nature methods

Article

https://doi.org/10.1038/s41592-025-02939-1

Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model

Received: 22 November 2024

Accepted: 26 October 2025

Published online: 24 November 2025

Check for updates

Felix Holst[1,7], Anthony M. Bolger[2,7], Felicitas Kindel[1,2,7], Christopher Günther[1], Janina Maß[3], Sebastian Triesch[1,4], Niklas Kiel[1,4], Nima Saadat[3,4], Oliver Ebenhöh[3,4], Björn Usadel[2,4,5], Rainer Schwacke[2], Andreas P. M. Weber[1,4], Marie E. Bolger[2,7] ✉ & Alisandra K. Denton[1,4,6]

The accurate identification of genes is vital for understanding biological function, yet this remains challenging across many newly sequenced or less-studied species. Here we present Helixer, an artificial intelligence-based tool for ab initio gene prediction that delivers highly accurate gene models across fungal, plant, vertebrate and invertebrate genomes. Unlike traditional methods, Helixer operates without requiring additional experimental data such as RNA sequencing, making it broadly applicable to diverse species. We show that Helixer's pretrained models achieve accuracy on par with or exceeding current tools, producing gene annotations that closely match expert-curated references across multiple evaluation metrics. Its design enables immediate use on genomes without retraining, providing an efficient, accessible solution for genome annotation in both research and applied settings. The tool is available as an open-source software for local installation via GitHub. An online web interface is also available as well as through the Galaxy ToolShed.

Major advances in genome sequencing and assembly[1] have led to a rapid increase in genomics data. Accurate and fast in silico models are needed to analyze these data, extract biological knowledge and thereby accelerate research progress in biology and bioengineering. Gene calling, or structural gene annotation, plays a critical role but has not seen comparable advances in recent years.

Historically, eukaryotic gene calling relied primarily on (generalized) hidden Markov models (HMMs) such as GeneMark-ES[2,3], FGENESH[4] or AUGUSTUS[5,6]. However, these models lack the capacity to fully model biological complexity on their own. As a result, they are currently used as parts of data integration pipelines such as MAKER2[7], PASA[8], TOGA[9] and BRAKER3[10]. These pipelines perform best with wet laboratory data such as RNA sequencing, other extrinsic evidence such as homologous proteins or databases with repetitive elements, and require substantial computational resources, often becoming the bottleneck in genome projects. The inconsistent availability of such resources leads to a heterogeneous quality of the resulting gene models. Errors are still being identified in gene models of extensively studied species such as human and mouse[11], while in less-studied species the lower overall annotation quality is disruptive in large-scale analyses and can necessitate project-specific reannotation[12]. Illustrating this further, many newly sequenced species lack any annotation, with only circa 24% of the latest eukaryotic assemblies on the National Center for Biotechnology Information (NCBI) database having an accompanying annotation, lagging far behind the 89% annotated in prokaryotes. There is a pressing need for improved tools than can

[1]Institute of Plant Biochemistry, Heinrich Heine University, Düsseldorf, Germany. [2]IBG-4 Bioinformatics, Forschungszentrum Jülich, Jülich, Germany. [3]Institute of Quantitative and Theoretical Biology, Heinrich Heine University, Düsseldorf, Germany. [4]Cluster of Excellence on Plant Sciences, Heinrich Heine University, Düsseldorf, Germany. [5]Institute for Biological Data Science, Heinrich Heine University, Düsseldorf, Germany. [6]Recursion, Valence Labs, Montreal, Quebec, Canada. [7]These authors contributed equally: Felix Holst, Anthony M. Bolger, Felicitas Kindel, Marie E. Bolger. ✉e-mail: m.bolger@fz-juelich.de

easily produce consistent high-quality annotations and strengthen the foundation for downstream applications such as target-gene characterization, transcriptomics, proteomics, genome-wide association studies and more.

Deep learning is a transformative technology where massive networks with the capacity to make extraordinarily complex and nonlinear fits are trained on large amounts of data. Deep learning has demonstrated remarkable performance across diverse fields, including language models (for example, GPT-3[13]), playing strategy games (for example, AlphaGo[14]), deciphering mathematical equations[15,16] and protein folding[17]. In recent years, deep learning networks have been employed with promising outcomes for some elements of gene calling such as predicting genes in prokaryotes[18], differentiating coding from noncoding RNAs[19], differentiating exons and introns in human sequences[20], assessing splice site type and strength[21,22] and predicting gene model elements such as start codons, splice sites and poly-adenylation sites[23,24]. Building on these developments, a recent addition to the field is Tiberius, a deep neural network specifically designed for annotating mammalian genomes[25]. In our proof-of-principle version of Helixer, we used deep learning to classify the genic class of each base pair[26], achieving substantial performance gains compared to an existing ab initio predictor and even exceeding the quality of some references in consistency with independent RNA-sequencing data.

Deep learning has already shown promising results in tasks related to eukaryotic gene annotation. This work advances the field by achieving finalized eukaryotic gene models. Here, we present Helixer as an application-ready tool that includes various improvements to our previous work, along with a hidden Markov model-based postprocessing tool named HelixerPost. This is integrated, and in a single command, the Helixer.py script takes the genome sequence as input (in FASTA format) and outputs structural annotations for primary gene models (in GFF3 format). Pretrained models are now provided for fungal, invertebrate and mammalian genomes, in addition to the previously available plant and vertebrate models. Helixer does not require extrinsic data nor species-specific retraining.

Results

Overview of Helixer

Helixer is a deep learning-based framework for eukaryotic gene annotation directly from genomic DNA. It uses a sequence to label neural network that predicts base-wise genomic features including coding regions, untranslated regions (UTRs) and intron–exon boundaries based solely on nucleotide sequence. The architecture integrates convolutional and recurrent layers to capture both local sequence motifs and long-range dependencies, followed by a biologically informed decoding step that assembles coherent gene models. Trained end-to-end on high-quality reference annotations, Helixer generalizes across species without requiring transcriptomic or homology-based evidence. This design enables consistent annotations while minimizing manual curation, providing a scalable solution for annotating newly sequenced genomes and supporting large-scale comparative genomics.

The released Helixer models show state-of-the-art performance compared to existing ab initio gene calling tools and previous Helixer models

As our previous work[26] set the state of the art for base-wise predictions, we first compared the latest models to models trained with the hand-selected six (vertebrate) or nine (plant) species used previously, as well as to all intermittently released models (Supplementary Figs. 1–8 and Supplementary Tables 1–4). The best models released here (vertebrate_v0.3_m_0080, invertebrate_v0.3_m_0100, land_plant_v0.3_a_0080 and fungi_v0.3_a_0100), had the highest median Genic F1 for their phylogenetic target range, and showed a more balanced performance across said range, compared to other Helixer models. Nevertheless, no single model was consistently better for all species, so all plotted models are released to allow researchers to select the model likely to perform best for their species of interest.

Table 1 | Base-wise phase F1 statistics for test species, summarized as the mean value per group

Group	HelixerBW	HelixerPost	GeneMark-ES	AUGUSTUS[a]
Fungi	0.9514	**0.9540**	0.9466	0.9221
Plants	0.8001	**0.8099**	0.4089	0.6300
Vertebrates	0.8641	**0.8829**	0.2326	0.6372
Invertebrates	0.8458	**0.8562**	0.7172	0.7914

The highest performing tool for each group is highlighted in bold. [a]For AUGUSTUS, pretrained models were only available for 4 of the 13 plant, 1 of the 11 vertebrate and 5 of the 15 invertebrate test species. All reported summary statistics reflect only those species.

While promising, the high base-wise performance of raw HelixerBW predictions does not necessarily imply that the performance is maintained through postprocessing. Therefore, we additionally computed the F1 for the phases (phase F1) (Supplementary Figs. 9–13 and Supplementary Table 5), the F1 for the subgenic classes (subgenic F1) (Supplementary Table 6) and the F1 for the genic classes (genic F1) (Supplementary Table 7) scores for the final gene models output by HelixerPost on the selected test species. Importantly, we see that predictions postprocessed via HelixerPost have very similar performance to the HelixerBW predictions, with a slight increase in 43 of 45 test species.

Additionally, we inferred gene models using two existing HMM tools, GeneMark-ES and—where trained models were available for the test species or a closely related species—AUGUSTUS. GeneMark-ES was used without repeat masking, while AUGUSTUS was used both with and without repeat masking.

The phase F1 of HelixerPost (Table 1 and Supplementary Table 5) was notably higher than GeneMark-ES and Augustus across both plants and vertebrates and still somewhat higher for invertebrates generally, although not for all species. All three tools showed similar performance for fungi, with HelixerPost having a slight margin of 0.007 overall. Similar patterns were found with subgenic and genic F1 metrics (Supplementary Figs. 14 and 15 and Supplementary Tables 6 and 7).

Moving from base-wise to feature- or protein-level evaluation, we found lower absolute precision, recall and F1 scores (Fig. 1, Table 2, Extended Data Fig. 1, Supplementary Fig. 16 and Supplementary Tables 8–11). Helixer scored highest for plants and vertebrates, but both HMM tools gained an edge in fungi. Helixer maintained a small advantage in invertebrates overall, although AUGUSTUS and GeneMark-ES were again strongest in some species. Overall, the gene precision and recall scores were lower for every tool than the exon scores, which is expected for the harder task. For most species, Helixer tends to have a higher recall score than precision.

When comparing Helixer and AUGUSTUS with softmasking enabled, Helixer outperforms AUGUSTUS in exon F1 score for 9 species, gene F1 score for 10 species and intron F1 score for 8 species out of a total of 16 tested (Supplementary Fig. 17 and Supplementary Tables 8–11).

The completeness of predicted proteomes[27] was quantified for the reference annotation and the three prediction tools (Extended Data Table 1, Supplementary Figs. 18–22 and Supplementary Table 12).

The reference annotations generally had the highest performance (39 of 45 species), which is unsurprising since they typically benefit from additional data sources and manual curation. The results of the prediction tools were broadly similar to the base-wise and feature assessments above, with HelixerPost leading strongly within plants and vertebrates, where it even approached the result of the reference. Invertebrates again varied by species, with Helixer leading by a smaller margin overall but GeneMark-ES performing best on several species.

https://doi.org/10.1038/s41592-025-02939-1

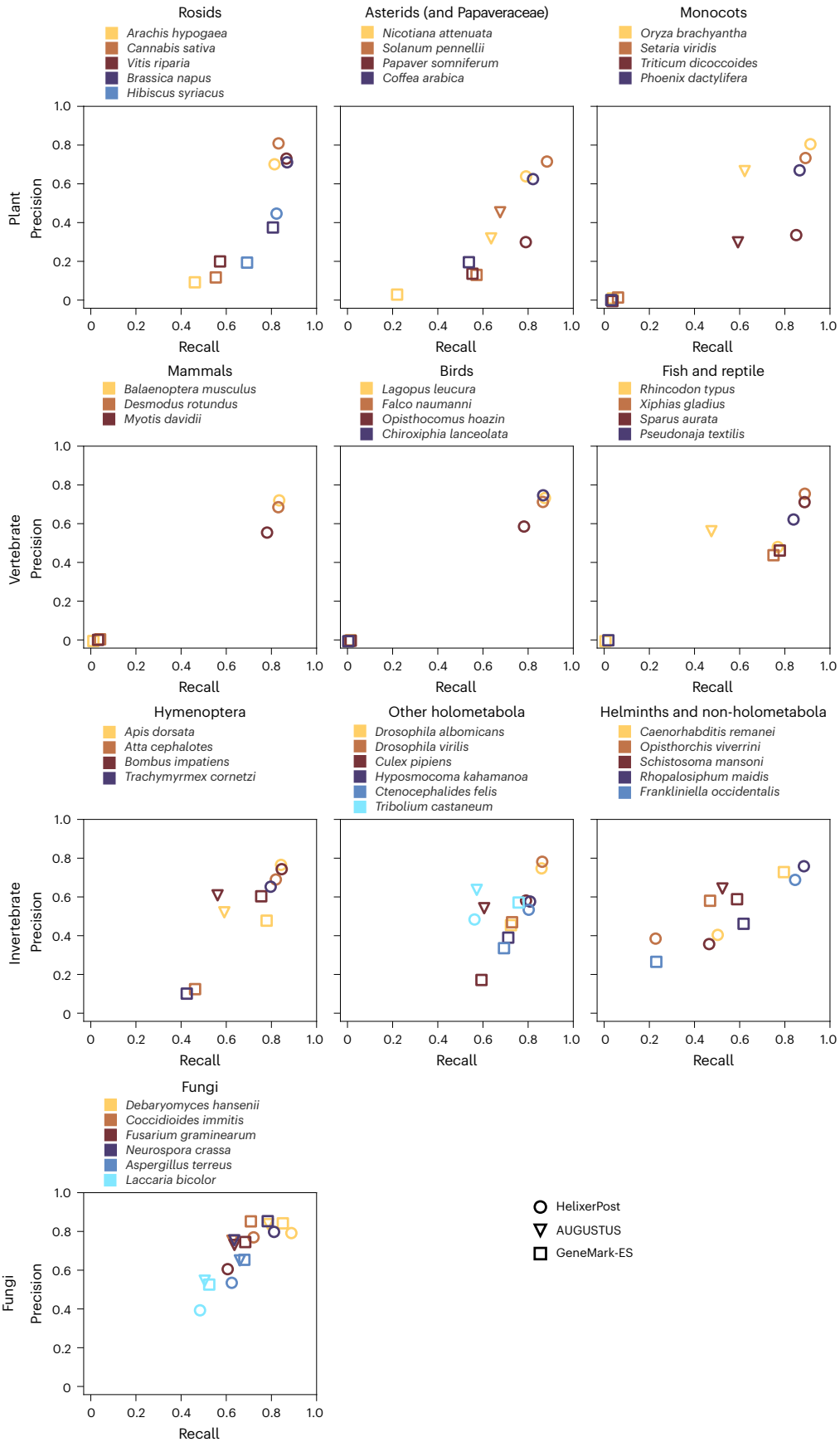


Fig. 1 | Exon precision and recall comparison. A comparison of precision and recall between HelixerPost (circles), AUGUSTUS (triangles) and GeneMark-ES (squares). The first row shows plants, the second vertebrates, the third invertebrates and the fourth fungi.

Table 2 | Feature F1 statistics for test species at the exon, intron, intron chain and transcript levels, summarized as the mean value per group and level

Group	Exon			Intron			Intron chain			transcript		
	Helixer	GeneMark-ES	AUGUSTUS[a]	Helixer	GeneMark-ES	AUGUSTUS[a]	Helixer	GeneMark-ES	AUGUSTUS[a]	Helixer	GeneMark-ES	AUGUSTUS[a]
Fungi	0.6678	**0.7240**	0.6743	0.6061	**0.7053**	0.6812	0.4431	**0.5210**	0.4615	0.5386	**0.5981**	0.5269
Plant	**0.7143**	0.1810	0.5044	**0.7232**	0.1861	0.5251	**0.4338**	0.0554	0.1792	**0.4618**	0.0934	0.2310
Vertebrate	**0.7405**	0.1078	0.5145	**0.6912**	0.1059	0.4966	**0.1740**	0.0130	0.0656	**0.1977**	0.0170	0.0823
Invertebrate	**0.6608**	0.4872	0.5786	**0.6416**	0.5142	0.6187	**0.2939**	0.1581	0.1475	**0.3066**	0.1783	0.1513

The highest F1 score per row is highlighted in bold. [a]For AUGUSTUS, pretrained models were only available for 4 of the 13 plant, 1 of the 11 vertebrate and 5 of the 15 invertebrate test species. All reported summary statistics reflect only those species.

Fungi was the most competitive clade, with Helixer leading by a small margin, and interestingly all tools outperformed the reference.

Interestingly, the three invertebrates where HelixerPost scored behind GeneMark-ES in phase F1 and Benchmarking Universal Single-Copy Orthologs (BUSCO) count had the overall lowest BUSCO count in the reference annotations. This hints at larger challenges related to annotating these genomes, be it exceptional divergence or simply a paucity of well-annotated genomes in close phylogenetic proximity to use either for training or for the homology mapping step of an annotation pipeline. However, it may also simply reflect that the invertebrate prediction models are less optimized.

Comparison with Tiberius

To compare Helixer to Tiberius, we trained a model focused on the mammalian clade. We used the three test species *Homo sapiens*, *Bos taurus* and *Delphinapterus leucas*. Even though we could improve Helixers prediction quality compared to our vertebrate model, Tiberius still outperforms Helixer (Fig. 2 and Supplementary Table 13). Especially gene recall and precision is consistently 20% higher. Regarding exon recall, Helixer and Tiberius are almost on par, with Tiberius still taking the lead. For exon precision Helixer shows 10–15% lower values. While we acknowledge that Tiberius outperforms Helixer in the Mammalia clade, we offer a more phylogenetically diverse range of models, especially for the often-underrepresented plant species.

Ablation analyses

To validate the importance of major changes made during development, we performed an ablation analysis by individually excluding each of the major changes, training a new model, then predicting and comparing prediction quality to the final model (Supplementary Fig. 23).

While none of the changes had a major effect on the overall genic F1 score (Supplementary Fig. 23a), this was not the target of most of the changes. Without transition weights (Supplementary Fig. 23e), the network displays high uncertainty immediately around start and stop codons, as well as acceptor and donor splice sites. All of the low, medium and final transition weights push the networks toward sharper transitions in an example gene, both right at the start codon (Supplementary Fig. 23e) and more widely helping to clear up uncertainty throughout the UTR (Supplementary Fig. 23d), with the final transition weights being most effective. To quantify this on a larger scale, we calculated the genic F1 specifically for the base pairs immediately before or after transitions. Higher transition weights resulted in the expected increase in transition genic F1, with the largest difference between low transition weights and none (Supplementary Fig. 23c). Moreover, the models with reduced transition weights show markedly more internal phase mistakes, that is, where both reference and predictions have a 0–2 codon phase, but the phases do not match (Supplementary Fig. 23b).

Interestingly, including the phase also improved the performance around transitions in the example (Supplementary Fig. 23d,e). In addition, phase predictions are useful signals for postprocessing into final gene models. The long short-term memory (LSTM) model and the hybrid model have similar performances. We acknowledge that this analysis could be made more conclusive by addressing the effect of random starting parameters by adding replicates or, where applicable (the number and arrangement of trainable parameters is only constant for the transition weight ablations), recording and reusing a random seed.

Helixer's annotations approach reference quality

The F1 metrics shown so far treat the references as the 'ground truth'; however, the provided references themselves are ultimately the output of a gene annotation pipeline, that is. data-supported predictions. Measuring performance against such references is particularly limited for understanding how good the absolute performance is as it approaches that of the provided reference. In the BUSCO analysis above we already observed that Helixer's performance was close to that of the reference. Therefore, we used two additional homology-based methods to evaluate in more detail the quality of Helixer and GeneMark-ES predictions versus the reference for the plant test species. AUGUSTUS predictions were omitted since they covered only four plant species.

First, we used the ability of a set of proteomes to form orthogroups as a reference-free indicator for quality, which is particularly relevant when considering annotation applicability for comparative genomic tasks. The more accurate annotation is expected to have more orthogroups with all 12 species represented and generally more orthogroups with higher numbers of species represented. The results are shown in Extended Data Fig. 2 and Supplementary Figs. 24–28.

Here, the reference performs best, with 0.38% of orthogroups containing all 12 species, and 36.5% containing two or more species, followed by Helixer at 0.26% and 31.1% for the same statistics, respectively, and GeneMark-ES at 0.019% and 21.2%, respectively. The reference has the most orthogroups with four or more species, followed by Helixer with two or three species and GeneMark-ES with only one species (Extended Data Fig. 2a). Notably, Helixer's performance by these metrics was more like that of the reference, that is, to extrinsic data-supported predictions, than to GeneMark-ES's predictions, which, comparable to Helixer's are made from the DNA sequence alone. Moreover, all of the reference annotations were respectable (minimum complete BUSCOs of 96.5%, and eight of the species above 99%).

Second, given that the orthogroup-based numbers could potentially be skewed by consistent cross-species mistakes, such as misannotation of transposons, we used the Mapman4 protein annotations (curated plant protein-coding gene families) to evaluate annotation quality. This further gives us a proxy for precision (the percentage of proteins that could be annotated) and recall (the percentage of gene families occupied by a protein). Measured this way, the reference had an average precision, recall and harmonic mean (of precision and recall) of 0.966, 0.931 and 0.948, followed by Helixer (0.878, 0.958 and 0.914, respectively) and GeneMark-ES (0.719, 0.742 and 0.724, respectively) (Extended Data Fig. 2b and Supplementary Table 14).

Thus, the ab initio annotations from Helixer show higher recall than the references for the plant species test set but are behind the references on precision, particularly for the specific species

https://doi.org/10.1038/s41592-025-02939-1

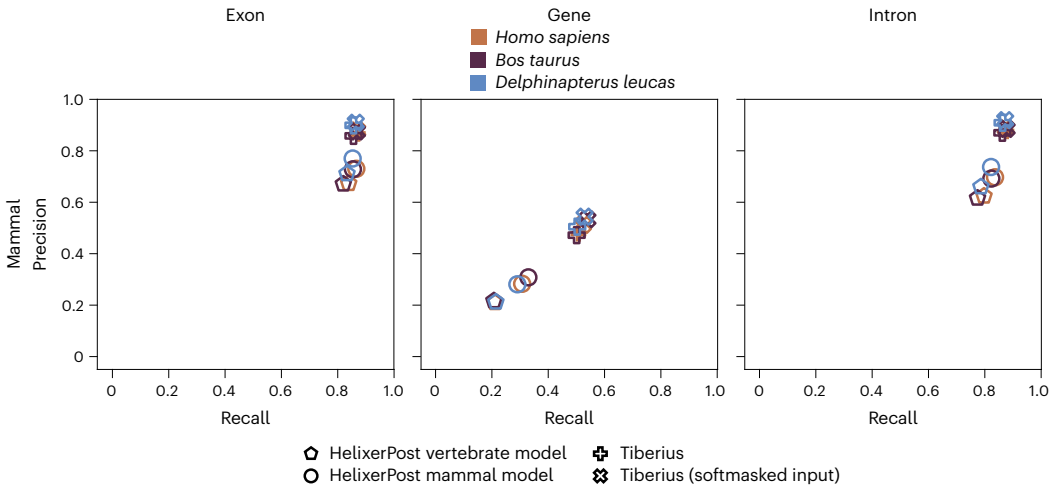


Fig. 2 | Precision and recall comparison. A comparison of precision and recall between the HelixerPost vertebrate model (pentagons), HelixerPost mammal model (circles), Tiberius (pluses) and Tiberius with softmasked input (crosses). The first column depicts exon, the second gene and the third intron metrics.

Papaver somniferum and *Triticum dicoccoides* (Extended Data Fig. 2b, Supplementary Figs. 24, 26 and 28, and Supplementary Table 14), which we note have the most fragmented assemblies, leading to an idea on how to further improve Helixer's predictions (Supplementary Information).

Filling gaps in *Arabidopsis thaliana*

We compared the annotations produced by Helixer for the model plant *A. thaliana*, to the established high-quality reference annotations TAIR10 and Araport11 (newest versions available in November 2022). When comparing Helixer to Araport11 (Extended Data Fig. 3a), the majority (~70%) of gene loci were found to have an exact match in Helixer. From the remaining ~30%, around 27% are highly similar but varied in their length. The unmatched sequences that remained from Helixer and Araport11 were considered as true annotations if the sequences were found to be conserved in at least 20 streptophyte species. For the Araport11 annotations, 93 of 1,401 were found to be true annotations, while 102 of 711 Helixer annotations were identified as true. Similar results were found when comparing Helixer to TAIR10 annotations (Supplementary Figs. 29 and 30).

From these 102 Helixer annotations not found in the Araport11 reference annotation, a notable example is the Phosphatidylinositol *N*-acetylglucosaminyltransferase γ subunit. This complex is known to be active in *A. thaliana*[28,29] and the γ subunit locus identified by Helixer shows expression (Extended Data Fig. 3b and Supplementary Fig. 31). However, this γ subunit is entirely missing from TAIR10, and has a chimeric annotation with the next gene in Araport11, resulting in a mere 4 bp of out-of-frame overlap between the resulting protein annotations. While the final postprocessed annotation from Helixer has truncations in at least UTR and introns relative to both the RNA-sequencing data and the raw predictions,the resulting protein was long enough for homology-based identification. This highlights the power of Helixer to complement and improve even the most polished reference annotations.

Benchmarking

As shown in Extended Data Fig. 4, Helixer is faster than the state-of-the-art tools AUGUSTUS and GeneMark-ES, and scales approximately linearly with genome size within phylogenetic groups. When comparing to GeneMark-ES and Augustus on the smallest and largest fungus test genomes and smallest test plant genome, Helixer required 6.2–20.1 fold lower wall time in single-threaded mode. Both AUGUSTUS and Helixer can be parallelized by splitting the genome into multiple fasta files. However, we acknowledge that the other tools make use of more complete multithreading than Helixer does. In single-threaded mode, Helixer.py can annotate the 263-Mbp *Oryza brachyantha* genome in 27 min and the 3.3-Gbp human genome in just under 8.5 h (Supplementary Methods). The exact speed will vary by system, particularly in regards to the speed of the disk storage and the GPU (Supplementary Fig. 32). A speed comparison to Tiberius was performed by the authors of Tiberius[25]. In their setup, Tiberius annotated the human genome in 1 h and 39 min and Helixer in 8 h and 54 min.

Discussion

Helixer represents a substantial advancement as a deep learning-based tool for producing full eukaryotic gene models and demonstrates what can be accomplished with a state-of-the-art ab initio gene caller. The primary gene models predicted by Helixer approach the quality of references produced via data-supported pipelines. This is achieved using a single tool, requiring no additional data beyond the DNA sequence, and only modest compute time.

Helixer performs a predictive task that no pre-existing tool is currently optimized for, namely going from an unmasked genome sequence to primary gene models for a phylogenetically diverse range of species, and all comparisons to existing tools were performed aligned to Helixer's task. We acknowledge that if comparative tools had been consistently selected and configured for optimal performance in a different scenario, such as gene prediction without project-specific RNA-sequencing data but incorporating repeat masking and protein homology, they would probably have achieved higher performances. However, such approaches also entail substantially greater computational costs and require more user expertise. While this does not constitute a conceptual difference in the type of input information, since Helixer's weights implicitly capture the extrinsic data, domain knowledge and compute resources used to construct the training references, it represents a major practical distinction owing to the greatly reduced requirements during inference.

In a comparison between Helixer and BRAKER2 on 54 crop plant species, Helixer matched BRAKER2 in overall coding sequence accuracy and outperforms it in detecting exon-containing regions and complete single-copy BUSCO genes, showing strength in identifying coding regions even under high GC content[30]. However, it was less precise at defining gene boundaries, particularly in fragmented genomes, and while BRAKER2 has higher gene-level recall, Helixer proved more consistent across plant clades, especially outperforming BRAKER2 on monocots.

While Helixer's overall prediction quality may not yet match predictions generated with extrinsic data-supported pipelines, the standardized output format enables easy incorporation in such pipelines. This integration has the potential to immediately improve the final quality of

https://doi.org/10.1038/s41592-025-02939-1

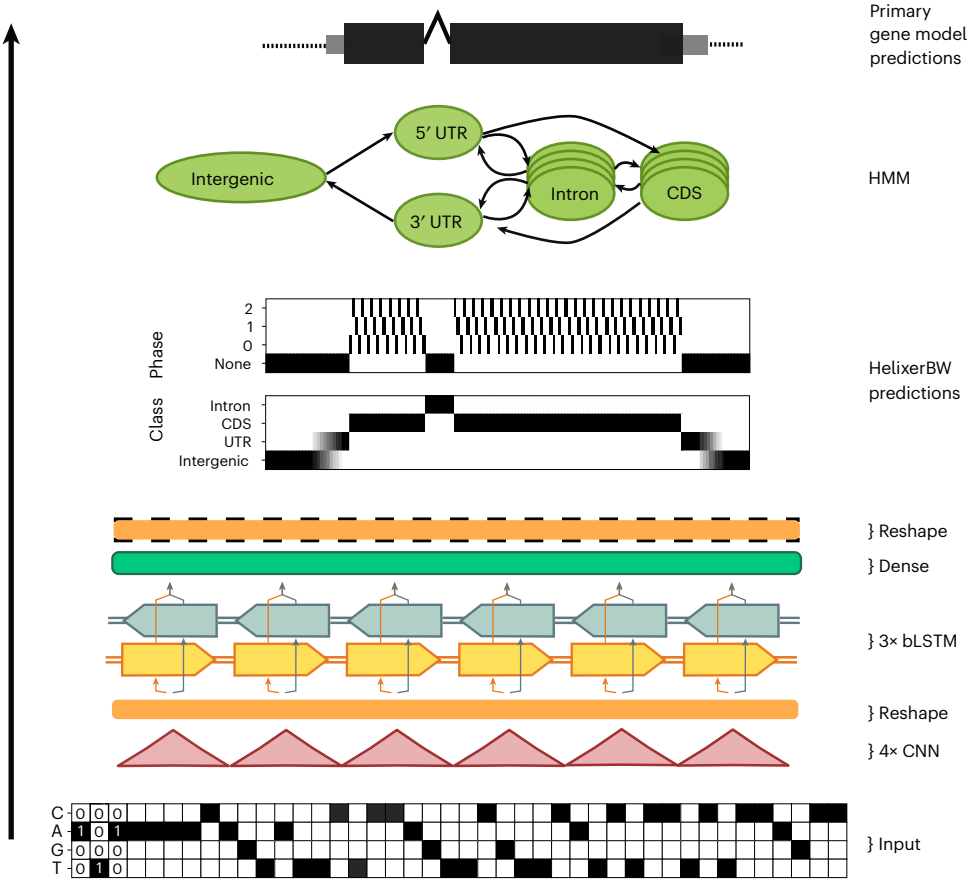


Fig. 3 | Visual summary of Helixer's predictive process. The depicted HMM state CDS is further broken down into 10 substates by phase and codon type (start, stop or regular), and the depicted intron state is further broken down into 60 substates by start or continuation of intron, splice motif and outer state. Not to scale. Hyperparameters are examples and can vary.

predictions. Notably, our research has demonstrated that Helixer's predictions can identify areas for improvement even in well-studied model species. Moreover, these annotations have been successfully applied for comparative genomics[31]. Helixer was also used for the ab initio genome annotation, either alone or in combination with other tools, of plants such as *Ribes nigrum* L.[32], *Camellia sinensis*[33], *Oxalis articulata*[34] and the arctic moss *Ptychostomum knowltonii*[35], as well as invertebrates such as *Coenonympha arcania*[36], two *Frankliniella* species[37,38] and the three economically harmful invasive species *Cydalima perspectalis*, *Leptoglossus occidentalis* and *Tecia solanivora*[39]. The combined attributes of sufficiently high quality and ease of generation make the structural annotations produced by Helixer valuable for many projects, especially for nonexperts.

The deep neural network Tiberius[25] is very similar to Helixer in approach and concept. Although it shows higher performance in mammals overall, including more balanced exon precision and recall values, it lacks the phylogenetic diversity of the Helixer models. While Tiberius' current focus is on mammals only (also predicting the longest gene isoform only), Helixer also provides fungi, plant, vertebrate and invertebrate models, which have already been adopted by the scientific community. Although Helixer represents a notable milestone as a fully applicable deep learning-based gene caller, it is not the pinnacle of what can be achieved by applying deep learning to structural gene annotation. We identify two primary areas for improvement: modeling and data handling.

In terms of modeling, there are several promising options that warrant exploration in future research. One such approach involves leveraging end-to-end prediction to harness the full potential of deep learning. Previous studies have successfully predicted transitions such as splice sites[22,40,41]. It is conceivable to encode a full gene structure as a series of transition tokens and positions, and predict structure with a many-to-many model architecture, similar to those used for large-scale language modeling[13,42]. This approach offers the advantage of being readily extensible for alternative splicing, but brings with it the challenge of an extremely sparse encoding. Sparse encoding coupled with erroneous or arbitrary labels (see below) may result in difficulties to get the model to converge during training. However, encoding the data in this manner would simplify the application of powerful techniques from language modeling, such as unsupervised pretraining[43,44]. Additionally, this would shorten the output length compared to base-wise encoding, thus reducing the memory required for low-bias transformer architectures that have demonstrated exceptional results across various domains[42,43,45].

Regarding data, there are both extensive challenges and opportunities for improvement. While there is generally no shortage of total data, obtaining high-quality data that represent a diverse and well-balanced selection of species across phylogenetic groups remains a major issue. We strongly believe that data quality is currently a limiting factor in network performance. This observation is supported by our findings (in ref. 26) and the present study, where we observed the lowest scoring predictions in the UTR, which is the class where the highest discrepancy between references and independent RNA-sequencing data was found, probably indicating errors in the reference annotations.

There are several options available to address data quality issues. The conceptually simplest—but poorly scaling—approach is to invest time, money and expertise to improve the quality of reference annotation for the training genomes. Indeed, ongoing efforts to improve annotations can be leveraged by retraining Helixer, with the automated selection of training species enabling Helixer to adapt to even large

changes in available input data. An intermediate option would be to use public extrinsic data to identify and differentially weight regions of higher or lower quality in the reference data. The network then learns more from the former than the latter, thereby reducing the noise the network sees while training.

Finally, there are additional modeling options that can be explored such as the unsupervised pretraining mentioned above[43,44] or pseudo-labeling[46]. These approaches could be combined and used in an iterative approach, where state-of-the-art ab initio predictions from Helixer are incorporated into data-supported pipelines. The resulting state-of-the-art references could, in turn, be used to train a deep learning ab initio caller. Although this approach would be effort intensive, it offers a highly reliable route to achieve major improvement. Another option is to train deep learning models to use RNA-sequencing, CAGE or other extrinsic information as input. This would probably boost performance when both training and inference are performed with the highest-quality extrinsic data. However, handling the potentially high variability of the extrinsic data at inference time will be a challenge.

The above ideas are neither exhaustive nor tested but they highlight the potential for performance gains. It is the hope of the authors that the power of deep learning techniques demonstrated here, and also in other genomics tasks[17,22,45,47], will promote research interest and improve available resources to the point where they reach a critical mass where modeling tasks previously considered untenable, such as achieving reference quality ab initio annotations, can be accomplished. This will free up time and resources that were previously consumed on using more cumbersome tools, extensive troubleshooting and manual validation and consequently accelerate research progress.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02939-1.

References

1. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Curr. Opin. Plant Biol.* **54**, 26–33 (2020).
2. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
3. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
4. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, 1–12 (2006).
5. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215–225 (2003).
6. Stanke, M. et al. Augustus: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439 (2006).
7. Holt, C. & Yandell, M. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
8. Haas, B. J. et al. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
9. Kirilenko, B. M. et al. Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107 (2023).
10. 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).
11. Oz-Levi, D. et al. Noncoding deletions reveal a gene that is critical for intestinal function. *Nature* **571**, 107–111 (2019).
12. Wang, M. F. et al. Uncovering transcriptional dark matter via gene annotation independent single-cell RNA sequencing analysis. *Nat. Comm.* **12**, 2158 (2021).
13. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
14. Silver, D. et al. Mastering the game of go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
15. Lample, G. & Charton, F. Deep learning for symbolic mathematics. Preprint at https://arxiv.org/abs/1912.01412 (2019).
16. Fawzi, A. et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53 (2022).
17. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
18. Amin, M. R., Yurovsky, A., Tian, Y. & Skiena, S. Deepannotator: genome annotation with deep learning. In *Proc. 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 254–259 (Association for Computing Machinery, 2018).
19. Hill, S. T. et al. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.* **46**, 8105–8113 (2018).
20. Singh, N., Nath, R. & Singh, D. B. Prediction of eukaryotic exons using bidirectional LSTM–RNN based deep learning model. *Int. J. Emerg. Trends Eng Res.* **9**, 275–278 (2021).
21. Louadi, Z., Oubounyt, M., Tayara, H. & Chong, K. T. Deep splicing code: classifying alternative splicing events using deep learning. *Genes* **10**, 587 (2019).
22. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).
23. Wei, C. et al. NeuroTIS: enhancing the prediction of translation initiation sites in mRNA sequences via a hybrid dependency network and deep learning framework. *Knowl.-Based Syst.* **212**, 106459 (2021).
24. Liu, Q. et al. DeepGenGrep: a general deep learning-based predictor for multiple genomic signals and regions. *Bioinformatics* **38**, 4053–4061 (2022).
25. 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).
26. Stiehler, F. et al. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* **36**, 5291–5298 (2020).
27. 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).
28. Lalanne, E. et al. SETH1 and SETH2, two components of the glycosylphosphatidylinositol anchor biosynthetic pathway, are required for pollen germination and tube growth in *Arabidopsis* [W]. *Plant Cell* **16**, 229–240 (2004).
29. Beihammer, G., Maresch, D., Altmann, F. & Strasser, R. Glycosylphosphatidylinositol-anchor synthesis in plants: a glycobiology perspective. *Front. Plant Sci.* **11**, 611188 (2020).
30. Abbas, Q., Wilhelm, M., Kuster, B., Poppenberger, B. & Frishman, D. Exploring crop genomes: assembly features, gene prediction accuracy, and implications for proteomics studies. *BMC Genomics* **25**, 619 (2024).
31. Triesch, S. et al. Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species. *Plant Biol. (Stuttg.)* **26**, 270–281 (2024).
32. Ziegler, F. M. R. et al. A full genome assembly reveals drought stress effects on gene expression and metabolite profiles in blackcurrant (*Ribes nigrum* L.). *Hortic. Res.* **12**, uhae313 (2025).

33. Tariq, A. et al. In-depth exploration of the genomic diversity in tea varieties based on a newly constructed pangenome of *Camellia sinensis*. *Plant J.* **119**, 2096–2115 (2024).
34. Yang, W. et al. A haplotype-resolved chromosomal-level genome assembly of *Oxalis articulata*. *Sci. Data* **12**, 856 (2025).
35. Ma, C. et al. Chromosome-level genome assembly and annotation of the arctic moss *Ptychostomum knowltonii*. *Genome Biol. Evol.* **17**, evae268 (2025).
36. Legeai, F. et al. Chromosome-level assembly and annotation of the pearly heath *Coenonympha arcania* butterfly genome. *Genome Biol. Evol.* **16**, evae055 (2024).
37. Song, W., Cao, L.-J., Chen, J.-C., Bao, W.-X. & Wei, S.-J. Chromosome-level genome assembly of the western flower thrips *Frankliniella occidentalis*. *Sci. Data* **11**, 582 (2024).
38. Song, W. et al. A chromosome-level genome for the flower thrips *Frankliniella intonsa*. *Sci. Data* **11**, 280 (2024).
39. Lombaert, E. et al. Draft genome and transcriptomic sequence data of three invasive insect species. *Peer Commun. J.* **5**, e65 (2025).
40. Zhang, Y., Liu, X., MacLeod, J. N. & Liu, J. Deepsplice: deep classification of novel splice junctions revealed by RNA-seq. In 2016 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 330–333 (IEEE, 2016).
41. Wang, R., Wang, Z., Wang, J. & Li, S. Splicefinder: ab initio prediction of splice sites using convolutional neural network. *BMC Bioinformatics* **20**, 652 (2019).
42. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5999–6009 (2017).
43. Devlin, J. et al. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (long and short papers), 4171–4186 (Association for Computational Linguistics, 2019).
44. Dai, A. M. & Le, Q. V. Semi-supervised sequence learning. *Adv. Neural Inf. Process. Syst.* **28**, 3079–3087 (2015).
45. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
46. Arazo, E. et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2020).
47. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).

Methods

Data

The foundation of every deep learning application is the data used to train it. Here, we expand the set of genomes to train and evaluate with from 237 in our previous work to 936, including many more plants and vertebrates as well as the new target groups, fungi and invertebrates (Supplementary Table 15). We acquired this data from RefSeq[48] and Phytozome13[49] and optimized the assignment of species to training or validation sets in each group (Supplementary Table 16 and Supplementary Methods). We reserved and set aside test set species, only using them for the final evaluation (Supplementary Methods). The exact accessions of all species used during development of Helixer are listed in the Supplementary Tables 17–21 and the different training–validation splits tested are listed in Supplementary Tables 22–25.

GeenuFF (https://github.com/weberlab-hhu/GeenuFF) is a tool to preprocess genome FASTA and annotation GFF3 files and store the cleaned annotations in a SQLite3 database. With GeenuFF, we identify partial gene models and other errors, before generating numerical encodings for the genomic sequence (C, A, T and G), the genic class (intergenic, UTR, coding DNA sequence (CDS) and intron) and coding phase (none, phase 0, phase 1 and phase 2 (indicating the number of base-pairs until the start of the next codon)). We use one-hot encodings, apart from the genomic sequence encoding, which supports ambiguity codes (Supplementary Table 26). Further, for training, we split both input and target data into subsequences of 21,384 bases and apply padding as necessary for shorter contigs or sequence ends. Finally, to prevent padded bases or regions GeenuFF identified as erroneous (Supplementary Methods) from affecting training or evaluation, we mask these base pairs from both the loss and metric calculations.

Architecture

Helixer (v0.3.0 used here) takes the genomic sequence as input and makes deep learning-based base-wise predictions (referred to as HelixerBW) for the genic class and coding phase. Next, our new HMM, HelixerPost, processes the HelixerBW predictions to generate primary gene models (Fig. 3). This combination allows us to benefit from the powerful pattern recognition of neural networks alongside the structured modeling of eukaryotic gene grammar enabled by HMMs. The probabilistic output of the deep learning stack is leveraged in this approach, capturing the inherent uncertainty associated with genetic structure and used to predict precise gene models.

The deep learning stack takes genomic sequence as input into a convolutional neural network followed by a bidirectional LSTN (bLSTM) hybrid model. This combines the powerful recognition of local patterns of convolutional neural networks with the ability of bLSTM to compress and conserve state over many thousands of inputs. This capability of bLSTMs also allowed us to use longer subsequences of 106,920 bp for plants and 213,840 bp for vertebrates and invertebrates compared to our standard length of 21,384 bp during inference, enabling the inclusion of most genes within a single subsequence (Supplementary Table 27).

The deep learning stack produces two simultaneous outputs: one for the genic class and the other for the coding phase (Fig. 3). We train the model to minimize by base-pair categorical cross-entropy loss but select both hyperparameters and the best model in each training run to maximize genic F1 ('Metrics' section) on the validation species. During inference, we predict (by default) on a sliding window of subsequences and use a mean ensemble to combine overlaps, as described in our earlier work[26].

HelixerPost

We implemented a postprocessing HMM tool, HelixerPost, that takes the genomic sequence and HelixerBW predictions of genic class and coding phase as input. Using this information, HelixerPost determines the most biologically plausible gene model consistent with these genic class and coding phase predictions for each locus, and outputs a final GFF3 file of primary gene models. HelixerPost first identifies candidate genic regions with consistent low intergenic probability with a sliding window. Within each candidate region, HelixerPost finds the gene model(s) that minimize a penalty for (1) discrepancy between the state of the HMM and Helixer's predictions and (2) for transitions at sequences conflicting with biological knowledge, such as a start codon that is not at ATG. For a more detailed description, see Supplementary Methods.

It should be noted that the role of the HMM in HelixerPost is relatively limited compared to established gene-finding HMMs. In these cases, the HMM or generalized HMM includes various trained state and state transition likelihoods. In contrast, HelixerPost aims to ensure that each state transition in HelixerBW is resolved at a biologically plausible point, and thus the HMM within HelixerPost needs to encode just well-established biological knowledge.

Metrics

Base wise. We use base-wise F1 metrics[26]. These F1 metrics are calculated as the weighted average of the F1 score for the genic classes UTR, CDS and intron (genic F1), the genic classes CDS and intron (subgenic F1) or the coding phase classes 0, 1 and 2 (phase F1). We calculate the base-wise metrics only on bases that are not padding and, additionally in this work, not masked by GeenuFF (for details, see Supplementary Methods).

For each training and validation genome, we calculate the F1 scores on a random sample of 800 subsequences, while for test genomes, we calculate F1 scores genome wide

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{3}$$

where TP is true positive, FP is false positive and FN is false negative.

Feature or full model. Complementary to the base-wise metrics, we employ both reference-centric and homology-based metrics to capture performance at the level of full features and gene models. With GffCompare (0.12.8)[50], we calculate feature and feature combination-level precise matches against the reference for the primary (longest protein) transcripts with all UTRs removed. We removed UTRs because GffCompare does score a true positive if every feature is exactly identical to the reference. All tools approached 0% exactly identical UTRs and therefore we excluded them.

The homology-based methods also support evaluation of the reference. We estimate proteome completeness with BUSCO (5.2.2)[27]. For the plant test genomes, we further assigned Mapman4 protein categories[51] with Mercator4 (5.0)[52] to quantify both completeness and annotatability by comparison to curated gene families. Additionally, we assessed the comparability between proteomes by clustering test plant proteomes into orthogroups with OrthoFinder (2.5.4)[53]. We focused on 12 plant test genomes (those that finished in 1 week for GeneMark-ES), and double-checked consistency for all 13 species with only the references and Helixer's predictions. Finally, we performed a detailed analysis for *A. thaliana* (for details, see Supplementary Methods).

Species selection

While developing Helixer, we found that the generalization performance of a trained model is highly dependent on both the quality and the quantity of the training genomes. Finding the optimal tradeoff

proved to be difficult and often a bit counterintuitive. To automatically obtain more generalizable, higher-quality and consistent models, we implemented a custom hyperparameter optimization for splitting training and validation genomes. This comprises selecting promising random sets of training genomes via twofold cross-validation and then remixing the best genomes from the folds for a final evaluation (Supplementary Methods and Supplementary Table 16).

Incorporation of biologically important features into the loss function

The loss function is the optimization criteria for the network and must thus be as close to the actual goal of a project as possible. Vanilla base-wise categorical cross-entropy treats all bases equally and does not proportionally reflect the importance of some mistakes (for example, those inducing frame shifts). We tuned the loss function to push the network toward more biologically plausible predictions. First, we implemented different weights for each class to be computed in the loss function. This counteracted the fact that the intergenic class is the most common, but also the one we care the least about. We then moved to increase these weights if they lie directly before or after one of four class transition sites (start codon, stop codon, donor splice site or acceptor splice site). This was done as the model had previously little incentive to output very sharp predictions around these transitions, which occur very infrequently in the data (see parameters class_weights and transition_weights in Supplementary Methods and Supplementary Tables 28–32). Finally, we added the phase output described above to capture phase directly in the loss and to support HMM postprocessing (see below). The overall loss is the weighted sum of the primary loss (weight of 0.8) and the categorical cross-entropy loss of the phase predictions (weight or 0.2).

Annotation quality comparison

We employed two established HMM tools to set baseline expectations for an ab initio gene calling tool. Where trained models were available, we used AUGUSTUS (3.3.2)[6] as a high-performance baseline (Supplementary Table 33), but for feasibility reasons, we did not retrain here. We did however perform an additional comparison running AUGUSTUS (3.5.0) with softmasking enabled and softmasked input genomes (Supplementary Fig. 17 and Supplementary Tables 8–11). Therefore, for the complete test set, we employed GeneMark-ES (4.71_lic)[2,3], which uses unsupervised training. Furthermore, we compared a Helixer to the similar DNN Tiberius version 1.1.4[25] with and without softmasking enabled. For this purpose, a mammal only model was trained using the same species as Tiberius (Supplementary Table 21). Helixer inference details can be found in Supplementary Tables 27 and 33 and Supplementary Methods. For comparing Helixer's prediction to alternative tools and the reference, we use only the splice variants producing the longest protein, unless otherwise specified.

Ablations

During ablations, we compared the best plant model to plant models trained without phase, with lower transition weights or with an LSTM instead of hybrid model. All training parameters differing from land_plant_v0.3_a_0080 as well as ablation-specific inference parameters are listed (Supplementary Tables 35 and 36, respectively). Ablations were evaluated on all plant test genomes except the very large *T. dicoccoides*.

Model releases

For this paper, we evaluated and documented all models released during development, including the training parameters (Supplementary Tables 28–32) and training species (Supplementary Tables 21–25). Released models are named as lineage_release_key_rank, where the lineage indicates the appropriate application lineage of the trained model; the release indicates the code version with which it is compatible; the key is a brief marker of 'a' for automatic species selection and 'm' for manual, which can be either entirely manual or automatic followed by manual tuning; and rank indicates relative model performance within the aforementioned categories (lower is better).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study were acquired from public databases. Accession numbers can be found in Supplementary Tables 17–20. The pretrained models are available via GitHub at https://github.com/usadellab/Helixer or via Zenodo at https://doi.org/10.5281/zenodo.10836346 (ref. 54).

Code availability

Helixer's source code is available via GitHub at https://github.com/usadellab/Helixer. The code is available via Zenodo at https://doi.org/10.5281/zenodo.17404832 (ref. 55). The source code for HelixerPost is available via GitHub at https://github.com/Usadellab/HelixerPost and also via Zenodo at https://doi.org/10.5281/zenodo.17414354 (ref. 56). Helixer's online web interface is available at https://www.plabipd.de/helixer_main.html.

References

48. O'Leary, N. A. et al. Reference sequence (refseq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, 733–745 (2016).
49. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, 1178–1186 (2012).
50. Pertea, G. & Pertea, M. Gff utilities: Gffread and gffcompare. *F1000Res.* **9**, 304 (2020).
51. Schwacke, R. et al. Mapman4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant* **12**, 879–892 (2019).
52. Bolger, M., Schwacke, R. & Usadel, B. in *Solanum tuberosum*. *Methods in Molecular Biology* Vol. 2354 (eds Dobnik, D. et al.) 195–212 (Humana, 2021); https://doi.org/10.1007/978-1-0716-1609-3_9
53. Emms, D. M. & Kelly, S. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
54. Denton, A. K. et al. Helixer–ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. Trained models. *Zenodo* https://doi.org/10.5281/zenodo.10836346 (2025).
55. Holst, F. et al. usadellab/Helixer: version 0.3.0 (v0.3.0). *Zenodo* https://doi.org/10.5281/zenodo.17404832 (2025).
56. Bolger, T., gglyptodon, Denon, A. usadellab/HelixerPost: v0.3.0 (v0.3.0). *Zenodo* https://doi.org/10.5281/zenodo.17414354 (2025).

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf and the IBG-4 Usadellab computing cluster at Forschungszentrum Jülich. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC-2048/1 project ID 390686111, by the DFG project 497667402 and by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (nos. 031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A and 031A532B). Helixer's integration into the Galaxy ToolShed (https://usegalaxy.eu/.) was kindly supported by the Galaxy Team. F.K. was supported by Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE HIDSS-004).

Author contributions

A.K.D., F.H., A.M.B. and A.P.M.W. conceived the study. A.K.D., A.P.M.W., B.U., O.E. and M.E.B. supervised the study. A.P.M.W. and B.U. acquired funding sources. A.K.D., F.H., A.M.B., J.M., F.K., C.G. and M.E.B. wrote software. A.K.D., F.H., J.M., C.G., A.M.B., N.K., R.S., N.S. and S.T. evaluated the results. A.K.D., F.H., A.M.B., F.K. and M.E.B. wrote the paper. All authors read, edited and approved this work.

Funding

Open access funding provided by Forschungszentrum Jülich GmbH.

Competing interests

A.K.D. is currently an employee at Valence Labs, part of Recursion Pharmaceuticals Inc. and has received real ownership interest in the company. The other authors declare no competing interests.

Additional information

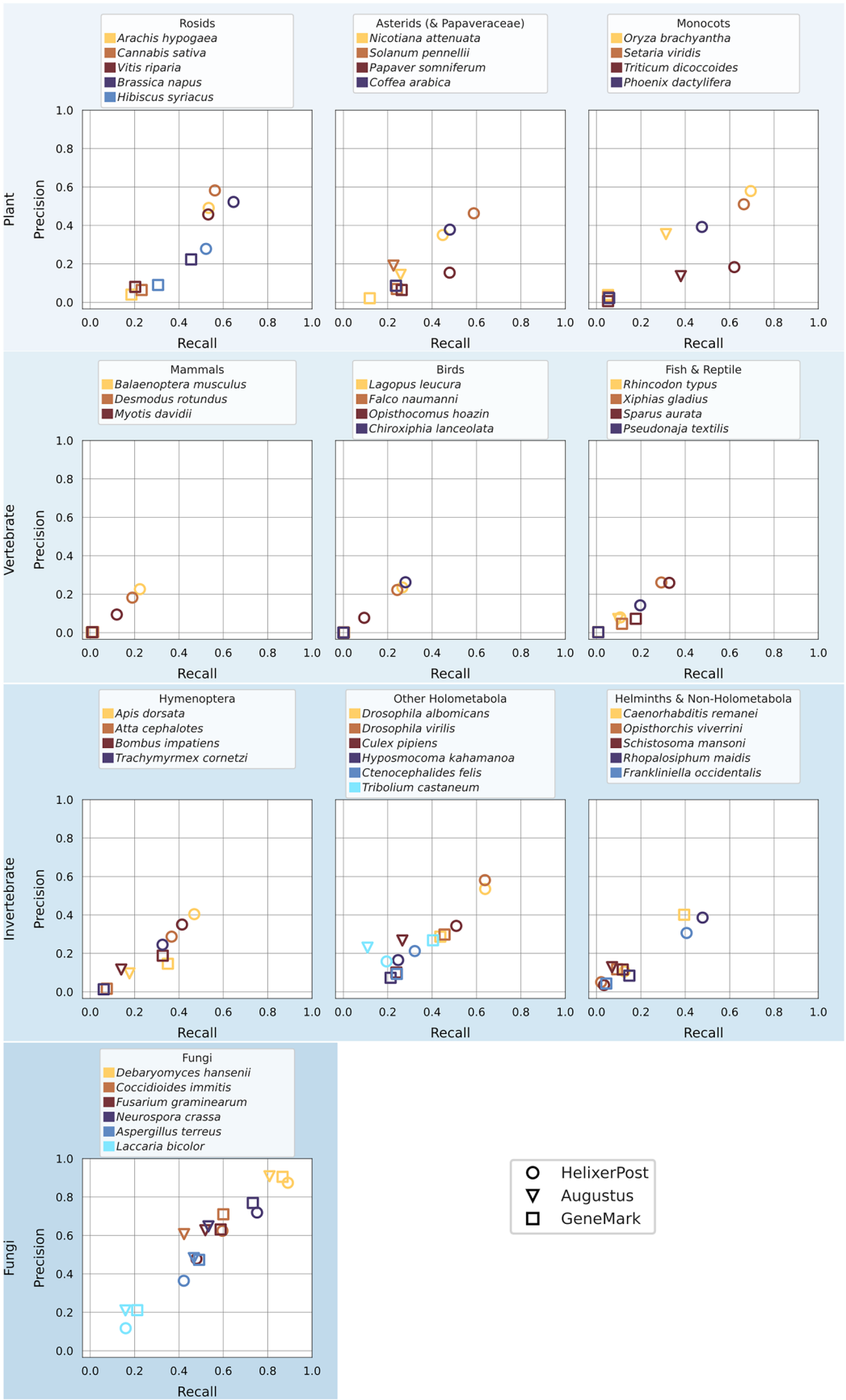
Extended data is available for this paper at https://doi.org/10.1038/s41592-025-02939-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02939-1.

Correspondence and requests for materials should be addressed to Marie E. Bolger.

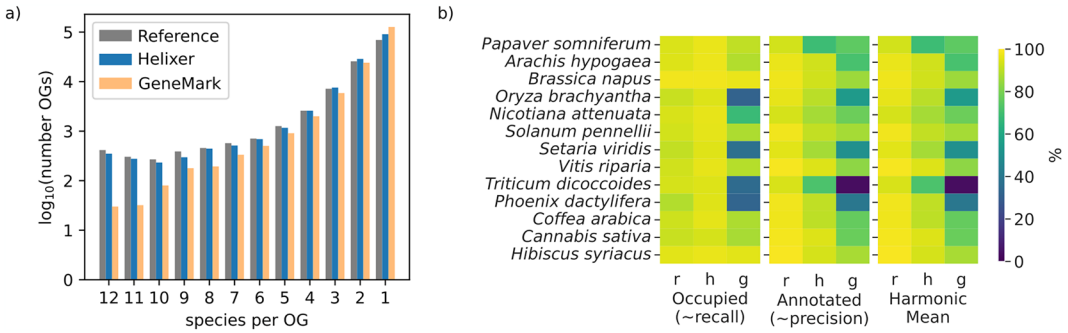
Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-025-02939-1



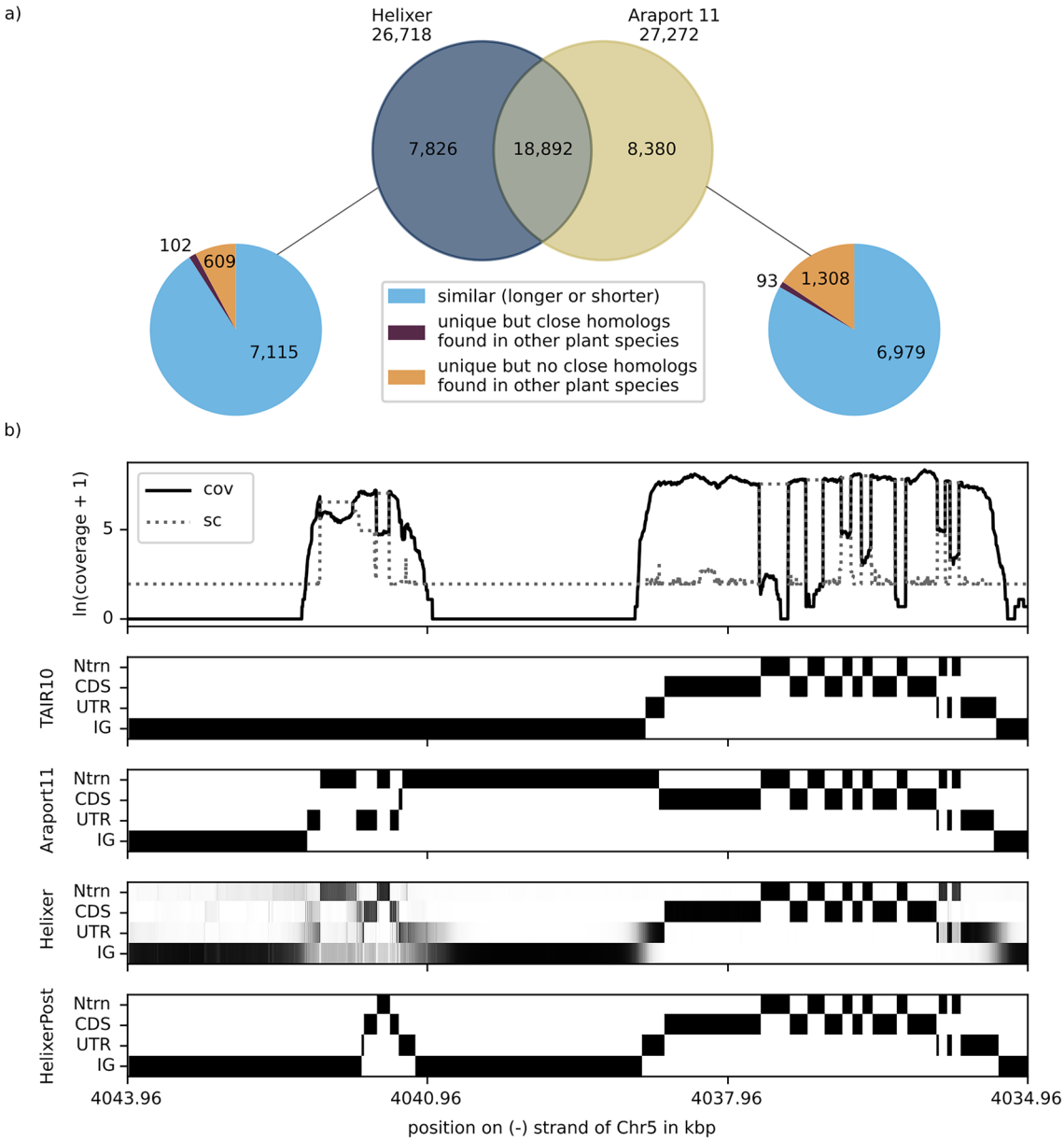
Extended Data Fig. 1 | Gene precision and recall. Gene precision and recall comparison between HelixerPost (circle), Augustus (triangle) and GeneMark (square). The first row shows plants, the second vertebrates, the third invertebrates and the fourth fungi.

https://doi.org/10.1038/s41592-025-02939-1



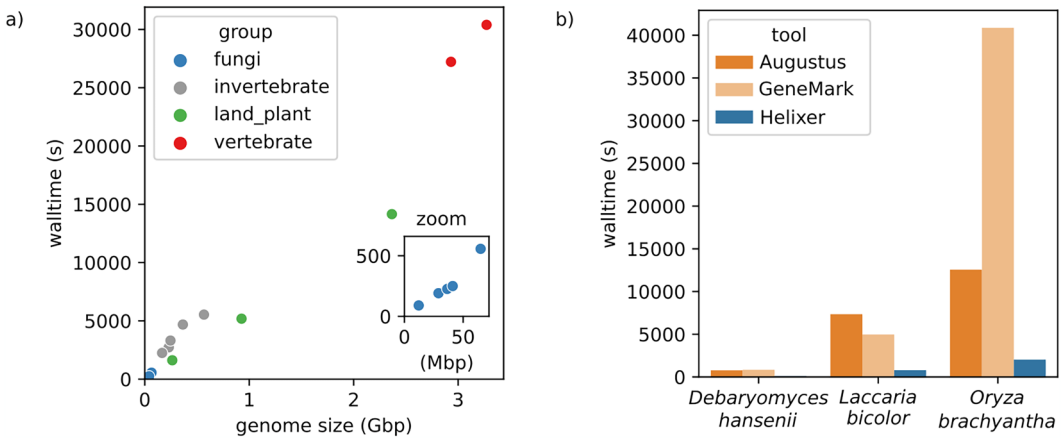
Extended Data Fig. 2 | Homology based evaluation of annotations. **a**) Orthogroup occupancy (number of species represented out of 12) for orthogroups based on the reference, Helixer's, and GeneMark's annotations (each clustered individually). *Triticum dicoccoides* was excluded from this analysis. **b**) Comparison of proteomes to curated Mapman4 protein annotations. r=reference, h=Helixer (post processed), g=GeneMark. The harmonic mean is that of the %Annotated and the %Occupied, and here is conceptually a proxy for the F1.

https://doi.org/10.1038/s41592-025-02939-1



Extended Data Fig. 3 | Comparison of the *Arabidopsis thaliana* proteome predicted by Helixer to the existing Araport11 annotation. a) Venn diagram and pie charts of the overlaps and differences between predicted proteomes. The pie charts categorize the unique sections from the Venn diagram further into three categories: 'similar (longer or shorter)', 'unique with homologs in other plant species' and 'unique without homologs in other plant species. The second category encompasses likely legitimate annotations not found in the other proteome. The third category encompasses likely artefacts not found in the other proteome. **b**) visualization of RNAseq expression (cov=coverage; sc=spliced coverage) and available annotations at the genomic loci of the Phosphatidylinositol N-acetylglucosaminyltransferase γ-subunit (Chr5, - strand, from 4043960 to 4034960 bp, via HelixerPost).

https://doi.org/10.1038/s41592-025-02939-1



Extended Data Fig. 4 | Benchmarking. a) Helixer walltime on test genomes and *Homo sapiens* on 'workstation A'; **b**) comparison in single-threaded walltime between ab initio tools.

https://doi.org/10.1038/s41592-025-02939-1

Extended Data Table 1 | BUSCO completeness statistics

Extended Data Table 1: BUSCO completeness statistics for the test species, summarized as the mean value per group assessing the reference and results from the annotation tools.

group	reference	helixer (post)	genemark	augustus*
Fungi	0.9659	**0.9903**	0.9848	0.9661
Plant	**0.9904**	0.9772	0.5554	0.7924
Vertebrate	**0.9424**	0.9046	0.2306	0.5786
Invertebrate	**0.9222**	0.8422	0.7600	0.7130

*For AUGUSTUS, pre-trained models were only available for 4 of the 13 plant, 1 of the 11 vertebrate and 5 of the 15 invertebrate test species. All reported summary statistics reflect only those species. The highest BUSCO score per row is highlighted in bold.

BUSCO completeness statistics for the test species, summarized as the mean value per group assessing the reference and results from the annotation tools.

nature portfolio

Corresponding author(s): Marie Bolger

Last updated by author(s): Oct 22, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No physical experimental data was generated for this study. The genome assemblies and annotations were downloaded from public sources. Restricted data (e.g. genomes released before their corresponding publication) were removed.
Data analysis	All code developed for this project is available, either from our public Github repositories or from others in the community. The most important code versions are listed below for clarity. Main: Helixer v0.3.0 HelixerPost v0.3.0 Helixer's code is available on GitHub https://github.com/usadellab/Helixer and the specific release of Version 0.3.0 is also available on Zenodo https://doi.org/10.5281/zenodo.17404831. All software requirements for Helixer can also be accessed via these links. HelixerPost is available on GitHub https://github.com/usadellab/HelixerPost and the specific release of Version 0.3.0 is also available on Zenodo https://doi.org/10.5281/zenodo.17414354. Software versions: GeenuFF v0.3.0 (GitHub tag) HelixerPost v0.3.0 GffCompare v0.12.8 BUSCO v5.2.2

Mercator4 v5.0
Orthofinder v2.5.4
AUGUSTUS v3.3.2 and v3.5.0
GeneMark-ES v4.71_lic
Tiberius v1.1.4

Software for plotting (only major python packages; Python v3.12.3 was used):
h5py v3.13.0
jupyter v1.1.1
matplotlib v3.10.3
matplotlib-venn v1.1.2
notebook v7.4.2
numpy v2.2.6
pandas v2.2.3
seaborn v0.13.2

For RNA-seq processing the following tools were used:
FastQC v0.11.5
Trimmomatic v0.36; extra parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:3:30:10:1:true MAXINFO:36:0.7 MINLEN:36,
Hisat v22.1.0; extra parameters: –max-seeds 8 –dta –pen-canintronlen G,-8,1.5 –pen-noncanintronlen G,-8,1.5
Samtools v1.6
PicardTools v52.0; extra parameters: STRAND=SECOND READ TRANSCRIPTION STRAND
MultiQC v1.8

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The pre-trained models are available via GitHub https://github.com/usadellab/Helixer or on Zenodo https://zenodo.org/records/10836346.

Fungus training, validation, and test genomes were acquired from RefSeq on March 4th, 2022; exact accessions can be found in Supplementary Table 17
Plant training and validation genomes were acquired from Phytozome13 on June 7th 2021, test plant genomes were acquired from RefSeq on July 14th 2022; exact accessions can be found in Supplementary Table 18
Vertebrate training, validation, and test genomes were acquired from RefSeq on May 6th, 2022; exact accessions can be found in Supplementary Table 19
Invertebrate training, validation, and test genomes were acquired from RefSeq on May 6th, 2022; exact accessions can be found in Supplementary Table 20
Mammal training, validation and test genomes were acquired from RefSeq on March 13th, 2025; exact accessions can be found in Supplementary Table 21
RNA-seq coverage of Arabidopsis thaliana shown in Extended Data Figure 3 was collected from these SRA accessions:
SRS3032258
ERS1647356
SRS1605924
ERS3438334
SRS2705778
ERS3438336
ERS2617740

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sampling was used to a very limited extent in this study. During the training process, for computational reasons, 800 blocks were randomly selected from validation genomes to provide estimates of accuracy metrics to select candidate models. These estimates are included in supplementary figures and tables to illustrate the consistency of prediction metrics across the relevant clades, but since they are based on the genomes used during training, should not considered as definitive metrics in any case. Independent test metrics were calculated with the full independent test genomes (and annotations).
Data exclusions	Genome assemblies and associated annotations were quality assessed using automated approaches such as BUSCO. Only the primary (longest) transcript was used from each gene model. During training, gene models which were incomplete (e.g lacking UTRs) or invalid (overlapping another model, invalid splicing, incorrect coding phase, missing start or stop codons etc) were masked.
Replication	Multiple training runs were performed using different random splits of the genomes into training and validation sets. The resulting models were assessed using the separate test data set and showed similar performance.
Randomization	Samples were not assigned to experimental groups in the traditional sense in this study. However, multiple training runs were performed using different random splits of the genomes into training and validation sets.
Blinding	Blinding was not relevant to this study since assessments of the various existing tools and newly trained models used the same deterministic, objective and fully automated process on pre-determined datasets. Random partitioning of genomes into training, validation and test datasets ensured data leakage during the training process thus preventing bias.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA