GeneCAD: Plant Genome Annotation with a DNA Foundation Model

Zong-Yan Liu[1], Ana Berthel[2], Eric Czech[3], Michelle Stitzer[2], Sheng-Kai Hsu[2], Matt Pennell[4], Edward S. Buckler[1,2,5†] and Jingjing Zhai[2†]

[1] Section of Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY USA 14853
[2] Institute for Genomic Diversity, Cornell University, Ithaca, NY USA 14853
[3] Open Athena AI Foundation, New York, NY, USA 10001
[4] Department of Computational Biology, Cornell University, Ithaca, NY USA 14853
[5] USDA-ARS; Ithaca, NY, USA 14853
[†] Corresponding authors: Edward S. Buckler (esb33@cornell.edu); Jingjing Zhai (jz963@cornell.edu)

Abstract

Accurate genome annotation remains a bottleneck in plants, where polyploidy and repeat-rich sequence confound homology- and RNA-based pipelines. We introduce GeneCAD, a sequence-only method that predicts complete plant gene models directly from DNA. GeneCAD couples representations from a plant DNA foundation model, PlantCAD2, with a lightweight ModernBERT encoder and a chromosome-wide conditional random field that enforces splice-phase and feature order, and applies a protein language-model screen to suppress repeat-driven open reading frames. To limit label noise, we rank and filter public annotations using a sequence-based masked-motif score and fine-tune on five phylogenetically diverse, high-quality references. Across five held out angiosperms, including the allotetraploid *Nicotiana tabacum*, GeneCAD improves transcript-level F1 by 8–10% on average over Helixer and BRAKER3, increases exact match transcripts, and sharpens boundaries at start/stop codons and splice junctions. By removing dependence on species-matched RNA-seq or proteomics while retaining cross-species accuracy, GeneCAD provides an accurate, scalable route to biologically coherent plant gene models from DNA alone.

Main

Coordinated long-read consortia, most notably the Earth BioGenome Project[1], have generated thousands of chromosome-scale reference genomes across plants and animals and are scaling toward tens of thousands more. Realizing the full value of these assemblies now hinges on precise annotation, which enables gene function prediction, anchors genotype–phenotype associations, and accelerates advances from basic biology to crop improvement. For newly assembled genomes, pipelines such as BRAKER3[2] integrate species-matched RNA-seq and homologous protein evidence to train generalized hidden Markov model (HMM) based gene finders (e.g., GeneMark-ETP[3] and AUGUSTUS[4]) and to reconcile predictions (e.g., TSEBRA[5]) for genome-wide gene calling. Homology detection, efficiently implemented with profile HMMs, provides a rapid and broadly applicable signal that complements the tissue- and stage-diverse transcriptomic and proteomic data required for optimal accuracy. However, RNA-seq can inflate false positives, including transposable-element (TE) fragments and pseudogenes[6], and substantial model inconsistency persists even in well-curated species like maize[7,8]. Proteome offers a stronger discriminator than RNA-Seq driven workflows, flagging about 28% of maize transcripts[9] as incorrectly annotated or nonfunctional while agreeing with classical and proteome-supported sets. Although proteomics yields high-confidence coding evidence, it remains expensive and low throughput, and many specimens provide only DNA with no accessible RNA, let alone proteome. Limited coverage therefore leads to false negatives, including missed exons and low abundance isoforms[10]. For wild or rare plant specimens, generating comprehensive RNA-seq and proteomes is often infeasible, and sequences are frequently fragmented. In the absence of these inputs, RNA- and protein-dependent pipelines can fail or return incomplete models, and homology support weakens with phylogenetic distance. Under such constraints, limited assay coverage, polyploidy, and repeat-rich genomes further hinder genome annotation.

To reduce dependence on experimental data, sequence-only deep learning tools such as Helixer[11], Tiberius[12], and ANNEVO[13] predict genes directly from DNA without additional transcriptomic or homology inputs. These approaches are attractive because they scale across species and circumvent the need for assay generation. Because supervised models require large labeled datasets to achieve cross-species performance, they are typically trained on public reference annotations from public databases such as Phytozome[14], Ensembl[15], and RefSeq[16]. However, these resources vary in quality and consistency and contain known misannotations, including pseudogenes, TE-derived open reading frames, fused or split genes, and junctions reflecting pre-spliced RNA, which can propagate label noise during training. We therefore hypothesize that stronger curation of training data is necessary to achieve consistent, biologically coherent performance across plants.

Recent advances in self-supervised foundation models have opened new possibilities for genome annotation at scale. Trained on raw DNA or protein, these models learn conservation-aware representations that transfer across species. Independent studies have likewise trained large nucleotide-level segmentation models on extensive annotations and reported strong transfer. For example, SegmentNT[17] built on the Nucleotide Transformer[18]. For DNA, representative models include PlantCAD[19], PlantCAD2[20], GPN[21], Nucleotide Transformer[18], and

Evo2[22]; for proteins, ESM2[23], ESM3[24]. PlantCAD2 was pretrained on 64 diverse plant genomes with an 8,192-bp context window, capturing long-range signals of evolutionary constraint across angiosperms, after which task-specific heads incorporate structural and functional priors. Using PlantCAD2, we demonstrate cross-species generalization with limited labels. However, nucleotide-level labeling alone does not constitute gene prediction; without structure-aware decoding, outputs can violate annotation rules and cannot be assembled into valid transcript models.

We introduce GeneCAD, a conservation-driven framework for plant genome annotation that works directly from DNA. GeneCAD couples single-nucleotide-resolution representations from PlantCAD2[20] with a ModernBERT[25] head and a chromosome-scale conditional random field (CRF)[26] that enforces gene-structure constraints. Single-nucleotide BILOU tags[27] capture boundaries for UTRs, introns, and coding segments, and the CRF converts those scores into coherent transcripts that respect feature order and splice rules. To limit label noise during fine-tuning, we curated five high-quality reference genomes with a zero-shot masked-motif score and trained GeneCAD on a compact, phylogenetically diverse subset. A protein language model screen removes phase-consistent but implausible open reading frames typical of repeat-rich regions. Together, these components connect fine-grained conservation signals to transcript-level coherence, producing accurate gene models directly from DNA and outperforming existing sequence-only baselines in cross-species benchmarks.

Results

Overview of the GeneCAD framework

Current genome-annotation approaches fall into two groups. RNA- and protein-supported pipelines, such as BRAKER3, achieve high accuracy but require extensive experimental data. Sequence-only deep-learning systems, such as Helixer and Tiberius, operate directly on DNA and scale across species, yet they often need large training sets and can fragment loci or blur boundaries (Fig. 1a). GeneCAD bridges these paths by using the pretrained PlantCAD2 DNA foundation model to encode evolutionary conservation across angiosperms, then fine-tunes a ModernBERT head with a chromosome-wide conditional random field to produce coherent, single-transcript gene models (Fig. 1b).

First, to supply cross-species context, assemblies are tiled into 8,192-bp windows and PlantCAD2 provides nucleotide-resolution embeddings. A ModernBERT head accepts these embeddings and produces three feature streams aligned to key biological signals, one for transcriptional boundaries, one for splice junctions, and one for coding frames. A gated MLP then fuses the streams and assigns BILOU labels across 5′ UTR, coding sequence, intron, and 3′ UTR. This design sharpens boundaries because splice sites and start/stop codons are governed by short, conserved motifs; the dedicated splice and frame streams concentrate probability at these positions, and the BILOU scheme localizes transitions to single bases. In contrast, UTRs lack a

single consensus motif, so evidence accumulates over longer spans and the fused output varies more gradually in these regions.

Second, to enforce biological coherence, we concatenate windowed outputs in genomic order and decode them with a linear-chain conditional random field whose transition matrix encodes legal feature order and splice-phase continuity. Viterbi decoding reduces illegal transitions and merges evidence into chromosome-consistent transcripts.

Third, to increase functional precision, two light refinements are applied. We reconnect loci that were split by long introns only when a single in-frame open reading frame with preserved splice phase is supported. We then remove repeat-driven open reading frames using a protein language-model plausibility score from ReelProtein, which increases precision with a small loss of recall.

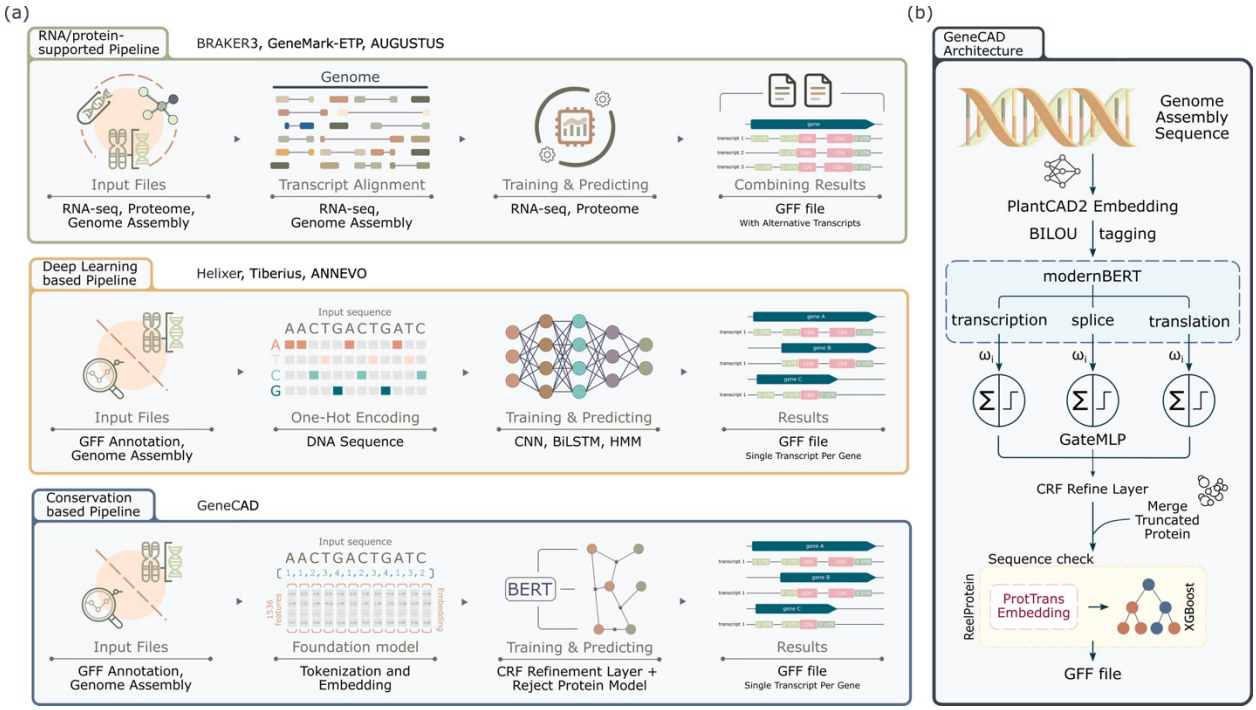

Figure 1 | GeneCAD architecture compared with existing annotation pipelines. (a) **Three pipeline families.** Evidence-supported tools (for example BRAKER3/AUGUSTUS) align RNA-seq and proteomic data to produce multi-isoform annotations. Ab initio deep-learning tools (for example Helixer, Tiberius) operate only on DNA and often blur feature boundaries. GeneCAD uses a conservation-aware, foundation-model strategy: it ingests PlantCAD2 embeddings and applies structured decoding to output one canonical, structurally coherent transcript per locus.

(b) **GeneCAD architecture.** Genome sequence is embedded with PlantCAD2 and labeled at single-nucleotide resolution with BILOU tags. An eight-layer ModernBERT encoder produces contextual states that feed three task streams for transcription boundaries, splicing, and translation. A gated MLP fuses the streams to yield per-base logits. A chromosome-wide CRF with empirically derived transition constraints performs Viterbi decoding, enforcing valid feature order and splice consistency. Post-processing reconnects window-split loci when a single in-frame ORF is supported, and predicted CDSs are screened with ReelProtein. The final output is a GFF3 with one canonical transcript per gene.

Evaluating and curating public genome annotations with masked-motif logistic regression (MMLR)

Public plant genome annotations vary in quality across species and releases, and supervised machine learning techniques rely on high-quality, error-free labels when training models. We hypothesized that a sequence-only score could quantify this variation in quality at the transcript level, allowing us to identify and remove poor quality transcripts while maintaining a high diversity of represented transcripts in our training data. We therefore developed the masked-motif logistic regression score (MMLR), which assigns a transcript-level value directly from DNA sequence. MMLR is derived from PlantCAD2 and measures how well conserved coding and splice motifs are recoverable from local context at translation start and stop codons and at splice donor and acceptor sites (Fig. 2a). PlantCAD2 is pretrained to predict masked nucleotides from the surrounding sequence. High recovery probabilities at these motifs indicate that the sequence patterns in a transcript resemble features that are frequent and conserved across species. We interpret higher MMLR values as stronger evidence for a biologically credible transcript and use MMLR as the transcript-level quality score in this study. We did not include transcription start or termination sites because plant boundaries are diffuse and lack a single consensus motif, as reported previously (Methods).

We applied MMLR to all transcripts in public annotations for four model plants: *Arabidopsis thaliana*, *Zea mays*, *Solanum lycopersicum*, and *Oryza sativa*. In *A. thaliana*, both peptide-supported and unsupported transcripts showed high MMLR values, consistent with the overall quality of the reference. In the other three species, peptide-supported transcripts scored higher on average, indicating that stronger sequence signals tend to align with independent proteomic evidence (Fig. 2b). The separation was modest, which is expected because the proteomic datasets are not exhaustive. Absence of peptide support reflects unknown status rather than poor quality and can result from low abundance or tissue-specific expression.

We then extended transcript MMLR to all 64 genomes used for PlantCAD2 pretraining. The resulting histogram showed broad, species-level variation in score distribution (Extended Data Fig. 1; Extended Data Table 1). For each species, we computed the mean of its transcript-level MMLR values and summarized the results as a histogram across the 64 species, which showed broad variation at the species level (Fig. 2c).

For GeneCAD fine-tuning, we prioritized references using a joint screen that included higher species-level mean MMLR, high assembly contiguity, BUSCO v5.4.3 completeness, and evidence-supported pipelines. This procedure selected five species. Within each selected species, we retained transcripts with MMLR greater than 0.5 and chose the top-scoring isoform at multi-isoform loci. Approximately 60 to 90% of transcripts passed this threshold for each species, yielding a compact and phylogenetically diverse corpus for training and evaluation (Fig. 2c).

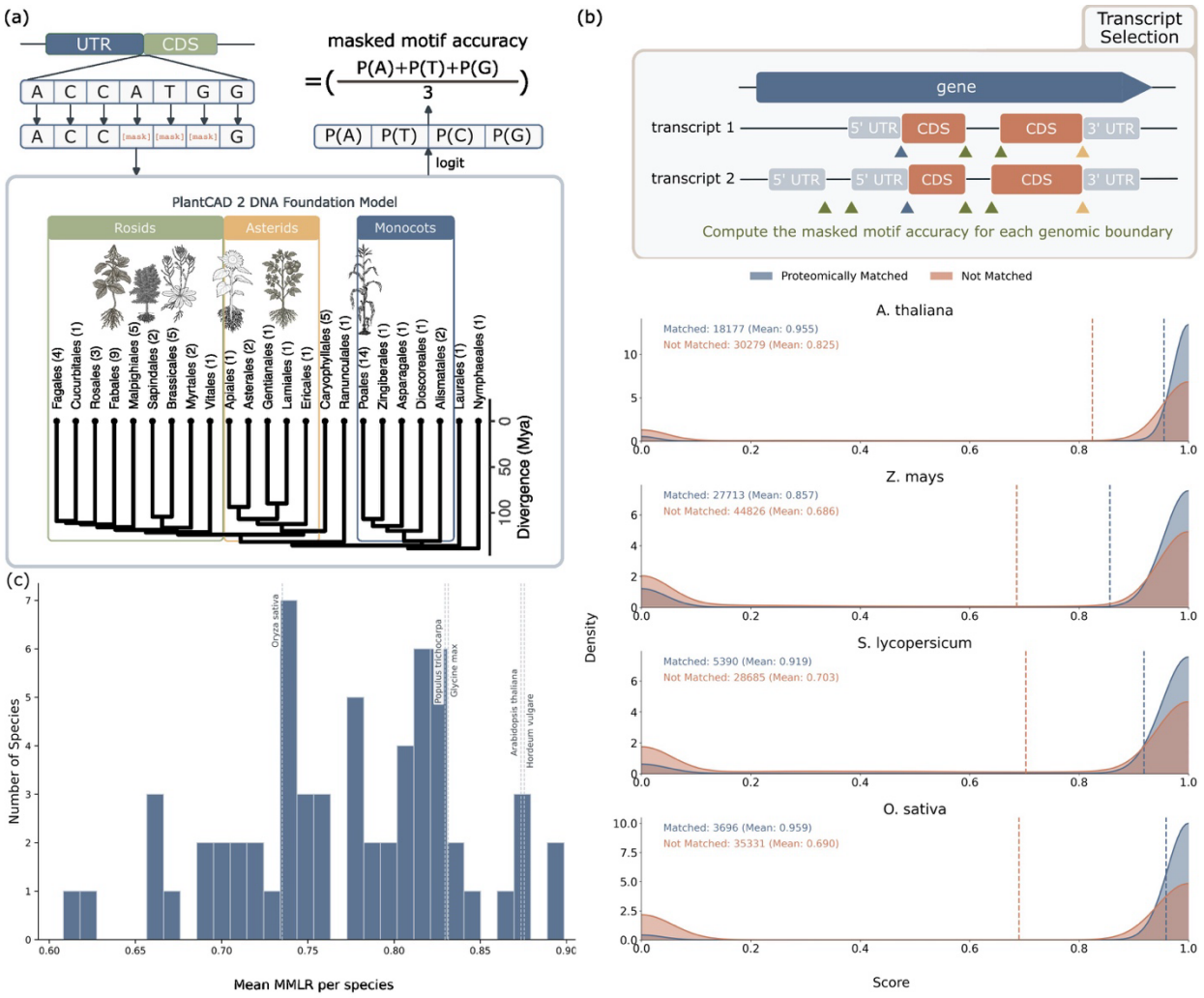

Figure 2 | Masked motif logistic regression (MMLR) enables quality assessment and curation of plant genome annotations.

(a) **Masked-motif scheme.** At each annotated boundary—translation start or stop and splice donor or acceptor—the focal nucleotides are masked. PlantCAD2 is queried zero-shot to recover the masked base from surrounding context. Per-motif probabilities are combined into a transcript-level masked-motif logistic regression score (MMLR).

(b) **Transcript-level scoring and proteome support.** For each transcript, masking is applied at all coding and splice boundaries to yield one MMLR value used for selection. Density plots compare MMLR distributions for peptide-supported (blue) and unsupported (red) transcripts in four species (*Arabidopsis thaliana*, *Zea mays*, *Solanum lycopersicum*, *Oryza sativa*). Vertical dashed lines mark group means; the curation threshold is indicated at right in each panel.

(c) **Species-level variation and reference selection.** Mean MMLR per species across the PlantCAD2 pretraining genomes shows broad variation. Dashed markers denote the five references selected for fine-tuning based on MMLR, assembly quality, and BUSCO completeness (*A. thaliana*, *Glycine max*, *Hordeum vulgare*, *O. sativa*, *Populus trichocarpa*).

Cross-species benchmarking shows higher accuracy and precision across five angiosperms

We benchmarked GeneCAD against the leading pipelines BRAKER3, Helixer, and Tiberius on five held-out angiosperm genomes that were not used for training: *Juglans regia*, *Coffea arabica*, *Phaseolus vulgaris*, *Nicotiana sylvestris* (diploid), and *Nicotiana tabacum* (allotetraploid). These references have strong experimental support, enabling evaluation at the nucleotide, exon, and full-transcript levels under exact-match criteria.

GeneCAD achieved higher accuracy than alternative tools across all metrics (Fig. 3a). Averaged across species, mean F1 at the base level was 0.76 with a range of 0.73 to 0.79, exon-level F1 was 0.70 with a range of 0.67 to 0.75, CDS-level F_1 was 0.58 with a range of 0.54 to 0.66, and full-transcript F1 was 0.44 with a range of 0.42 to 0.49. The advantage held in both diploid and polyploid genomes. In *N. tabacum*, which is challenging to annotate because of duplicated subgenomes and repeat-rich sequence, transcript-level F1 reached 0.42, indicating robust performance in complex genomes (Fig. 3a, Extended Data Table 2).

Boundary profiles showed sharper, higher-confidence transitions at conserved landmarks such as initiation codons and splice junctions (Fig. 3b). Predictions across 5′ and 3′ UTRs displayed smoother gradients, consistent with diffuse transcription start and termination boundaries observed in plants. In several loci, GeneCAD predicted UTRs do not present in the reference. These count as false positives under the benchmark but likely reflect more complete models.

Visual inspection of representative regions highlighted differences in specificity (Fig. 3c). *Ab initio* baselines frequently produced multiple gene-like calls in intergenic sequence, often overlapping annotated pseudogenes or transposable-element fragments. By requiring concordant evidence for transcription, splicing, and translation before emitting a gene, GeneCAD reduced these spurious calls and achieved higher precision at comparable recall.

DNA foundation model and curated data drive GeneCAD performance

To understand why GeneCAD outperforms existing tools, we ran ablation tests starting from the full system and sequentially removing each component one at a time. We focused on CDS-level evaluation as the primary metric because UTR annotations in public references remain heterogeneous, and plant transcription start and termination sites are diffuse in biology.

Beginning with the complete system, we first assessed the two lightweight post-prediction refinements. Because inference runs in 8,192 bp windows, some genes with long introns can be split across adjacent segments; reconnecting loci when a single in-frame open reading frame with preserved splice phase is supported restores full-length structures without altering local boundaries. The protein language-model screen removes coding-like predictions that lack protein

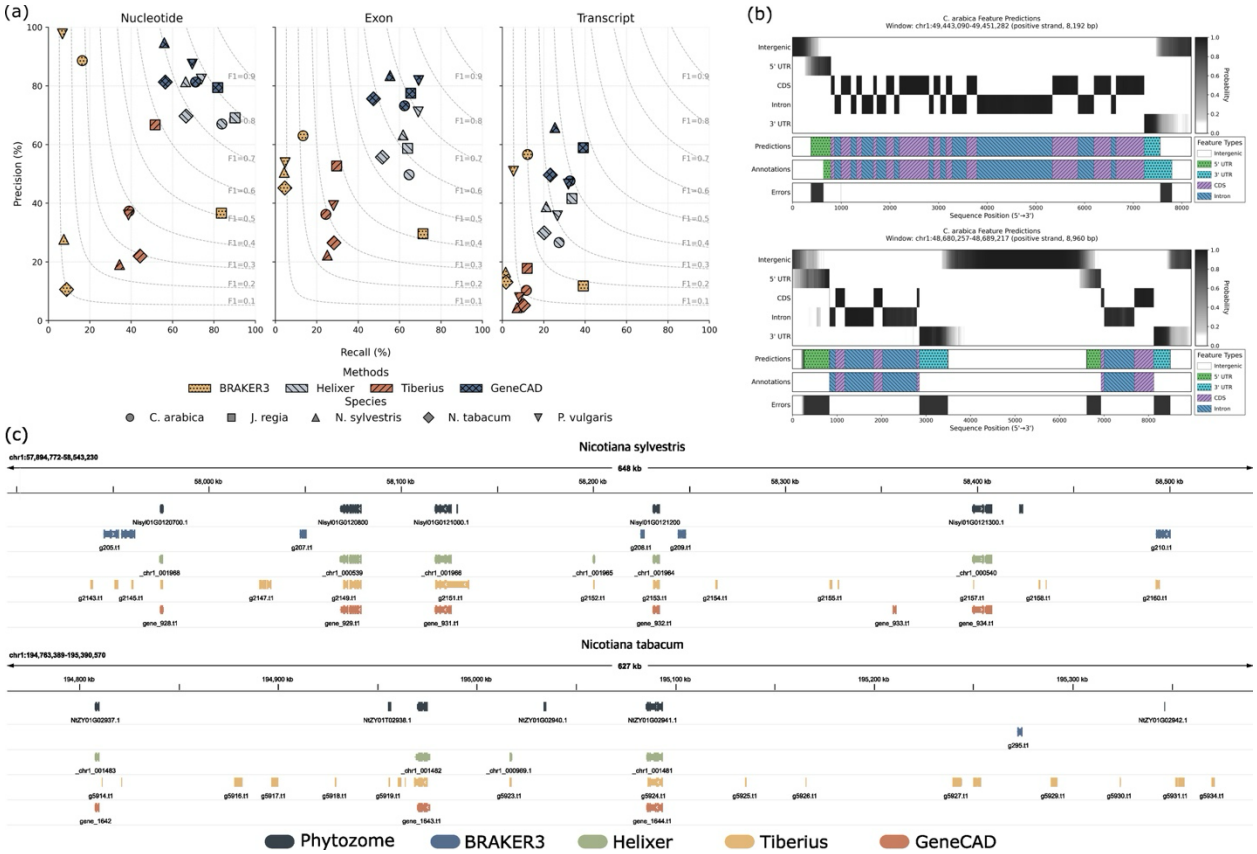

Figure 3 | Benchmarking GeneCAD against evidence-supported and sequence-only pipelines
(a) **Precision–recall across evaluation levels.** Precision–recall plots at nucleotide, exon, and transcript levels compare BRAKER3, Helixer, Tiberius, and GeneCAD on five held-out angiosperms (*Juglans regia*, *Coffea arabica*, *Phaseolus vulgaris*, *Nicotiana sylvestris*, *Nicotiana tabacum*). Points are colored by species; F1 iso-contours are shown. GeneCAD occupies the best precision–recall region across panels, with the largest gains at the transcript level.
(b) **Boundary profiles in an example window.** Per-base class probabilities for GeneCAD are shown above the reference track. Probabilities peak at start/stop codons and splice junctions, while 5′ and 3′ UTRs appear as broader gradients. The error track marks differences relative to the reference.
(c) **Representative loci in *Nicotiana*.** Browser views in N. sylvestris and N. tabacum compare predictions from Phytozome (reference), BRAKER3, Helixer, Tiberius, and GeneCAD. GeneCAD closely matches curated models and reduces intergenic calls relative to ab initio baselines, including in repeat-rich regions and across subgenomes.

-like regularities. Dropping these refinements reduced precision, with the protein screen yielding a modest gain while recall changed little. Larger declines arose when structure-aware decoding and conservation features were removed. Without the chromosome-scale CRF, feature transitions became incoherent and long loci fragmented. Without PlantCAD2 conservation features, the ModernBERT head failed to learn robust cross-species rules. In both cases, CDS-level F1 fell sharply across validation species. On average across species, removing the CRF reduced CDS-level F1 by around 5%, and removing PlantCAD2 reduced it by 30%. Together, these tests show that evolutionary priors from PlantCAD2, structure-aware decoding, and the two

refinements provide complementary gains, with PlantCAD2 and the CRF accounting for most of the improvement (Fig. 4a; Extended Data Figure 2; Extended Data Table 2).

We next asked how much curated data the model requires. Starting from the five curated species selected by the MMLR screen, we reduced training diversity from five to two species, retaining *Oryza sativa* and *Arabidopsis thaliana*. CDS-, exon-, and transcript-level F1 values decreased only modestly, and the two-species model retained most of the accuracy of the five-species model. Precision–recall points shifted slightly toward lower recall but followed similar F1 contours, indicating that conservation features transfer well and that a compact, phylogenetically diverse training set is sufficient for high performance (Fig. 4b; Extended Data Table 2).

Overall, conservation features, CRF decoding, and light post-prediction checks yield coherent, protein-plausible gene models from DNA alone, and MMLR enables compact and quality-aware training.

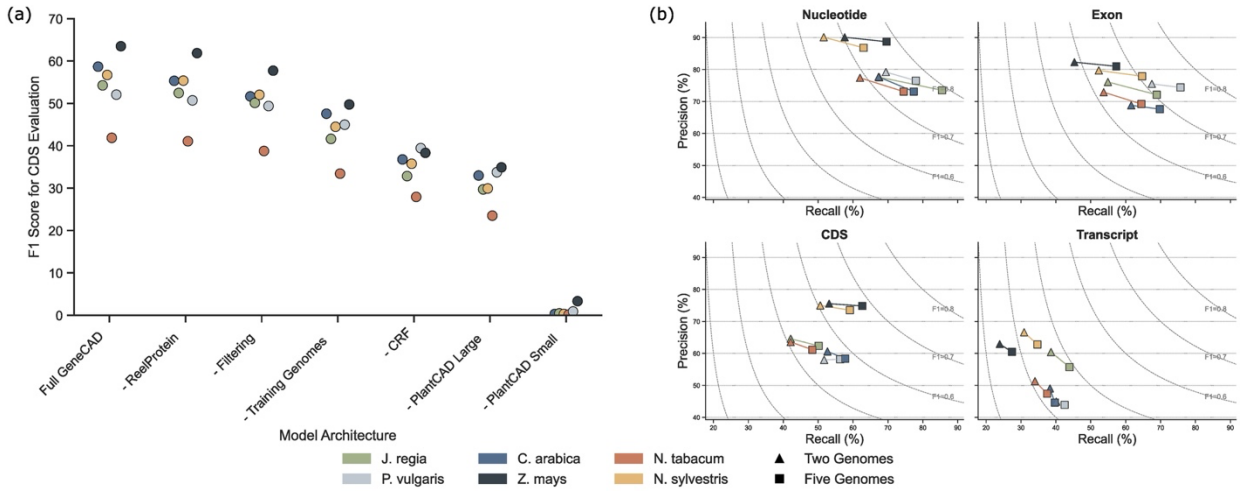

Figure 4 | Contributions of model components and curated data to GeneCAD performance
(a) **Ablation on CDS accuracy.** CDS F1 scores for each test species are shown as components are removed from the full system (left to right): Full GeneCAD, without the protein plausibility screen (ReelProtein), without post-prediction filtering, trained on two rather than five curated genomes, without the chromosome-wide CRF, without PlantCAD2 Large embeddings, and without PlantCAD2 Small embeddings. Points are colored by species (*Juglans regia*, *Coffea arabica*, *Phaseolus vulgaris*, *Nicotiana tabacum*, *Nicotiana sylvestris*; *Zea mays* shown for analysis only). Removing PlantCAD2 features and the CRF produces the largest drops, indicating that conservation features and structure-aware decoding drive most of the gain.
(b) **Precision–recall trade-offs by training diversity.** Precision–recall plots at nucleotide, exon, CDS, and transcript levels compare models trained on two curated genomes (triangles) versus five curated genomes (squares). Curves show F1 iso-contours. Across species, the five-genome model shifts toward higher recall with similar precision, whereas the two-genome model retains most of the accuracy, especially at the CDS level.

Discussion

GeneCAD shows that accurate and scalable plant gene annotation is possible without species-matched transcriptomics or proteomics. By coupling pretrained plant DNA representations from PlantCAD2[20] with structure-aware decoding and a protein-level plausibility check[9], GeneCAD converts evolutionary grammar in sequence into biologically coherent gene models. Across held-out angiosperm genomes that include diploids and an allotetraploid, the method delivers higher accuracy at exon, CDS, and full-transcript levels than current pipelines while using only DNA sequence as its input.

Toward conservation-driven, sequence-only annotation

RNA- and protein-supported pipelines such as BRAKER3[2] remain highly effective in well-studied species, but they are hard to scale when experimental resources are sparse or uneven across taxa. To reflect conditions typical for wild or collected specimens, our benchmark provided ten RNA-seq libraries per genome spanning multiple tissues and stages and no proteomics (Fig. 1a). BRAKER3 usually benefits from deeper RNA-seq and protein evidence, so this configuration likely underestimates its best-case accuracy. Even so, precision was high, yet many transcripts lacked sufficient evidence and were missed, consistent with elevated false negatives when coverage is narrow.

In practice, BRAKER3 training also demands substantial compute time, and UTR modeling remains challenging, which can lower exact transcript matches in our evaluation (Fig. 3a). *Ab initio* deep-learning methods such as Helixer[11] and Tiberius[12] lessen the dependence on assays, but they typically require large supervised datasets and can struggle to maintain long-range coherence and biologically realistic gene structures."

GeneCAD uses a conservation-based approach. Pretrained plant DNA representations from PlantCAD2 capture evolutionary grammar in sequence, and structure-aware decoding[26] with a protein-level plausibility check[9] converts those signals into coherent gene models (Fig. 1b). Because GeneCAD fine-tunes a pretrained DNA foundation model, it trains effectively on a small, carefully curated corpus and still achieves strong cross-species performance. In contrast, prior *ab initio* systems usually rely on much larger annotation sets to reach comparable generalization. This data efficiency is important in plants, where high-quality labels are uneven and many clades lack deep resources.

Requiring concordant transcription, splicing, and translation signals, and screening predicted proteins, also reduces intergenic false positives in both diploids and polyploids, indicating robustness to genome size and repeat content[10,28].

In practice, conservation-driven annotation complements RNA and protein supported workflows. It provides strong baseline gene models where assays are limited, and in data-rich

species helps triage loci for targeted validation. As high-contiguity plant genomes proliferate, anchoring annotation in evolutionary conservation via PlantCAD2 offers a scalable route to generate consistent cross-species gene models while keeping training demands modest.

Quality-aware learning as a complementary factor

Public plant annotations are heterogeneous[6,8,9]. Even well-studied species mix high-confidence genes with questionable models, and unfiltered labels can blunt boundary resolution and inflate intergenic false positives. We therefore used a sequence-only metric, the masked-motif logistic regression score (MMLR) derived from PlantCAD2, to score each transcript and select high-confidence training examples across phylogenetically diverse species (**Fig. 2a**). MMLR aligns with independent peptide support on average[10]. It also highlights a related issue with proteomics, since some biologically real transcripts may lack detectable peptides when expression is low or tissue specific[29]. In these cases, MMLR can retain credible models that proteomics does not capture (Fig. 2b).

Curating by MMLR produced a smaller but cleaner corpus and improved performance, with gains most evident at CDS and exon levels and with fewer spurious intergenic calls. These improvements reflect sharper boundary delineation and better splice-phase consistency (Fig. 3b). Our experiments focus on curated, high-confidence references, so we limit claims to this setting; the impact of including lower-quality annotations and the optimal choice of threshold remain to be tested. Together with conservation-aware representations from PlantCAD2, quality-aware curation supplies cleaner supervision during fine-tuning and contributes to precise and biologically coherent models.

Biological realism and interpretable outputs

Beyond overall accuracy, GeneCAD is designed for biological realism and interpretability. The model emits a gene only when sequence signals for transcription, splicing, and translation are concordant, including promoter and UTR context, canonical splice motifs, and an in-frame start and stop codon with preserved phase. This suppresses gene-like calls that are not consistent with the central dogma. Chromosome-scale decoding with a linear-chain CRF enforces legal feature order and splice-phase continuity across long genomic spans. The CRF's transition matrix penalizes invalid boundary transitions, including donor and acceptor mismatches, which reduces incoherent transitions and fragmentation in gene-dense or repeat-rich regions (Fig. 3b). In ablation tests, removing the CRF reduced CDS-level F1 and introduced incoherent feature transitions, confirming its contribution to structure and coherence (Fig. 4a).

GeneCAD's probability outputs provide an interpretable view of sequence signals. Probability profiles rise and fall sharply at initiation codons and splice junctions, while gradients across 5′ and 3′ UTRs are smoother[30], reflecting diffuse transcription boundaries in plants (Fig. 3b). These landscapes quantify uncertainty at base resolution and help curators target additional evidence where it is most informative: for example, at weak splice contexts or uncertain UTR

limits. In several loci, GeneCAD proposes UTRs that are absent from the reference. Under our exact-match scoring, such additions are counted as false positives, yet they often reflect plausible extensions rather than spurious calls. In practice, these loci are good candidates for targeted validation with orthogonal assays such as long-read Iso-Seq, CAGE, or 3′ end profiling, and for iterative curation in evidence-rich species (Fig. 3a).

A complementary protein-level screen adds an orthogonal check on coding plausibility[9]. Using embeddings from a protein language model with a pretrained classifier (ReelProtein), we remove predicted coding sequences that receive low plausibility scores, indicating atypical amino-acid co-occurrence and weak long-range dependency patterns characteristic of nonfunctional or repeat-derived ORFs. This conservative step increases precision with minimal impact on recall, producing call sets that are better suited for downstream analyses and that reduce manual curation burden (Extended Data Fig. 2). Together, sequence-level constraints through the CRF and protein-level plausibility provide a two-tier safeguard. Conservation signals learned by PlantCAD2[20] are translated into coherent gene structures, and candidate proteins are vetted for realism, yielding outputs that are both accurate and biologically credible.

Implications for plant biology and beyond

GeneCAD provides a practical route to high-confidence annotations directly from genome assemblies. For model species such as *Arabidopsis* and rice, it can reduce curation burden and increase consistency across assemblies and releases, yielding more uniform gene models. For orphan crops and wild species, where RNA and proteomic resources are limited, it offers a realistic starting point for functional genomics and comparative analyses. In pan-genome studies, conservation-guided predictions help flag conserved and lineage-specific genes and prioritize loci for validation and breeding.

Ablation tests show that GeneCAD generalizes well with limited supervision (Fig. 4b). Starting from a pretrained DNA foundation model allows effective fine-tuning on a small, carefully curated corpus while retaining most of the accuracy observed with broader training. In practice, users who trust their own annotations can fine-tune GeneCAD on two to five high-quality, closely related genomes and obtain strong performance without assembling large heterogeneous datasets or extensive compute. MMLR further streamlines this process by ranking and filtering candidate transcripts so that fine-tuning focuses on high-confidence labels.

The approach is not plant specific. We expect that replacing PlantCAD2 with an appropriate animal or microbial DNA foundation model and applying the same quality-aware curation and structure-aware decoding will extend the workflow to other lineages, pending systematic evaluation.

Limitations and future directions

GeneCAD leads genome annotation pipelines in accuracy across nucleotides, exon, CDS and transcript metrics (Fig. 3a). Gains are most stable at the CDS level, which is less confounded by isoform choice and by diffuse UTR boundaries. Because the system emits a single canonical transcript per locus, full-transcript matching can undercount loci with genuine alternative splicing, and variable or incomplete reference UTRs can depress transcript scores. Inference over 8,192-base windows also limits effective context. Very long introns may be split across adjacent windows and appear as an intergenic sequence; our merge step reconnects many such cases when splice phase is preserved, but rare edge effects remain.

Several directions can close these limitations. First, by extending effective context and continuity with longer-range attention, stitching decoding across windows, and training in a chromosome-aware manner, long introns and dense gene clusters can remain intact. Second, moving from a single-isoform output to probabilistic isoform modeling would allow scoring of alternative splice chains and selection of a small set of plausible transcripts per locus. Third, we can make supervision more evidence-aware. Rather than adding hard labels, we would refine the pretrained predictor with sequencing-derived rewards that favor concordance with splice junction support from short-read RNA-seq, full-length isoforms from long-read data, transcription start and end profiles, ribosome occupancy over in-frame CDS, and peptide evidence. In this setting, offline reinforcement learning or preference-based optimization[31,32] would learn a reward over candidate transcripts without new sequencing, and contextual bandit updates would prioritize which loci or assays to query next.

Training-data curation can also be broadened. MMLR improves label quality but may exclude valid atypical transcripts; adaptive thresholds or a learned quality score may retain informative outliers while limiting noise. With respect to proteins, expanding the plausibility model to plant-specific and lineage-expanded families is expected to reduce repeat-driven open reading frames while avoiding penalties on novel but credible sequences. Taken together, longer effective context, isoform-aware decoding, evidence-guided refinement, and broader supervision build on the conservation-based foundation of PlantCAD2 and move GeneCAD toward complete, isoform-aware gene models at scale.

Materials and Methods

Source Data and Phylogenetic Scaffolding

We began with the 64 plant genomes used to pretrain PlantCAD2 and applied a sequence-only screen to prioritize references for fine-tuning. For each annotated transcript, we queried PlantCAD2[20] in a zero-shot setting to obtain recovery probabilities at the translation start codon, the translation stop codon, and the splice donor and acceptor dinucleotides. We combined these motif-level probabilities into a transcript-level masked-motif logistic regression score (MMLR) The logistic model was trained with positive–unlabeled learning, using classical *Zea mays* genes as positives and all other transcripts as unlabeled (Extended Data Note 1). We ranked species by mean MMLR and also considered assembly contiguity, BUSCO v5.4.3 completeness[33], and whether annotations were produced by evidence-supported pipelines that yield consistent models. This process selected five references: *Arabidopsis thaliana*, *Glycine max*, *Hordeum vulgare*, *Oryza sativa*, and *Populus trichocarpa* (Extended Data Table 1). MMLR was computed in a zero-shot manner and was used only for species and transcript selection. None of the selected species appears in the held-out benchmarking set. For the external check, we assigned proteome support in *A. thaliana*[34], *Z. mays*[9,34], *S. lycopersicum*[35], and *O. sativa*[36] from public peptide evidence (Fig. 2b). A transcript was considered supported when at least one peptide uniquely mapped to its encoded protein sequence from the reference FASTA. These labels were used only for external check and were not used for GeneCAD training.

Training Data Generation and Curation

Within each selected species, we retained transcripts with MMLR greater than 0.5. At loci with multiple isoforms, we designated the highest-scoring isoform as canonical. We standardized GFF3 files to harmonize feature hierarchies. We excluded transcripts that contained internal stop codons, non-canonical splice forms, or same-strand overlaps with neighboring genes. Retained models were converted to a tabular schema of genes, mRNAs, exons, coding segments, and UTRs, and were paired with genomic coordinates to yield a distilled, non-redundant corpus for fine-tuning and validation (Fig. 2b,c). MMLR was computed as described above and in Extended Data Note 1. Per-species mean MMLR and retention counts are reported in Extended Data Table 1.

Independent Test Set for Rigorous Evaluation

To assess generalization across diverse evolutionary histories and genome architectures, we assembled an independent five-species test set that was not used for GeneCAD training: *Juglans regia* (walnut), *Coffea arabica* (coffee), *Phaseolus vulgaris* (common bean), *Nicotiana sylvestris* (diploid tobacco), and *Nicotiana tabacum* (allotetraploid tobacco) [37–41]. These species span major angiosperm clades and include a diploid–allotetraploid pair, enabling evaluation across ploidy levels. Reference annotations were used as released and were not re-filtered with MMLR. Each assembly is supported by extensive experimental evidence, including RNA-seq,

long-read Iso-Seq, and proteomics. Approximate separations from the nearest training lineage range from tens to more than 100 million years (Extended Data Table 1). We also evaluated *Zea mays* as an additional held-out case. *Z. mays* was not used to fine-tune GeneCAD and provides a high-evidence reference that is informative for model behavior. We used maize to quantify absolute performance and to analyze the effects of pretraining and curated training diversity, including ablations (Fig. 4). To avoid circularity, we did not include maize in the primary cross-tool benchmark because it appears in the training data of some comparator pipelines.

Data representation and labeling

Genomic FASTA files for the curated training set and the independent test set were tokenized to integer IDs with a single-base vocabulary (A, C, G, T, N). Forward and reverse-complement token streams were stored in Zarr format to permit random access during training.

Per-base labels were derived from curated canonical transcripts. Four feature types were annotated at nucleotide resolution: CDS, intron, 5′ UTR and 3′ UTR. Intergenic sequence was assigned to a single Outside (O) class. For each contiguous span of a feature, we used the BILU scheme: the first base was tagged B, internal bases I, the last base L, and single-base spans U. These tags were applied separately for CDS, intron, 5′ UTR and 3′ UTR; all remaining bases were labeled Outside. In total, the labeling scheme defines 17 classes: four feature types, each with four BILU tags, and one Outside class [27].

GeneCAD Model Architecture

Input Representation

Chromosomes were tiled into 8,192-bp windows with a 4,096-bp stride, yielding 50% overlap. At each position, the input vector combined a frozen contextual embedding from PlantCAD2 with a trainable token embedding for the local base (A, C, G, T, or N). PlantCAD2 hidden states were projected linearly and then concatenated with the token embedding to produce a 768-dimensional per-base representation for the head encoder (Fig. 2b). We used two backbone variants that differ only in the size of the frozen PlantCAD2 encoder but yield the same 768-dimensional input. In GeneCAD-Small, the 1,536-dimensional PlantCAD2 state was projected to 640 dimensions and concatenated with a 128-dimensional token embedding. In GeneCAD-Large, the 3,072-dimensional PlantCAD2 state was likewise projected to 640 dimensions and concatenated with the same 128-dimensional token embedding. In both cases the PlantCAD2 encoder remained frozen during GeneCAD training, and only the projection layer and the token embedding were learned to adapt local motif cues to the annotation task. The resulting sequences of 768-dimensional vectors were passed to the ModernBERT head encoder and the downstream structured decoder (Fig. 1b).

ModernBERT Encoder

The head encoder was an eight-layer ModernBERT stack with hidden size 768 and eight attention heads per layer. Each layer applied multi-head self-attention with rotary positional

encodings, followed by a feed-forward block with GELU and an intermediate width four times the hidden size[42]. Residual connections and layer normalization were used throughout, with dropout 0.1. The encoder consumed the 768-dimensional per-base inputs from the PlantCAD2 projection and token embedding, preserved sequence length, and returned 768-dimensional contextual states for every position, which were passed to the classifier and then to the chromosome-scale decoder (Fig. 2b). To expose biological priors, we formed three lightweight feature streams from the encoder outputs: one emphasized transcriptional boundaries, one targeted splice junctions, and one focused on coding-frame cues. The streams were learned end-to-end and were fused before per-base classification.

Gated MLP Classification Head

Final per-base classification used a gated MLP matched in width to the feed-forward block. For each position, a linear projection produced a signal vector and a gate vector. The signal was activated and multiplied element-wise by the gate, followed by dropout 0.1 and a linear map to the 17 output classes that covered CDS, intron, 5′ UTR and 3′ UTR in BILU tags, plus the shared Outside class. The resulting logits were provided to the chromosome-scale decoder.

Model Training and Optimization

Models were implemented in PyTorch with PyTorch Lightning and trained with mixed-precision. We used AdamW (learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$, and a weight decay of 0.01). The schedule applied a linear warm-up over 10% of steps and then held the learning rate constant for the remainder of training. Dropout (0.1) was applied in attention, MLP, and classifier layers. Mini-batches contained eight windows per GPU, and training ran with distributed data parallelism across multiple GPUs.

The objective was token-level cross-entropy over $C = 17$ labels, computed only on supervised positions. Let $p_{i,e}$ be the logit for class c at token i, y_i the label, and $m_i \in \{0, 1\}$ a mask that excludes padded and unlabeled tokens. The loss was

$$l_i = -log \frac{exp(p_{i,y_i})}{\sum_{i=1}^{C} exp(p_{i,c})}, L = \frac{\sum_i m_i l_i}{\sum_i m_i}$$

Training was monitored by token-level F1 and by interval-level precision, recall, and F1 for exons, CDS, and transcripts on validation windows. We did not use early stopping.

Chromosome-Wide Conditional Random Field (CRF)

We concatenated per-base logits in genomic order and decoded with a linear-chain CRF to recover the most probable sequence of annotation states across each chromosome (Intergenic, 5′ UTR, CDS, 3′ UTR, Intron). The transition matrix was fixed, with entries estimated from genome-wide label transitions in the *A. thaliana* and *O. sativa* training genomes; it encoded legal feature order and splice-phase continuity by assigning near-zero probability to illegal jumps.

Viterbi decoding produced state-consistent labels, which were then collapsed into segments and converted to gene models.

Post-Prediction Refinement and Validation

After chromosome-scale decoding, we applied two lightweight refinements. We first reconnected fragments that arose when introns exceeded the 8,192 bp inference window. Candidate merges were considered only for immediately adjacent, same-strand predictions with no intervening gene prediction. For each candidate, CDS segments were ordered by transcript direction (with reverse-complement handling on the minus strand), splice phase offsets were respected per exon, and the concatenated coding sequence was retained only if it formed a single open reading frame (ORF) with an ATG start, an in-frame TAA/TAG/TGA stop, and no internal stops. If any criterion failed, fragments were left unmerged. UTR coordinates were not altered by this step.

We then screened the translated CDSs with ReelProtein (reelGene), which embeds proteins with a frozen ProtTrans model and scores plausibility using a pretrained XGBoost classifier fit on maize gene sets. We did not retrain this model. For each species, a decision threshold was selected on a held-out validation split to maximize gene-level F1, and predictions below the threshold were removed. Because this screen operates on translated sequences, it does not modify UTR boundaries.

Performance Evaluation Metrics

We exported GeneCAD predictions as standard GFF3 files and compared them to reference GFF3s obtained from public repositories or source publications, without re-filtering beyond coordinate normalization. Exon and transcript matches were computed with the tool gffcompare using its exact-match criteria (identical intron chain and exon boundaries on the same strand). Nucleotide-level metrics were derived by rasterizing both GFF3 files to per-nucleotide labels and counting a nucleotide as correct only when the predicted class matched the reference label; masked or undefined positions were excluded. CDS agreement was assessed with a custom evaluator that declared a true positive only when all coding segments matched exactly in coordinates, order, count, and splice phase on the same strand; any disagreement yielded a false positive for the prediction and a corresponding false negative for the reference. Precision, recall, and F1 were computed in the standard way, micro-averaged across chromosomes within each species and then averaged across species for reporting.

Acknowledgments

This work was supported by the U.S. Department of Agriculture Agricultural Research Service through a cooperative agreement with Cornell University, the National Science Foundation (grant no. 2240888), and by the National Institutes of Health (grant no. 5R35GM151348). We gratefully acknowledge the SCINet project, the Texas Advanced Computing Center at The University of Texas at Austin, and the BioHPC Cloud provided by the Bioinformatics Facility at Cornell University (RRID: SCR_021757) for providing the computational resources that enabled large-scale experimentation. We also thank Lambda (lambda.ai) in partnership with Open Athena AI Foundation, Inc. for the generous GPU infrastructure grant that enabled validation, containerization, and deployment of our pipeline for the community.

Conflict of interest

The authors declare no competing interests.

Code availability

The source code for GeneCAD is available at https://github.com/plantcad/genecad. The complete masked motif accuracy scores for all analyzed genomes are available in Supplementary Tables.

References

1. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).
2. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
3. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
4. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–9 (2006).
5. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
6. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
7. Jiao, Y. *et al.* Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
8. Tello-Ruiz, M. K. *et al.* Double triage to identify poorly annotated genes in maize: The missing link in community curation. *PLoS One* **14**, e0224086 (2019).
9. Schulz, A. J. *et al.* Fishing for a reelGene: evaluating gene models with evolution and machine learning. *Plant J.* **123**, e70483 (2025).
10. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).
11. Stiehler, F. *et al.* Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* **36**, 5291–5298 (2021).
12. Gabriel, L., Becker, F., Hoff, K. J. & Stanke, M. Tiberius: end-to-end deep learning with an HMM for gene prediction. *Bioinformatics* **40**, btae685 (2024).
13. Ye, K. *et al.* Highly accurate ab initio gene annotation with ANNEVO. (2025).
14. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).
15. Bolser, D., Staines, D. M., Pritchard, E. & Kersey, P. Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods Mol. Biol.* **1374**, 115–140 (2016).
16. Haft, D. H. *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).
17. de Almeida, B. P. *et al.* Annotating the genome at single-nucleotide resolution with DNA foundation models. *Nat. Methods* 1–15 (2025).
18. Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
19. Zhai, J. *et al.* Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2421738122 (2025).
20. Zhai, J. *et al.* PlantCAD2: A Long-Context DNA Language Model for Cross-Species Functional Annotation in Angiosperms. *bioRxiv* (2025) doi:10.1101/2025.08.27.672609.

21. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2311219120 (2023).
22. Nguyen, E. *et al.* HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv [cs.LG]* 43177–43201 (2023) doi:10.48550/arXiv.2306.15794.
23. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
24. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
25. Warner, B. *et al.* Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv [cs.CL]* (2024).
26. Lafferty, J., McCallum, A. & Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML* 282–289 (2001).
27. Ratinov, L. & Roth, D. Design challenges and misconceptions in named entity recognition. in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL '09* 147–155 (Association for Computational Linguistics, Morristown, NJ, USA, 2009).
28. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**, 329–342 (2012).
29. Smaczniak, C. *et al.* Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues. *Nat. Protoc.* **7**, 2144–2158 (2012).
30. Le, N. T. *et al.* Epigenetic regulation of spurious transcription initiation in Arabidopsis. *Nat Commun* **11**, 3224 (2020).
31. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
32. Rafailov, R. *et al.* Direct Preference Optimization: Your language model is secretly a reward model. *Neural Inf Process Syst* **abs/2305.18290**, 53728–53741 (2023).
33. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
34. Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Res.* **34**, D655–8 (2006).
35. Liu, M. *et al.* Candidate regulators of drought stress in tomato revealed by comparative transcriptomic and proteomic analyses. *Front. Plant Sci.* **14**, 1282718 (2023).
36. Lin, F. *et al.* Proteomics coupled with metabolite and cell wall profiling reveal metabolic processes of a developing rice stem internode. *Front. Plant Sci.* **8**, 1134 (2017).
37. Martínez-García, P. J. *et al.* The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J.* **87**, 507–532 (2016).
38. Zan, Y. *et al.* The genome and GeneBank genomics of allotetraploid Nicotiana tabacum provide insights into genome evolution and complex trait regulation. *Nat. Genet.* **57**, 986–996 (2025).
39. Carrère, S. *et al.* First whole genome assembly and annotation of a European common bean cultivar using PacBio HiFi and Iso-Seq data. *Data Brief* **48**, 109182 (2023).
40. Medrano, J. F. *et al.* De novo whole-genome assembly and annotation of Coffea arabica var. Geisha, a high-quality coffee variety from the primary origin of coffee. *G3 (Bethesda)* **15**, jkae262 (2025).

41. Portwood, J. L., 2nd *et al.* MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* **47**, D1146–D1154 (2019).
42. Shazeer, N. GLU Variants Improve Transformer. *arXiv [cs.LG]* (2020).