GAP-MS: Automated validation of gene predictions using integrated mass spectrometry evidence

Qussai Abbas[1], Mathias Wilhelm[2,3], Bernhard Kuster[3,4], & Dimitri Frishman[1, §]

[1] Chair of Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany.
[2] Computational Mass Spectrometry, TUM School of Life Sciences, Technical University of Munich, Freising, Germany.
[3] Munich Data Science Institute, Technical University of Munich, Garching, Germany.
[4] Chair of Proteomics and Bioanalytics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany.

[$] Corresponding author: dimitri.frischmann@tum.de

Keywords: Proteomics, Genomics, Mass Spectrometry, Gene prediction, Genome annotation, Proteogenomics, Crops, Plants

1. Abstract

Accurate genome annotation is fundamental to modern biology, yet distinguishing authentic protein-coding sequences from prediction artifacts remains challenging, particularly in complex plant genomes where automated methods are error-prone and manual curation is rarely feasible due to prohibitive time and costs. Here, we present GAP-MS (**G**ene model **A**ssessment using **P**eptides from **M**ass **S**pectrometry), an automated proteogenomic pipeline that leverages mass spectrometry evidence to systematically validate the protein-level accuracy of predicted gene models. Applied across 9 major crop species, GAP-MS consistently improved prediction precision for four widely used gene prediction tools. In addition to filtering erroneous models, the pipeline identified hundreds of previously missing gene models from current standard reference annotations. These peptide-supported loci were further verified by transcriptional evidence, well-supported functional annotations, and high coding-potential scores. Together, these results demonstrate that direct proteomic evidence provides a robust framework for resolving annotation ambiguities, defining high-confidence reference proteomes, and uncovering overlooked protein-coding genes, while facilitating the identification of sequences that may require further investigation. GAP-MS is freely available as a web interface at https://webclu.bio.wzw.tum.de/gapms/.

2. Introduction

High-throughput sequencing technologies and advances in assembly algorithms have led to a rapid increase in the number and quality of genome assemblies across the tree of life. Large-scale initiatives such as the Earth BioGenome Project [1] and affiliated efforts, including Darwin Tree of Life [2], the 10,000 Plant Genome Sequencing Project (10KP) [3], and the European Reference Genome Atlas [4], aim to generate reference genomes for all known eukaryotic species.

In contrast, genome annotation has progressed more slowly, and accurate prediction of protein-coding genes, their structures and isoforms remains a major bottleneck in turning raw assemblies into biologically interpretable resources [5, 6]. Computational gene prediction has therefore become a central focus, with a growing ecosystem of tools that combine *ab initio* models, comparative genomics and transcriptomic evidence to infer gene models. Early gene prediction tools for eukaryotes, including GeneMark [7], Genscan [8], and later Augustus [9], relied primarily on intrinsic sequence signals such as coding potential, splice-site motifs, and other statistical properties of genomic DNA to identify candidate coding regions. As annotation strategies evolved, pipelines began incorporating external evidence sources, including protein homology (e.g., orthologous genes) and experimental transcript data. Initially derived from expressed sequence tags (ESTs) and later from RNA-seq experiments, these data substantially improved gene model accuracy. Modern annotation frameworks such as Maker [10], Braker1/2/3 [11-13], the Gnomon system [14], and the Ensembl gene annotation pipeline [15] combine these heterogeneous evidence types to generate more reliable annotations. More recently, deep learning–based approaches, including Helixer [16] and Annevo [17], have emerged, applying neural network models to directly learn complex patterns in genomic sequences and further improve gene structure prediction.

Despite substantial progress, current pipelines still suffer from false positives and false negatives, fragmented gene models and incompletely resolved alternative splice variants, particularly in complex eukaryotic genomes [18]. As a result, many predicted genes lack direct experimental support at the protein level, and a significant fraction of the proteome remains unannotated or misannotated [19]. These inaccuracies accumulate rapidly in public databases, propagating errors across subsequent analyses and tools that rely on these resources, which underscores the need for orthogonal validation strategies to filter and refine predicted gene models systematically.

At the same time, mass spectrometry–based proteomics is undergoing a parallel expansion in both scale and coverage, driven by large coordinated data-generation initiatives. One such effort is the crop proteome atlas within the “Proteomes that Feed the World” project [20], which aims to chart the proteomes of the 100 most important crops for human nutrition, alongside the continued accumulation of public datasets coordinated through the ProteomeXchange consortium [21].

Proteomics data offers a powerful and complementary line of evidence to address genome annotation errors. By directly detecting peptides derived from translated proteins, proteomics can provide physical confirmation of protein expression, validate predicted coding sequences and refine gene structures, including exon–intron boundaries and alternative start/stop sites [22]. Proteogenomic approaches have already demonstrated their value in prokaryotes, where relatively compact genomes facilitate comprehensive searches and genome-wide validation of gene models [23]. Early studies also showed that this approach can be applied to large, complex eukaryotic genomes, enabling direct protein identification from raw genomic sequence data [24]. Nevertheless, in eukaryotes, its applications have largely been restricted to individual species or focused studies [25, 26], and systematic large-scale integration of proteomics into routine genome annotation remains uncommon.

This imbalance highlights a critical methodological gap. Genome sequencing and mass spectrometry are now both high throughput and widely accessible, but the field still lacks robust, scalable and automated pipelines to exploit proteomics data for genome annotation at scale. Developing such pipelines, capable of updating gene models in an automated but quality-controlled manner, will be essential to fully realize the potential of proteomics for improving genome annotation in eukaryotic organisms.

Here, we introduce GAP-MS, an automated proteogenomic pipeline designed to systematically evaluate the protein-level accuracy of predicted gene models by integrating mass spectrometry–derived peptide evidence. We apply GAP-MS across 9 major crop species, enabling large-scale, comparative assessment of genome annotations and providing direct experimental support for translated gene products.

3. Methods

3.1 Data retrieval

3.1.1 Plant genomic data

Genome assemblies and annotation data for the top 100 crops essential for human nutrition were compiled from GenBank [27], following the procedures outlined in our previous work [28]. Briefly, the list of major food crops was obtained from the Food and Agriculture Organization (FAO) of the United Nations [29], ranked by annual production, and for each crop the most recent genome assembly and associated metadata were retrieved.

From this initial set of 100 crops, species were selected only when a RefSeq assembly and its matching GFF reference annotation were available, and the assembly status was either complete or at chromosome level. To ensure better accuracy of the collected genomes and their annotations, we only retained genomes sequenced using long-read sequencing (PacBio or Oxford Nanopore), with assemblies released after 2020. Gap content was constrained to ≤1%, calculated as ((total sequence length – total ungapped length) / total sequence length × 100%). Only species annotated using the NCBI Eukaryotic Genome Annotation Pipeline version ≥9.0 were retained. Assembly and gene-model completeness were evaluated using BUSCO v5.2.2 [30] with the embryophyta_odb10 lineage dataset, representing land plants, and both the assembly and its annotation were required to achieve ≥98% complete (single-copy + duplicated) BUSCOs. We further assessed intergenic and repetitive-region contiguity using the LTR Assembly Index (LAI) [31], applying a threshold ≥10. Genomes failing any criterion were excluded.

This filtering strategy retained 18 crops with high-quality assemblies and annotations (Supplementary Table 1), from which we selected 9 crops with available mass spectrometry data at the time of the study. These 9 species include both monocots (e.g., maize and barley) and dicots (e.g., tomato and apple), span substantial phylogenetic diversity across the plant kingdom (Supplementary Fig. 1), and served as our benchmark datasets for evaluating gene prediction tools.

3.1.2 Gene prediction

Gene prediction was conducted using four modern computational tools. Braker2 [12], Galba [32], and Helixer [16] were selected based on their superior accuracy in our previous benchmarking study [28]. Additionally, we included Annevo [17], a recently developed *ab initio* gene prediction tool, due to its reported high prediction accuracy. Braker2 and Galba utilize diverse protein databases to align proteins to the genome, thereby generating essential hints for Augustus [9] to predict gene models and infer multiple isoforms per gene. In contrast, Helixer employs a deep neural network with a bidirectional long short-term memory (LSTM) architecture to predict genes, typically yielding a single isoform per gene. Annevo applies a genomic language model that captures long-range sequence dependencies and joint evolutionary relationships across diverse genomes. All tools were executed using their default parameter settings.

3.1.3 Proteomic data acquisition

Raw label-free MS/MS datasets for 9 crops were obtained from the Crop Proteome Atlas [20] and analyzed using the FragPipe v23.1 pipeline [33]. Protein FASTA databases corresponding to (i) RefSeq reference annotations and (ii) predicted proteomes from Braker2, Galba, Helixer, and Annevo were prepared as independent search spaces using Philosopher v5.1.2 [34], including

reversed-sequence decoys and common LC–MS/MS contaminants from the cRAP database [35]. Spectra were searched with MSFragger v4.3 [36] using trypsin as protease and allowing up to two missed cleavages. Peptide–spectrum matches (PSM) were rescored with MSBooster v1.3.17 [37], leveraging Koina [38] to obtain Prosit deep-learning predictions for fragment ion intensity (Prosit_2020_intensity_HCD) and retention time (Prosit_2019_irt) [39]. PSM validation was performed with Percolator v3.7.1 at a 1% False Discovery Rate (FDR) threshold [40], followed by protein inference and FDR filtering using Philosopher with a 1% protein-level FDR.

3.1.4 Evaluation of peptide identification

To evaluate how well predicted proteomes derived by different gene annotation tools support peptide identification, each predicted proteome from Braker2, Galba, Helixer and Annevo was tested independently as a search space for peptide identification. The resulting peptide sets were then compared to those obtained using the reference RefSeq annotation. For each species and annotation, we quantified (i) the total number of identified peptides, (ii) the number and fraction of peptides shared with RefSeq, and (iii) the number of peptides uniquely identified using the predicted proteins (Fig. 1).

3.2 GAP-MS pipeline

We developed GAP-MS, a pipeline that integrates mass-spectrometry–based proteomic evidence with in-silico predictions to evaluate and filter gene models. The pipeline is designed to work across different gene prediction tools and takes as input predicted gene models in GTF/GFF format together with predicted proteins sequences in FASTA format, as well as the list of identified peptides from mass spectrometry experiments in TXT format. Using these inputs, GAP-MS assigns confidence levels to the provided protein sequences based on the integrated peptide translation evidence, thereby distinguishing authentic coding sequences from probable annotation artifacts. GAP-MS further enables comparative analyses against reference annotations and the identification of novel protein-coding candidates, as well as the correctness of misannotated gene models.

3.2.1 Peptide mapping and feature engineering

GAP-MS begins by mapping the identified peptides from the mass spectrometry experiments to the protein sequences of the predicted gene models using Proteomapper v1.5 [41]. For each protein, a feature set is compiled that integrates sequence attributes with proteomic evidence. Sequence features include protein length, the number of splice sites, and the number of annotated isoforms. Proteomic features include the number of mapped peptides, sequence coverage (the percentage of residues overlapped by identified peptides, typically 10 to 40% and occasionally above 70% for highly expressed proteins), peptide uniqueness at the protein and gene levels, and the position of each peptide within the protein. Peptides are classified as gene-specific (mapping to any isoform of the same gene), protein-specific (mapping uniquely to a single protein isoform), internal (entirely within one exon), splice (spanning an exon–exon junction), N-terminal (mapping to the start of the protein), and C-terminal (mapping to the region immediately upstream of the stop codon). The resulting feature matrix is used to train the classification model described below.

3.2.2 Classification of proteins into confidence groups

Based on the extracted features, gene models were partitioned into three sets: high-confidence, low-confidence, and unlabeled. High-confidence proteins were defined by robust proteomic support , meeting at least one of the following stringent criteria: the presence of two or more unique

protein-specific peptides, sequence coverage of at least 80%, detection of both N-terminal and C-terminal peptides, or complete support for all annotated splice sites by corresponding splice peptides. In contrast, low-confidence proteins were characterized by limited peptide evidence, specifically exhibiting fewer than two mapped peptides and an absence of gene-specific peptides. Proteins with intermediate evidence were designated as unlabeled.

To resolve the status of gene models with intermediate unlabeled evidence, a multivariate XGBoost classifier [42] was employed , capturing non-linear relationships between protein length, sequence coverage, peptide uniqueness, and peptide location that rule-based filtering would systematically misclassify. Accordingly, the high-confidence and low-confidence sets served as positive and negative training examples, respectively. The resulting model was applied to the unlabeled set to assign a final status of verified or dismissed. The classifier was optimized using randomized hyperparameter search with stratified 5-fold cross-validation and ROC-AUC scoring. Details on model training, performance evaluation, and score cutoffs are provided in Supplementary File 1.

3.2.3 Model interpretation and validation

To elucidate the contributions of individual features to protein classification, SHAP (SHapley Additive exPlanations) analysis was employed [43]. This analysis revealed the most impactful features influencing model predictions, thereby validating the biological relevance of our chosen metrics.

3.4 Assessment of GAP-MS impact on gene prediction accuracy

The accuracy of Braker2, Galba, Helixer, and Annevo (both with and without applying GAP-MS) was evaluated at the exon, intron chain, and transcript levels by comparing their boundaries to those in the RefSeq reference annotations of 9 crops using Gffcompare v0.12.10 [44]. Specifically, exon level accuracy required exact matching of start and end coordinates for individual coding exons. Intron chain level accuracy evaluated the precise sequence of internal splice junctions within a transcript, disregarding the outer transcription start and end sites. Transcript level accuracy imposed the strictest criteria, requiring an identical match for all exon boundaries throughout the entire gene model [44]. Accuracy metrics included recall (the ability to detect true positive genes), precision (the ability to eliminate false positives), and F1 score (the harmonic mean of recall and precision).

Due to the non-normal distribution of the data, as confirmed by the Shapiro-Wilk test, non-parametric statistical methods were employed. Specifically, the Kruskal-Wallis test was used to compare the distribution of each metric across the tools. For pairwise comparisons, the Mann-Whitney U test was applied to identify significant differences between individual tools.

3.5 Identification of novel gene loci

Gene models predicted by gene-prediction tools and supported by peptide evidence in GAP-MS, but absent from the RefSeq reference annotations, were subjected to additional validation and characterization. To identify high-confidence novel coding regions, we cross-referenced the GAP-MS supported gene models against the RefSeq models using gffcompare. Transcripts assigned class code 'u' (intergenic) or 'x' (antisense) were designated as novel gene loci, while those with class code 'j' were identified as novel isoforms of annotated genes.

3.5.1 Transcriptional validation with RNA-Seq

RNA-seq datasets for the 9 crops were retrieved from the Sequence Read Archive [45] using VARUS [46]. Reads were aligned to the corresponding genome assemblies with HISAT2 [47]. Per-base coverage and exon–intron support were quantified with BEDTools [48]. For each novel peptide-supported gene, we summarized the coverage across the coding sequence.

3.5.2 Assessment of protein coding potential

To further validate the coding potential of the novel gene models, the protein sequences were evaluated using PSAURON [49]. This tool utilizes a machine learning model trained on over 1,000 plant and animal genomes to assign a probability score reflecting the likelihood that a query sequence represents a genuine protein-coding region. Predicted proteins were classified based on the tool's binary output, where sequences labeled as true sequences were considered high-quality candidates distinguishable from spurious open reading frames.

3.5.3 Functional annotation of protein sequences

Functional annotation of the novel peptide-supported proteins was performed with InterProScan v5 [50]. We recorded InterPro signatures, predicted domains, and associated Gene Ontology (GO) terms, and we flagged sequences lacking any domain assignment.

3.5.4 Visualization of gene models

Gene models were visualized and inspected in the IGV genome browser [51]. Reference genome assemblies were loaded together with GFF annotation tracks for standard RefSeq gene models, raw gene prediction outputs, and the final GAP-MS–filtered gene sets. Transcriptional support was evaluated using HISAT2 BAM alignments to visualize read coverage depth and splice junctions. Proteomic evidence produced by GAP-MS (BED format) was mapped to genomic coordinates and displayed as custom tracks.

4. Results

4.1 Benchmarking of gene prediction tools using proteomics data

In this study, we evaluated how well the predicted proteins derived from gene-prediction tools perform as search spaces for peptide identification in proteomics analyses. Parallel database searches were performed for nine crop species, matching raw MS/MS spectra against peptide sequences derived from these predicted protein sets. This assessment quantifies the concordance between peptides identified using predicted sequences from four gene prediction tools (Braker2, Galba, Annevo, and Helixer) versus those derived from standard RefSeq reference annotations. A high degree of overlap confirms that the prediction accurately captures the expressed proteome. Based on this comparison, the set of identified peptides were categorized into three distinct groups: **shared peptides** (consensus support), **reference-specific peptides** (coding regions missed by the tool), and **prediction-specific peptides** (potential novel coding sequences absent from the RefSeq annotation) (Fig. 1a).

The predicted proteomes generally showed high concordance with RefSeq reference datasets (Fig. 1b); shared peptides typically accounted for >90% of the total identified spectra. Helixer consistently demonstrated superior predictive accuracy, achieving the highest peptide recovery rates and the lowest number of missed reference peptides across all species. While “prediction-specific” peptides (potential novel coding sequences) typically ranged between 3,000 and 6,000 peptides per species, “reference-specific” peptides (missed by prediction) were generally more numerous, ranging from 5,000 to 12,000.

However, we observed some species-specific variation. Braker2 predictions for *Zea mays*, and *Hordeum vulgare* diverged significantly from the reference, exhibiting a marked loss of reference-specific peptides (ranging from 43,000 to 63,000), indicating reduced Braker2 predictive quality for these species. Conversely, *Malus domestica* exhibited a dramatic expansion in prediction-specific peptides (53,000–58,000) across all four tools. This suggests systemic incompleteness in the current *Malus domestica* RefSeq annotation rather than an artifact of the prediction tools. These findings highlight the utility of mass spectrometry data as an orthogonal metric for both benchmarking prediction tools and identifying gaps in existing reference annotations.

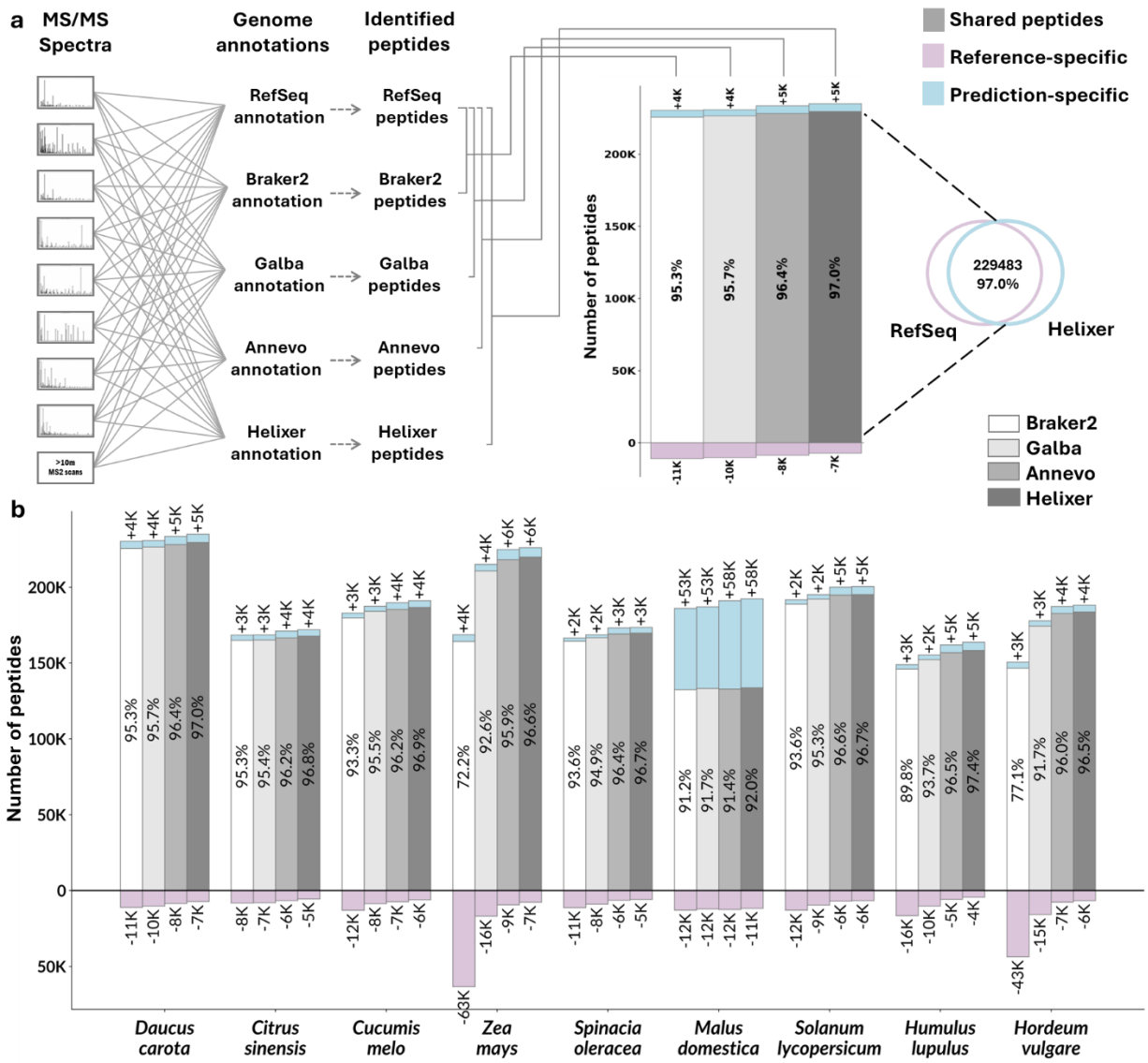


Figure 1. Comparative peptide identification rates of gene prediction tools across 9 crop species. (a) Schematic of the evaluation workflow: Raw MS/MS spectra are searched in parallel against the RefSeq annotation and predicted gene models (Braker2, Galba, Annevo, Helixer). The resulting identified peptides are compared to categorize results: Shared peptides (gray; consensus), Reference-specific peptides (pink; missed by prediction), and Prediction-specific peptides (blue; putative novel genes). The Venn diagram illustrates this overlap for a representative comparison. **(b)** Stacked bars represent the total peptide yield for each prediction tool across 9 species. Gray bars indicate agreement with the reference; pink bars (negative y-axis) indicate missed reference peptides; and blue bars indicate peptides unique to the prediction. Numeric labels denote peptide counts (in thousands, K), and percentages within the bars denote the recovery rate, calculated as $(N_{Shared} / N_{Reference}) \times 100\%$, where $N_{Reference}$ is the sum of shared and reference-specific peptides. The four bars for each species correspond to Braker2, Galba, Annevo, and Helixer, respectively (left to right).

4.2 Development of a proteomic-guided classification model (GAP-MS)

To distinguish authentic protein-coding regions from annotation artifacts, we developed GAP-MS, a pipeline that integrates genomic features with direct proteomic evidence (Fig. 2a). For each predicted gene model, GAP-MS extracts a feature set combining sequence attributes with quantitative mass spectrometry metrics. High-confidence training sets were established for each species by selecting gene models with robust peptide evidence, while low-confidence sets comprised gene models with limited peptide evidence. An XGBoost classifier trained independently for each species using the species-specific high- and low-confidence sets as positive and negative training data was then used to predict binary status (verified or dismissed) for ambiguous proteins with intermediate proteomic evidence (see *Methods*). The final list of peptide-supported proteins comprises both high-confidence and model-verified predictions. This classification is illustrated using *Daucus carota* (Fig. 2b), which displays the varying proportions of high-confidence, verified, dismissed, and low-confidence proteins across the four prediction tools. The analysis of *Daucus carota* highlights major differences in reliability among the four prediction tools. Although all tools identify a similar core of 15,000 to 16,000 high-confidence proteins, their total outputs vary significantly. Braker2 and Galba generate the largest initial protein sets but also produce the highest volume of non-validated models, with 31,743 and 24,198 proteins relegated to the low-confidence category. In contrast, Annevo and Helixer achieve higher precision with fewer dismissed models. Annevo is the most conservative, yielding only 6,931 low-confidence proteins.

Expanding this analysis to the full dataset of 9 crops, proteomic support rates varied substantially across the prediction tools. The support percentage was calculated by dividing the total number of peptide-supported proteins by the total number of predicted proteins. This evaluation revealed that deep learning-based tools maintained relatively balanced distributions, with 63% and 51% of the total predictions classified as supported for Annevo and Helixer, respectively. Conversely, Braker2 and Galba exhibited heavily skewed distributions, with only 14% and 20% of their gene models receiving proteomic support (Supplementary Fig. 2).

This difference is primarily driven by the total number of predicted proteins. Braker2 and Galba produced substantially larger predicted proteins sets (averaging 150k and 100k per species, respectively) compared to the more conservative outputs of Annevo and Helixer (averaging 29k and 36k per species, respectively). However, despite the lower support percentages, the absolute number of supported proteins yielded by Braker2 and Galba (average ~20,000) was comparable to that of Annevo and Helixer (average ~19,000–20,000). These patterns indicate that while Helixer and Annevo prioritize precision, homology-based approaches like Braker2 and Galba generate a substantial excess of unsupported potential artifacts to capture a similar core of authentic proteins (Fig. 2c).

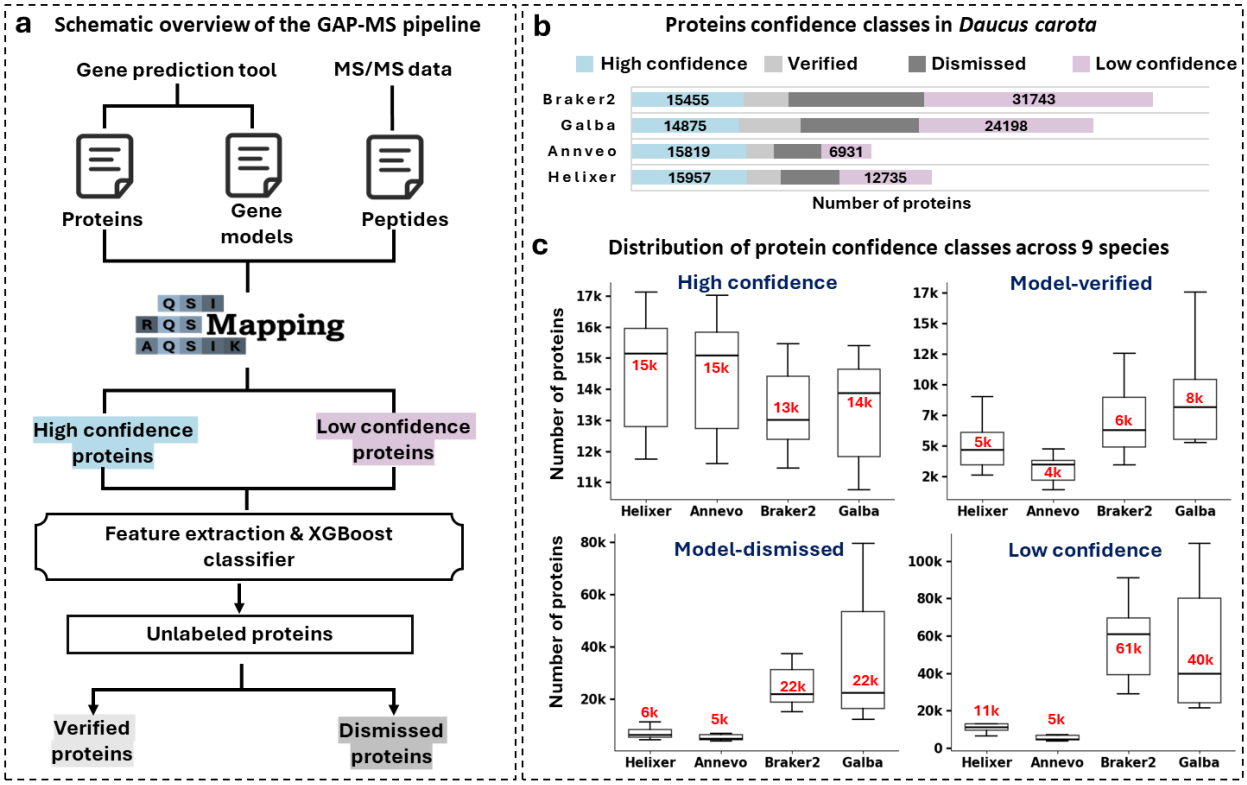


Figure 2. Design, training, and validation of the GAP-MS pipeline. (a) Schematic overview of the GAP-MS pipeline. Gene predictions are integrated with MS/MS peptide data to define training sets (high and low confidence) and train a species-specific XGBoost classifier for determining the support status of unlabeled proteins. (b) Classification profile for *Daucus carota*. Stacked bars display the absolute number of proteins assigned to high-confidence, verified, dismissed, and low-confidence categories for each prediction tool. (c) Distribution of protein confidence classes across 9 species. Box plots visualize the variation in protein counts for each classification category across the evaluated prediction tools. Center lines represent medians.

Analysis of the peptide evidence underlying these classifications revealed that supported models were substantiated by high-quality features. On average, 15,094 proteins per crop were validated by protein-specific peptides across the four prediction tools, while 11,085 proteins received support from splice-spanning peptides. Complementary evidence was provided by terminal-specific peptides, with an average of 595 proteins supported by N-terminal peptides and 1899 proteins exhibiting C-terminal peptide support per crop (Fig. 3a). Notably, Braker2 showed a clear deficiency in C-terminal peptide validation, with an average of only 447 supported proteins per crop, compared with approximately 2,000 in Galba, Annevo, and Helixer. This deficit at the 3′ end of gene models suggests that Braker2 has difficulty accurately defining stop codons or maintaining reading-frame consistency across terminal introns. Together, these results demonstrate that GAP-MS serves not only as a filtering approach but also as a practical diagnostic tool for identifying systematic structural errors in genome annotation pipelines.

To understand which specific attributes most strongly influenced the distinction between verified and dismissed models, SHAP analysis was conducted, revealing that sequence coverage and the presence of protein-specific peptides are consistently shown as the primary drivers of the classification model across all species (one example in *Daucus carota* is shown in Fig. 3b).

Moreover, to evaluate the impact of proteomic depth on gene model validation by GAP-MS, a saturation analysis was performed on *Malus domestica,* selected for its comprehensive availability of 23 distinct tissue datasets. By cumulatively merging these tissue datasets (from 1 to 23), the count of peptide-supported gene models was tracked. The resulted curve shows a sharp initial increase between 1 and 5 tissues, where the inclusion of distinct datasets resulted in substantial gains in gene support. This confirms that multi-tissue atlases are essential for detecting hidden genes. Importantly, this trend decelerated after the 15th tissue, with the curve approaching a plateau that indicates proteomic saturation (Fig. 3c).

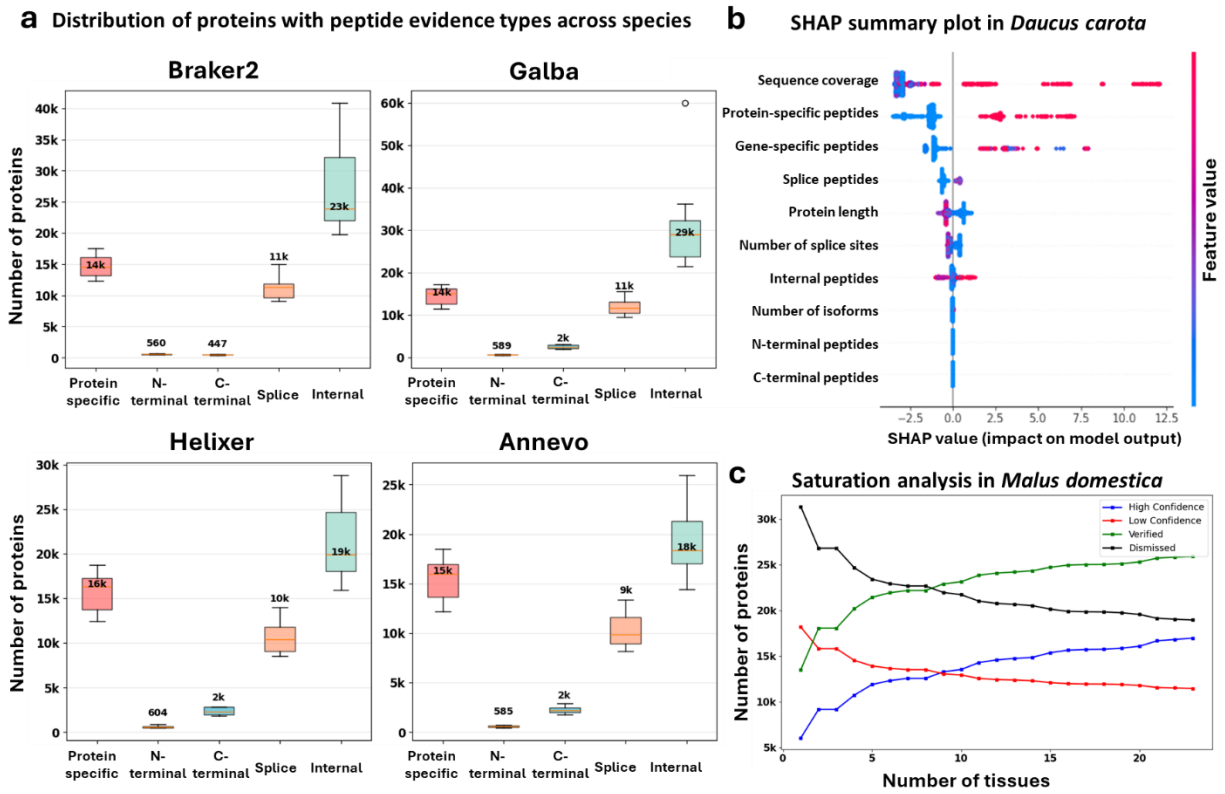


Figure 3. Characterization of peptide evidence and model interpretation. (a) Distribution of peptide evidence types across 9 crops. Box plots quantify the number of proteins supported by protein-specific, N-/C-terminal, splice-spanning, and internal peptides for each prediction tool. Center lines represent medians; box limits indicate the interquartile range. **(b)** SHAP feature importance analysis for Helixer predictions in *Daucus carota*. Features are ranked by their contribution to the model output. Each dot represents a protein, with color indicating the feature value (red = high, blue = low) and x-axis position denoting the impact on the classification probability (positive values favor "verified" status). **(c)** Proteomic saturation analysis in *Malus domestica*. The cumulative count of classified proteins is plotted against the number of merged tissue datasets. The trajectory shows a rapid initial accumulation of supported models followed by a plateau after approximately 15 tissues.

4.3 Improving precision through proteomic filtering

The performance of the four selected gene prediction tools (Braker2, Galba, Helixer, and Annevo) was evaluated across 9 crop species at the exon, intron chain, and transcript levels, both prior to and following filtration with GAP-MS. Across all tested tools, the application of GAP-MS consistently enhanced precision, albeit an associated expected tradeoff in recall (Fig. 4).

The most substantial improvements were observed in Braker2 and Galba. These tools, which exhibited lower baseline precision (average initial transcript F1 scores of 18–20%), showed a marked increase in precision following filtration - approximately 33% for Braker2 and 27–29% for Galba across all feature levels. This gain in precision came at the cost of sensitivity, with a reduction in recall of approximately 8–10%. Consequently, the overall F1 score improved at all levels, specifically transcript-level F1 scores increased by 4.9% for Braker2 and 3.4% for Galba, indicating that the removal of false positives outweighed the loss of true positives for these tools (Fig. 4). Deep-learning-based tools demonstrated higher baseline accuracies, with initial average transcript F1 scores of ~25% for Annevo and ~28% for Helixer. While GAP-MS filtration successfully increased precision for these tools by 4-8% across all measured levels, the associated penalty in recall was more pronounced, leading to a net decrease in F1 scores.

Statistical analysis confirmed that the shifts in the accuracy scores post-GAP-MS were significant across all tools (Mann-Whitney U test, $p < 0.05$). Collectively, these data demonstrate that GAP-MS serves as a rigorous specificity filter, most effectively enhancing the performance of tools prone to high false-positive rates.

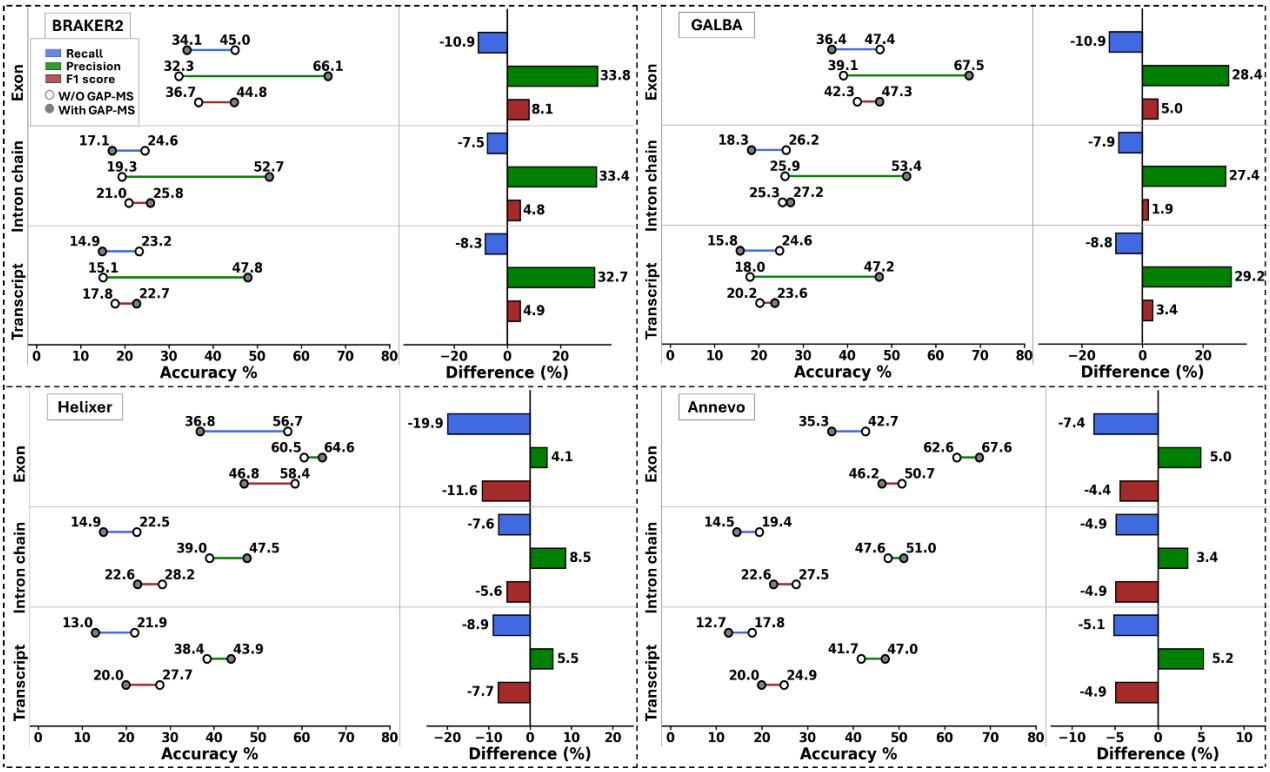


Figure 4. Comparative assessment of gene model accuracy before and after GAP-MS integration. Average accuracy metrics for Braker2, Galba, Helixer, and Annevo at three genomic levels (Exon, Intron chain, and Transcript) across 9 crop species. Left panels: Dumbbell plots illustrate the performance shift from the raw prediction (open white circles) to the GAP-MS refined set (filled grey circles). Colors denote specific metrics: Recall (blue), Precision (green), and F1 Score (red). Right panels: Bar charts quantify the performance differential, calculated as (GAP-MS metric – raw metric). Positive values (right-extending) indicate an increase in the metric, while negative values (left-extending) denote a reduction attributable to filtering.

4.4 Validation of novel peptide-supported genes

Reference genome annotations frequently miss genuine genes due to factors such as the absence of detectable homology, divergent codon usage, lineage-specific gene expansions, or the masking of low-complexity regions [52, 53]. Additionally, micro-exons of 30 bp or less are particularly difficult to detect as they often fall below statistical thresholds or are discarded during splice graph construction [54]. Consequently, novel gene models predicted by gene prediction tools and supported by peptide evidence via GAP-MS likely represent authentic coding sequences missing from current reference annotations.

The recovery of these sequences is facilitated by GAP-MS through a comparative analysis between supported predictions and user-defined reference annotations. This process identifies specific discrepancies, including entirely novel gene models as well as refinements to existing models where proteomic evidence supports alternative translation start codons, stop codons, or splice sites. To enable visualization and manual investigation via genome browsers, genomic coordinates of mapped peptides are generated in BED format. To illustrate such cases and demonstrate the correction of annotation errors by GAP-MS, Helixer was selected, as it consistently demonstrated the highest prediction accuracy and peptide recovery rates among the evaluated tools.

Across the 9 crop species, a total of 9,152 novel peptide-supported proteins were identified within the Helixer predictions. The abundance of these newly discovered loci exhibited substantial inter-species variation, ranging from a minimum of 371 novel proteins in *Solanum lycopersicum* to a maximum of 2,891 in *Cucumis melo*. To investigate potential factors influencing this distribution, the relationship between the number of RefSeq protein-coding genes and the number of novel Helixer peptide-supported proteins was analyzed. A Pearson correlation coefficient of -0.36 (p = 0.34) was observed, indicating that the detection of novel genes was not significantly correlated with the size of the annotated proteome (Supplementary Fig. 3).

To independently validate novel Helixer-predicted genes absent from RefSeq annotations, transcript evidence was assessed using RNA-seq read coverage breadth. As indicated in Figure 5a, a substantial proportion of these proteins demonstrated extensive transcriptional support (coverage ≥ 50%), with high-coverage models outnumbering low-coverage ones in most species. This RNA-seq coverage analysis provides independent validation that these novel gene models correspond to actively transcribed and translated loci. Analysis of transcript coverage density distributions revealed that while low-confidence proteins peaked near zero, novel peptide-supported proteins displayed a profile closely resembling the high-confidence standard, characterized by a skew toward full coverage (100%). This trend is exemplified in *Solanum lycopersicum* (Supplementary Fig. 4), where 287 novel proteins exhibited high transcript coverage compared to 84 with low coverage. Observed discrepancies, such as high transcript coverage of low-confidence models, likely reflect MS detection limits, or instances where expression occurred only under specific conditions not sampled. In contrast, some novel proteins exhibited low RNA-seq coverage, which might explain why these specific loci were overlooked and excluded from prior RefSeq annotation.

Furthermore, the coding potential of these novel sequences was independently verified using Psauron, a machine learning-based tool designed to distinguish genuine protein-coding regions from spurious open reading frames. Across the 9 crops, an average of 95% of the novel peptide-supported proteins were classified as true high-quality coding sequences (Supplementary Fig. 5), reinforcing that GAP-MS recovers authentic protein-coding sequences.

Finally to assess the biological validity of the novel gene models recovered by GAP-MS, a systematic functional annotation of all novel Helixer-predicted, peptide-supported proteins across the 9 crop species was performed. Remarkably, over 95% of these novel sequences exhibited strong evidence of functional activity, characterized by significant homology to known protein domains or families (Fig. 5b). This high rate of functional conservation confirms that GAP-MS recovers authentic, evolutionarily conserved proteins.

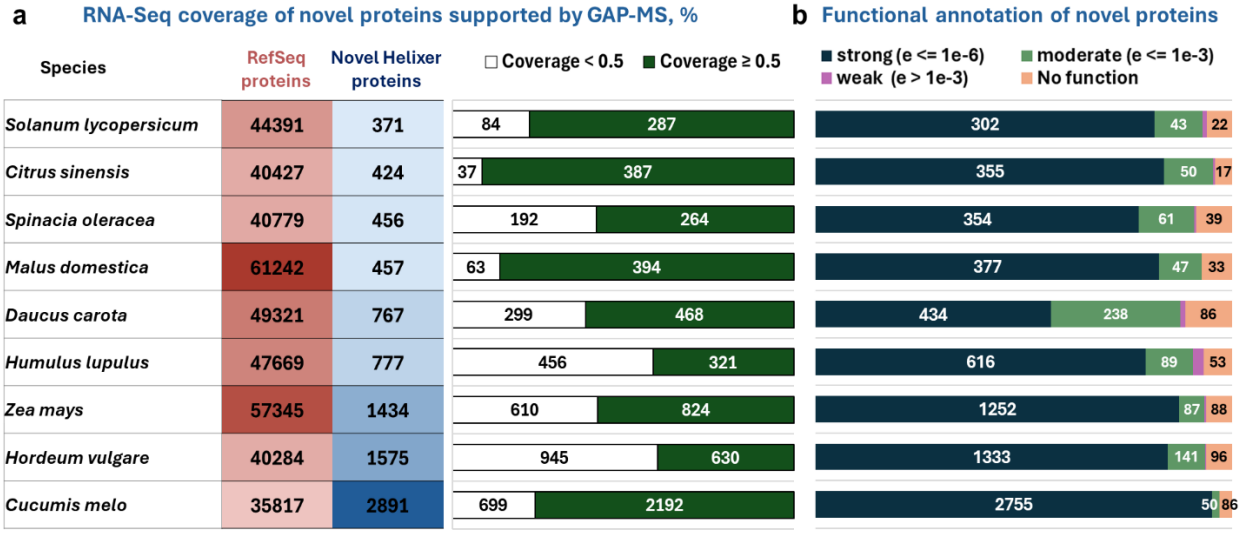


Figure 5. Transcriptional validation and functional characterization of novel peptide-supported gene models. (a) RNA-seq support analysis across 9 crop species. The heatmap compares the abundance of reference proteins (red) against novel peptide-supported Helixer predictions (blue). Horizontal bars quantify RNA-seq coding sequence coverage of the novel sets: high support (≥50% breadth, dark green) versus low support (<50% breadth, white). **(b)** Functional annotation profile. Stacked bars display the classification of novel proteins based on sequence homology to known domains, stratified by statistical significance (E-value): strong ($\leq 10^{-6}$), moderate ($\leq 10^{-3}$), weak ($> 10^{-3}$), or no functional assignment.

4.5 Recovery of missing functional genes

To explore the biological significance of the recovered genes in greater depth, we focused on the *Malus domestica* proteome, where we identified over 50,000 peptides absent from the RefSeq annotation. Out of the 44,900 Helixer-predicted gene models, GAP-MS validated 25,153 peptide-supported models, which included 635 models with N-terminal peptide support and 2,804 with C-terminal support. Moreover it recovered 457 novel, peptide-supported gene models previously missing from the current RefSeq proteome.

The pipeline's utility in resolving annotation gaps is exemplified by the recovery of a novel gene model functionally annotated as a C3HC4-type RING finger protein, a putative E3 ubiquitin ligase. While this specific gene is missing from the standard NCBI RefSeq annotation, GAP-MS identified a robust gene model supported by 88.3% RNA-seq breadth and three unique splice-spanning peptides (Fig. 6a). Crucially, this protein sequence overlaps with an existing entry in the Ensembl annotation, allowing GAP-MS to act as an independent proteomic validator that confirms the Ensembl model is correct and that the RefSeq exclusion represents a false negative.

A further example of a biologically locus recovered by GAP-MS includes the TIR-NBS-LRR disease resistance protein (Fig. 6b). Accurate annotation of R-genes is often hindered by their characteristically low expression levels, which prevents RNA-seq datasets from providing sufficient evidence for automated gene prediction [55]. Furthermore, NB-LRR loci are frequently masked by annotation pipelines utilizing transposable element databases because these resistance genes are often misidentified as repetitive sequences [52]. Consequently, this specific gene is absent from the NCBI RefSeq annotation. GAP-MS identified proteomic support for the locus, including 13 internal peptides and one splice junction peptide. Domain analysis of the protein product confirmed a high prevalence of Leucine-Rich Repeats and NB-ARC signaling domains required for pathogen recognition. This discovery represents a broader trend across the dataset; systematic functional analysis of all novel gene models revealed a non-random distribution, significantly enriched for Gene Ontology (GO) terms associated with stress response and defense mechanisms (Supplementary Fig. 6).

Beyond discovering novel loci, GAP-MS demonstrates the capacity to correct structural errors in existing annotations, such as falsely merged genes. In one instance, the RefSeq annotation included a single continuous gene model "LOC103438048", whereas Helixer predicted two distinct, adjacent genes (Fig. 6c). GAP-MS provided definitive evidence supporting the split model by identifying a peptide mapping precisely to the N-terminus of the second predicted gene. The genomic sequence upstream of this peptide lacked the Arginine (R) or Lysine (K) residues required for tryptic cleavage, confirming the peptide represents a native biological N-terminus rather than a digestion artifact. This proves the existence of an independent translation initiation site in the middle of the RefSeq gene.

Structural refinement was also achieved for the gene model "LOC103414217", where C-terminal peptides supported a gene-split correctly predicted by Helixer but merged in RefSeq (Fig. 6d). While the stop codon at the end of the protein sequence was supported by a C-terminal peptide, another internal stop codon in the middle of the frame was confirmed by a second C-terminal peptide. Crucially, this internal peptide lacked a subsequent downstream tryptic cleavage site (not followed by R or K), providing physical evidence for an independent translation termination site. This confirms the Helixer prediction that the locus encodes two separate proteins rather than the single chimeric model presented in the standard reference.

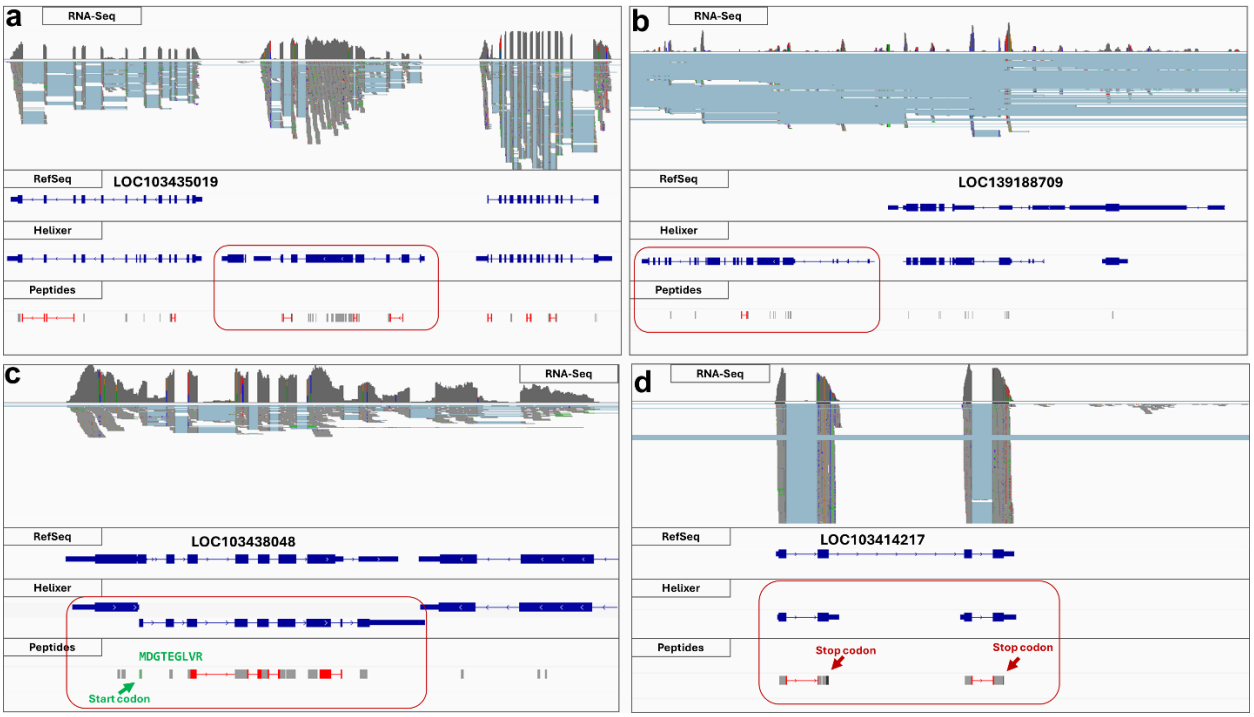


Figure 6. Proteogenomic validation and correction of RefSeq gene annotations in *Malus domestica*. Genome browser views showing loci where missing or mis-annotated RefSeq gene models are resolved using the GAP-MS proteogenomic pipeline. The top track displays RNA-seq read coverage. Gene models of RefSeq and Helixer annotations are shown in blue, mapped peptides in gray, and splice-junction peptides in red (bottom track). **(a)** GAP-MS recovers a functional gene absent from the reference RefSeq annotation, predicted by Helixer and supported by both peptide evidence and extensive RNA-seq coverage. **(b)** Identification of another gene missing from the NCBI RefSeq annotation: a TIR-NBS-LRR disease resistance protein containing an NB-ARC signaling domain required for pathogen recognition. **(c)** Correction of a falsely merged gene model. Helixer predicts two distinct adjacent gene models, whereas RefSeq incorrectly merges them into a single chimeric gene. GAP-MS identifies a peptide mapping precisely to the N-terminus of the downstream gene. The absence of an upstream tryptic cleavage site (R or K) confirms an independent translation start site. **(d)** Proteomic support for a gene split via C-terminal peptide mapping. An internal stop codon is confirmed by a C-terminal peptide lacking a subsequent tryptic cleavage site, corroborating the presence of two separate proteins rather than the merged RefSeq model.

5. Discussion

In this study, we introduce GAP-MS, a proteogenomic pipeline that integrates mass spectrometry evidence to evaluate and refine gene model predictions. By utilizing a high-quality subset from the forthcoming Crop Proteome Atlas, we applied this framework to 9 major crop species, demonstrating that integrating proteomic data allows us to confirm the translation of authentic coding sequences and point out possible prediction artifacts. To our best knowledge, this is the first pipeline to utilize mass spectrometry data to strictly filter gene models based on direct peptide translation evidence, moving beyond homology-based validation.

To evaluate the pipeline performance, we benchmarked GAP-MS against standard RefSeq annotations. While GAP-MS consistently improved precision across all tested tools, achieving substantial gains of up to ~32%, it reduced recall when compared to reference annotations. However, standard reference annotations are not absolute ground truth and often contain errors

such as pseudogenes, fragmented models, or transcriptionally silent regions. Thus, the observed drop in recall might reflect the correction of false positives in the reference rather than a loss of true genes. Moreover, the inherent limitations of mass spectrometry must be considered when interpreting “un-supported” gene models. Proteomics data is biased toward abundant proteins, making it difficult to detect those with low expression levels, or high tissue or condition specificity. Consequently, a genuine gene might be classified as unsupported simply because it falls below the detection limit of the instrumentation or was not expressed in the specific tissues sampled. By increasing proteomic depth through deep proteome sequencing, such as multi-protease digestion and extensive fractionation, researchers can achieve the sequence coverage required to bypass standard sensitivity limits. These workflows enable the global detection of protein isoforms and sequence variants typically missed by conventional shotgun methods. Integrating such comprehensive datasets into the GAP-MS framework directly enhances validation sensitivity and allows for the identification of novel gene models currently inaccessible to standard workflows [56].

Finally, GAP-MS was applied to identify novel coding regions, resulting in the recovery of functional genes missing from RefSeq (specifically those involved in stress response) showing its utility in improving genome annotation. Notably, across 9 crops more than 9,000 novel Helixer-predicted, peptide-supported gene models absent from current RefSeq annotations were identified. GAP-MS also demonstrated the capacity to correct structural errors in existing RefSeq reference annotations, such as resolving falsely merged chimeric models by identifying independent translation initiation and termination sites through specific N- and C-terminal peptides. To facilitate manual investigation, the pipeline also generates genomic coordinates of these mapped peptides in standard formats for seamless visualization in genome browsers.

Although not demonstrated in this work, GAP-MS could be extended with additional orthogonal evidence to improve sensitivity while preserving the distinction between inferred and directly observed proteins. Our ongoing work integrates transcriptomic data to provide evidence for transcript presence and to prioritize open reading frames that may be translationally repressed or fall below current proteomic detection limits. In parallel, advances in protein language models (PLMs) may help identify gene models with strong structural or evolutionary support despite lacking experimental protein evidence. Combining these orthogonal data types would create a more robust system that balances the strict specificity of proteomics with the broader sensitivity of transcriptomics and PLMs sequence analysis. Furthermore, de novo peptide sequencing offers a promising avenue to bypass the search space constraints of database-dependent proteogenomics. By deriving sequences directly from MS/MS spectra without a reference, this approach identifies translation products overlooked by initial gene prediction tools. Crucially, these de-novo-derived peptides can be seamlessly integrated into the GAP-MS input to facilitate the discovery of further novel genes.

Overall, GAP-MS represents a valuable advancement in the field of genome annotation, serving as an independent quality control filter. By validating gene models with physical evidence, the pipeline refines structural annotations and recovers biologically critical loci that are systematically missed by standard methods. The identification of overlooked genes, which are frequently fragmented or discarded due to their repetitive nature, highlights the immediate practical utility of this approach. In an era where accurate genomic resources are essential for modern breeding programs, the systemic loss of such agronomic traits in reference annotations compromises crop improvement strategies. Therefore, as the demand for high-quality functional genomics grows, we are convinced that integrating direct proteomic evidence through pipelines like GAP-MS will be essential for defining high-confidence reference proteomes and unlocking the full functional potential of crop genomes.

6. Data Availability

All computational scripts and datasets generated in this study are available on GitHub: https://github.com/qussai96/GAP-MS.

GAP-MS is freely available as a web interface at https://webclu.bio.wzw.tum.de/gapms/.

Identified peptides from mass spectrometry data of 9 crops together with gene models predicted with Braker2, Galba, Helixer and Annevo are uploaded to Zenodno https://zenodo.org/uploads/18445783.

7. Authors' contributions

Q.A., B.K. and D.F. jointly conceived the study. Q.A. wrote the main manuscript text, developed the pipeline, conducted the benchmarking of the tools, and created all plots and figures. D.F., M.W. and B.K. acquired the financial support for this project, conceptualized and supervised the work, in addition to reviewing and refining the manuscript. M.W. supervised the proteomics analyses, reviewed and edited the manuscript. B.K. supported the project with his expertise in proteomics and contributed to the revision process. All authors read and approved the final manuscript.

8. Conflict of interest

M.W. and B.K. are a co-founders of MSAID GmbH and scientific advisors to Momentum Biotechnologies, but have no operational role in either company. The other authors have no competing interests to declare.

9. Acknowledgements

The authors are grateful to all members of the 'Proteomes that Feed the World' international program for their essential contributions to the cultivation, sample preparation, and mass spectrometry data acquisition for the 9 crop species utilized in this study. Their foundational work is part of a coordinated effort to establish the Crop Proteome Atlas.

10. Funding

This research was funded by Elitenetzwerk Bayern (grant number F-6-M5613.6.K-NW-2021-411/1/1).

599 # 11. References

600 1. Lewin, H.A., et al., *Earth BioGenome Project: Sequencing life for the future of life.*
601 Proceedings of the National Academy of Sciences, 2018. **115**(17): p. 4325-4333.

602 2. *Darwin Tree of Life*. Available from: https://www.darwintreeoflife.org/.

603 3. Twyford, A.D., *The road to 10,000 plant genomes.* Nature Plants, 2018. **4**(6): p. 312-313.

604 4. *The European Reference Genome Atlas (ERGA) initiative*. Available from:
605 https://www.erga-biodiversity.eu/.

606 5. Guigó, R., *Genome annotation: From human genetics to biodiversity genomics.* Cell
607 Genom, 2023. **3**(8): p. 100375.

608 6. Vuruputoor, V.S., et al., *Welcome to the big leaves: Best practices for improving genome*
609 *annotation in non-model plant genomes.* Applications in Plant Sciences, 2023. **11**(4).

610 7. Brůna, T., A. Lomsadze, and M. Borodovsky, *GeneMark-EP+: eukaryotic gene prediction*
611 *with self-training in the space of genes and proteins.* NAR Genomics and Bioinformatics,
612 2020. **2**(2).

613 8. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J
614 Mol Biol, 1997. **268**(1): p. 78-94.

615 9. Stanke, M., et al., *AUGUSTUS: a web server for gene finding in eukaryotes.* Nucleic Acids
616 Research, 2004. **32**(suppl_2): p. W309-W312.

617 10. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database*
618 *management tool for second-generation genome projects.* BMC Bioinformatics, 2011.
619 **12**(1): p. 491.

620 11. Hoff, K.J., et al., *BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with*
621 *GeneMark-ET and AUGUSTUS.* Bioinformatics, 2016. **32**(5): p. 767-9.

622 12. Brůna, T., et al., *BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+*
623 *and AUGUSTUS supported by a protein database.* NAR Genomics and Bioinformatics,
624 2021. **3**(1).

625 13. Hoff, K.J., et al., *Whole-Genome Annotation with BRAKER.* Methods Mol Biol, 2019. **1962**:
626 p. 65-95.

627 14. *The NCBI Eukaryotic Genome Annotation Pipeline*. Available from:
628 https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/.

629 15. Aken, B.L., et al., *The Ensembl gene annotation system.* Database (Oxford), 2016. **2016**.

630 16. Holst, F., et al., *Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining*
631 *Deep Learning and a Hidden Markov Model.* bioRxiv, 2023: p. 2023.02.06.527280.

632 17. Ye, K., et al., *Highly accurate ab initio gene annotation with ANNEVO*. 2025.

633 18. Djossou, A., W. Ouedraogo, and A. Ouangraoua, *An overview of computational methods*
634 *for gene prediction in eukaryotes: strengths, limitations, and future directions.* Bioinform
635 Adv, 2025. **5**(1): p. vbaf222.

636 19. Meyer, C., et al., *Understanding the causes of errors in eukaryotic protein-coding gene*
637 *prediction: a case study of primate proteomes.* BMC Bioinformatics, 2020. **21**(1): p. 513.

638 20. *The Proteomes that Feed the World*. Available from: https://www.ls.tum.de/crop-
639 proteomes/the-proteomes-that-feed-the-world/.

640 21. Deutsch, E.W., et al., *The ProteomeXchange consortium in 2026: making proteomics data*
641 *FAIR.* Nucleic Acids Research, 2025. **54**(D1): p. D459-D469.

642 22. Datta, K.K., A.K. Madugundu, and H. Gowda, *Proteogenomic Methods to Improve*
643 *Genome Annotation.* Methods Mol Biol, 2016. **1410**: p. 77-89.

644 23. Ansong, C., et al., *Proteogenomics: needs and roles to be filled by proteomics in genome*
645 *annotation.* Briefings in Functional Genomics & Proteomics, 2008. **7**(1): p. 50-62.

646 24. Küster, B., et al., *Mass spectrometry allows direct identification of proteins in large*
647 *genomes.* Proteomics, 2001. **1**(5): p. 641-50.

648 25. Wright, J.C., et al., *Improving GENCODE reference gene annotation using a high-*
649 *stringency proteogenomics workflow.* Nature Communications, 2016. **7**(1): p. 11778.

650 26. Bock, T., et al., *An integrated approach for genome annotation of the eukaryotic*
651 *thermophile Chaetomium thermophilum.* Nucleic Acids Res, 2014. **42**(22): p. 13525-33.

652 27. *The NIH genetic sequence database (GenBank)*. Available from:
653 https://www.ncbi.nlm.nih.gov/genbank/.

654 28. Abbas, Q., et al., *Exploring crop genomes: assembly features, gene prediction accuracy,*
655 *and implications for proteomics studies.* BMC Genomics, 2024. **25**(1): p. 619.

656 29. *Food and Agriculture Organization of the United Nations (FAO)* . Available from:
657 https://www.fao.org/faostat/en/#data/QCL.

658 30. Seppey, M., M. Manni, and E.M. Zdobnov, *BUSCO: Assessing Genome Assembly and*
659 *Annotation Completeness.* Methods Mol Biol, 2019. **1962**: p. 227-245.

660 31. Ou, S., J. Chen, and N. Jiang, *Assessing genome assembly quality using the LTR*
661 *Assembly Index (LAI).* Nucleic Acids Res, 2018. **46**(21): p. e126.

662 32. Brůna, T., et al., *GALBA: Genome Annotation with Miniprot and AUGUSTUS.* bioRxiv,
663 2023.

664 33. *FragPipe: A complete proteomics pipeline with the MSFragger search engine at heart*.
665 Available from: https://fragpipe.nesvilab.org/.

34. da Veiga Leprevost, F., et al., *Philosopher: a versatile toolkit for shotgun proteomics data analysis.* Nature Methods, 2020. **17**(9): p. 869-870.

35. *The common Repository of Adventitious Proteins, cRAP*. Available from: https://www.thegpm.org/crap/.

36. Kong, A.T., et al., *MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics.* Nature Methods, 2017. **14**(5): p. 513-520.

37. Yang, K.L., et al., *MSBooster: improving peptide identification rates using deep learning-based features.* Nature Communications, 2023. **14**(1): p. 4539.

38. Lautenbacher, L., et al., *Koina: Democratizing machine learning for proteomics research.* Nature Communications, 2025. **16**(1): p. 9933.

39. Gessulat, S., et al., *Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning.* Nature Methods, 2019. **16**(6): p. 509-518.

40. Käll, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nature Methods, 2007. **4**(11): p. 923-925.

41. Mendoza, L., et al., *Flexible and Fast Mapping of Peptides to a Proteome with ProteoMapper.* J Proteome Res, 2018. **17**(12): p. 4337-4344.

42. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.

43. Lundberg, S. and S.-I. Lee *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv:1705.07874 DOI: 10.48550/arXiv.1705.07874.

44. Pertea, G. and M. Pertea, *GFF Utilities: GffRead and GffCompare.* F1000Res, 2020. **9**.

45. Leinonen, R., et al., *The Sequence Read Archive.* Nucleic Acids Research, 2010. **39**(suppl_1): p. D19-D21.

46. Stanke, M., et al., *VARUS: sampling complementary RNA reads from the sequence read archive.* BMC Bioinformatics, 2019. **20**(1): p. 558.

47. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.

48. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

49. Sommer, Markus J., Aleksey V. Zimin, and Steven L. Salzberg, *PSAURON: a tool for assessing protein annotation across a broad range of species.* NAR Genomics and Bioinformatics, 2025. **7**(1).

700 50. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.*
701 Bioinformatics, 2014. **30**(9): p. 1236-40.

702 51. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

703 52. Bayer, P.E., D. Edwards, and J. Batley, *Bias in resistance gene prediction due to repeat*
704 *masking.* Nature Plants, 2018. **4**(10): p. 762-765.

705 53. Dimonaco, N.J., et al., *No one tool to rule them all: prokaryotic gene prediction tool*
706 *annotations are highly dependent on the organism of study.* Bioinformatics, 2022. **38**(5): p.
707 1198-1207.

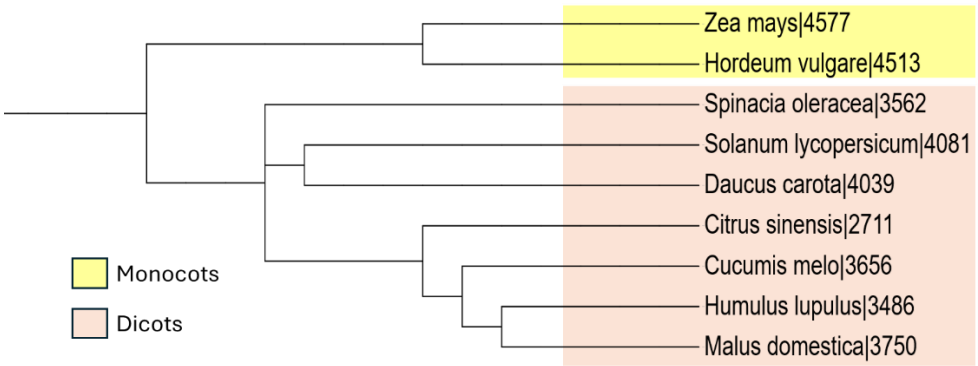
708 54. Parada, G.E., et al., *MicroExonator enables systematic discovery and quantification of*
709 *microexons across mouse embryonic development.* Genome Biology, 2021. **22**(1): p. 43.

710 55. Andolfo, G., J.C. Dohm, and H. Himmelbauer, *Prediction of NB-LRR resistance genes*
711 *based on full-length sequence homology.* Plant J, 2022. **110**(6): p. 1592-1602.

712 56. Sinitcyn, P., et al., *Global detection of human variants and isoforms by deep proteome*
713 *sequencing.* Nat Biotechnol, 2023. **41**(12): p. 1776-1786.

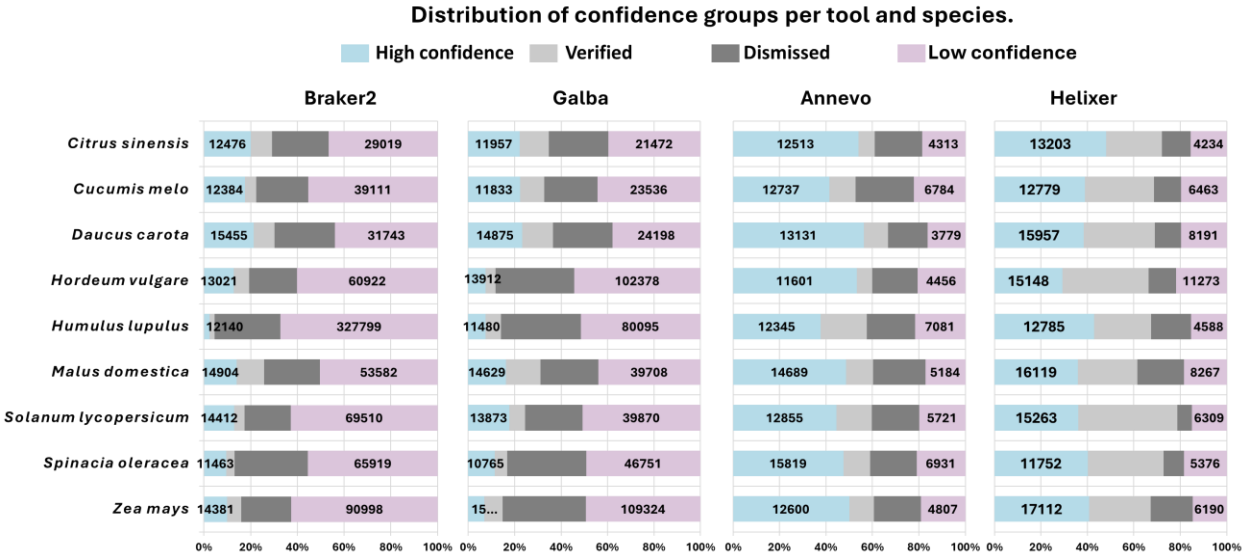
714

715 # 12. Supplementary figures



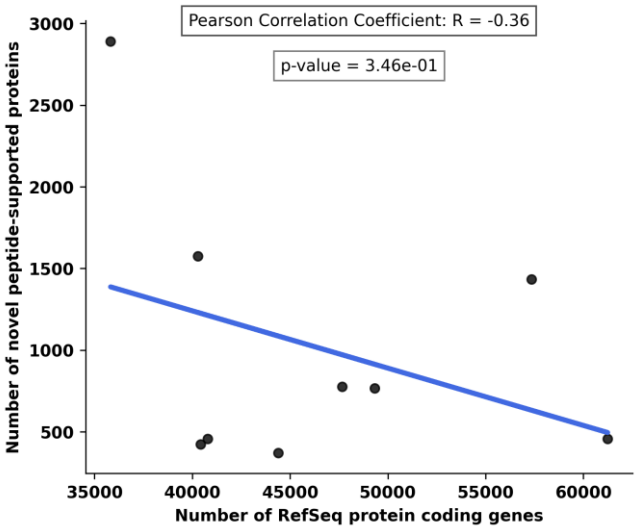
716

717 **Supplementary Fig 1. Phylogenetic tree of the selected 9 crop species (Species | TaxID).** Selected
718 species span diverse taxonomic groups and plant classes (monocots and dicots). The tree was
719 visualized and annotated using the interactive Tree of Life (iTOL) tool using NCBI taxonomy IDs.

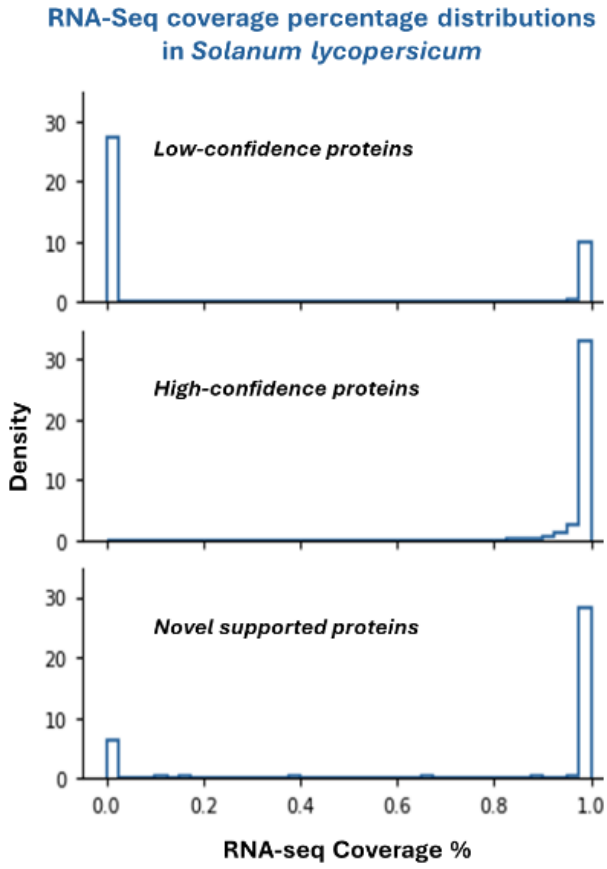


720

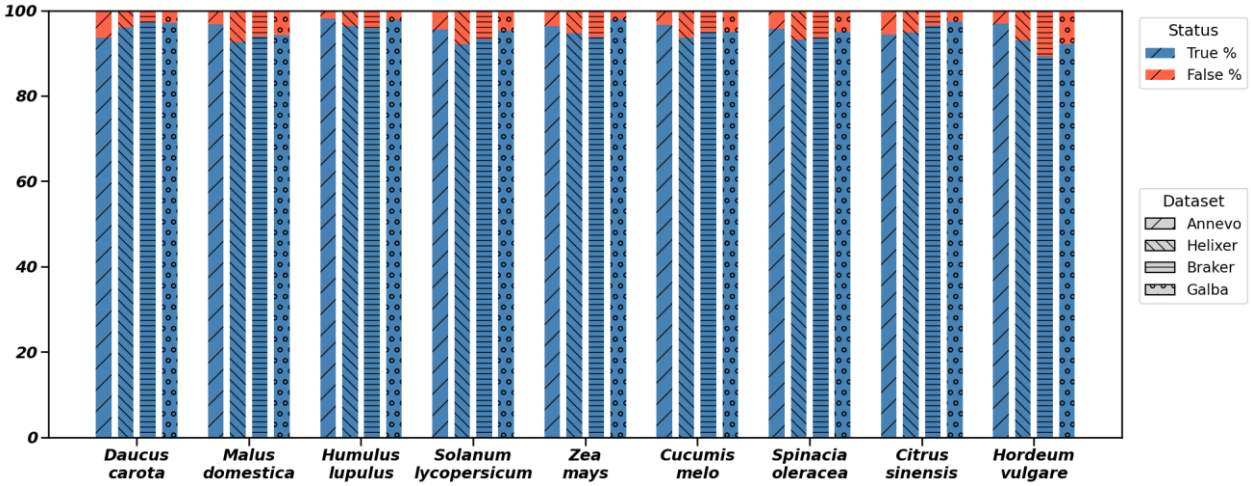
721 **Supplementary Fig 2. Distribution of confidence groups per species.** Stacked bars show the
722 proportion of proteins assigned to high-confidence (blue) and low-confidence (pink) training sets, as
723 well as the unlabeled proteins subsequently classified as supported (light gray) or unsupported (dark
724 gray) across 9 crops and 4 prediction tools.



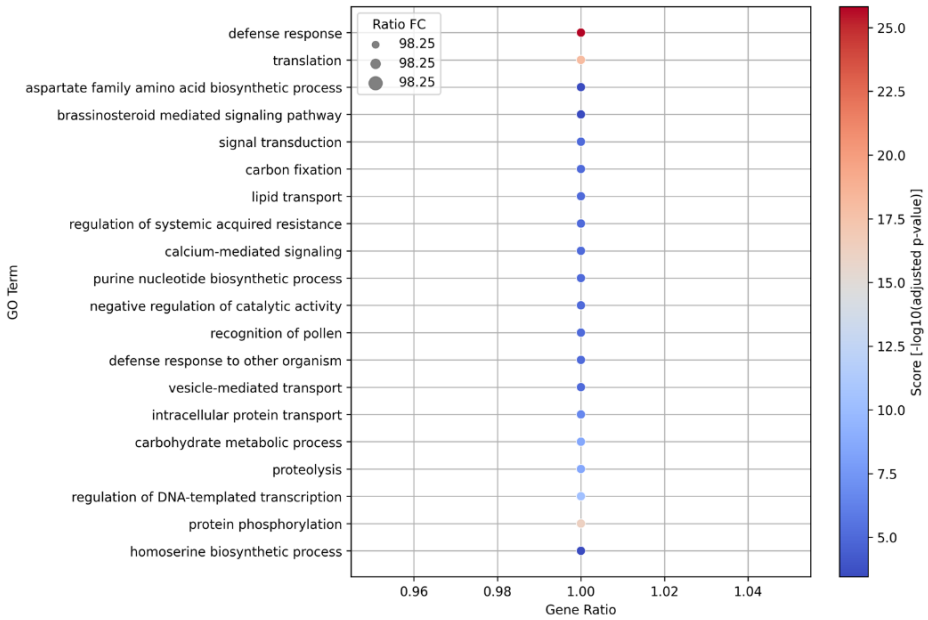
Supplementary Fig 3. Correlation between the number of RefSeq protein coding genes and the number of novel peptide-supported Helixer genes. A p-value of 0.34 indicated no significant correlation.



Supplementary Fig 4. Density distributions of RNA-seq coverage percentages in *Solanum lycopersicum* (tomato). The panels compare the coverage profiles of high-confidence reference proteins (top), low-confidence reference proteins (middle), and the novel peptide-supported proteins (bottom), highlighting the similarity in distribution between the novel and high-confidence sets.



Supplementary Fig 5. Quality assessment of novel peptide-supported proteins. Proportion of novel Helixer-predicted proteins (supported by GAP-MS but absent from RefSeq) that were classified as valid coding sequences (True) versus invalid (False) by Psauron.



Supplementary Fig 6. GO enrichment analysis of novel helixer peptide supported genes in *Malus domestica*. The y-axis lists enriched GO biological processes, showing that defense response emerging as the top enriched process, which indicated that the gene set is dominated by stress- or immunity-related functions