

Confidence-graded, proteome-wide identification of candidate antimicrobial peptide precursors in the Ghaf tree (*Prosopis cineraria*) using a preregistered, decoy-calibrated pipeline

Kiran Zahid

kzahid1@hct.ac.ae

Higher Colleges of Technology

Muhammad Riaz

Higher Colleges of Technology

Research Article

Keywords: antimicrobial peptides, *Prosopis cineraria*, plant defensins, snakins, non-specific lipid transfer proteins, proteome mining, preregistration, empirical calibration

Posted Date: September 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10927781/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background. Antimicrobial peptides are a candidate source of agents against resistant pathogens, and plant proteomes are routinely mined for them. Such surveys commonly report candidates without a control establishing the classifier's background positive rate, without calibrated significance for short-peptide similarity, and without a statement of how strongly each candidate is supported. *Prosopis cineraria*, the Ghaf, is an arid-zone mimosoid legume with a sequenced genome, documented antimicrobial activity in crude extracts and one characterised antifungal seed protein, but its antimicrobial peptide precursor complement has not been surveyed systematically.

Methods and results. We searched the annotated Ghaf proteome under a protocol registered before any sequence data were retrieved (DOI 10.17605/OSF.IO/RG9HA). Quality control and isoform reduction gave 29,222 primary proteins. Five prespecified evidence axes were applied: a machine-learning classifier, homology to curated reference peptides and precursors, antimicrobial peptide family domain signatures, secretion signals requiring agreement between two prediction tools, and an independent prediction applied to mature regions inferred under registered conditions. Short-peptide similarity was assigned empirical significance against 1,000 composition-preserving decoys per reference peptide rather than a fixed E-value threshold; 418 of 1,803 tested pairs survived calibration, with acceptance ranging from 8 of 86 pairs at 10–19 residues to 109 of 160 at 61–120 residues. Candidates were assigned to confidence tiers by an ordered decision rule fixed in advance. Of 489 shortlisted proteins, 18 reached the highest tier under both the registered rule and a stricter sensitivity analysis: six defensins with a knottin scaffold, five snakins or GASA proteins, four non-specific lipid transfer proteins, two knottins and one defensin. These candidates are short, cysteine-rich and predicted to be secreted, with a median full-precursor cysteine content of 10.1% against 3.7% in the next tier.

Conclusions. Eighteen candidate antimicrobial peptide precursors in the Ghaf proteome are supported by convergent evidence from prespecified criteria with calibrated homology significance. All are computational predictions and none has been shown to possess antimicrobial activity. We also report two control results of general relevance: a composition-preserving null cannot test discrimination for a classifier built on compositional features, and a peptide classifier applied beyond its stated size range saturates in a way that is easily mistaken for a property of the tool.

1. Introduction

Antimicrobial resistance has narrowed the range of effective treatments for common bacterial and fungal infections. Bacterial resistance was associated with an estimated 4.95 million deaths worldwide in 2019, of which 1.27 million were attributed directly to it (Murray et al., 2022); both are modelled estimates, and the authors emphasize that surveillance data are sparse in the settings where the burden is heaviest. Most antibacterial agents entering clinical use in recent decades are derivatives of existing scaffolds rather than compounds acting on new targets (Guilhelmelli et al., 2013; Moravej et al., 2018), which limits their durability against organisms already resistant to the parent class. The search for antimicrobials with distinct modes of action has consequently widened to include compound classes previously peripheral to drug development.

Antimicrobial peptides are one such class. They occur across all kingdoms as components of innate immunity, and many act by disrupting or perturbing microbial membranes rather than by inhibiting a single enzyme (Bahar & Ren, 2013; Zhang et al., 2021). That mechanism is thought to make resistance slower to emerge than for target-specific agents, although bacteria are not without countermeasures against it (Bechinger & Gorr, 2017; Moravej et al., 2018). Plants are a substantial and incompletely surveyed source. A sessile organism cannot escape a pathogen, and the constitutive and inducible peptide defences that plants deploy instead have produced a chemical diversity reflecting a long history of selection under attack (Bakare et al., 2022).

Plant antimicrobial peptides are conventionally grouped into families defined by cysteine spacing and disulfide-stabilised folds (Bakare et al., 2022). The lengths given here refer to mature peptides after removal of the signal peptide and any propeptide. Defensins, typically 45 to 54 residues with eight cysteines forming four disulfide bonds, are the largest and best characterised; they adopt a cysteine-stabilised $\alpha\beta$ motif and show broad activity against fungal and bacterial pathogens, in some cases through binding to fungal glucosylceramides or to the cell wall precursor lipid II. Non-specific lipid transfer proteins are longer, approximately 90 to 115 residues, and carry a conserved eight-cysteine motif. Hevein-like chitin-binding peptides of 30 to 43 residues bind chitin in the fungal wall through a type 1 chitin-binding domain. Knottins and cyclotides are shorter, commonly around 30 residues though the family spans a wider range, with three disulfide bonds forming a cysteine knot; cyclotides are additionally distinguished by a circular backbone. Thionins span roughly 45 to 50 residues with three or four disulfide bonds. Snakins and GASA proteins, at 63 to 104 residues with twelve cysteines, are the most heavily cross-linked of the group. What unites these families is a compact, cysteine-rich architecture that is stable outside the cell; many antimicrobial members are also cationic at physiological pH, and these properties together are what make the families recognisable computationally.

That recognisability has made genome and proteome mining a standard route to new candidates. Machine-learning classifiers trained on curated peptide databases can score an entire proteome in minutes (Fingerhut et al., 2020), and comparative assessments of such tools are

now available (Gabere & Noble, 2017; Xu et al., 2021; Dong et al., 2026). Related genome-wide studies have catalogued individual families: 28 defensin genes in durum wheat (Gamas et al., 2025) and 27 genes encoding 31 protein isoforms in soybean (Yuan et al., 2026). These are family-specific inventories built from domain searches and phylogenetic analysis rather than proteome-wide screens for antimicrobial peptides as a class, and the two approaches answer different questions.

The methods used in proteome-wide surveys deserve more scrutiny than they usually receive, on three counts.

The first concerns classifier behaviour at scale. Independent benchmarking has shown that predictive performance varies substantially between tools and that some carry high false-positive rates (Gabere & Noble, 2017; Xu et al., 2021). Applied to twenty or thirty thousand proteins, even a modest false-positive rate yields a substantial absolute number of positives, and a survey reporting no control cannot separate that contribution from genuine signal.

The second concerns similarity searching. Although alignment statistics account for query and database length (Karlin & Altschul, 1990; Altschul et al., 1997), very short peptide queries contain limited sequence information and are correspondingly sensitive to amino acid composition and to scoring assumptions. Conventional fixed thresholds may therefore give inadequate discrimination for short antimicrobial peptide sequences, which motivates empirical calibration against matched decoys.

The third concerns how candidates are reported. Prioritisation strategies do exist: the soybean study used selection-pressure ratios to identify genes under purifying selection (Yuan et al., 2026), and the durum wheat study used interaction networks and docking (Gamas et al., 2025). What is less common is a statement, fixed before the results are seen, of how strongly each candidate is supported by the antimicrobial evidence itself, so that a protein supported by several independent lines can be distinguished from one resting on a single marginal hit.

Prosopis cineraria, the Ghaf, is a mimosoid legume native to arid regions of the Arabian Peninsula and South Asia, where it tolerates prolonged drought, high salinity and extreme heat and contributes to soil stabilisation and nitrogen fixation in desert ecosystems (Gupta, 2024; Sharma & Singh, 2025). It has extensively documented traditional medicinal use, and crude extracts of bark, leaf, pod and stem have been reported to inhibit a range of Gram-positive and Gram-negative bacteria and several fungi (Giustra et al., 2022; Gupta, 2024). Extract-level activity in a plant may derive from secondary metabolites as readily as from peptides, and we treat it as background rather than as evidence of a peptide basis.

Protein-level evidence does exist in this species. A 38.6 kDa antifungal protein has been purified and characterised from Ghaf seed (Solanki et al., 2018). It is roughly an order of magnitude larger than any member of the peptide families described above, so while it establishes that Ghaf tissue contains antimicrobial protein activity, it does not identify an antimicrobial peptide, nor the gene encoding it. In the wider genus, peptides released by enzymatic hydrolysis of *Prosopis laevigata* cotyledon proteins inhibit *Staphylococcus aureus* (Sánchez-Jiménez et al., 2026); these are hydrolysis products of storage proteins rather than natively encoded antimicrobial peptides and are noted here as indirect context only.

A reference genome and annotation for *P. cineraria* are available (Sudalaimuthuasari et al., 2022), and the genomic resources for the species have recently been reviewed (Sharma & Singh, 2025). The antimicrobial peptide precursor complement encoded in that genome has not, to our knowledge, been systematically surveyed.

This study reports a survey conducted under a protocol registered before any sequence data were retrieved. Five evidence axes were combined: a machine-learning classifier, homology to curated reference peptides and precursors, antimicrobial peptide family domain signatures, predicted secretion, and an independent prediction applied to inferred mature regions. Two features distinguish the approach from the surveys cited above. Short-peptide similarity was assigned empirical significance against composition-matched decoy sequences rather than a fixed threshold, and candidates were assigned to grade confidence tiers by an ordered decision rule fixed in advance, so that the strength of support for each candidate is stated rather than left to be inferred. We report the null and negative results of the pipeline alongside the positive ones, including a control that proved uninformative and a prediction tool that saturated when applied outside its design range, because both bear on how surveys of this kind should be calibrated.

2. Materials and Methods

2.1 Preregistration, reproducibility and claim boundaries

Protocol version 2.2 was preregistered on the Open Science Framework on 9 August 2026 (DOI 10.17605/OSF.IO/RG9HA), before any sequence data were downloaded; the registration timestamp precedes the proteome download recorded in Section 2.2 on the same day. The final protocol version is 2.8, and all versions are deposited with the registration. Six subsequent amendments were recorded in the registration with their dates and reasons. Three were made before the analyses they governed had been run: two concerned properties of the reference

data that could not have been known in advance, and one concerned the source of the precursor reference set. Three were made after the relevant step had been executed. Two corrected statistical or implementation errors identified during internal checking, and one followed a structured methodological review of the protocol and manuscript draft conducted with AI assistance. In the latter case, the implementation had departed from the registered mature-region rule; the departure, correction and consequences are described in Section 2.8. None of the amendments was motivated by the biological content of a result, and every affected analysis was rerun.

All analyses were run in R 4.6.1 on a single Windows workstation, with package versions frozen using renv and a fixed random seed (20260808) for every stochastic step. Scripts, the package lockfile, the decision log and all intermediate tables are deposited with the registration.

Every candidate reported here is a computational prediction. We have not demonstrated antimicrobial activity for any Ghaf protein, and nothing in this study should be read as doing so.

2.2 Proteome acquisition and quality control

We obtained the *Prosopis cineraria* proteome from NCBI RefSeq (assembly GCF_029017545.1, annotation release GCF_029017545.1-RS_2023_03) on 9 August 2026 and verified each downloaded file against the published MD5 checksums before use. This is a Gnomon annotation of the assembly reported by Sudalaimuthuasari et al. (2022), not the annotation described in that paper, so protein counts differ between the two and are not directly comparable.

The download contained 50,190 proteins across 30,645 GeneIDs. We retained the longest valid isoform of each gene, breaking ties by the lexicographically first stable accession, which gave 29,222 primary sequences. Every protein mapped to a GeneID, and no accession was duplicated.

Sequences containing undetermined residues were excluded, since an ambiguous position leaves classifier featurisation, alignment scoring and physicochemical descriptors undefined, and substituting a residue would amount to inventing data. This removed 1,423 proteins. Each was the sole isoform of its gene, so 1,423 genes, 4.6% of the annotated total, are absent from the analysis entirely rather than represented by an alternative product. Their median length was 543 amino acids, well above the size of a plant antimicrobial peptide precursor, but 94 fall within the plausible range and we report these individually as not assessable rather than as negative results.

2.3 Reference peptide and precursor sets

The mature-peptide reference set was drawn from the 2024 sequence release of the Antimicrobial Peptide Database (Wang et al., 2026), plant subset, and metadata were inherited from the DRAMP general dataset by exact sequence match. Inclusion rests on APD curation, which admits natural peptides with a defined sequence and demonstrated antimicrobial activity. We had originally intended to grade each reference from the primary experimental report, but the assembled set would have required reading of the order of two hundred publications, and the available metadata carried a quantitative activity value for only one of 177 matched entries. We therefore inherited the APD standard and measured its reliability instead, by verifying a random sample of thirty included peptides against the primary literature. Records annotated in DRAMP as predicted or inferred from homology were excluded. This yielded 257 curated peptides, which we describe throughout as peptides curated by APD6 as natural antimicrobial peptides with reported activity, and not as experimentally validated peptides.

Of the thirty peptides drawn, twenty-nine were assessed; one could not be traced to a source record. Eighteen were confirmed on both criteria, being natural plant peptides with a reported antimicrobial assay. Five carried a documented antimicrobial assay but had been obtained by chemical synthesis or heterologous expression rather than isolated from plant tissue. We count these as natural plant gene products, on the grounds that the route by which material was obtained for assay does not bear on whether the peptide occurs naturally, which gives a verification rate of 79% (23 of 29). Under the stricter reading, in which only peptides isolated directly from plant tissue are counted, the rate is 62% (18 of 29). Two records could not be resolved either way.

Four peptides were confirmed as natural plant products but had no antimicrobial assay in the cited source; the reported assays were haemolytic, insecticidal, structural or cytotoxic. All four are cyclotides. This is not a random failure. Cyclotides are frequently catalogued on the strength of the family's documented activity while individual members are characterised for other properties, and because cyclotides are among the larger families in the mature reference set we report this as a specific limitation rather than folding it into an overall error rate.

A random 20% sample of the curated reference set (51 peptides) was independently reclassified without reference to the recorded classification, on three judgements: whether the peptide should be included, whether it is a mature peptide or a precursor, and its source species. Concordance with the recorded classification was 51 of 51 for inclusion, 50 of 51 for mature versus precursor status, and 36 of 37 for source species among peptides where both records named an organism. The two discordant records are informative. One 92-residue sequence recorded as a mature peptide was reclassified as ambiguous, its length and composition supporting neither assignment

confidently. In the second, the recorded source was the kingdom Plantae rather than an organism, and the independent classification identified *Zea mays*, which the primary literature confirms. Both discrepancies originate in the inherited metadata rather than in the analysis.

The registered protocol specified that this audit report Cohen's kappa. That statistic cannot be computed for these data. The recorded classification carries no variance: every peptide in the sample is recorded as included and as a mature peptide, because those values were assigned by the inherited curation standard rather than by an independent curator, and Cohen's kappa is undefined when one set of classifications is invariant, there being no chance agreement to correct for. The audit is therefore reported as a concordance check, and the deviation is recorded in the protocol amendment log. High concordance is the expected result for a classification that follows a deterministic rule, and we do not present it as evidence that the framework is robust; what it establishes is that the inherited classification is internally consistent and that its two errors, both in source metadata, are of a kind that does not propagate into the analysis. A preliminary systematic classification of the same sample had been generated and reviewed before the reclassification reported here; the reclassification was blind to the recorded first-pass values but not naive to that preliminary pass, and this is stated as a limitation. No result reported in this study depends on the audit.

Neither APD nor DRAMP catalogues full-length precursors, and DRAMP's Swiss-Prot cross-references point predominantly at mature-peptide records: of 104 accession-peptide pairs we retrieved, 59 returned the mature peptide itself. We therefore assembled the precursor set by querying UniProtKB directly for reviewed Viridiplantae entries carrying antimicrobial, antibiotic or antifungal keywords, and retained those with protein-level or transcript-level evidence, an annotated signal peptide, a length of 40 to 600 residues and no non-standard residues. The evidence filter is not discretionary here: calibrating a homology search against references that were themselves identified by homology would be circular. It reduced the set from 373 to 152 entries and, as a consequence, reduced the *Arabidopsis thaliana* share from 76% to 46%. That residual over-representation of a single well-annotated model species remains a limitation of the precursor set.

The two reference sets therefore rest on different curation standards, and they feed two separate analyses with separate thresholds. Thirty of the 152 precursors contain a curated mature peptide, so the two sets have limited sequence overlap. Limited overlap does not make the resulting evidence axes statistically independent, and Section 2.11 sets out the dependencies that remain.

Both reference sets were clustered at 90% identity using Clusterize in DECIPHER 3.7.1 (Wright, 2024; cutoff 0.10, method "shortest", minimum coverage 0.90), retaining one representative per cluster chosen by strongest documented evidence, then greatest length, then first accession. Cluster membership is not transitive under these settings, so members of a cluster are not necessarily pairwise 90% identical. Clustering gave 204 mature and 139 precursor representatives.

2.4 Classifier screen

We scored all 29,222 primary proteins with the precursor model of ampir 1.1.0 (Fingerhut et al., 2020), using prespecified probability thresholds of 0.50, 0.80 and 0.90. Ninety-six proteins reached 0.80.

Calibrating this axis required a control, and our first choice was uninformative. A composition-preserving shuffle of the whole proteome produced almost exactly the observed positive rate at every threshold, with observed and shuffled scores correlated at $r = 0.76$. Shuffling preserves amino acid composition, which is what the classifier's decision boundary uses, so the null cannot test discrimination; we report it as a property of the method rather than as evidence about the classifier. A positive control settled the question. Scoring the 139 curated precursors recovered 95.7% at a threshold of 0.80, against a proteome background of 0.3%, an enrichment of 291-fold. Because ampir's training data draw on the same public repositories that supplied our reference sets, this figure is an upper bound on sensitivity rather than an unbiased estimate.

2.5 Homology searches and empirical calibration

All searches used BLAST + 2.17.0 (Camacho et al., 2009) with a fixed effective database size of 10^8 and low-complexity filtering disabled, the latter because SEG masking removes the cysteine- and glycine-rich composition that characterises the target families. Fixing the database size matters because E-values scale with it, so a constant E-value threshold applied across proteomes of different sizes would impose a different stringency in each.

Precursor references were searched against the Ghaf proteome with standard blastp. We accepted hits at $E \leq 10^{-5}$ with at least 30% identity and 70% query coverage, and report hits between 10^{-5} and 10^{-3} separately as a sensitivity band. This gave 673 accepted alignments across 284 proteins.

Short mature peptides required different treatment. The blastp-short task is parameterised for queries below roughly 30 residues, so we split the reference set at that boundary and searched each part under its own task, keeping the same matrix, database size and coverage filter within each group. Searches were run permissively so that the full score distribution survived for calibration.

For each mature reference we generated 1,000 composition-preserving shuffles, searched them under the identical task, and retained only the highest bit score for each shuffled query. Decoys returning no hit above the coverage threshold contributed a score of zero and remained in the null; discarding them would truncate the distribution and make subsequent P values anticonservative. Null distributions were pooled within length stratum and task, never across tasks. Each observed reference-protein pair received an empirical P value of $(1 + k)/(1 + N)$, where k was the number of pooled decoy scores at least as high as the observed score and N was the number of decoy scores in that group, followed by Benjamini-Hochberg correction across all 1,803 tested pairs. Pairs with $P \leq 0.01$ and $q \leq 0.05$ were accepted, giving 418 pairs across 77 proteins.

Pooling improves resolution but assumes approximate exchangeability of the best-decoy score distributions among reference peptides within each length stratum despite differences in amino-acid composition. The resulting values are therefore interpreted as stratum-calibrated rather than query-specific empirical probabilities. The per-reference decoy scores are deposited with the study registration, so a query-specific null can be reconstructed and the accepted-pair set recomputed under each peptide's own distribution without regenerating decoys.

An empirical P value computed this way cannot fall below $1/(1 + N)$, and 304 of the 418 accepted pairs reached that bound. For those pairs the P value is a bound rather than an estimate, and we report it as such. Greater resolution would require more decoys, not a different test.

2.6 Domain annotation

The registered protocol specified local HMMER3 searches against a frozen Pfam-A release as the primary domain route, with InterPro used for confirmation. No local HMMER3 search was performed. Domain annotation rests entirely on the InterPro sequence search (release 109.0), which was applied to the full shortlist. All 489 sequences were submitted in batches within a single recorded session so that member-database versions could not change between batches (Blum et al., 2025; Paysan-Lafosse et al., 2025). Of these, 480 returned annotations; the nine that did not were recorded as unannotated rather than as lacking domains. This departure from the registered domain route reduces exact reproducibility of the F-axis calls and is treated as a limitation.

We defined the domain axis narrowly, counting only canonical plant antimicrobial peptide families: defensins and defensin-like proteins, knottins, hevein-like chitin-binding proteins, non-specific lipid-transfer proteins, snakins and GASA proteins, and low-molecular-weight cysteine-rich proteins. Seventy proteins carried at least one such signature. Thaumatin-like and PR-5 proteins, which contributed the largest single block of annotations, and Kunitz and Bowman-Birk protease inhibitors are defence-related but were not counted as antimicrobial peptide families; they are reported separately. The definition is deliberately conservative and may exclude proteins with genuine antimicrobial activity. The complete dictionary, listing every InterPro and Pfam accession counted as an antimicrobial peptide family together with the family it maps to, is provided as Supplementary Table S4. Accessions excluded from the axis, including the thaumatin and protease inhibitor entries, are listed alongside so that the boundary of the definition is explicit.

Snakin and GASA proteins were absent from the initial family dictionary and were added following the AI-assisted methodological review described in Section 2.1, increasing the F-positive count from 52 to 70; they account for 16 of the 70 proteins. All downstream evidence matrices and tier assignments were recomputed after this addition. Four candidates moved from Tier C to Tier B as a result; no candidate entered or left Tier A, and five Tier A candidates gained a family assignment they had previously lacked. This change illustrates the sensitivity of a family-based axis to the completeness of the dictionary that defines it, a general limitation of the approach rather than a property of this dataset.

2.7 Secretion signals and membrane topology

Signal peptides were predicted with SignalP 6.0 (Eukarya, fast mode; Teufel et al., 2022), subcellular targeting with TargetP 2.0 (Plant networks; Almagro Armenteros et al., 2019), and transmembrane topology with DeepTMHMM 1.0.57 (Hallgren et al., 2022). SignalP cannot distinguish a secretory signal peptide from a plastid or mitochondrial transit peptide, and in a photosynthetic tree such peptides are abundant, so we required both tools to agree before recording positive evidence of secretion. Three hundred and fifteen proteins met this criterion and thirty-six were recorded as conflicting.

The three tools returned 316, 350 and 347 signal regions respectively. We report that spread rather than reconciling it, since it indicates how much a signal peptide call depends on the method chosen. Among proteins where SignalP and TargetP agreed on the presence of a signal peptide, they placed the cleavage site identically in 90% of cases and within three residues in 98%. Notably, no protein called a signal peptide by SignalP was assigned a transit peptide by TargetP, so the concern that motivated the requirement did not materialise in this dataset, although the check remains necessary.

2.8 Mature regions and physicochemical characterisation

A mature region was assigned only where permitted by the registered protocol: where a decoy-calibrated alignment from a curated mature reference peptide defined the corresponding region directly, or where an antimicrobial peptide family domain began at or after the predicted

signal-peptide cleavage site, making the domain boundaries compatible with cleavage. Alignment evidence took precedence over domain evidence where both were available, because an alignment to a characterised peptide is stronger evidence of a boundary than a domain model. Where a candidate carried more than one calibrated alignment, the region was taken from the alignment with the highest bit score. Where the boundary derived from domain signatures and more than one signature was present, the region spanned the earliest signature start to the latest signature end. For every assignment, precursor coordinates, the inference source and the selected boundary evidence were recorded in the mature-region inference table. The protocol also permitted inference from a conserved, literature-supported processing motif; this could not be automated and contributed no assignments. Where no condition applied, the mature region was recorded as not applicable and the precursor characterised separately. Regions shorter than 10 residues were rejected, since no meaningful physicochemical description can be derived from them. No upper length bound was applied: the longest region inferred under these conditions was 95 residues, so any such bound would have been inactive.

Eighty-three of 489 candidates met these conditions, with a median inferred length of 64 residues (range 15–95).

We report a correction to an earlier implementation of this step. The analyses used in a previous manuscript draft and in protocol versions up to v2.7 treated the entire signal-cleaved sequence returned by SignalP as the mature region and applied none of the conditions above. For a small defensin the two approaches nearly coincide, but for a longer protein the earlier approach produced a putative mature region of 200 residues or more with no biological counterpart; across the 315 candidates affected, the median was 189 residues. The error was identified during the AI-assisted methodological review described in Section 2.1, and the affected analyses were rerun. It did not alter tier membership, but it changed the physicochemical description of candidates and generated the spurious independent-predictor behaviour described in Section 2.9.

The 83 inferred mature regions were rescored with ampir's mature model, which is the model intended for mature-peptide sequences; in our control, the precursor model recovered only 23% of curated mature peptides and would yield systematically lower scores here. Net charge, isoelectric point, hydrophobicity, Boman index, aliphatic index, instability index and residue composition were calculated with the Peptides package (Osorio et al., 2015). The same descriptors were computed for the curated reference peptides to permit direct descriptive comparison.

These properties describe and prioritise candidates. They are not an evidence axis and take no part in tier assignment. They also cannot provide independent validation of the tiers because the classifier axis is itself based on compositional and physicochemical features; agreement is interpreted as internal coherence rather than external corroboration.

2.9 Independent mature-peptide prediction

Inferred mature regions were submitted to the CAMPR4 natural-peptide prediction server, which presented three models: random forest, support vector machine and artificial neural network (Gawde et al., 2023). A strict majority of the three models was required for a positive M-axis call. Of the 83 candidates with an inferable mature region, 75 were positive and eight negative; the remaining 406 candidates were recorded as not applicable because no mature region could be inferred.

The earlier analysis based on the erroneous regions described in Section 2.8 returned 311 positives among 315 candidates, with the support vector machine assigning a probability of at least 0.999 to 246. We initially attributed this saturation to application beyond the classifier's stated 3–100-residue training range and restricted the axis accordingly. That attribution was incorrect: the classifier had been supplied with whole signal-cleaved precursors rather than inferred mature peptides. After correction, all 83 regions fell within the stated range; the support vector machine saturated for 17 of 83, and all three models agreed for 78% of candidates rather than 57%. The earlier length restriction was therefore withdrawn.

CAMPR4's training data draw on the same public repositories as ampir and the reference sets used here. The M axis is therefore not independent of the classifier or homology axes and is treated as corroborating evidence only.

2.10 Comparative analysis

We compared candidates against three legume proteomes: *Neltuma alba* (syn. *Prosopis alba*, GCF_004799145.1), *Glycine max* (GCF_000004515.6) and *Phaseolus vulgaris* (GCF_000499845.2). Each accession was verified as current on the download date. The primary-isoform rule applied to the Ghaf proteome was applied identically to all three, reducing them by between 29% and 37%, and every search used the same fixed database size. Without both measures, apparent differences in conservation would reflect annotation style and proteome size rather than biology. Homologues were accepted at $E \leq 10^{-5}$ with at least 30% identity and 70% query coverage.

The *N. alba* annotation dates from 2019 and the two *Papilionoideae* annotations are more mature than the 2023 Ghaf release. Absence of a detectable homologue may therefore reflect annotation completeness rather than true absence, and we do not describe any candidate as species-specific on this evidence.

2.11 Confidence tiers

Candidates were assigned to one of three confidence tiers by an ordered decision table evaluated from top to bottom, stopping at the first matching rule and terminating in a default so that every candidate received exactly one tier. The rule that fired was recorded for each candidate; a tier that cannot be traced to a rule index constitutes a protocol violation.

Tier A required a classifier probability of at least 0.80, homology or domain evidence, no major annotation flag, no unresolved contradiction, and either a positive independent prediction or no inferable mature region. Tier B admitted candidates with homology or domain evidence supported by at least one further axis, and separately admitted candidates whose only positive axis was calibrated homology. The latter rule reflects that a composition-based classifier can miss specific cysteine-rich families and that decoy-calibrated homology alone is stronger than a classifier score alone. All remaining candidates fell to Tier C.

An unresolved contradiction was defined before examining how many candidates it would affect as any of: a predicted transmembrane helix, because a classical secreted antimicrobial peptide precursor should not be a membrane protein; disagreement between SignalP and TargetP; or an antimicrobial peptide domain in a protein too long to be that family's precursor. Missing evidence was not treated as a contradiction. A protein without a detected signal peptide, an inferable mature region, an applicable independent prediction or a matching domain under the deliberately narrow family definition was recorded as lacking that evidence rather than as conflicting. One hundred and one candidates carried at least one contradiction.

Rule R2 permits Tier A assignment when the independent prediction is either positive or not applicable, a branch intended to avoid penalising a candidate when the tool cannot be applied. Because only 83 of 489 candidates had an inferable mature region, this branch can admit candidates without independent-prediction evidence. We therefore performed a sensitivity analysis in which a positive prediction was required. Tier A contained 23 candidates under the registered rule and 18 under the sensitivity rule. The five candidates that differed were 146–229 residues long, lacked an antimicrobial peptide family domain and had no inferable mature region. They are reported separately from the robust 18-candidate subset throughout.

The evidence axes are not statistically independent. The classifier, reference sets and independent predictor draw on overlapping public repositories, while domain and homology evidence frequently derive from the same underlying similarity. Convergence across axes is therefore described qualitatively and axis-level probabilities are not combined.

2.12 Declaration of AI assistance

AI tools were used in two roles, both disclosed here. First, analysis scripts were drafted with AI assistance, then reviewed, executed, and verified by the authors; all outputs were checked against the source data. Second, AI tools assisted with language and presentation during manuscript preparation. AI was not used to generate data, results, or references, and was not used as a source of factual or scientific content. The author reviewed and verified all scripts, outputs, and text, and takes full responsibility for the work.

3. Results

3.1 The Ghaf proteome, reference sets and candidate shortlist

The analysis pipeline and the counts at each stage are summarised in Fig. 1. The downloaded annotation contained 50,190 proteins. Sequence quality control and deterministic isoform selection produced 29,222 analysable primary proteins, one per GeneID. A further 1,423 GeneIDs are unrepresented in the analysis because their only annotated product contained undetermined residues; 94 of these fall within the size range of a plant antimicrobial peptide precursor and are recorded individually as not assessable rather than as negative.

Filtering and clustering of the reference data produced 204 mature-peptide and 139 precursor representatives (Table 1). The two sets rest on different curation standards and feed the two homology analyses separately.

The shortlist contained 489 proteins. Most entered on a single axis: 290 through calibrated homology alone, 76 through a classifier probability of at least 0.80 alone, and 84 through an intermediate classifier score alone. Only 20 proteins carried both a high classifier score and homology evidence, and 19 carried an intermediate score with homology.

Table 1

Composition of the antimicrobial peptide reference sets Sources, filtering and clustering of the two reference sets used to identify candidate antimicrobial peptide precursors in the Ghaf proteome. The two sets rest on different curation standards and were used in two separate analyses (Steps 7A and 7B) with separate thresholds.

Source	Mature peptide set	Precursor set
	Antimicrobial Peptide Database, 2024 sequence release, plant subset	UniProtKB/Swiss-Prot
Retrieval	Direct download	Query: reviewed entries, Viridiplantae (taxonomy ID 33090), keywords antimicrobial, antibiotic or antifungal
Date retrieved	9 August 2026	9 August 2026
Metadata source	DRAMP general dataset, matched by exact sequence	UniProtKB annotation
Curation standard inherited	APD curation: natural peptide, defined sequence, demonstrated antimicrobial activity	Swiss-Prot review
Records retrieved	268	552
Excluded: outside eligible length	6 (4 below 10 aa; 2 above 120 aa, routed to the precursor set)	89 outside 40–600 aa
Excluded: non-standard residues	0	14
Excluded: evidence annotated as predicted or inferred from homology	5	233
Excluded: no annotated signal peptide	not applicable	64
Retained after filtering	257	152
Exact duplicate sequences removed	0	0
Clustering	DECIPHER 3.7.1 Clusterize: cutoff 0.10, method “shortest”, minimum coverage 0.90, terminal gaps included	identical parameters
Cluster representatives used	204	139
Length range of representatives	10–120 aa	61–557 aa
Largest cluster	8 members	3 members

Membership overlap between the classifier and homology axes was therefore limited. Twenty proteins were shared between the 96 with a classifier probability of at least 0.80 and the 329 with homology evidence, a Jaccard index of 0.049; taking all 199 proteins at 0.50 or above gives 39 shared and an index of 0.080 (Fig. 2). Limited membership overlap is not the same as statistical independence, and we do not treat it as such: the classifier, the reference sets and the independent predictor all draw on overlapping public repositories. What it does establish is that the axes are not redundant, and that a candidate satisfying several of them has been reached by more than one route.

Upper panel, the number of candidates in each intersection; lower left, the number positive on each axis; lower right, the intersection matrix, in which filled circles indicate the axes satisfied. Axes are A, classifier probability at least 0.80; H, calibrated homology; F, antimicrobial peptide family domain; S, secretion signal; M, independent prediction. The largest intersections are single axis rather than multi-axis: 290 candidates entered the shortlist on homology alone and 20 carried both a high classifier score and homology evidence, a Jaccard index of 0.049. Limited membership overlap does not establish statistical independence, as the classifier, the reference sets and the independent predictor all draw on overlapping public repositories.

3.2 Classifier calibration

Ninety-six of the 29,222 primary proteins reached a classifier probability of 0.80 or above. Our first attempt to calibrate this axis was uninformative, and we report it because the reason is general rather than particular to this dataset. A composition-preserving shuffle of the entire proteome produced almost exactly the observed positive rate at every threshold and observed and shuffled scores correlated at $r = 0.76$. Shuffling destroys residue order while preserving amino acid composition and many of the physicochemical features the classifier uses.

It therefore cannot provide an informative negative null for a classifier driven largely by those properties, and the result characterises the method rather than the proteome.

Positive control answered the question the null could not. Scoring the 139 curated reference precursors recovered 133 of them (95.7%) at a threshold of 0.80, against 96 of 29,222 background proteins (0.3%), an enrichment of 291-fold (Fig. 3). This recovery should be read as a potentially optimistic estimate rather than an unbiased measure of sensitivity, since the repositories that supplied our reference set may overlap the classifier's training data.

Distribution of ampir precursor-model probabilities for four sets of sequences: curated reference precursors ($n = 139$), curated reference mature peptides ($n = 204$), the Ghaf primary proteome ($n = 29,222$), and a composition-preserving shuffle of that proteome ($n = 29,222$). Violins show the full distribution and boxes the interquartile range with median. The dashed line marks the prespecified threshold of 0.80. Reference precursors were recovered at 95.7% against a proteome background of 0.3%, an enrichment of 291-fold. The shuffled proteome reproduces the observed distribution almost exactly, because shuffling preserves the compositional and physicochemical features on which the classifier's decision boundary rests while destroying only residue order; it therefore cannot serve as an informative negative control for this classifier. Reference mature peptides score lower than reference precursors because the precursor model is applied to them outside its intended input.

3.3 Homology under empirical calibration

Searching the curated precursor references against the Ghaf proteome produced 673 accepted alignments across 284 proteins at the prespecified thresholds.

The mature-peptide searches were run permissively and yielded 19,761 raw alignments, reduced to 1,803 reference–protein pairs after the coverage filter. Empirical calibration against 1,000 composition-preserving decoys per reference peptide accepted 418 of these pairs, distributed across 77 proteins.

Acceptance varied sharply with peptide length. Eight of 86 tested pairs were accepted in the 10–19 aa stratum, 35 of 773 at 20–39 aa, 266 of 784 at 40–60 aa and 109 of 160 at 61–120 aa. Longer reference peptides therefore produced alignments that shuffled sequence rarely reproduced at the required coverage, while short peptides produced alignments that the decoys frequently matched. Short-peptide similarities were therefore substantially more likely than longer ones to be reproduced by composition-matched decoys, which is the behavior the calibration was designed to expose.

Of the 418 accepted pairs, 304 (72.7%) scored higher than every decoy in their group and therefore reached the minimum attainable empirical P value, $1/(1 + N)$. These values cannot distinguish between matches more extreme than the simulated null; additional decoys would provide greater resolution, but a different test would not.

3.4 Domain annotation, secretion signals and independent prediction

Domain annotation returned signatures for 480 of the 489 shortlisted proteins; the nine that returned none are recorded as unannotated rather than as lacking domains. Seventy proteins carried at least one signature from the six counted antimicrobial peptide families: non-specific lipid transfer proteins, defensins and defensin-like proteins, knottins, snakins and GASA proteins, hevein-like chitin-binding proteins, and low-molecular-weight cysteine-rich proteins. Thaumatin-like and PR-5 proteins contributed the largest single block of annotations to the shortlist but are not counted here and are reported separately with the Kunitz and Bowman-Birk protease inhibitors.

Requiring agreement between two prediction tools, 315 proteins carried positive evidence of secretion and 36 were recorded as conflicting. The three tools applied returned 316, 350 and 347 signal regions respectively, a spread we report rather than reconcile. Among proteins where both tools called a signal peptide, the predicted cleavage site was identical in 90% of cases and within three residues in 98%. No protein called as carrying a signal peptide by SignalP was assigned to an organellar transit-peptide class by TargetP, so the specific misassignment that motivated the two-tool requirement did not arise in this dataset.

A mature region could be inferred under the registered conditions for 83 candidates, and independent prediction was assessable only for these. Seventy-five were positive and eight negatives. An earlier analysis that supplied whole signal-cleaved precursors to the predictor rather than inferred mature regions returned a positive call for 311 of 315 inputs; correcting the input reduced that saturation substantially and brought every submitted sequence within the predictor's stated size range.

3.5 Confidence tiers

Under the registered decision rule the 489 candidates resolved into 23 Tier A, 229 Tier B and 237 Tier C. Under the sensitivity analysis, in which an inapplicable independent prediction no longer satisfies the Tier A condition, the counts were 18, 234 and 237. All 101 candidates carrying an unresolved contradiction fell to Tier C.

The 18 candidates supported under both rules comprise six defensins with a knottin scaffold, five snakins or GASA proteins, four non-specific lipid transfer proteins, two knottins without a defensin call, and one defensin without a knottin call (Fig. 4, Table 2). All 18 carry domain evidence, secretion evidence and a positive independent prediction; 15 carry calibrated homology evidence.

Rows are the 18 candidates assigned to Tier A under both the registered decision rule and the sensitivity analysis, ordered by classifier probability and labelled with accession and RefSeq annotation. Columns are the evidence axes, with the homology axis shown separately for its precursor and mature-peptide arms. Filled cells indicate positive evidence. All 18 candidates carry domain, secretion and independent-prediction evidence, and all satisfy the classifier threshold; 15 carry calibrated homology evidence, by either arm. Variation across the panel is confined to the two homology arms, several candidates being supported by precursor homology without a calibrated mature-peptide alignment, or the reverse.

Table 2

Tier A candidate antimicrobial peptide precursors The 18 highest-confidence candidate antimicrobial peptide precursors identified in the *Prosopis cineraria* proteome. All are Tier A under both the registered tier rule and the sensitivity analysis. Candidates are ordered by classifier probability. Axis abbreviations: A, classifier (ampir precursor model); H, calibrated homology, shown separately for the precursor and mature-peptide arms; F, antimicrobial peptide family domain; S, secretion signal (SignalP and TargetP agreement); M, independent prediction (CAMPR4 majority of three models). Conservation gives the number of the three comparative legume proteomes in which a homologue was detected.

Accession	RefSeq annotation	Length (aa)	Mature (aa)	Family	A	H prec.	H mat.	F	S	M	Charge	Cys (%)	Cons.
XP_054810874.1	defensin NsD7-like	78	49	Defensin, knottin	1.000	+	+	+	+	+	+ 7.68	16.3	3/3
XP_054825515.1	defensin Ec-AMP-D1-like isoform X1	80	56	Defensin, knottin	1.000	+	+	+	+	+	+ 6.60	14.3	3/3
XP_054822800.1	defensin-like protein 1	74	47	Defensin, knottin	0.998	+	+	+	+	+	+ 6.78	17.0	3/3
XP_054809759.1	defensin Ec-AMP-D1-like	77	47	Defensin, knottin	0.998	+	+	+	+	+	+ 8.69	17.0	3/3
XP_054799467.1	protein RSI-1	96	59	Snakin/GASA	0.997	-	+	+	+	+	+ 9.34	20.3	3/3
XP_054811879.1	defensin-1-like	81	48	Knottin	0.996	+	+	+	+	+	+ 0.66	14.6	0/3
XP_054804942.1	defensin-like protein 1	82	45	Defensin, knottin	0.996	-	+	+	+	+	+ 2.69	17.8	3/3
XP_054799468.1	peamaclein	88	60	Snakin/GASA	0.991	+	+	+	+	+	+ 7.35	20.0	3/3
XP_054810497.1	defensin-like protein 2	78	58	Defensin, knottin	0.979	+	+	+	+	+	+ 5.69	13.8	3/3
XP_054799935.1	putative defensin-like protein 128	91	43	Defensin	0.966	-	-	+	+	+	-1.13	18.6	2/3
XP_054804865.1	gibberellin-regulated protein 12-like	96	59	Snakin/GASA	0.950	-	+	+	+	+	+ 4.35	20.3	3/3
XP_054813856.1	defensin-like protein 1	81	49	Knottin	0.947	+	+	+	+	+	-0.16	14.3	1/3
XP_054819093.1	putative lipid-transfer protein DIR1	101	75	nsLTP	0.937	-	-	+	+	+	+ 10.50	10.7	3/3
XP_054782801.1	non-specific lipid-transfer protein 2	92	37	nsLTP	0.908	-	+	+	+	+	+ 5.75	10.8	3/3
XP_054789817.1	gibberellin-regulated protein 4-like	107	66	Snakin/GASA	0.883	+	+	+	+	+	+ 10.34	18.2	3/3
XP_054799942.1	gibberellin-regulated protein 2-like	93	63	Snakin/GASA	0.848	-	+	+	+	+	+ 6.35	19.1	3/3

Accession	RefSeq annotation	Length (aa)	Mature (aa)	Family	A	H prec.	H mat.	F	S	M	Charge	Cys (%)	Cons.
XP_054804013.1	protein MEN-8	96	62	nsLTP	0.834	–	–	+	+	+	–1.40	12.9	3/3
XP_054797586.1	non-specific lipid-transfer protein 1-like	118	90	nsLTP	0.831	+	+	+	+	+	+4.50	8.9	3/3

The five candidates admitted under the registered rule alone are three P21-like and two PR-4-like proteins. All five lack a counted antimicrobial peptide family domain and have no inferable mature region, which is why the independent prediction was not applicable to them, and their median length of 227 aa is approximately 2.5 times the 89.5 aa median of the 18 (Table 3). We report them separately throughout and do not include them in the headline count.

Table 3

Candidates admitted to Tier A under the registered rule only Five candidates assigned to Tier A by the registered decision rule but to Tier B under the sensitivity analysis. Rule R2 admits a candidate when the independent prediction is positive **or** not applicable; these five qualify through the second branch, having no inferable mature region and therefore no applicable independent prediction. They are reported separately from the 18 candidates in Table 2 throughout this study.

Accession	RefSeq annotation	Length (aa)	Mature region	Family domain	H prec.	H mat.	Charge	Cys (%)	Cons.
XP_054813049.1	protein P21-like	227	not inferable	none assigned	+	+ 5.10	7.1	7.9	3/3
XP_054823413.1	pathogenesis-related protein PR-4-like	146	not inferable	none assigned	+	–0.16	4.8	4.9	3/3
XP_054813702.1	protein P21-like	227	not inferable	none assigned	+	+ 2.10	7.1	7.9	3/3
XP_054792011.1	pathogenesis-related protein PR-4-like	146	not inferable	none assigned	+	+ 5.84	4.8	4.9	3/3
XP_054813275.1	protein P21-like	229	not inferable	none assigned	+	+ 6.98	7.9	8.8	3/3

3.6 Sequence characteristics and conservation

Full-precursor cysteine content most clearly distinguished the 18 sensitivity-supported Tier A candidates from the remainder of the shortlist (Fig. 5, Table 4). Their interquartile range, 9.77 to 11.11%, scarcely overlaps that of Tier B, 1.87 to 6.15%: only three of 229 Tier B candidates fall inside the Tier A interquartile range. Tier A candidates are also short and tightly distributed in length, whereas the other tiers span an order of magnitude.

Distributions of length, cysteine content, net charge at pH 7 and cationic residue content, computed on the full precursor for all 489 shortlisted candidates. Tier A is shown split into the 18 candidates supported under both the registered rule and the sensitivity analysis and the 5 supported under the registered rule alone. Boxes show the interquartile range with median; whiskers extend to 1.5 times the interquartile range and points beyond are plotted individually. Cysteine content distinguishes the 18 most clearly: their interquartile range of 9.77 to 11.11% scarcely overlaps that of Tier B, 1.87 to 6.15%, and only three of 229 Tier B candidates fall within it. Net charge separates the groups far less clearly, the Tier C median of + 3.06 approaching the Tier A median of + 6.99. Composition is reported on the full precursor so that no tier is represented by a subset; mature-region values could be computed for only 59 of 229 Tier B and 6 of 237 Tier C candidates. These sequence characteristics took no part in tier assignment, but agreement between them and the tiers is internal coherence rather than independent validation, since the classifier axis is itself built on compositional features.

Table 4

Characteristics of the confidence tiers Evidence and sequence characteristics of the 489 shortlisted candidates by confidence tier. Tier A is shown split into the 18 candidates assigned to Tier A under both the registered decision rule and the sensitivity analysis, and the 5 assigned by the registered rule alone (Table 3). Composition is reported on the full precursor for every candidate, so that no tier is represented by a subset; mature-region values are given separately below and only where a region could be inferred under the registered conditions. Values are medians, with interquartile ranges in parentheses where given, or percentages where stated. None of the sequence characteristics took any part in tier assignment.

Candidates	Tier A (18)	Tier A (5)	Tier B	Tier C
	18	5	229	237
Evidence axes (% positive)				
A – classifier ≥ 0.80	100	100	0	30
H – calibrated homology	83	100	99	35
F – AMP family domain	100	0	20	3
S – secretion signal	100	100	88	38
M – independent prediction, positive	100	0	22	3
Unresolved contradiction	0	0	0	43
Precursor composition (all 489 candidates)				
Length (aa)	89.5	227	221	127
Cysteine content (%)	10.13 (9.77–11.11)	7.05 (4.79–7.05)	3.73 (1.87–6.15)	3.85 (2.50–7.55)
Net charge at pH 7	+ 6.99 (2.54–7.69)	+ 5.10 (2.10–5.84)	+ 0.34 (– 3.31–3.34)	+ 3.06 (– 0.06–6.01)
Cationic residues (%)	13.92	9.25	8.62	11.11
Inferred mature region				
Candidates with an inferable region	18 / 18	0 / 5	59 / 229	6 / 237
Mature region length (aa)	57.0	–	69.0	66.5
Net charge at pH 7	+ 6.05	–	+ 1.94	+ 5.76
Cysteine content (%)	16.68	–	9.72	10.67
Four or more cysteines, even count (%)	89	–	54	67
Conservation				
Candidates assessed comparatively	18 / 18	5 / 5	229 / 229	155 / 237
Homologue in all three legumes (% of assessed)	83	100	94	83

Net charge separated the groups far less clearly. The Tier C median of + 3.06 approaches the Tier A median of + 6.99 and the distributions overlap substantially, and cationic residue content distinguishes Tier A from Tier B but not from Tier C. We therefore rest no claim on charge.

These comparisons are internal coherence rather than independent validation. The classifier axis, which every Tier A candidate must satisfy, is itself built on compositional and physicochemical features, so some agreement between tier assignment and composition is expected by construction. What the comparison establishes is that the tiers are associated with distinct sequence profiles, and not only with the number of evidence axes a candidate satisfies.

Comparative searching against three legume proteomes did not discriminate between the tiers. A homologue was detected in all three comparators for 15 of 18 sensitivity-supported Tier A candidates (83%), all five registered-only candidates (100%), 216 of 229 Tier B candidates (94%) and 129 of 155 assessed Tier C candidates (83%). The comparative search was run on the 407 candidates carrying classifier, homology or domain evidence, so 82 Tier C candidates have no comparative result; expressing the Tier C figure over all 237 would treat those as lacking a homologue and would understate conservation as 54%. Conservation is therefore reported as characterisation and not as evidence.

One Tier A candidate, XP_054811879.1, returned no detectable homologue in any of the three comparators. It also carries seven cysteine residues, an odd number and unusual for a disulfide-rich antimicrobial peptide family, and a net charge of only + 0.66. We describe it as unusual within this set rather than as restricted to *Prosopis cineraria*: the *Neltuma alba* annotation dates from 2019 and is less complete than the two *Papilionoideae* annotations, so absence may reflect annotation coverage rather than true absence.

4. Discussion

4.1 Principal findings and contribution

Applying a preregistered pipeline to the annotated *Prosopis cineraria* proteome identified 489 candidate antimicrobial peptide precursors, of which 18 met the Tier A decision rule under both the registered analysis and a stricter sensitivity analysis. Those 18 fall into families that a plant biologist would expect from a legume: six defensins carrying a knottin scaffold, five snakins or GASA proteins, four non-specific lipid transfer proteins, two knottins and one defensin. They are short, cysteine-rich and predicted to be secreted, and most are conserved across the three comparative legume proteomes.

The candidate set is not, however, a set of discoveries. Nine of the 18 already carry a RefSeq annotation naming them as a defensin or defensin-like protein, and several of the remainder are annotated as gibberellin-regulated or lipid-transfer proteins. What this study contributes is not the existence of these genes but their systematic identification under criteria fixed in advance, with calibrated significance attached to the homology evidence and a graded confidence assignment that separates well-supported candidates from weakly supported ones. Accordingly, the principal contribution lies in evidence-calibrated prioritisation rather than in the first annotation of these gene families.

Comparison with published work is complicated by differences in scope. Most genome-wide studies of plant antimicrobial peptides are family inventories rather than proteome-wide screens: 28 defensin genes in durum wheat (Gamas et al., 2025), 27 genes encoding 31 isoforms in soybean (Yuan et al., 2026). Others proceed in the opposite direction, isolating a peptide first and then searching sequence databases to identify it, as in the characterisation of the hevein-like peptide HEV-CANN from bell pepper leaves (Games et al., 2016). Neither design is directly comparable with a proteome-wide screen for candidates across families, and we therefore make no claim that the shortlist reported here is larger or smaller than is typical. What differs is that the basis for placing a candidate in the highest tier is stated in advance rather than implied by a threshold.

4.2 The candidate repertoire in context

The families recovered are consistent with what is known of legume defence peptide repertoires. The soybean defensin family comprises 27 loci encoding 31 protein isoforms, larger than the 15-member *Arabidopsis* family and comparable to those reported in other legumes (Yuan et al., 2026). That figure is an exhaustive family inventory and is not directly comparable with the seven defensin-family candidates reported here, which are the survivors of a filter requiring several independent lines of evidence rather than a complete census of the family.

Where expression has been examined in a legume, defensins are broadly but variably distributed. In soybean, qRT-PCR showed one subfamily expressed constitutively across tissues and another enriched at likely points of pathogen entry, with one member over 40-fold enriched in roots relative to leaves and another expressed strongly in pods (Yuan et al., 2026). We have no expression data for Ghaf, and that pattern is relevant to what should be measured next rather than to what has been found here.

Defensins and knottin-scaffold peptides dominate the highest-confidence group, which is unsurprising given that these families are both well represented in the reference databases and structurally distinctive enough for a compositional classifier to recognise. The five snakin and GASA proteins are worth noting separately, because they entered the domain axis only after that family was added to our dictionary during revision. Had we not corrected that omission, five of the 18 would have carried no family assignment, and the apparent composition of the candidate set would have been different.

That episode illustrates a general point about family-based evidence axes. Their output is bounded by the completeness of the dictionary that defines them, and a dictionary assembled by hand will reflect the assembler's familiarity with the literature. We report the correction rather than quietly incorporating it because the sensitivity of this axis to dictionary completeness is a property of the approach, not an accident of our implementation.

One candidate, XP_054811879.1, returned no detectable homologue in any comparator proteome. It is tempting to describe such a candidate as species-restricted, and we resist that description. The *Neltuma alba* annotation dates from 2019 and is less complete than the two *Papilionoideae* annotations, so absence of a detected homologue is at least as consistent with incomplete annotation as with genuine absence. The same candidate also carries seven cysteine residues, an odd number in a family whose members are defined by paired cysteines, and a net charge of only + 0.66. We report it as unusual within the set and would not build a claim on it.

4.3 What the controls revealed

Two control results are worth reporting, because both concern how computational antimicrobial peptide prediction is usually calibrated rather than anything specific to Ghaf.

The first is that a composition-preserving null was uninformative for the classifier we used. Shuffling every protein in the proteome while preserving its amino acid composition produced almost exactly the observed positive rate at every threshold. The reason is straightforward once stated: the classifier's decision boundary rests on compositional features, and shuffling preserves precisely those features while destroying only residue order. A null that preserves the signal it is meant to remove cannot test discrimination. We report this because a shuffled-sequence null is a plausible choice, and because its failure here was not obvious to us in advance.

The positive control answered the question the null could not. Curated reference precursors were recovered at 95.7% against a proteome background of 0.3%, an enrichment of 291-fold. That figure should be read as a potentially optimistic estimate of recovery rather than an unbiased estimate of sensitivity, since the public repositories that supplied our reference set overlap the classifier's training data. This concern is general rather than particular to our design: benchmarking studies have shown that similarity between training and evaluation sequences causes random data splitting to overestimate predictive performance, and that most predictors are trained on only one or a few public databases and perform less well on genuinely independent data (Gabere & Noble, 2017; Xu et al., 2021; Dong et al., 2026). The dependency runs through the whole of our design. The classifier, the reference sets and the independent predictor all draw on the same handful of public databases, so convergence across those axes is partly convergence on a shared origin. We describe the axes as non-redundant rather than independent, and we do not combine their probabilities.

The second control result is a caution about applying tools beyond their stated range. Our independent predictor returned a positive call for 311 of 315 inputs when supplied with whole signal-cleaved precursors, and the support vector machine returned a probability of at least 0.999 for 246 of them. We initially attributed this saturation to the tool. It was in fact a consequence of our own error in inferring mature regions, and correcting the input brought every submitted sequence within the predictor's design range and reduced the saturation substantially. The general lesson is that an implausibly uniform output should first prompt checks for a mismatch between the tool and its input, before it is interpreted as a property of either. Length dependence of this kind is documented: classifier behaviour reflects the length distribution of the sequences on which the model was trained, commonly 10 to 80 residues (Xu et al., 2021).

4.4 Calibrated homology and the value of tiering

Assigning empirical significance to short-peptide alignments changed what the homology axis reported. Of 1,803 tested reference–protein pairs, 418 survived calibration against composition-matched decoys. Acceptance varied sharply with peptide length: 8 of 86 pairs in the 10–19 aa stratum against 109 of 160 at 61–120 aa. Short peptides produced alignments that shuffled sequence reproduced frequently, and a fixed E-value threshold applied across these strata would have accepted many of them.

The principle is not new. Decoy databases constructed from shuffled or reversed sequences are standard practice in mass-spectrometry proteomics, where they are used to estimate false discovery rates in spectrum-to-peptide matching, and the approach and its limitations have been examined extensively (Choi & Nesvizhskii, 2008; Couté et al., 2020). What appears uncommon is the application of the same principle to sequence-similarity calibration in plant antimicrobial peptide mining. None of the studies we examined reports it. Those studies also illustrate substantial variation in the thresholds and search strategies used for defensin-family identification: durum wheat used known wheat defensins as queries at $E < 10^{-50}$ followed by domain confirmation (Gamas et al., 2025), while soybean combined *Arabidopsis* queries at $E < 10^{-5}$ with hidden Markov model profiling (Yuan et al., 2026). Because their query sets, databases and workflows differed, these thresholds are not directly comparable; their variation nonetheless reinforces the value of dataset-specific empirical calibration.

The need has been noticed before. Games et al. (2016) observed that computational tools and databases are not adjusted for short sequences and that statistical tests in large databases require adjustment, but did not implement a calibration.

The tier assignment did work that a simple candidate count would not. Candidates entering on a single axis behaved differently from those entering on several, and the difference is visible in sequence characteristics that took no part in the assignment. Full-precursor cysteine content separates the 18 Tier A candidates from Tier B almost completely, with only three of 229 Tier B candidates falling inside the Tier A interquartile range. Net charge, by contrast, does not separate the groups, and we rest no claim on it.

We are careful to describe this as internal coherence and not as validation. The classifier axis, which every Tier A candidate must satisfy, is itself built on compositional features, so some agreement between tier and composition follows from the construction. What it does show is that the tiers are associated with distinct sequence profiles rather than only with the number of criteria satisfied, and that a homology hit alone is weak evidence of antimicrobial peptide identity in a plant proteome: Tier B carries homology evidence for 99% of its members yet has a median classifier probability of 0.07 and a median precursor length two and a half times that of Tier A.

There is a further reason to prefer graded tiers to a ranked list of scores. A classifier score is a ranking device and need not be a calibrated probability: as one benchmarking study puts it, an ROC curve measures the quality of a ranking without addressing whether 80% of the predictions assigned a posterior probability of 80% are in fact correct (Gabere & Noble, 2017). Tiers defined by which evidence a candidate satisfies do not inherit that problem, because they make no probabilistic claim.

4.5 Limitations

No antimicrobial activity has been demonstrated for any candidate reported here. Every result is a computational prediction, and the tier labels rank the strength of computational evidence rather than any likelihood of biological function. In this respect the study is not unusual: the durum wheat and soybean surveys likewise report predicted gene families with expression profiling but without functional assays, and both note that protein-level validation is required (Gamas et al., 2025; Yuan et al., 2026).

The reference set rests on inherited curation. Of a random sample of thirty peptides drawn for verification against the primary literature, twenty-nine could be assessed; 79% of these were confirmed under our reading, and 62% under a stricter one that counts only peptides isolated directly from plant tissue. The failures were not randomly distributed: all four peptides lacking a documented antimicrobial assay were cyclotides, a family catalogued partly on the reputation of its members' relatives. Since cyclotides are among the larger families in our mature reference set, this is a specific weakness rather than a diffuse error rate. The underlying problem is general. Public peptide datasets are assembled from heterogeneous sources without unified standards for collection, screening criteria or experimental background, which limits both curation quality and cross-study comparison (Xiaoying et al., 2026).

Our reference sets also carry a taxonomic bias, with *Arabidopsis thaliana* contributing 46% of the precursor set after evidence filtering. This too reflects a wider pattern, in which prediction tools trained on a small number of public databases perform less well on independent data (Xu et al., 2021).

The domain axis uses a deliberately narrow definition. Thaumatin-like and PR-5 proteins, and the Kunitz and Bowman-Birk protease inhibitors, are excluded although several have documented antifungal activity. This choice keeps the axis interpretable but will exclude proteins with genuine antimicrobial function, and the five candidates admitted to the highest tier only under the registered rule are all pathogenesis-related proteins that the narrow definition does not count.

Annotation completeness constrains several results. The RefSeq annotation used here differs from the annotation published with the genome, and 1,423 genes, 4.6% of the annotated total, are absent from the analysis entirely because their only product contained undetermined residues. Ninety-four of those fall within the size range of a plant antimicrobial peptide precursor. In the comparative analysis, differing annotation maturity between the four proteomes means that no statement about absence of a homologue can be read as a statement about absence of a gene.

Finally, the protocol was amended six times after registration, three of those amendments after the affected analysis had run. One corrected a departure from the registered protocol in which mature regions were inferred without applying the registered conditions. We have documented each amendment, its date and its consequences, and re-run every affected analysis. Readers should nonetheless weight a preregistration with this history differently from one that required no amendment.

4.6 Conclusions and further work

Eighteen candidate antimicrobial peptide precursors in the Ghaf proteome are supported by convergent computational evidence drawn from five prespecified axes, and belong to families with documented antimicrobial function in other plants. Not every candidate satisfies every axis: fifteen of the eighteen carry calibrated homology evidence, and all eighteen carry domain, secretion and independent-prediction evidence. They are a defensible starting point for experimental work, which is the only route by which any of them can be shown to have antimicrobial activity.

The inferred mature regions range from 37 to 90 residues, which spans more than one established route to material. Peptides at the shorter end are within comfortable reach of solid-phase synthesis. Heterologous expression provides an alternative for longer candidates: recombinant potato snak-in-1 has been produced in a soluble state in baculovirus-infected insect cells after bacterial, yeast and synthetic systems proved problematic (Almasia et al., 2017), and snak-ins produced in heterologous systems have been reported to retain antibacterial and antifungal activity. Antimicrobial activity would nonetheless have to be established for any expressed Ghaf product; the recombinant material in that study was used to raise antibodies rather than assayed. Purification from tissue is a third route, and was how the 40-residue bell pepper hevein-like peptide HEV-CANN was obtained and assayed (Games et al., 2016). We would suggest a panel of four to six candidates spanning the represented families rather than the highest-scoring alone, followed by growth-inhibition assays against defined bacterial and fungal targets, with haemolysis and cytotoxicity assessed for any peptide showing activity. Expression profiling under salinity,

drought and pathogen challenge would address the environmental question this analysis cannot, and there is precedent for it: soybean defensin subfamilies show tissue-specific expression consistent with distinct defensive roles (Yuan et al., 2026).

Beyond the candidate list, we would draw attention to three methodological points that generalise. Composition-preserving nulls do not test discrimination for classifiers built on composition. Empirical calibration of short-peptide similarity changes which hits survive, and does so most where sequence is shortest; the principle is borrowed from proteomics rather than invented here, but it appears little used in this application. And an evidence-tier system with an explicit, ordered decision rule makes the strength of support for each candidate legible to a reader, which a ranked list of scores does not.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests. The author declares no competing interests.

Ethics approval. Not applicable. This study analysed publicly available genomic and proteomic sequence data and involved no human participants, animals, or field collection of plant material.

Data availability. All primary data analysed in this study are publicly available. The *Prosopis cineraria* proteome was obtained from NCBI RefSeq (assembly GCF_029017545.1, annotation release GCF_029017545.1-RS_2023_03), and the comparative proteomes from GCF_004799145.1 (*Neltuma alba*), GCF_000004515.6 (*Glycine max*) and GCF_000499845.2 (*Phaseolus vulgaris*). Reference peptides were obtained from the Antimicrobial Peptide Database 2024 sequence release and the DRAMP general dataset, and reference precursors from UniProtKB/Swiss-Prot. Derived data, including the full candidate table with evidence assignments for all 489 shortlisted proteins, the curated reference sets, the domain family dictionary and all supplementary tables, are deposited with the study registration on the Open Science Framework (DOI 10.17605/OSF.IO/RG9HA).

Code availability. All analysis scripts, the package lockfile recording every dependency version, the decision log and the full protocol version history are deposited with the study registration (DOI 10.17605/OSF.IO/RG9HA). Analyses were run in R 4.6.1 with a fixed random seed (20260808).

Preregistration. The analysis protocol was registered on the Open Science Framework on 9 August 2026, before any sequence data were downloaded (DOI 10.17605/OSF.IO/RG9HA). Protocol version 2.2 was registered; the final version is 2.8. All six amendments, their dates, their justifications and their effects on the reported results are documented in the registration record and summarised in the Methods.

Declaration of AI assistance. See Section 2.12. AI tools assisted with the analysis scripts and manuscript preparation; all outputs were verified by the authors, who takes full responsibility for the work.

Author contributions. Author contributions. Kiran Zahid conceived and designed the study, registered the protocol, performed the analyses, verified the results, and wrote the manuscript. Muhammad Riaz contributed to the design and running of the candidate screening, analysis, and interpretation, and reviewed and approved the manuscript. Both authors approved the final version.

References

1. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H (2019) Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance* 2(5):e201900429
2. Almasia NI, Molinari MP, Maroniche GA, Nahirňak V, Barón B, Taboga MP, O.A. and, Rovere V (2017) C., Successful production of the potato antimicrobial peptide Snakin-1 in baculovirus-infected insect cells and development of specific antibodies. *BMC Biotechnology*, 17(1), p.75
3. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402
4. Bahar AA, Ren D (2013) Antimicrob peptides *Pharmaceuticals* 6(12):1543–1575
5. Bakare OO, Gokul A, Fadaka AO, Wu R, Niekerk LA, Barker AM, Keyster M, Klein A (2022) Plant antimicrobial peptides (PAMPs): features, applications, production, expression, and challenges. *Molecules*, 27(12), p.3703
6. Bechinger B, Gorr SU (2017) Antimicrobial peptides: mechanisms of action and resistance. *J Dent Res* 96(3):254–260
7. Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale

- DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A (2025) InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* 53(D1):D444–D456
8. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1), p.421
 9. Choi H, Nesvizhskii AI (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res* 7(1):47–50
 10. Couté Y, Bruley C, Burger T (2020) Beyond target–decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics. *Anal Chem* 92(22):14898–14906
 11. Dong X, Yang K, Wu T, Li P, Gao L (2026) A systematic benchmark for peptide property prediction. <https://doi.org/10.64898/2026.02.09.704773>. bioRxiv
 12. Fingerhut LC, Miller DJ, Strugnell JM, Daly NL, Cooke IR (2020) ampir: an R package for fast genome-wide prediction of antimicrobial peptides. *Bioinformatics* 36(21):5262–5263
 13. Gabere MN, Noble WS (2017) Empirical comparison of web-based antimicrobial peptide prediction tools. *Bioinformatics* 33(13):1921–1929
 14. Gamas N, Smaoui F, Ben Romdhane W, Wiszniewska A, Baazaoui N, Bouteraa MT, Chouaibi Y, Ben Hsouna A, Kačániová M, Kluz MI, Garzoli S (2025) Genome-wide identification of the defensin gene family in *Triticum durum* and assessment of its response to environmental stresses. *Biology*, 14(4), p.404
 15. Games PD, daSilva EQG, Barbosa MO, Almeida-Souza HO, Fontes PP, deMagalhães-Jr MJ, Pereira PRG, Prates MV, Franco GR, Faria-Campos A, Campos SVA, Baracat-Pereira MC (2016) Computer aided identification of a hevein-like antimicrobial peptide of bell pepper leaves for biotechnological use. *BMC Genomics* 17(Suppl 12):999
 16. Gawde U, Chakraborty S, Waghu FH, Barai RS, Khanderkar A, Indraguru R, Shirsat T, Idicula-Thomas S (2023) CAMPR4: a database of natural and synthetic antimicrobial peptides. *Nucleic Acids Res* 51(D1):D377–D383
 17. Giustra M, Cerri F, Anadol Y, Salvioni L, Abella A, Prosperi T, Galli D, P. and, Colombo M (2022) Eco-luxury: making sustainable drugs and cosmetics with *Prosopis cineraria* natural extracts. *Frontiers in Sustainability*, 3, p.1047218
 18. Guilhelmelli F, Vilela N, Albuquerque P, Derengowski LDS, Silva-Pereira I, Kyaw CM (2013) Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. *Front Microbiol* 4:353
 19. Gupta A (2024) *Prosopis cineraria* pharmacological properties and health benefits: a review study of a potential herb of the Thar Desert. *Int J Pharmacognosy Life Sci* 5(2):91–103
 20. Hallgren J, Tsirigos KD, Pedersen MD, Almagro Armenteros JJ, Marcatali P, Nielsen H, Krogh A, Winther O (2022) DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. bioRxiv, pp.2022–04. 10.1101/2022.04.08.487609
 21. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA*, 87(6), pp.2264–2268
 22. Moravej H, Moravej Z, Yazdanparast M, Heiat M, Mirhosseini A, Moosazadeh Moghaddam M, Mirnejad R (2018) Antimicrobial peptides: features, action, and their resistance mechanisms in bacteria. *Microb Drug Resist* 24(6):747–767
 23. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, Han C, Bisignano C, Rao P, Wool E, Johnson SC (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399(10325):629–655
 24. Osorio D, Rondón-Villarreal P, Torres R (2015) Peptides: a package for data mining of antimicrobial peptides. *R J* 7(1):4–14
 25. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, Salazar GA, Bileschi ML, Llinares-López F, Meng-Papaxanthos L, Colwell LJ, Grishin NV, Schaeffer RD, Clementel D, Tosatto SCE, Sonnhammer E, Wood V, Bateman A (2025) The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res* 53(D1):D523–D534
 26. Sanchez-Jimenez O, Huerta-Rodriguez E, Rocha-Pizaña MDR, Luna-Vital DA, Mata-Gómez MA (2026) Bioactive properties of peptides obtained from the enzymatic hydrolysis of mesquite (*Prosopis laevigata*) cotyledon proteins. *Foods*, 15(8), p.1399
 27. Sharma M, Singh CP (2025) Functional genomics of *Prosopis cineraria* (L.) Druce: recent advances and new prospects. *J Plant Biochem Biotechnol* 34(1):102–118
 28. Solanki DS, Kumar S, Parihar K, Tak A, Gehlot P, Pathak R, Singh SK (2018) Characterization of a novel seed protein of *Prosopis cineraria* showing antifungal activity. *Int J Biol Macromol* 116:16–22
 29. Sudalaimuthuasari N, Ali R, Kottackal M, Rafi M, Al Nuaimi M, Kundu B, Al-Maskari RS, Wang X, Mishra AK, Balan J, Chaluvadi SR (2022) The genome of the mimosoid legume *Prosopis cineraria*, a desert tree. *International Journal of Molecular Sciences*, 23(15), p.8503
 30. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* 40(7):1023–1025

31. Wang G, Schmidt C, Li X, Wang Z (2026) APD6: the antimicrobial peptide database is expanded to promote research and development by deploying an unprecedented information pipeline. Nucleic Acids Res 54(D1):D363–D374
32. Wright E (2024) Accurately clustering biological sequences in linear time by relatedness sorting. Nat Commun 15(1):3047
33. Xu J, Li F, Leier A, Xiang D, Shen HH, Lago M, Li TT, Yu J, D.J. and, Song J (2021) Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. Brief Bioinform 22(5):083. .bbab
34. Yuan R, Zhang C, Gao T, Liu X, Wang X, Zhang F, Zhao K, Xia X, Lamlom SF, Ren H (2026) Genome-wide identification, characterization, and expression analysis of the plant defensin (PDF) gene family in soybean. BMC Plant Biol. <https://doi.org/10.1186/s12870-026-09177-3>
35. Zhang QY, Yan ZB, Meng YM, Hong XY, Shao G, Ma JJ, Cheng XR, Liu J, Kang J, Fu CY (2021) Antimicrobial peptides: mechanism of action, activity and clinical potential. Military Medical Research, 8(1), p.48

Figures

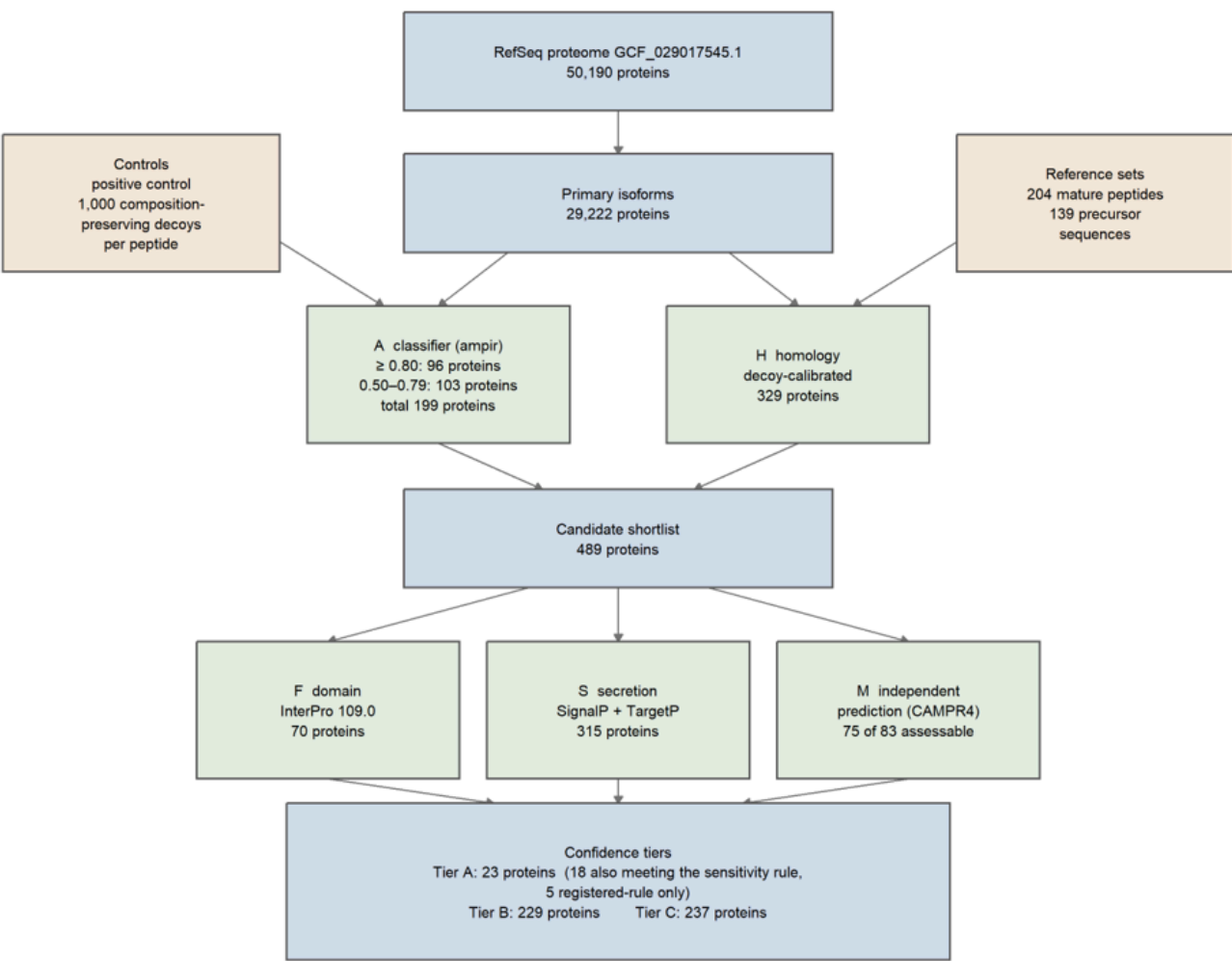


Figure 1

Analysis pipeline. The annotated *Prosopis cineraria* proteome was reduced to one primary isoform per gene. Two evidence axes were applied to all 29,222 primary proteins: **A**, a machine-learning classifier (ampir precursor model), calibrated against a positive control of curated reference precursors; and **H**, homology to curated reference peptides and precursors, calibrated against 1,000 composition-preserving shuffled decoys per reference peptide. The candidate shortlist is the union of proteins positive on either axis, taking the classifier at a probability of 0.50 or above ($199 + 329 - 39 \text{ shared} = 489$). Three further axes were applied to the shortlist alone: **F**, antimicrobial peptide family domain annotation; **S**, secretion signal, requiring agreement between SignalP and TargetP; and **M**, independent prediction, assessable only for the 83 candidates with a mature region inferable under the registered conditions. Arrows from F, S and M denote contribution of evidence to tier assignment and not sequential filtering: no candidate is removed from the shortlist by these axes. Tiers were assigned by an ordered decision rule fixed in advance. Counts shown are the number of proteins positive on each axis.

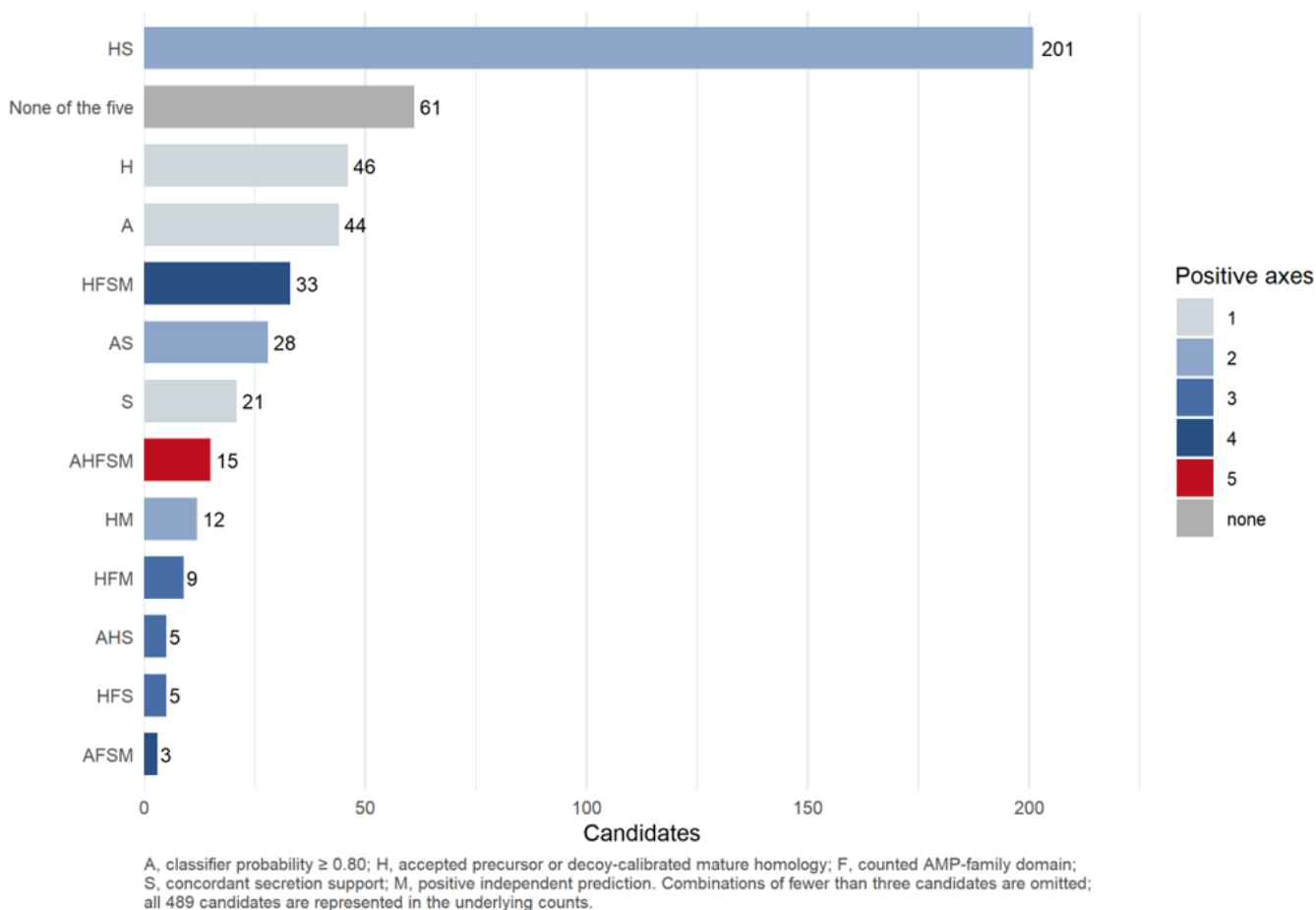


Figure 2

Intersections between the five evidence axes across the 489 shortlisted candidates.

Upper panel, the number of candidates in each intersection; lower left, the number positive on each axis; lower right, the intersection matrix, in which filled circles indicate the axes satisfied. Axes are A, classifier probability at least 0.80; H, calibrated homology; F, antimicrobial peptide family domain; S, secretion signal; M, independent prediction. The largest intersections are single axis rather than multi-axis: 290 candidates entered the shortlist on homology alone and 20 carried both a high classifier score and homology evidence, a Jaccard index of 0.049. Limited membership overlap does not establish statistical independence, as the classifier, the reference sets and the independent predictor all draw on overlapping public repositories.

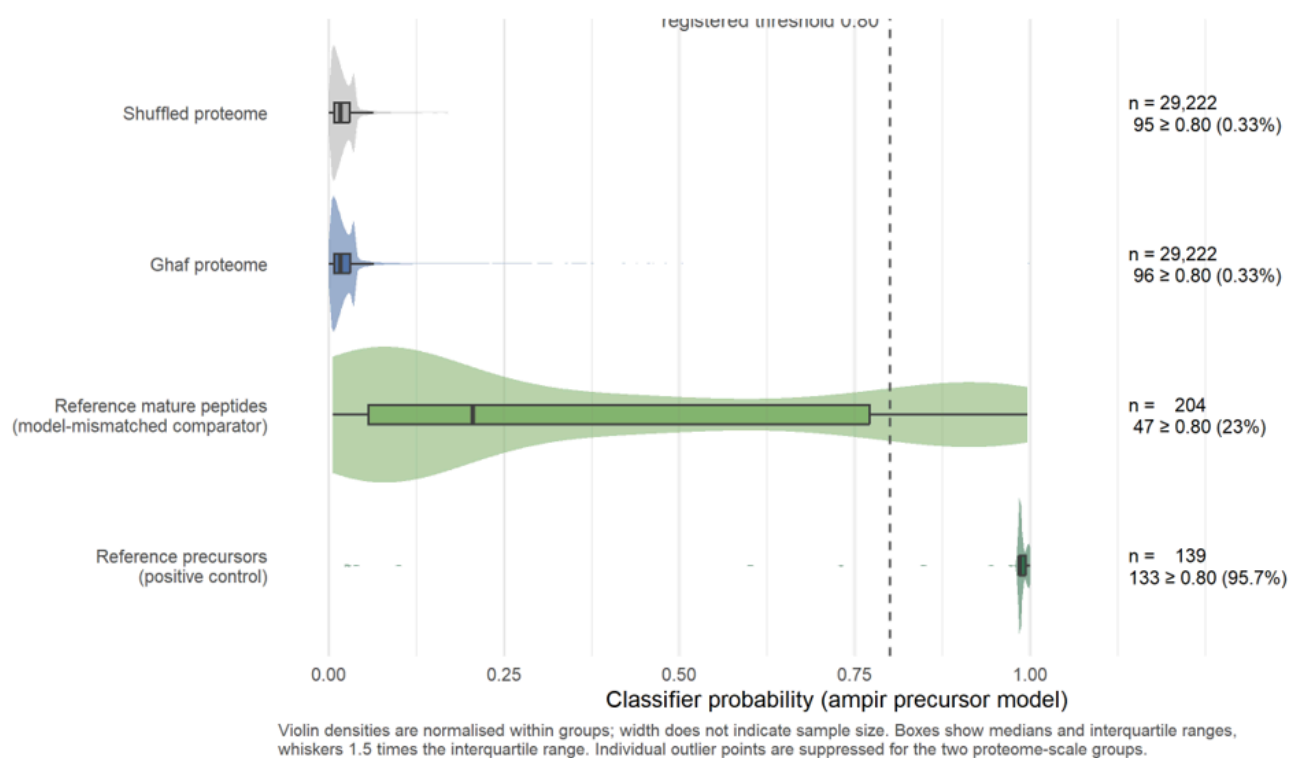


Figure 3

Classifier calibration against positive and null controls.

Distribution of ampir precursor-model probabilities for four sets of sequences: curated reference precursors ($n = 139$), curated reference mature peptides ($n = 204$), the Ghaf primary proteome ($n = 29,222$), and a composition-preserving shuffle of that proteome ($n = 29,222$). Violins show the full distribution and boxes the interquartile range with median. The dashed line marks the prespecified threshold of 0.80. Reference precursors were recovered at 95.7% against a proteome background of 0.3%, an enrichment of 291-fold. The shuffled proteome reproduces the observed distribution almost exactly, because shuffling preserves the compositional and physicochemical features on which the classifier's decision boundary rests while destroying only residue order; it therefore cannot serve as an informative negative control for this classifier. Reference mature peptides score lower than reference precursors because the precursor model is applied to them outside its intended input.



Figure 4

Evidence supporting the 18 Tier A candidates.

Rows are the 18 candidates assigned to Tier A under both the registered decision rule and the sensitivity analysis, ordered by classifier probability and labelled with accession and RefSeq annotation. Columns are the evidence axes, with the homology axis shown separately for its precursor and mature-peptide arms. Filled cells indicate positive evidence. All 18 candidates carry domain, secretion and independent-prediction evidence, and all satisfy the classifier threshold; 15 carry calibrated homology evidence, by either arm. Variation across the panel is confined to the two homology arms, several candidates being supported by precursor homology without a calibrated mature-peptide alignment, or the reverse.

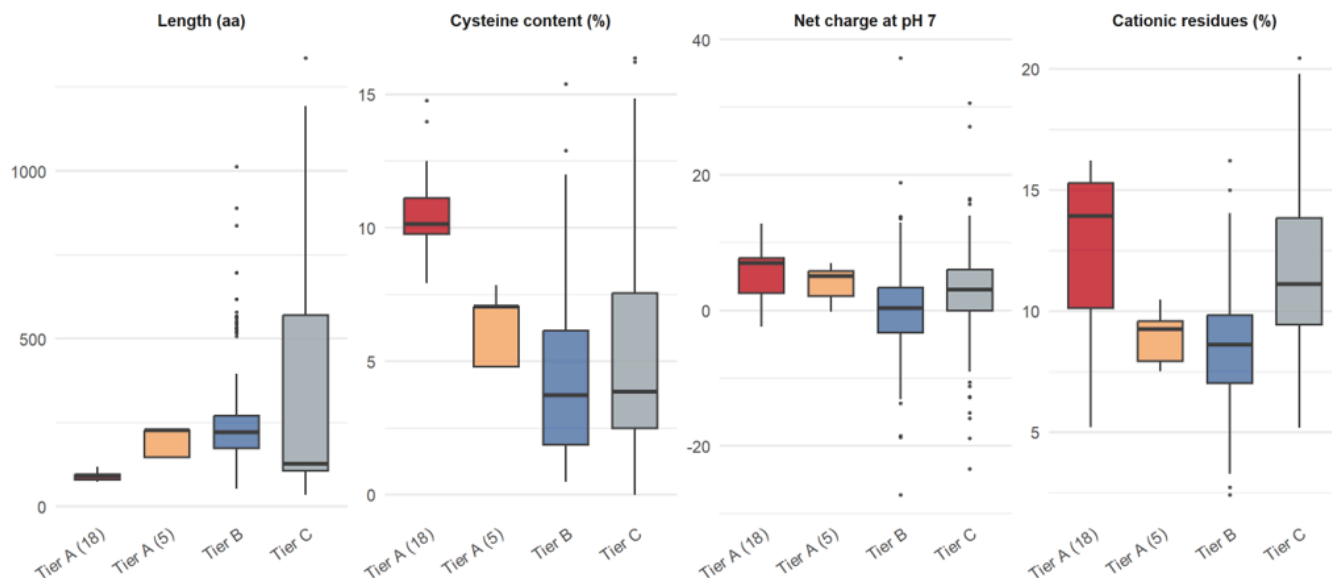


Figure 5

Precursor composition by confidence tier.

Distributions of length, cysteine content, net charge at pH 7 and cationic residue content, computed on the full precursor for all 489 shortlisted candidates. Tier A is shown split into the 18 candidates supported under both the registered rule and the sensitivity analysis and the 5 supported under the registered rule alone. Boxes show the interquartile range with median; whiskers extend to 1.5 times the interquartile range and points beyond are plotted individually. Cysteine content distinguishes the 18 most clearly: their interquartile range of 9.77 to 11.11% scarcely overlaps that of Tier B, 1.87 to 6.15%, and only three of 229 Tier B candidates fall within it. Net charge separates the groups far less clearly, the Tier C median of +3.06 approaching the Tier A median of +6.99. Composition is reported on the full precursor so that no tier is represented by a subset; mature-region values could be computed for only 59 of 229 Tier B and 6 of 237 Tier C candidates. These sequence characteristics took no part in tier assignment, but agreement between them and the tiers is internal coherence rather than independent validation, since the classifier axis is itself built on compositional features.