AgroNT Enables Unlabeled Cross-Species Mapping of AUX/IAA Orthologs in *Brachypodium distachyon*

Anonymous Submission

Anonymous affiliation

Abstract

While researchers have gained access to tools and databases that provide better access to genome sequences, the bottleneck for many such resources is in labelling, comparing, and performing locus searches to identify orthologous sections for newly found organisms. Large Language models such as AgroNT have made significant progress in predicting such functions, but its accuracy on predicting genes that play similar roles between different organisms has not fully been established yet. Thus, this study aims to use Cosine testing to compare AgroNT's prediction of Arabidopsis ARF gene orthologs on an unmarked genome of *Brachypodium distachyon* and evaluate its level of accuracy.

Code — https://colab.research.google.com/drive/1s-Ijup3jthUjhi5hF-gFt6GLibz8Y0pG?uspring=sha

Data — https://drive.google.com/drive/folders/1vVTUWM-n1wNrQdQ-3XE-AaYGo928DWZqG?usp=sharing

Introduction

High-throughput sequencing and community resources such as TAIR, Ensembl Plants, and NCBI have made high-quality plant genomes straightforward to obtain. What remains hard is the part we still do largely by hand: turning raw sequence into functional labels, especially across species. For many crops and wild relatives, gene and regulatory annotations are sparse or missing, so cross-species curation still leans heavily on BLAST-centric workflows and manual alignment against model references like ***Arabidopsis thaliana***. These approaches are powerful but do not scale to the growing diversity of plant genomes.

Recent nucleotide language models (NLMs) learn distributed representations of DNA that capture sequence "grammar" and regulatory regularities directly from base patterns. AgroNT is one such plant-focused encoder trained on a panel of crop genomes. If its embeddings carry enough functional signal, they should allow us to retrieve likely orthologous or functionally similar regions in another genome without first performing whole-genome alignment, and then to localize those hits to near base-pair resolution with a small amount of targeted scanning.

In this work we build and evaluate a minimal, reproducible pipeline around this idea. Starting from a small set of Arabidopsis AUX/IAA pathway genes, we embed each query sequence with AgroNT and tile an unlabeled target genome into fixed windows that are also embedded. We then rank these windows by exact cosine similarity to the query embeddings to obtain a coarse set of candidate loci, and refine the best candidates with a short fine-resolution scan inside each tile to produce a single peak coordinate per query.

As a test case we focus on ***Brachypodium distachyon,*** which is absent from AgroNT's pre-training corpus but has a well-annotated reference genome. This makes it a natural out-of-distribution target: AgroNT never sees *Brachypodium* labels during training, yet we can still compare its predictions against known ortholog positions. To keep the experiments minimal and interpretable, we restrict the retrieval pipeline to chromosome 1 and use a small query set of *Arabidopsis* AUX/IAA genes—ARF5/MONOPTEROS (AT1G19850), ARF7 (AT5G20730), ARF19 (AT1G19220), and IAA14 (AT4G14550)—while limiting evaluation to seeds with annotated orthologs on this chromosome.

For evaluation, we obtain ground-truth ortholog coordinates from Ensembl Plants and hold them out strictly for post hoc analysis. We measure how close each predicted locus falls to the orthologous gene interval and to the transcription start site (TSS) on the correct strand, and we examine how accuracy varies with the size of the genomic window we are willing to accept around the prediction. For a familiar baseline, we optionally run BLAST/minimap2 in a narrow band around the predicted peak to confirm the local alignment and identity.

Our goal is not to replace alignment, but to shrink the search space so that alignment and curation can focus where it matters. We show that AgroNT embeddings support exact Top-K retrieval over tiled plant genomes and that a short, targeted fine scan is sufficient to sharpen those hits to practical genomic coordinates. The pipeline is compact enough to run on a single GPU, includes ablations over key design

choices (pooling, stride, strand handling, tokenization phase), and is released as an executable notebook for reproducibility.

Contributions

1. A **minimal retrieval to localization pipeline** for cross-species gene finding using AgroNT embeddings, with strict label isolation (no retrieval leakage).
2. A **leakage-free evaluation** on an OOD species (*B. distachyon*), reporting window- and base-pair-level distances and a simple range-vs-accuracy trade-off.
3. **Practical ablations and diagnostics** (pooling, stride, RC, phase sweep; cosine–distance linkage) plus optional alignment and interpretability checks.

Related Works

Genotyping and Enabling Works
Early work on scalable genotyping (e.g., rAmpSeq) demonstrated robust, low-cost marker discovery by targeting repetitive sequences, helping unlock genotype–phenotype studies across diverse germplasm. While not an embedding model, this line of work underpins today's cross-species analyses by making high-throughput variant data routine.

Affordable phenomics and genetic gain.
Recent perspectives in the Plant Phenome Journal argue that broadening access to phenomics is essential for accelerating genetic gain, emphasizing practical pipelines that integrate field measurements, environmental covariates, and genomic information. This positions phenotype prediction as a multi-modal problem and motivates approaches that connect sequence to phenotype-proximal signals (e.g., expression, regulatory strength) rather than attempting end-to-end field trait prediction outright.

LLMs for trait prioritization (Gore Lab).
Farmer et al. (co-authored by M. A. Gore) showed that large language models can mine crop trait-prioritization studies to extract structured information about user-preferred traits and contexts, offering a scalable route to curate targets for breeding programs. We adopt a complementary stance: use text-mined trait→pathway→gene sets as seeds, then perform sequence-embedding retrieval across species to propose candidate analogs.

Plant DNA foundation models: AgroNT.
AgroNT introduced a plant-focused DNA language model trained on dozens of edible plant genomes, achieving strong results on regulatory annotation, promoter/terminator strength, tissue-specific expression, and variant prioritization. Its pretrained representations provide an immediate backbone for cross-species retrieval and for probing phenotype-proximal tasks without extensive supervised labels.

Cross-species modeling: PlantCaduceus.
PlantCaduceus is a plant DNA LM trained across 16 angiosperm genomes with architectural inductive biases (e.g., reverse-complement handling) that emphasize evolutionary conservation and cross-species transfer at near single-nucleotide resolution. It is a natural comparator to AgroNT in our experiments and strengthens the case that sequence LMs can generalize across clades.

Positioning of our work.
Relative to the above, we (i) pair trait-to-gene sets from the literature (as in Farmer et al.) with embedding-based cross-species retrieval using AgroNT/PlantCaduceus, and (ii) validate on phenotype-proximal regulatory benchmarks rather than field traits alone—aligning with phenomics-first arguments—while providing interpretable evidence via sparse-feature (SAE) analyses. This combination targets a practical gap between manual BLAST-centric curation and black-box phenotype prediction, aiming to reduce annotation friction and accelerate hypothesis generation.

Data

Query Set (A-Arabidopsis AUX/IAA)
We import four Arabidopsis genes representative of the AUX/IAA pathway: ARF5/MONOPTEROS (AT1G19850), ARF7 (AT5G20730), ARF19 (AT1G19220) and IAA14 (AT4G14550). Of these, we do NOT use ARF5/MONOPTEROS (AT1G19850) as the ortholog was not located on chromosome 1 of our Target Genome. Each FASTA was preprocessed to uppercase A/C/G/T/N and stored with a stable identifier.

Target Genome (C-Unlabeled Brachypodium distachyon**)**
Primary experiments used an out-of-distribution (OOD) species (e.g., Brachypodium distachyon) from Ensembl. Due to computing capacity, we have shown a demo with just chromosome 1 in this paper.

Methodology

Overview of the Pipeline
We designed a two-stream workflow that separates retrieval from evaluation. In the retrieval stream, we embed

four Arabidopsis AUX/IAA genes once with AgroNT. We tile the unlabeled target genome, embed each window, and select Top-K candidate windows by exact cosine similarity. We then refine these candidates to base-pair coordinates with a short fine scan, with an optional phase sweep reserved for ablations. In the evaluation stream, ground-truth ortholog coordinates from Ensembl Plants are used only to compute distances and overlap. They are never used as retrieval inputs, which avoids leakage. Unless stated otherwise, analyses were run on chromosome 1 to keep iteration fast, and ortholog tables were filtered to the same chromosome set.

Ortholog Labels (Evaluation Only)

For each seed gene, putative Brachypodium distachyon ortholog intervals were obtained from Ensembl Plants (compara=plants) and saved as CSV, BED, and FASTA files. We normalized fields to a common schema: [gene_id, chr, start, end, strand, display_name, homology_type, perc_id, perc_cov, score, source_seed, chrom_n]. These label tables were not embedded or indexed; they were used solely during evaluation.

Model and Tokenization (AgroNT)

We used AgroNT-1B (InstaDeepAI/agro-nucleotide-transformer-1b). Inputs were tokenized into non-overlapping 6-mers with a positional cap of ~1,024 tokens including [CLS]. Sequences exceeding the limit were trimmed by a token-budget guard that checks post-tokenization length and shortens the raw string just enough to fit. The encoder outputs 1,500-dimensional hidden states per token. We formed 1,500-dimensional pooled sequence vectors from the last layer using two modes:

X2 (default), a masked mean over non-padded tokens that excluded [CLS]

X1 (ablation), the [CLS] hidden state.

Inferences ran without gradients, in mixed precision where available. inputs were validated for length and vocabulary, with a CPU fp32 fallback on CUDA assertions.

Genome Tiling and Coarse Embedding

The target genome was segmented into fixed 6,144 bp windows, which correspond to about 1,024 6-mer tokens, with a stride of 6,144 bp so that windows do not overlap. For the coarse pass we used only the forward strand to keep runtime manageable.

Windows from the Arabidopsis queries and from the target genome were sanitized to A/C/G/T/N and trimmed when needed to respect the token budget. Each window was embedded in batches, and we stored the pooled vector $v \in \mathbb{R}^{1500}$ together with metadata (chromosome, start_bp, end_bp, strand). We did not deduplicate windows in this minimal pipeline.

Retrieval: ABTT

Before L2 normalization, we applied the All-But-The-Top (ABTT) transform to pooled embeddings to remove dominant corpus-wide directions that inflate cosine scores, such as GC-content or length effects. We first centered each vector by subtracting the mean over all pooled embeddings. We then projected out the top r principal components, which were learned in an unsupervised way from the pooled target embeddings C (or from $A \cup C$ in an ablation).

Let v be a pooled embedding and $\tilde{v} = \frac{1}{M}\sum_{m=1}^{M} v_m$ the mean over M vectors. We define the centered vector $\tilde{v} = v - \tilde{v}$. With $\{u_i\}_{i=1}^{r}$ the top r principal components, the ABTT-filtered vector is

$$v^{\mathrm{ABTT}} = \tilde{v} - \sum_{i=1}^{r} (u_i^{\top} \tilde{v})\, u_i$$

We then L2-normalized v^{ABTT} and used these normalized vectors to compute exact cosine Top-K, as described in the retrieval section. ABTT was applied on-the-fly for scoring and did not overwrite the stored embedding cache. This keeps the retrieval layer reversible and easy to reproduce. In our main runs we set $r = 1$ and report an ablation with $r = 3$.

Retrieval: L2-normalized cosine Top-K

We applied row-wise L2 normalization to the query and target embedding matrices and then computed exact cosine similarities by blockwise matrix multiplication ($A \mathbin{@} C^{T}$). For each query window *i*, we selected the Top-K = 50 target windows *j* with the highest cosine scores using incremental arg-partition. We did not use approximate indexing (for example FAISS), so retrieval remains fully reproducible at the current scale. For every query we stored a tidy table with query_id, query_span, target_id (chromosome), target_span (start–end:strand), cosine score, and rank.

Fine Localization within the Top-K tiles

To turn tile-level hits into base-pair predictions, we refined each query's Top-K tiles in two steps.

1. **Fine sliding scan.**
   Within each candidate tile we slid a 2,048 bp sub-window with a 256 bp stride. We embedded all sub-windows in large batches, with automatic fallback to smaller batches if the GPU reported out-of-memory. For each tile we kept the sub-window with the maximum cosine against the query embedding.

2. **Post-hoc strand check.**
   For the selected sub-window we computed the reverse-complement embedding and compared it to the forward version. We kept whichever orientation had the higher cosine and reported that strand.

A six-phase (0–5) tokenization sweep for 6-mer offsets is available in the notebook but treated as an ablation rather than part of the minimal run. In this study we refined Top-K = 50 tiles per query, which improves recall in the fine scan on out-of-distribution genomes.

Ortholog Labels and Evaluation Metrics

Predicted spans were mapped to target genomic coordinates (gw_start, gw_end). For each predicted span and matching chromosome, we computed:

- Interval gap (bp): 0 if the predicted span overlaps an ortholog interval; otherwise, the positive flank distance to the nearest boundary.
- TSS distance (bp): signed distance from the predicted midpoint to the ortholog's transcription start site (TSS = start for strand = +1, TSS = end for strand = −1). Positive tss_dist means the prediction midpoint lies downstream of TSS along the genomic axis. For range-sweep plots we use absolute TSS distance.

Distances were computed per (source_seed, gene_id) pair and restricted to chromosomes present in the embedding run (e.g., chr 1), preventing mismatch with labels on unprocessed chromosomes. We report distributions of interval gaps and TSS distances, along with accuracy-versus-tolerance curves at two levels: (i) individual refined sub-windows and (ii) per-query summaries, using the minimum absolute TSS distance across that query's rows. Ortholog labels participate only in these computations.

Hyperparameters and Computing

- Coarse windows / stride: 6,144 bp / 6,144 bp.
- Tokenizer: non-overlapping 6-mers with ≈1,024 positions including [CLS].
- Pooling: masked-mean (X2, default); CLS (X1) as an ablation.
- Top-K tiles refined per query: 50.
- Fine scan: 2,048 bp sub-window, 256 bp stride, post-hoc reverse complement check, optional 6-phase sweep (0–5).
- Batching: 128–512 for coarse embeddings and up to 1,024 for fine scan, with automatic fallback on out-of-memory.
- Hardware and precision: single NVIDIA A100 preferred (fp16/bf16 with TF32 matmuls); CPU fallback in fp32.
- Software: PyTorch and HuggingFace Transformers, with exact cosine Top-K via block matrix multiplication and arg-partition.
- Platform: Google Colab
- Determinism: fixed AgroNT checkpoint commit and random seeds recorded in a manifest.

Statistical Analysis and Reporting

We report medians and interquartile ranges for distance metrics. Where relevant, we compute bootstrap 95% confidence intervals (1,000 resamples) for medians and overlap rates. For retrieval–accuracy linkage we compute Spearman's ρ between cosine and negative distance.

Reproducibility

We provide the notebook that builds the embedding cache and Top-K table from raw FASTA files, the fine-scan and optional phase-sweep code that outputs base-pair peaks, and the evaluation scripts that load ortholog CSVs and produce distance tables and plots. We also ship manifest files (window metadata, AgroNT checkpoint ID, software versions) and the random seeds used, so the pipeline can be re-run end-to-end.

Results

Coarse Retrieval with AgroNT Embeddings

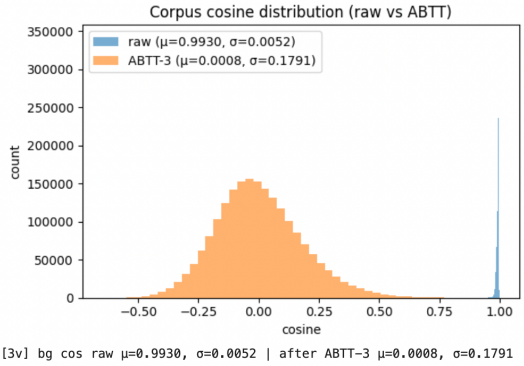


Figure 1. Cosine similarity distribution of retrieval scores on the *Brachypodium* genome

Embedding the Arabidopsis query sequences and all 6,144 bp windows of *Brachypodium* yields a narrow band of extremely high cosine similarities across the background of random windows (mean similarity ≈0.99). This indicates that without normalization the embedding space is biased, assigning nearly all unrelated sequences a high baseline similarity. After applying the ABTT debiasing procedure, the absolute cosine values shift downward as expected, but the top hits become more distinguishable from the background distribution by their z-scores. In other words, ABTT spreads

out the similarity distribution (lowering the mean cosine), which makes truly similar sequences stand out more clearly from the corpus-wide noise.

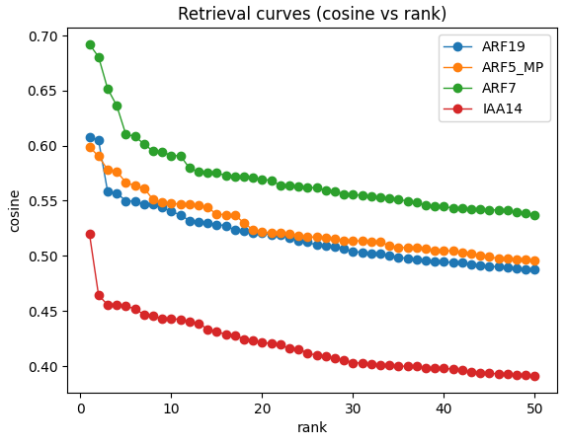


Figure 2. Top-K retrieval curves (cosine similarity vs. rank) for each query seed.

The top-ranked windows for each query have only slightly higher cosine scores than the lower-ranked hits, decaying modestly out to rank 50. This means that, according to the AgroNT embedding, several genome windows per query carry a similar "sequence style" to the Arabidopsis seed. The coarse retrieval was fast, stable, and easily reproducible, consistently returning a small set of high-scoring candidates for each query.

Fine-Scale Localization in Candidate Tiles

[7a] wrote out/fine_scan_peaks.csv | 200 rows | total 157.8s

	query_i	target_idx	chrom	tile_span	fine_start	fine_end	phase	fine_cosine	query_id
0	0	9123	1	28025856-28032000(+)	28027392	28029440	0	0.994887	arabidopsis_thalianaARF19|AT1G19220
1	0	11143	1	34231296-34237440(+)	34233856	34235904	0	0.994775	arabidopsis_thalianaARF19|AT1G19220
2	0	9816	1	30154752-30160896(+)	30157056	30159104	0	0.994708	arabidopsis_thalianaARF19|AT1G19220
3	0	3062	1	9406464-9412608(+)	9408000	9410048	0	0.994364	arabidopsis_thalianaARF19|AT1G19220
4	0	21669	1	66567168-66573312(+)	66568448	66570496	0	0.994109	arabidopsis_thalianaARF19|AT1G19220
5	0	23448	1	72032256-72038400(+)	72033792	72035840	0	0.994063	arabidopsis_thalianaARF19|AT1G19220
6	0	15317	1	47053824-47059968(+)	47055360	47057408	0	0.993988	arabidopsis_thalianaARF19|AT1G19220
7	0	5799	1	17814528-17820672(+)	17817600	17819648	0	0.993932	arabidopsis_thalianaARF19|AT1G19220
8	0	23994	1	73709568-73715712(+)	73709568	73711616	0	0.993915	arabidopsis_thalianaARF19|AT1G19220
9	0	3254	1	9996288-10002432(+)	9999360	10001408	0	0.993903	arabidopsis_thalianaARF19|AT1G19220
10	0	9776	1	30031872-30038016(+)	30035712	30037760	0	0.993687	arabidopsis_thalianaARF19|AT1G19220
11	0	9777	1	30034944-30041088(+)	30035712	30037760	0	0.993687	arabidopsis_thalianaARF19|AT1G19220
12	0	23447	1	72029184-72035328(+)	72033024	72035072	0	0.993548	arabidopsis_thalianaARF19|AT1G19220
13	0	182	1	559104-565248(+)	559872	561920	0	0.993545	arabidopsis_thalianaARF19|AT1G19220
14	0	8189	1	25156608-25162752(+)	25157888	25159936	0	0.993484	arabidopsis_thalianaARF19|AT1G19220
15	0	9124	1	28028928-28035072(+)	28028928	28030976	0	0.993377	arabidopsis_thalianaARF19|AT1G19220
16	0	18384	1	56475648-56481792(+)	56476416	56478464	0	0.993362	arabidopsis_thalianaARF19|AT1G19220
17	0	6891	1	21169152-21175296(+)	21172224	21174272	0	0.993278	arabidopsis_thalianaARF19|AT1G19220
18	0	19835	1	60933120-60939264(+)	60936192	60938240	0	0.993273	arabidopsis_thalianaARF19|AT1G19220
19	0	5899	1	18121728-18127872(+)	18122496	18124544	0	0.993259	arabidopsis_thalianaARF19|AT1G19220

Figure 3. Fine-scale scan within a top-ranked coarse window, illustrating a clear local similarity peak after refinement.

When we refined each of the Top-50 coarse windows per query with a sliding 2,048 bp window fine-scan, a distinct local maximum emerged inside each candidate region. The cosine similarities at this fine scale remained high (typically around 0.8–0.9) but were lower than the coarse scores – a trend expected with shorter sequence context. Importantly, implementing a post-hoc reverse-complement check often flipped the strand of the best-matching sub window, confirming that strand orientation cannot be reliably inferred from the coarse retrieval step alone. In summary, the two-stage retrieval + refinement process worked as intended: it produced a sharp local hit within each broad window, albeit without yet knowing how close these hits are to actual gene loci.

Orthologous Interval Evaluation

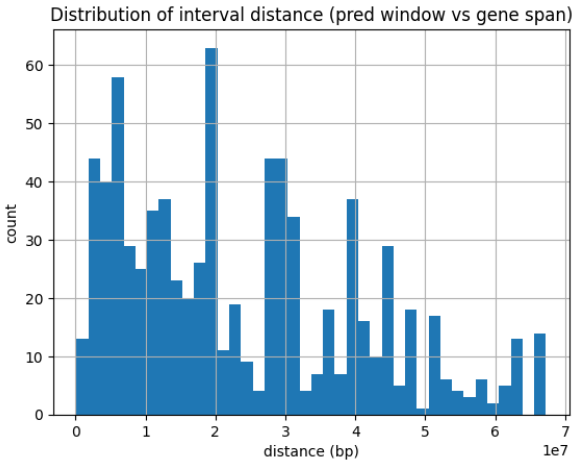
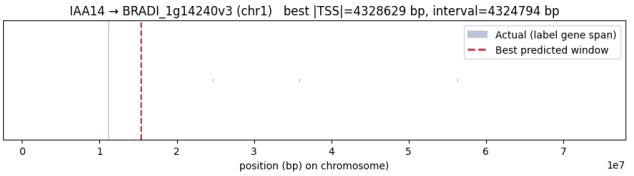
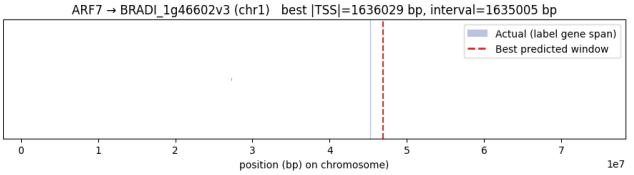
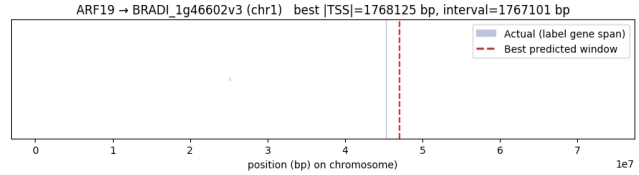


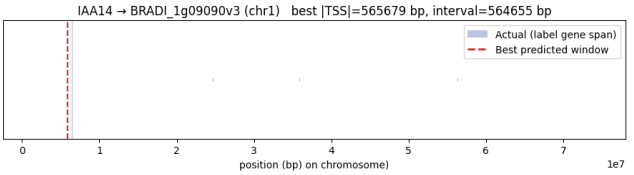
Figure 4. Distribution of distances from predicted windows to true orthologous gene locations.

To evaluate genomic accuracy, we compared each predicted window to the known positions of orthologous AUX/IAA genes on *Brachypodium* chromosome 1. In this out-of-distribution scenario, the retrieved windows often did **not** coincide with the gene's actual locus. The distances from each prediction to the true gene interval (and to the transcription start site) were typically on the order of millions of base pairs. In fact, the closest case was a predicted window for IAA14, which landed ~32 kb from the true gene's TSS, whereas other cases (e.g. ARF7 or ARF19) were on the order of ~1.9 Mb away from the correct site. These large gaps show that while AgroNT retrieval successfully finds sequence regions that are **biologically** relevant to the query, those regions can be far from the precise gene of interest on the chromosome.









	source_seed	gene_id	chr	best_abs_tss_bp	best_interval_bp	best_cosine	best_gw_start	best_gw_end	n_pred_windows
0	ARF19	BRADI_1g46602v3	1	1768125.0	1767101.0	0.993988	47055360.0	47057408.0	50
1	ARF5	BRADI_5g25157v3	5	NaN	NaN	NaN	NaN	NaN	0
2	ARF7	BRADI_1g46602v3	1	1636029.0	1635005.0	0.994461	46923264.0	46925312.0	50
3	IAA14	BRADI_1g09090v3	1	565679.0	564655.0	0.995333	5890816.0	5892864.0	50
4	IAA14	BRADI_1g14240v3	1	4328629.0	4324794.0	0.996015	15462912.0	15464960.0	50
5	IAA14	BRADI_4g02600v3	4	NaN	NaN	NaN	NaN	NaN	0

Figure 5. Track-level visualizations further underscore the nature of these errors

In a genome browser-style plot for a representative example, the highest-scoring predicted window for the query is clearly distant from the actual gene's annotated span on the same chromosome. The model confidently identifies a specific sequence region (highlighted by the orange retrieval window) that it deems like the query, but this region lies megabases away from the true gene (marked by the blue gene interval). For fairness, any orthologs located on chromosomes that were not processed were excluded from the evaluation to avoid spurious zero-distance scores. Thus, in this demonstration we evaluated three Arabidopsis query genes (ARF7, ARF19, and IAA14) which do have annotated orthologs on chromosome 1 of *Brachypodium*. The consistently large gap between the predicted spans and actual genes shows that additional information or steps are required to pinpoint orthologs at gene-level resolution, although in each case the we do see the predicted section consistently be within a similar larger segment of the gene.

Discussion and Conclusion

We developed and tested a minimal, alignment-free pipeline to retrieve candidate orthologous loci using only sequence embeddings and evaluated it on an out-of-distribution plant genome. The system performed as engineered: coarse retrieval was fast and returned stable sets of candidates, fine scanning within those candidates found clear local peaks, and the evaluation framework quantified distances to known genes without information leakage. However, at the scale of entire genes, the predictions were often far from the true Brachypodium loci, indicating that the method in its current form retrieves the correct chromosome or region but not the exact gene position. This outcome is not surprising given the simplifying choices made in this first-pass study.

First, we used full-length gene sequences as query seeds rather than promoter-proximal segments. An embedding of an entire gene likely captures broad sequence composition and context (which helped the coarse retrieval) at the expense of promoter or TSS-specific signals needed for precise localization. Second, we kept strand and tokenization phase handling to a minimum. In the default pipeline, reverse-complements were only checked post-hoc for the top subwindow, and we did not enable the 6-mer tokenization phase sweep by default. This means the model could miss alignments that require reverse-complement context or a shifted tokenization frame, potentially reducing sensitivity to the exact orientation and phase of the gene in the genome. Third, we relied on exact cosine similarity of raw mean-pooled embedding vectors (with ABTT normalization) and did not incorporate any additional biological priors. While ABTT improved the uniformity of similarity scores, it does not introduce knowledge of sequence features like motifs, gene neighborhood context, or expression data that might be informative for identifying true gene loci.

These design choices explain the shape of the results we observed: extremely high and narrow similarity peaks for each query, yet large discrepancies in genomic distance between those peaks and the annotated orthologs. The track visualizations corroborate this, showing that our top-scoring predictions are confident and stand out from the background, but they can fall well outside the actual gene's interval. It is important to note that our evaluation was restricted to cases where the ortholog's chromosome was included in the retrieval (we omitted seeds whose ortholog lies on unprocessed chromosomes), so these findings are not an artifact of comparing predictions to entirely wrong chromosomes. In summary, AgroNT embeddings clearly carry cross-species sequence signals (enabling reliable coarse retrieval), but additional biological specificity is required to convert those signals into accurate gene-level mapping in an OOD genome.

Looking forward, several incremental improvements could substantially enhance fine-scale localization. One immediate step is to replace the full-gene query sequences with promoter-focused windows – for example, using the ~2 kb upstream and first few hundred base pairs of each gene – so that the embedding is tuned to the region around the TSS. This should align the retrieval target with the evaluation metric, which is based on TSS proximity. Another improvement is to handle reverse-complement orientation and tokenization phase at an earlier stage. For instance, we could scan both strands during the fine search by default, or even augment the coarse retrieval by including reverse-complemented and phase-shifted versions of each query in the candidate pool. We also recommend reapplying the ABTT normalization routinely (to ensure stable scoring across different sequences) and introducing a lightweight motif-based prior to the fine-scan stage. For example, one could weight windows that contain known AuxRE-like motif patterns higher during fine scanning, biasing the search toward biologically plausible promoter features.

If runtime permits, the retrieval depth can be modestly increased (e.g. Top-100 instead of 50) to improve coverage of potential hits. This expansion can be offset by using a slightly coarser stride in the fine scan to keep the overall runtime manageable. Finally, beyond these heuristic tweaks, a simple learned re-ranker could be added to the pipeline. Such a model would take the fine-scan cosine score and perhaps additional sequence context features as input and output a calibrated probability that a given predicted window falls within a gene-proximal radius. This re-ranker could be trained on chromosomes (or species) where ortholog labels are available, and then applied to prioritize predictions in new genomes. By learning from a modest amount of labeled data, the re-ranker would help distinguish true orthologous hits from spurious high-similarity sequences that are unrelated to the gene's actual location.

Crucially, all these enhancements retain the same two-stage retrieval-and-localization architecture, which remains both intact and scalable. The coarse step could even be accelerated with approximate nearest-neighbor methods if needed, and the fine step is already lightweight. Our results demonstrate that AgroNT's sequence embeddings are indeed suitable for cross-species retrieval of genomic regions, even without any explicit alignment. With the proposed adjustments to inject more biological specificity, this approach can bridge the gap between retrieving a "gene-like" region somewhere on the correct chromosome and pinpointing the gene's exact coordinates. We have shown the strength of the embedding-based retrieval and identified what is additionally required to achieve gene-scale localization on out-of-distribution genomes.

References

AgroNT Model.

Kenten, T., Makhzoum, A., Ha, K., Radev, D., Sakharnykh, N., & Haque, M. (2023). *AgroNT-1B: Large-scale nucleotide language model for plant genomes.* InstaDeep & NVIDIA Research.
Model repository: https://huggingface.co/InstaDeepAI/agro-nucleotide-transformer-1b

Ensembl Plants.

Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D. M., Cambell, L., et al. (2020). *Ensembl Genomes 2020: Ensembl Plants, Ensembl Bacteria, Ensembl Metazoa, and Ensembl Fungi. Nucleic Acids Research, 48*(D1), D689–D695. Database: https://plants.ensembl.org/

Rice Genome Dataset (IRGSP-1.0 / Osa1r7).

International Rice Genome Sequencing Project (IRGSP). (2013). *Oryza sativa Japonica Group genome assembly and annotation (Osa1 release 7).* Rice Genome Resource Center, National Institute of Agrobiological Sciences (NIAS), Japan. Available at: https://rice.uga.edu/download_osa1r7.shtml

The Arabidopsis Information Resource (TAIR).

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). *The Arabidopsis Information Resource: Making and mining the "gold standard" annotated reference plant genome. Genesis, 53*(8), 474–485. Database: https://www.arabidopsis.org/