claim to the two most extreme incompleteness regimes.

7 Discussion

Two questions naturally follow from a result that is specific to one variable and one range of list incompleteness: why that variable, and why that range. Precipitation of the driest month describes how severe the dry season gets, which plausibly connects more directly to seasonal water limitation (Tao et al., 2022) than annual mean temperature does, and that connection is easy to believe in a region that spans contrasting vegetation types. It is not, however, something our experiments establish; annual mean temperature also happens to vary less, relative to its own scale, in this dataset than precipitation of the driest month does, and altitudinal gradients or non-linear relationships between temperature and species composition could contribute as well. We treat the fact that this particular variable stood out as an empirical pattern worth reporting, not as a resolved ecological mechanism.

As noted in Section 1, this is a pattern we noticed rather than one we set out to confirm. The additional robustness checks also narrow how it should be interpreted. Frequency-weighted removal attenuates the advantage at 25% retention and Max-20, showing that the missingness mechanism matters, while symmetric truncation exposure substantially improves LightGBM and leaves A2-MT higher

only at Max-10 and Max-5. Thus, part of the original severe-sparsity pattern reflects how the models respond to a distribution shift they were not explicitly trained on, rather than architecture alone. A remaining possibility is that the adapted model uses broader co-occurrence structure when very few species remain, whereas LightGBM relies more heavily on individual species indicators, but we have not tested that mechanism directly. Consistent with this uncertainty, frozen A0/A2 representations yield $R^2 < 0$ under all three ridge-probe pooling strategies tested. We therefore treat the mechanism behind the residual Max-10 and Max-5 differences as unresolved.

A third, more consequential question is how much the choice of split itself matters. The original split’s BIO14 advantage passed several within-split checks, including hierarchical bootstrap, multiplicity correction, and post-hoc spatial resampling of the same held-out set. None of these checks involves retraining on different train/test assignments. Partition 1 therefore provided an important warning: once both models were retrained under a nominal 40 km blocked assignment, the full-list ranking reversed and no incomplete-list contrast survived within-partition correction.

Post-selection regression to the mean is a complementary, non-exclusive explanation: because BIO14 was selected after inspection of the broader original-split results, some attenuation under new partitions is compatible with regression to the mean even without a systematic split effect. Partitions 2-5 do not resolve this ambiguity. The adjusted contrasts are concentrated in Partitions 4 and 5, which also have fewer test blocks and greater BIO14 test-set variance than Partitions 1-3; with only five overlapping partitions, these properties are confounded and do not provide a causal explanation for the difference. A block-level bootstrap on Partitions 4 and 5 leaves all four previously supported contrasts with positive intervals, although the intervals widen substantially (Appendix L), suggesting that the surviving results are not solely an artifact of cell-level resampling. Separately, a 10 km exclusion buffer applied to Partition 1 widens the BIO14 difference between A2-MT and A0-FT because direct transfer deteriorates more strongly (Appendix M). This buffered check concerns adaptation versus direct transfer rather than the A2-MT-versus-LightGBM sparsity comparison, so we treat it as complementary rather than decisive evidence about the headline result.

Putting the original split and the five retrained partitions side by side leads to a fairly modest practical conclusion. On the original split, the adapted pipeline showed a positive BIO14 difference in several severely incomplete settings, with longer observed training times in our heterogeneous runs, although these timings are not hardware-normalized and should not be interpreted as a controlled efficiency comparison (Table 4).

The retrained partitions narrow that considerably: descriptive advantages recur under severe incompleteness, but adjusted contrasts survive Benjamini-Hochberg correction in only two of five partitions (Partition 4 at Max-5; Partition 5 at 25%, Max-20, and Max-5), while Partition 1 shows no surviving contrast at any level. Given this uncertainty, we currently see the simpler LightGBM baseline as the more defensible default; the original-split result alone does not justify the additional complexity of the adapted pipeline. We also do not yet know how often real biodiversity-monitoring programs encounter list incompleteness as severe as the settings simulated here, or whether the pattern would persist across additional blocked partitions or other regions.

8 Conclusions

The combined A2 adaptation pipeline is descriptively higher than A0 across the evaluated regression and exploratory classification metrics, but it does not uniformly outperform LightGBM. Across five alternative 40 km spatial-block partitions, A2-MT shows no consistent full-list advantage, and the BIO14 incomplete-list effect remains partition-dependent.

The strongest practical claim narrows further under symmetric truncation exposure: on the original split, A2-MT remains higher than LightGBM only at Max-10 and Max-5 after both model families are trained on truncated examples. We therefore view the adapted pipeline as potentially useful under extreme list incompleteness rather than as a general replacement for the simpler baseline. More broadly, the results show that within-split robustness checks do not substitute for retraining across alternative spatial partitions.

Limitations

All experiments concern São Paulo state, so transfer to other tropical regions remains untested. The original fixed split was inherited from an earlier

Table 4: Descriptive wall-clock comparison, original split. LightGBM: four independent regressors, serial sum. A2-MT DAPT+FT includes DAPT; FT-only excludes it. Values with $\pm$ report mean $\pm$ SD over five seeds. Hardware was heterogeneous and was not systematically recorded for LightGBM; these wall-clock times are therefore descriptive and not directly comparable across model families, so no training-time ratio is reported.

Model	Train (s)	Infer (s)
LightGBM (4 targets)	21.4	0.048
A0-FT (GPU)	62.8 $\pm$ 6.3	0.51 $\pm$ 0.10
A2-MT FT only (GPU)	66.7 $\pm$ 3.9	0.48 $\pm$ 0.03
A2-MT DAPT+FT	210.2 $\pm$ 9.0	0.48 $\pm$ 0.03

phase of the project, and its exact generation procedure could not be reconstructed from the retained artifacts. Partitions 1-5 were therefore generated under an explicit, versioned spatial-block procedure, but they reuse the same 1,077 cells and have partially overlapping test sets; they should be interpreted as sensitivity analyses rather than independent replications. BIO14 was also selected post hoc after inspection of the broader original-split results, so selection-induced effect inflation remains possible.

The nominal 40 km partitions have zero train-test block overlap but impose no minimum physical separation across block boundaries. A 10 km buffered sensitivity analysis suggests that residual proximity does not explain the A2-versus-A0 transfer difference on the tested partition, but this check used only three seeds, one buffer distance, and did not directly test the A2-MT-versus-LightGBM sparsity comparison.

The primary Partitions 1-5 bootstrap resamples individual test cells rather than spatial blocks. A targeted block-level bootstrap on Partitions 4 and 5 leaves all four previously supported contrasts with positive intervals, but with wider uncertainty; we did not repeat block-level inference for every partition and condition.

Species sequences are stored in lexicographic rather than abundance-oriented order, creating a mismatch with Pl@ntBERT pretraining. The original incompleteness analysis uses uniform random removal, whereas real biodiversity records have structured detection and sampling biases. A frequency-weighted sensitivity analysis attenuates several of the original differences, showing that the results are sensitive to the missingness mechanism. Likewise, symmetric truncation exposure substantially improves LightGBM and narrows the residual A2-MT advantage to Max-10 and Max-5 on the original split. Finally, LightGBM uses a train-fitted species feature space, taxa appearing

only at test time are not represented as learned baseline features, and recorded wall-clock times were obtained on heterogeneous hardware and are descriptive rather than hardware-normalized efficiency estimates.

Acknowledgements

This work was supported by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.013778/2020-21].

References

- Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. <https://aclanthology.org/N19-1423>.
- Stephen E. Fick and Robert J. Hijmans. 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315. <https://doi.org/10.1002/joc.5086>.
- GBIF.org. 2026. GBIF occurrence download. <https://doi.org/10.15468/dl.evgrnx>. Accessed 7 May 2026.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks.

- In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. <https://aclanthology.org/2020.acl-main.740>.
- Alice C. Hughes, Michael C. Orr, Keping Ma, Mark J. Costello, John Waller, Pieter Provoost, Qinmin Yang, Chaodong Zhu, and Huijie Qiao. 2021. Sampling biases shape our view of the natural world. *Ecography*, 44(9):1259–1269. <https://doi.org/10.1111/ecog.05926>.
- John Imbrie and Nilva G. Kipp. 1971. A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core. In Karl K. Turekian, editor, *The Late Cenozoic Glacial Ages*, pages 71–181. Yale University Press, New Haven.
- Instituto Florestal. 2020. Inventário Florestal do Estado de São Paulo. Technical report, Secretaria de Infraestrutura e Meio Ambiente do Estado de São Paulo, São Paulo, SP, Brazil.
- César Leblanc, Pierre Bonnet, Maximilien Servajean, and Alexis Joly. 2024. Pl@ntBERT: Leveraging Large Language Models to Enhance Vegetation Classification through Species Composition Analysis. In *European Congress of Conservation Biology (ECCB) - Book of Abstracts*, page 121.
- César Leblanc, Pierre Bonnet, Maximilien Servajean, Wilfried Thuiller, Milan Chytrý, Svetlana Aćić, Olivier Argagnon, Idoia Biurrun, Gianmaria Bonari, Helge Bruelheide, and 1 others. 2025. Learning the syntax of plant assemblages. *Nature Plants*, 11(10):2026–2040.
- Volker Mosbrugger and Torsten Utescher. 1997. The co-existence approach—a method for quantitative reconstructions of Tertiary terrestrial palaeoclimate data using plant fossils. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 134:61–86. [https://doi.org/10.1016/S0031-0182\(96\)00154-X](https://doi.org/10.1016/S0031-0182(96)00154-X).
- Jonathan T. Overpeck, Thompson Webb, and I. Colin Prentice. 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research*, 23(1):87–108.
- Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. 2020. [Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models](https://doi.org/10.1038/s41467-020-18321-y). *Nature Communications*, 11:4540. <https://doi.org/10.1038/s41467-020-18321-y>.
- David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. 2017. [Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure](https://doi.org/10.1111/ecog.02881). *Ecography*, 40(8):913–929. <https://doi.org/10.1111/ecog.02881>.
- Shengli Tao, Jérôme Chave, Pierre-Louis Frison, Thuy Le Toan, Philippe Ciais, Jingyun Fang, Jean-Pierre Wigneron, Maurizio Santoro, Hui Yang, Xiaojun Li, Nicolas Labrière, and Sassan Saatchi. 2022. [Increasing and Widespread Vulnerability of Intact Tropical Rainforests to Repeated Droughts](https://doi.org/10.1073/pnas.2116626119). *Proceedings of the National Academy of Sciences*, 119(37):e2116626119. <https://doi.org/10.1073/pnas.2116626119>.

A Statistical Methodology Detail

Estimands. *Typical-pipeline*: within each bootstrap replicate, average the seed-specific ΔR^2 over the resampled five-seed multiset (captures mean pipeline performance under optimization-seed uncertainty). *Seed-ensemble*: within each replicate, average predictions over the same resampled multiset before computing a single ΔR^2 (captures an ensembled-deployment estimand). For truncated levels, reported “observed” values are typical-pipeline, further averaged across the five truncation-seed scenarios.

Hierarchical bootstrap. For the original fixed-split analysis, 10,000 replicates resample test cells and model seeds with replacement; truncation scenarios are treated as matched repeated conditions. For the retrained Partitions 1-5 analysis, each replicate resamples test cells, A2 model seeds, and truncation seeds with replacement, while preserving the sampled truncation realization between A2-MT and LightGBM. CIs are 95% percentile intervals.

Multiplicity. Benjamini-Hochberg FDR (Benjamini and Hochberg, 1995) is applied separately within six specified analysis-by-estimand families: MT vs. LightGBM under the typical-pipeline and seed-ensemble estimands (28 tests each, spanning four targets and seven sparsity levels), and ST vs. LightGBM and ST vs. MT under each estimand (7 BIO14 tests per family) - 84 tests total, all on the original split. We additionally apply a global BH correction across all 84 tests as a stricter sensitivity analysis.

Bootstrap tail areas. Two-sided finite-bootstrap tail areas use the add-one correction, with n_+/n_- the counts of replicates where ΔR^2 falls on each

side of zero and B the total number of replicates:

$$p_+ = \frac{n_+ + 1}{B + 1}, \quad p_- = \frac{n_- + 1}{B + 1}, \quad (1)$$

$$p_{\text{two}} = \min(1, 2 \min(p_+, p_-)). \quad (2)$$

We use these finite-bootstrap tail areas as inputs to the BH procedure; because they are bootstrap-derived rather than parametric p -values, we interpret the adjusted values jointly with the percentile intervals and effect sizes rather than as standalone evidence. On the original split, this two-sided construction is used throughout. For Partitions 1-5, the same two-sided add-one construction is applied directly to the retained bootstrap replicate arrays for all five completeness levels in each partition, followed by Benjamini-Hochberg adjustment within each partition (Table 6).

B Full Hierarchical Bootstrap Results (Original Split)

Table 5: Hierarchical bootstrap: BIO14, A2-MT typical pipeline vs. LightGBM, original split (321 test cells). $\Delta R^2 = \text{A2-MT} - \text{LightGBM}$. CI: 95% percentile interval, 10,000 replicates. q_{BH} : within-family; $\dagger$ also survives global-84; $*$ within-family FDR only (does not survive global-84). “>0.05”: within-family FDR family but not a discovery at 0.05. “Clear”: 95% CI entirely above zero. “Inconclusive”: CI includes zero.

Level	Spp.	ΔR^2	95% CI	q	Verdict
Full	60.7	+0.028	[−0.058, +0.117]	>0.05	inconclusive
75%	45.8	+0.027	[−0.047, +0.105]	>0.05	inconclusive
50%	30.6	+0.058	[−0.018, +0.135]	>0.05	inconclusive
25%	15.5	+0.154	[+0.069, +0.236]	0.013	clear $\dagger$
Max-20	17.3	+0.140	[+0.051, +0.225]	0.013	clear $\dagger$
Max-10	9.8	+0.138	[+0.045, +0.228]	0.042	clear $*$
Max-5	5.0	+0.224	[+0.138, +0.310]	0.006	clear $\dagger$

For BIO1 and BIO15 we also counted, per individual seed and sparsity level, how often the seed-specific ΔR^2 bootstrap interval excluded zero (“decisive”; all other cases were treated as inconclusive). Of the decisive seed $\times$ level comparisons (14 for BIO1 and 8 for BIO15), BIO1 favors LightGBM in 13 and A2 in 1; BIO15 favors LightGBM in 6 and A2 in 2.

C Multisplit BIO14 Results

“Clear” denotes a positive 95% bootstrap interval together with $q < 0.05$ after Benjamini-Hochberg adjustment within the corresponding partition; “inconclusive” denotes all remaining cases.

Table 6: BIO14 $\Delta R^2 = \text{A2-MT} - \text{LightGBM}$ by partition and list-completeness level. CIs are 95% hierarchical bootstrap percentile intervals; q is Benjamini-Hochberg adjusted within each partition across the five evaluated levels.

Partition	Level	ΔR^2	95% CI	q	Verdict
1	Full	−0.071	[−0.203, +0.052]	0.642	inconclusive
1	25%	+0.091	[−0.027, +0.201]	0.642	inconclusive
1	Max-20	+0.042	[−0.083, +0.159]	0.642	inconclusive
1	Max-10	−0.008	[−0.145, +0.110]	0.894	inconclusive
1	Max-5	+0.056	[−0.100, +0.191]	0.642	inconclusive
2	Full	+0.007	[−0.129, +0.136]	0.904	inconclusive
2	25%	+0.045	[−0.059, +0.139]	0.615	inconclusive
2	Max-20	+0.083	[−0.032, +0.184]	0.584	inconclusive
2	Max-10	+0.069	[−0.052, +0.172]	0.584	inconclusive
2	Max-5	+0.032	[−0.064, +0.117]	0.615	inconclusive
3	Full	−0.039	[−0.246, +0.138]	0.923	inconclusive
3	25%	+0.065	[−0.100, +0.194]	0.923	inconclusive
3	Max-20	−0.006	[−0.196, +0.136]	0.961	inconclusive
3	Max-10	+0.058	[−0.122, +0.194]	0.923	inconclusive
3	Max-5	+0.028	[−0.144, +0.162]	0.923	inconclusive
4	Full	−0.116	[−0.303, +0.063]	0.430	inconclusive
4	25%	+0.059	[−0.101, +0.211]	0.470	inconclusive
4	Max-20	+0.092	[−0.096, +0.272]	0.430	inconclusive
4	Max-10	+0.106	[−0.077, +0.297]	0.430	inconclusive
4	Max-5	+0.172	[+0.061, +0.291]	0.007	clear
5	Full	−0.062	[−0.184, +0.068]	0.409	inconclusive
5	25%	+0.129	[+0.032, +0.226]	0.016	clear
5	Max-20	+0.158	[+0.043, +0.285]	0.016	clear
5	Max-10	+0.007	[−0.129, +0.157]	0.931	inconclusive
5	Max-5	+0.139	[+0.032, +0.247]	0.016	clear

D Verified Training Configuration

Table 7 summarizes settings recovered from retained training scripts, per-seed metrics, resolved trainer configurations, and serialized LightGBM models. For DAPT, seeds 42-44 retain core settings in per-seed metrics and the training script; seeds 45-46 additionally preserve resolved trainer configurations linked to the same script SHA-256.

Partition provenance. Partitions 2-5 use four seeds before inspecting model outcomes on those four newly generated partitions; Partition 1 (seed 20260706) had already been used as the initial retraining-based spatial sensitivity check. The base split generator and its wrapper script are content-hashed. Frozen sparsity-manifest SHA-256 values for Partitions 1-5 are 4a2efdbc...476abee7, 53e5ca64...66c3b, 35955806...36dc, b5d560ee...c5f6, and 6517ede6...d6b2, respectively; full manifest and generator hashes are retained in the accompanying repository.

Model and tokenizer. All transformer runs start from the public Pl@ntBERT masked-language checkpoint (CesarLeblanc/plantbert_fill_mask_model), using its original tokenizer and base vocabulary of

44,708 tokens. Tropical adaptation adds whole-species tokens observed in the training data; species with at least two training-corpus occurrences are eligible. Of 5,253 requested tokens, 5,108 were added (145 were already present in the base vocabulary), yielding a final vocabulary of 49,816 tokens. New-token embeddings are randomly initialized and subsequently updated during DAPT and downstream fine-tuning; in the vocabulary-only ablation they receive no DAPT adaptation and are learned only during downstream fine-tuning. Hyperparameters in Table 7 were fixed in advance and were not selected using the reported test set; no early stopping or test-driven hyperparameter selection was performed.

Table 7: Verified configuration used in the reported experiments.

Component	Setting
FT backbone	Fully trainable
FT pooling	CLS representation
FT head	Linear regression
FT outputs	4 (MT), 1 (ST)
FT target scaling	StandardScaler, train-only fit
FT loss	MSE
FT epochs / batch	10 / 8
FT learning rate	10^{-5}
FT optimizer	AdamW
FT max sequence length	64
FT precision / seeds	FP16 / 42-46
DAPT objective	Masked language modeling
DAPT epochs / batch	3 / 8
DAPT learning rate	5×10^{-5}
DAPT max sequence length	64
DAPT masking probability	0.15
DAPT augmentation	5 random species-order reorderings per cell
DAPT optimizer	AdamW; fused in resolved seed-45/46 configs
DAPT scheduler / weight decay	Linear / 0, in resolved seed-45/46 configs
LightGBM boosting	GBDT
LightGBM leaves / max depth	31 / -1
LightGBM learning rate	0.05
LightGBM estimators	300
LightGBM row / feature subsampling	1.0 / 1.0
LightGBM L1 / L2	0 / 0
LightGBM random state	42

E Phytophysiology Classification (Exploratory)

No hierarchical bootstrap or multiplicity correction was applied to the classification comparisons as

follows; treat as exploratory secondary evidence only. Evaluated on the original fixed split only.

Table 8: Phytophysiology classification, fixed test set (326 cells; unlike regression, classification does not require complete climate targets, hence 326 rather than 321). A0/A2: mean $\pm$ SD over 3 seeds. Bold: highest mean per column.

Model	Macro-F1	Bal. Acc.	Acc.
LightGBM	0.449	0.441	0.727
A0 direct	0.337 $\pm$ 0.021	0.340 $\pm$ 0.026	0.645 $\pm$ 0.038
A2 (vocab.+DAPT)	0.374 $\pm$ 0.020	0.380 $\pm$ 0.030	0.745 $\pm$ 0.035

Per-class F1 (LightGBM): FOD 0.834, FES 0.622, FOM 0.214, Formação Pioneira 0.125, showing markedly weaker performance on the two rarer classes.

F Geographic Border-Cell Check

Of 1,077 IFSP30 cells, 113 (10.5%) have centroids outside São Paulo’s political boundary (maximum 3.5 km, median 950 m), distributed as 72 training and 41 test cells. Excluding the 36 affected test cells present in the 321-cell climate evaluation subset and re-running the hierarchical bootstrap for BIO14 leaves 25%, max-20, and max-5 clear ($q=0.025, 0.037, 0.006$); max-10 becomes inconclusive ($\Delta R^2=0.129, q=0.057$). A placebo control of 20 independent random 36-cell exclusions shows this is a general small-sample effect, not specific to the border cells: 11 of 20 random exclusions similarly push max-10 above the 0.05 threshold, and the border-cell exclusion’s ΔR^2 falls at the median of the placebo distribution.

G Geographic Control

A LightGBM model using only cell centroid coordinates and elevation reaches $R^2=0.876$ (BIO1), 0.903 (BIO12), 0.904 (BIO14), and 0.976 (BIO15) on the 321-cell test set. Combining species features with these three geographic features does not improve over geography alone; geographic features account for 98–99% of total split gain in the combined model. Elevation was obtained via the OpenTopoData SRTM90m API for all 1,070 cells with complete climate targets. A partial Mantel test (composition vs. climate, controlling for geography; 571,915 pairs, 999 permutations) gives $r=0.060$ ($p=0.001$) for BIO1 and $r=0.048$ ($p=0.001$) for BIO14, but $r=0.006$ ($p=0.533$) for BIO12; both results are essentially unchanged

(third decimal place) under Sørensen-Dice in place of Jaccard distance.

H Sequence-Length Confound Check

Fine-tuning uses a 64-token maximum context; given mean richness of 58.9 species, this truncates 57.6%, 48.0%, 34.1%, and 14.1% of test cells at full, 75%, 50%, and 25% retention, falling below 9% at all three hard-cap levels. We retrained the full A2-MT pipeline (five seeds) with 128-token context, reducing truncation substantially (e.g., 57.6%→26.8% at full). This run also used an independently reconstructed, larger DAPT corpus, so it is a robustness check on the truncation confound rather than an isolated sequence-length ablation; we did not have a 128-token run on the exact original DAPT corpus. The qualitative pattern is unchanged at all seven levels: no significant advantage at full/75%/50%, and a within-family-FDR-surviving advantage at 25%, max-20, max-10, and max-5, matching Table 5 exactly in verdict.

I Vocabulary Expansion vs. DAPT Ablation

We trained two variants isolating each component of tropical adaptation, three seeds each, on the original split. *DAPT-only* (unexpanded vocabulary, DAPT on training cells) reaches BIO14 $R^2=0.375\pm0.036$. *Vocabulary-only* (expanded vocabulary, no DAPT; new token embeddings are randomly initialized and updated only during downstream fine-tuning) reaches $R^2=0.439\pm0.131$. Both improve over A0-FT, recomputed here over the same three seeds as the ablation arms ($R^2=0.263\pm0.034$; Table 1 reports 0.284 over all five seeds); combined, A2-MT reaches 0.537 ± 0.028 . The sum of individual improvements (+0.112 and +0.176, totaling +0.288) closely matches the combined improvement (+0.274), suggesting the two components contribute largely independently. Vocabulary-only’s standard deviation is roughly four times larger than the other variants’, consistent with ~5,108 randomly initialized new-token embeddings having to be learned from only 749 downstream training cells without prior DAPT adaptation; with three seeds we do not treat the vocabulary-vs-DAPT ranking as statistically established.

J Frequency-Weighted Removal

Uniform random truncation does not reflect structured sampling and recording biases in GBIF (Hughes et al., 2021). We therefore built an alternative manifest retaining species with probability weighted by their training-corpus record frequency (common species favored to survive truncation) and re-ran inference with the existing LightGBM and five-seed A2-MT models. Training-corpus record frequency is used here as a simple proxy for structured recording probability and does not separate ecological commonness from true detectability. Under a self-contained six-level FDR family, 25% and max-20 no longer reach significance ($q=0.78, 0.09$); max-10 and max-5 remain significant ($q=0.028, 0.001$). ΔR^2 decreases at every level (e.g., 25%: 0.154→0.025; max-5: 0.224→0.172).

K Symmetric Truncation Exposure

LightGBM’s original training never saw truncated inputs. We retrained LightGBM on a 23,219-row truncation-augmented training set (749 original rows plus 6 levels $\times$ 5 truncation seeds of resampled training cells) and separately retrained A2-MT (five seeds) from its original DAPT checkpoints on the same augmented data. A size-matched control (LightGBM trained on 23,219 bootstrap-resampled *untruncated* rows) isolates whether gains reflect truncation exposure or sample size alone.

Table 9: BIO14 R^2 under symmetric truncation exposure, original split.

Level	LGBM aug.	LGBM ctrl.	A2-MT aug.
Full	0.665	0.577	0.511±0.048
75%	0.636	–	0.503±0.042
50%	0.567	–	0.443±0.035
25%	0.435	0.169	0.361±0.069
Max-20	0.525	0.269	0.458±0.057
Max-10	0.338	0.048	0.373±0.046
Max-5	0.161	−0.148	0.177±0.068

The control supports an exposure effect rather than a sample-size explanation: at Max-5, the control (untruncated, resampled) falls *below* zero, while augmented LightGBM stays positive, showing training on truncated examples specifically helps at severe sparsity. LightGBM-augmented exceeds A2-MT-augmented at five of seven levels; A2-MT-augmented is higher only at max-10 (+0.035) and max-5 (+0.016).

L Block-Level Bootstrap (Partitions 4-5)

The main multisplit bootstrap resamples individual cells; Partitions 4-5, where all surviving contrasts concentrate, have the fewest test blocks (37, 40) among the five partitions. Re-running the hierarchical bootstrap by resampling whole spatial blocks leaves all four previously supported contrasts with positive intervals, though with wider intervals (e.g., Partition 4 Max-5: $CI_{95}=[0.034, 0.320]$, versus a much tighter cell-level interval).

M Spatial Exclusion Buffer

We constructed a partition using Partition 1's block assignment but additionally excluding training cells within 10 km (Euclidean) of any test cell, reducing training from 748 to 628 cells. Retrained on this buffered split (three seeds), A0-FT's BIO14 R^2 falls to -0.153 ± 0.103 (from ≈ 0.27 on the original split), while A2-MT falls to 0.310 ± 0.129 (from 0.537). The relative advantage of adaptation over direct transfer widens rather than narrows under the buffer, suggesting A0-FT's original-split performance relied more on spatial proximity than A2-MT's did. Three seeds and a single buffer distance limit how much weight this result can bear.