*Article*

# Computational Mapping of Maize NLR Sequence Space Reveals Conserved Architectures Supporting Receptor Engineering Prioritisation

**Anyela V Camargo Rodriguez**

Fiscrop, Cambridge TF10 8JZ, UK; fiscrop_qa@fiscrop.com.

## ABSTRACT

Plant nucleotide-binding leucine-rich repeat receptors (NLRs) are key components of plant innate immunity and important targets for crop disease resistance engineering. Here, we develop an integrated computational framework combining protein language model representations with sequence, structural, phylogenetic, motif, domain, and biochemical features to characterise maize NLR sequence space and prioritise candidate receptor architectures. The framework revealed structured patterns of conservation and diversification across maize NLR proteins, including conserved NB-ARC domains and highly variable leucine-rich repeat (LRR) regions. Integrated feature analysis identified complementary relationships between protein representations, domain organisation, structural similarity, and biochemical properties. To assess whether the scoring framework captured general principles of NLR compatibility, we benchmarked predictions against experimentally characterised receptor engineering systems, including Gpa2/Rx1 domain swaps, R13 LRR recombination constructs, and Pikp-1 HMA interface variants. These independent validation datasets showed that feature-based scoring prioritised experimentally compatible configurations and detected differences associated with recognition-module compatibility. This framework provides a quantitative approach for exploring maize NLR diversity and prioritising candidate NLR variants and engineered architectures for downstream experimental testing in plant immunity and crop improvement.

**KEYWORDS:** plant immunity; NLR receptors; protein language models; ESM2; maize; disease resistance; structural biology; computational biology; protein engineering



Received: 03 Jun 2026
Accepted: 25 Aug 2026
Published: 08 Sep 2026



## INTRODUCTION

Plant immune systems rely on multilayered defence mechanisms to recognize invading pathogens and activate responses that are often associated with localised host cell death. Among the most important intracellular immune receptors are nucleotide-binding leucine-rich repeat receptors (NLRs), which mediate effector-triggered immunity through recognition of pathogen-derived molecules [1,2]. NLR proteins are

widespread across plants and constitute a gene family exhibiting heterogeneous, multi-dimensional diversity across sequence, structural, and genomic scales [3–5].

Plant NLRs typically consist of three major regions: an N-terminal signalling domain, commonly coiled-coil (CC) or Toll/interleukin-1 receptor (TIR); a central nucleotide-binding adaptor shared by APAF-1, resistance proteins, and CED-4 (NB-ARC) domain; and a C-terminal leucine-rich repeat (LRR) region associated with recognition specificity . The NB-ARC domain functions as a molecular switch controlling receptor activation, whereas the LRR region is frequently associated with pathogen recognition and specificity determination [7]. Despite extensive sequence diversification driven by host–pathogen coevolution many NLRs retain conserved structural and mechanistic features, including motifs such as the P-loop, GLPL, and MHD motifs . Understanding how conserved signalling architectures coexist with diversification in recognition-associated regions remains a central question in plant immunity.

Some recent pangenomic analyses in *Arabidopsis Thaliana* have shown that NLR diversity extends across multiple axes, including sequence variation, copy number, and genomic organisation, highlighting that NLR evolution cannot be fully captured by sequence similarity alone and instead reflects a continuum of conserved and highly diversified states across populations [9]. Another study revealed that NLR diversity is organised into a limited number of rapidly evolving hvNLR clades with distinct genomic and regulatory features, motivating the need for representations that can unify and compare immune receptor diversity beyond clade-based classification [10].

In maize, NLR receptors such as Rxo1, Rp3-1, and Rp1-D have been extensively characterised and provide a useful system for studying the relationship between conserved signalling architecture and diversification in recognition specificity [11,12]. Previous studies have also demonstrated that domain exchange and recombination within NLR loci can generate novel resistance specificities and, in some cases, autoactive phenotypes [12] suggesting potential for rational exploration of NLR sequence space in engineering contexts.

More broadly, recent work on plant NLR evolution in pangenomic contexts has shown that NLR diversity is distributed across multiple genomic and structural dimensions, with variation arising not only from point mutations but also from structural rearrangements and locus-level complexity [4]. These observations motivate approaches that move beyond linear sequence comparison toward integrated representations of protein space.

Recent advances in computational biology, particularly protein language models trained on large sequence corpora, provide embeddings that encode biochemical, evolutionary, and structural constraints in a unified representation [13,14]. In parallel, structural prediction and comparison tools now enable large-scale integration of sequence- and

structure-derived information [15,16]. However, how these representations jointly reflect functional organisation in rapidly evolving immune receptor families, such as NLRs, remains incompletely understood.

Here, we present an integrated computational framework for analysing plant NLR proteins. Using maize NLRs as a model system, we investigate how signalling-associated constraints and recognition-associated variability are reflected across complementary computational representations of protein space. We further explore how embedding-based representations, and other metrics, relate to structural and domain-level organisation, providing a basis for systematic exploration of NLR diversity and likely expansion to other domains. With this work, we aimed to establish a unified computational framework for exploring NLR sequence space by integrating evolutionary, structural, biochemical, and protein representation features. Rather than relying on individual measures of similarity, this framework combines complementary signals to evaluate receptor conservation, perturbation, and compatibility, providing a systematic approach for prioritising candidate NLR variants and engineered architectures for future experimental analysis.

## MATERIALS AND METHODS

### Dataset Collection and Preprocessing

Protein sequences corresponding to maize NLR-associated proteins were downloaded in FASTA format from UniProt [17]. Sequences containing NBS-LRR and disease resistance annotations were retained for downstream analysis. Proteins from the CCoAOMT2 family were also added to the analysis, they physically interact and negatively regulate NLR immune receptors such as Rp1-D21, a candidate target in this analysis. Protein identifiers, sequence names, and annotations were extracted from FASTA headers using regular expression parsing (Table S1).

Out of the target cohort, three biologically relevant target proteins were selected for focused engineering-oriented analyses:

Rxo1 (UniProt: Q5BMB3), Rp3-1 (UniProt: Q6PW75) and Rp1-D-like protein (UniProt: Q9AT73).

These targets were selected based on their established association with disease resistance in maize and their representation within the analysed NLR landscape [8]. A fourth protein, Rcg1, was not included in the analysis for the lack as associated information.

### Protein Language Model Embeddings

Protein embeddings were generated using the ESM2 [15] transformer-based protein language model implemented through the Hugging Face Transformers framework [18]. Specifically, the model facebook/esm2_t6_8M_UR50D was used for sequence embedding generation.

Sequences exceeding 1022 amino acids were truncated to comply with model tokenisation constraints. Tokenised sequences were processed on GPU hardware when available. Embeddings were generated by averaging the final hidden-state representation across sequence positions:

$$E = \frac{1}{n}\sum_{i=1}^{n} h_i$$

where ($h_i$) represents the hidden state embedding vector at amino acid position ($i$), and ($n$) denotes sequence length.

Generated embeddings were subsequently used for clustering, similarity analysis, mutational perturbation scoring, and engineering-oriented prioritisation analyses.

### Clustering and Dimensionality Reduction

Embedding vectors were clustered using K-means clustering. Principal component analysis (PCA) was performed to reduce embedding dimensionality for visualisation purposes.

Embedding similarity between proteins was calculated using cosine similarity:

$$cosine\ similarity(A, B) = \frac{A \cdot B}{||A||\ ||B||}$$

where ($A$) and ($B$) represent embedding vectors.

### HMMER Domain Annotation

Protein domain annotation was performed using HMMER [19] hmmscan searches against the Pfam-A hidden Markov model database [20]. Conserved NLR-associated domains, including NB-ARC and multiple LRR subclasses, were identified using domain-specific hits.

Parsed HMMER results included: domain identifiers, protein identifiers, E-values, and domain scores.

Domain-level analyses were subsequently used for receptor decomposition, engineering-oriented compatibility analyses, and chimera generation.

### Motif Discovery

Canonical NLR-associated motifs were identified using regular expression-based scanning. Analysed motifs included: P-loop motif, kinase-2 motif, GLPL motif, and MHD motif.

Motif presence or absence was recorded for each protein sequence.

### Phylogenetic Analysis

Multiple sequence alignment was performed using MAFFT with automatic parameter selection [21]. Phylogenetic tree inference was conducted using FastTree [22].

Phylogenetic analyses were used to evaluate evolutionary relationships among maize NLR proteins and contextualise the three target resistance proteins within the broader receptor landscape.

### Structural Comparison Analyses

Structural similarity analyses were conducted using Foldseek [23]. Predicted protein structures in PDB format were compiled into Foldseek databases and subjected to all-versus-all structural comparison. Reported structural metrics included: alignment TM-score, local distance difference test (lDDT), probability score, and sequence identity. Pairwise structural similarity matrices were generated and compared with embedding-derived similarity matrices.

### Embedding–Structure Correlation Analysis

To evaluate relationships between embedding similarity and structural conservation, pairwise cosine similarities between embeddings were compared with Foldseek-derived structural similarity scores. Pearson correlation coefficients were calculated between embedding similarity vectors, and structural similarity vectors.

This analysis was designed to assess whether protein language model embeddings capture biologically meaningful structural information among maize NLR proteins.

### Negative Control Analyses

Two classes of negative controls were generated. First, random protein sequences matched to the average length distribution of target NLR proteins were generated using uniform amino acid sampling. Second, shuffled-sequence controls were generated by randomly permuting amino acid positions within validated NLR sequences while preserving amino acid composition. Embedding similarity between validated NLR proteins and control sequences was subsequently evaluated. In addition to these artificial controls, biologically relevant negative controls were included using four maize LRR-containing proteins lacking canonical NLR NB-ARC architecture (A0A1D6F5N5, Q940E8, P0DL10, Q5GAP9). These LRR proteins represent immune-associated proteins with partial functional

similarity to NLRs but distinct receptor mechanisms. Protein language model embeddings were generated for all sequences, and cosine similarity, PCA projection, and K-means clustering were performed to evaluate whether NLRs could be distinguished from sequence-randomized and biologically related non-NLR proteins.

### Domain Extraction and Comparison

Proteins were decomposed into approximate CC, NB-ARC, and LRR regions using domain annotations and heuristic coordinate estimation. Separate embeddings were generated for each extracted domain. Domain-level cosine similarity scores were subsequently calculated between target proteins.

### Chimera Generation and Compatibility Prioritisation

Engineering-oriented chimera candidates were generated across maize NLR proteins using domain recombination strategies informed by embedding- and domain-level similarity. Experimentally characterised receptors (Rxo1, Rp3-1, and Rp1-D) were subsequently highlighted within the broader compatibility landscape.

All combinatorial domain configurations were evaluated computationally. Candidate protein pairs were ranked according to an embedding-based compatibility score integrating global embedding similarity, domain-level similarity, and signalling-domain conservation.

Compatibility scoring prioritised candidate combinations predicted to preserve signalling-associated structural constraints.

### LRR Mutational Perturbation Analysis

Systematic LRR perturbation analyses were performed across maize NLR proteins to evaluate mutational tolerance within embedding space. Experimentally characterised receptors were used as representative case studies for detailed interpretation. Mutated sequences were embedded using ESM2, and perturbation scores were calculated based on cosine similarity relative to original sequence embeddings. Mutations preserving high embedding similarity were interpreted as potentially tolerated perturbations within learned sequence constraints.

### Localisation-Associated Features

Simple sequence-derived localisation-associated features were evaluated, including nuclear localisation signal-like motifs, and transmembrane-like hydrophobic enrichment.

These analyses were exploratory and intended to provide additional functional context.

### Statistical Analysis

Pearson correlation analyses were performed using to measure similarity. Statistical significance set to $p < 0.05$. All analyses were implemented in Python using: PyTorch, Transformers, NumPy, pandas, scikit-learn, Biopython, SciPy, Matplotlib, FoldSeek.

### Control Validation

To evaluate whether the chimera prioritisation framework captures biologically meaningful constraints in NLR domain we benchmarked it against experimentally characterised NLR engineering systems. We used the Gpa2/Rx1 domain swap constructs (Gpa2CN/Rx1L and Rx1CN/Gpa2L) and the systematic R13 leucine-rich repeat (LRR) recombination series described in Slootweg et al. (2013) [24], which provide experimentally validated examples of functional and non-functional domain configurations in potato. Figure 1A and 3A of Slootweg et al. (2013) [24] was used to guide the engineering of all chimeras.

Full-length GPA2 and RX1 sequences were segmented into canonical NLR domains (CC, NB-ARC, and LRR) based on reported boundary definitions. Candidate chimeras and recombination variants were evaluated using the same region-aware compatibility scoring function as in the main framework, without retraining or parameter adjustment on experimental labels.

Benchmark performance was assessed by verifying (i) correct relative ordering of functional versus non-functional Gpa2/Rx1 swap constructs and (ii) structured response of scores across the graded R13 recombination series, reflecting sensitivity to progressive LRR perturbation

### Structural Case Study: Pikp–OsHIPP19 HMA Engineering System

To assess whether the framework can resolve biologically meaningful changes occurring at protein–protein interaction interfaces, we analysed the experimentally characterised Pikp–OsHIPP19 HMA engineering system [25] in rice. Unlike domain-scale chimera validation using Gpa2/Rx1, this system provides a test of local interface perturbations through both domain replacement and single-residue engineering.

Full-length Pikp-1 (UniProt accession E9KPB5) and OsHIPP19 (UniProt accession Q01IL6) sequences were retrieved from UniProt. Experimentally defined HMA domain boundaries were used to extract the relevant regions (Pikp-HMA: residues 186–263; OsHIPP19-HMA: residues 2–71) (residuesGly186-Asp264) [26]. The S258E substitution was selected because the equivalent glutamate residue in OsHIPP19 HMA contributes to AVR-Pik effector recognition, providing a structurally informed gain-of-function engineering perturbation.

A domain replacement construct was generated by substituting the native Pikp-HMA region with the OsHIPP19-HMA domain. In addition, the experimentally characterised Pikp-HMA Ser258Glu (S258E) gain-of-function mutation was introduced as a single amino-acid perturbation. The S258E substitution was selected because the equivalent glutamate residue in OsHIPP19 HMA contributes to AVR-Pik effector recognition, providing a structurally informed gain-of-function engineering perturbation.

The wild-type Pikp-HMA/AVR-PikC complex structure (PDB: 7QPX) was used as structural reference. Protein language model embedding similarity was calculated between wild-type Pikp and engineered variants to assess global sequence-context conservation. To evaluate local functional compatibility, structure-guided interface analysis was performed by identifying HMA residues within 5 Å of AVR-PikC residues. Interface conservation was calculated by comparing the preservation of experimentally defined binding residues in engineered variants.

For the S258E variant, residue-level interaction analysis was performed to identify potential changes introduced at the effector-binding interface. Atomic distances between the engineered Glu258 side chain and AVR-PikC interface residues were calculated, with distances compatible with hydrogen-bond formation considered indicative of potential stabilising interactions.

## RESULTS

### Embedding Analyses Reveal a Globally Conserved NLR Landscape

Protein language model embeddings (ESM2) revealed strong clustering among maize NLR proteins in a low-dimensional embedding space. PCA analysis (Figure 1) shows coherent grouping of disease resistance receptors, with target proteins (Rxo1, Rp3-1, Rp1-D) embedded within dense NLR clusters. K-means clustering confirmed structured partitioning of the embedding space. For example, cluster two, the furthest apart from the rest, represents CCoAOMT2 proteins. They primarily function in lignin biosynthesis and plant defence, and play a vital role in reinforcing cell walls and regulating immune responses [27] (Table S2).

Benchmarking analysis further supported this structure, with high intra-cluster similarity (0.950) and comparatively lower inter-cluster similarity (0.871), indicating a globally conserved but structured representation of NLR diversity. Two proteins closer to our target proteins were LOC606485 and MRPR1, both NBS-LRR type disease resistance proteins.

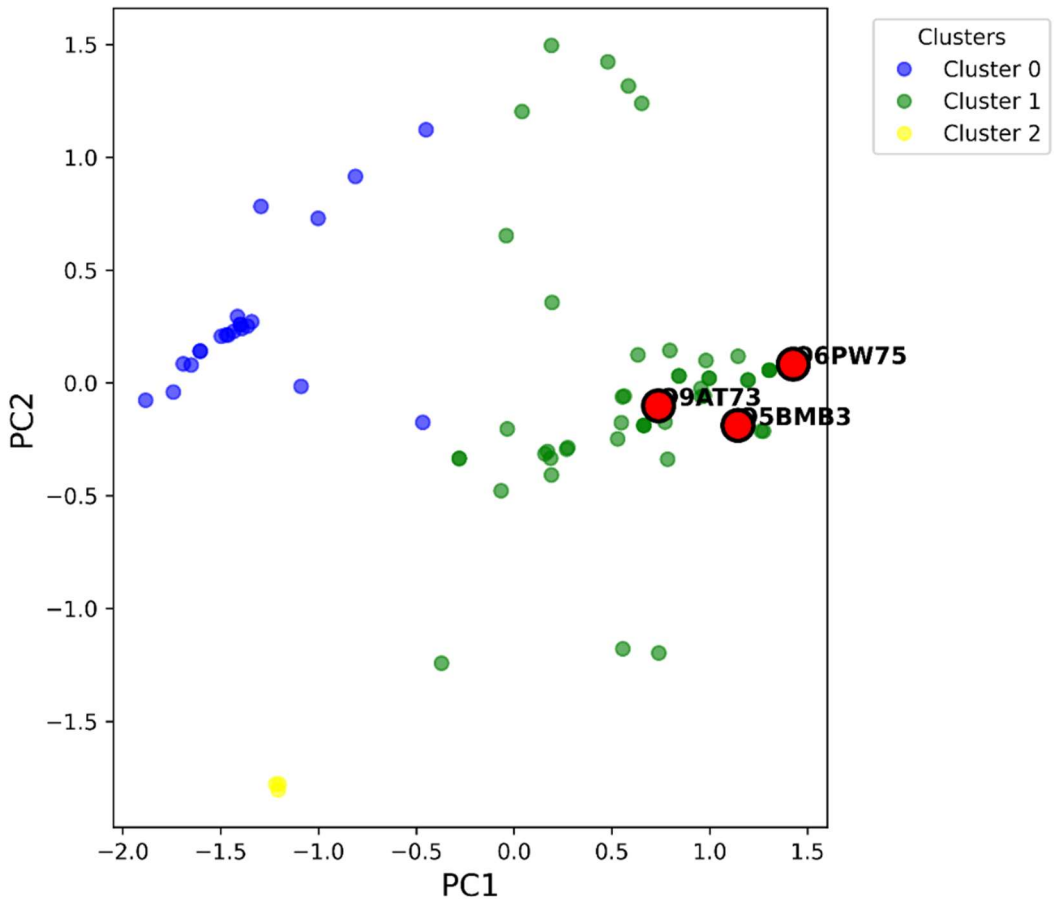


**Figure 1.** Global maize NLR embedding landscape. PCA projection of ESM2 embeddings across all proteins. Proteins are coloured by cluster assignment. Proteins highlighted in red are Rp3-1(Q6PW75), Rxo1(Q5BMB3), Rp1-D (Q9AT73).

### Maize NLR Proteins Exhibit Conserved Domain Architectures

HMMER-based annotation identified widespread conservation of canonical NLR-associated domains across maize proteins, with a total of 1049 significant hits across the dataset. Conserved NB-ARC domains were consistently detected with low E-values, indicating strong preservation of nucleotide-binding signalling machinery.

Multiple leucine-rich repeat (LRR) subclasses, including LRR_4, LRR_8, LRR_14, and LRR_RPS2, were detected across proteins annotated as disease resistance analogues, RPP13-like proteins, RPS2-associated proteins, and RGA4-associated proteins.

The coexistence of a highly conserved NB-ARC core with diversified LRR architecture suggests a modular receptor design in which signalling domains remain constrained while recognition-associated regions exhibited comparatively reduced conservation (Figure 2).

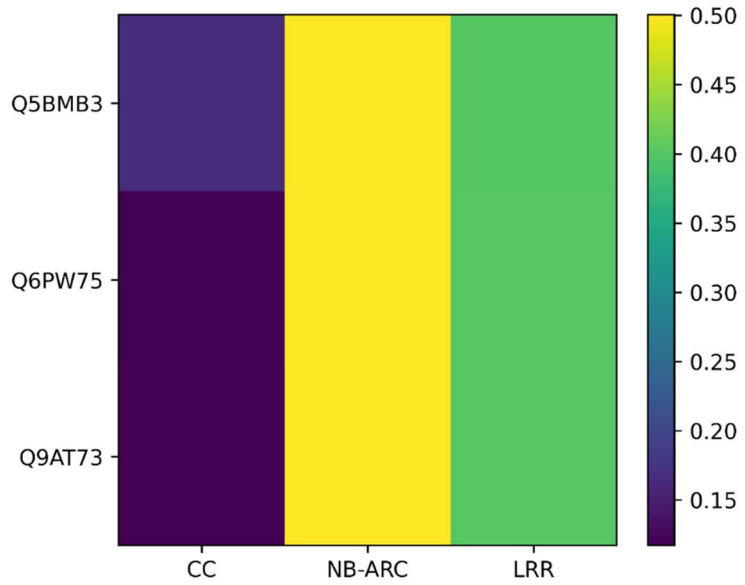


**Figure 2.** Domain architectures and domain compositions.

### Conserved NLR-associated Motifs Are Retained across Maize Receptors

Motif analysis revealed consistent preservation of canonical NLR signatures across proteins. P-loop motifs were universally detected, supporting conserved ATP-binding functionality. GLPL motifs were present in multiple receptors, whereas MHD motifs showed partial conservation. Kinase-2 motifs were absent or weakly detected in several proteins.

### Phylogenetic Analysis

Maximum-likelihood phylogenetic reconstruction identified multiple distinct clades within the maize NLR repertoire, consistent with substantial sequence diversification across the family. Experimentally characterised receptors (Rxo1, Rp3-1 and Rp1-D) were distributed within larger groups of related NLRs rather than forming isolated branches, indicating that these receptors are embedded within broader evolutionary lineages. Branch length variation across the tree further suggests heterogeneous rates of diversification among maize NLRs, consistent with the dynamic evolutionary history reported for plant immune receptors (Figure 3).

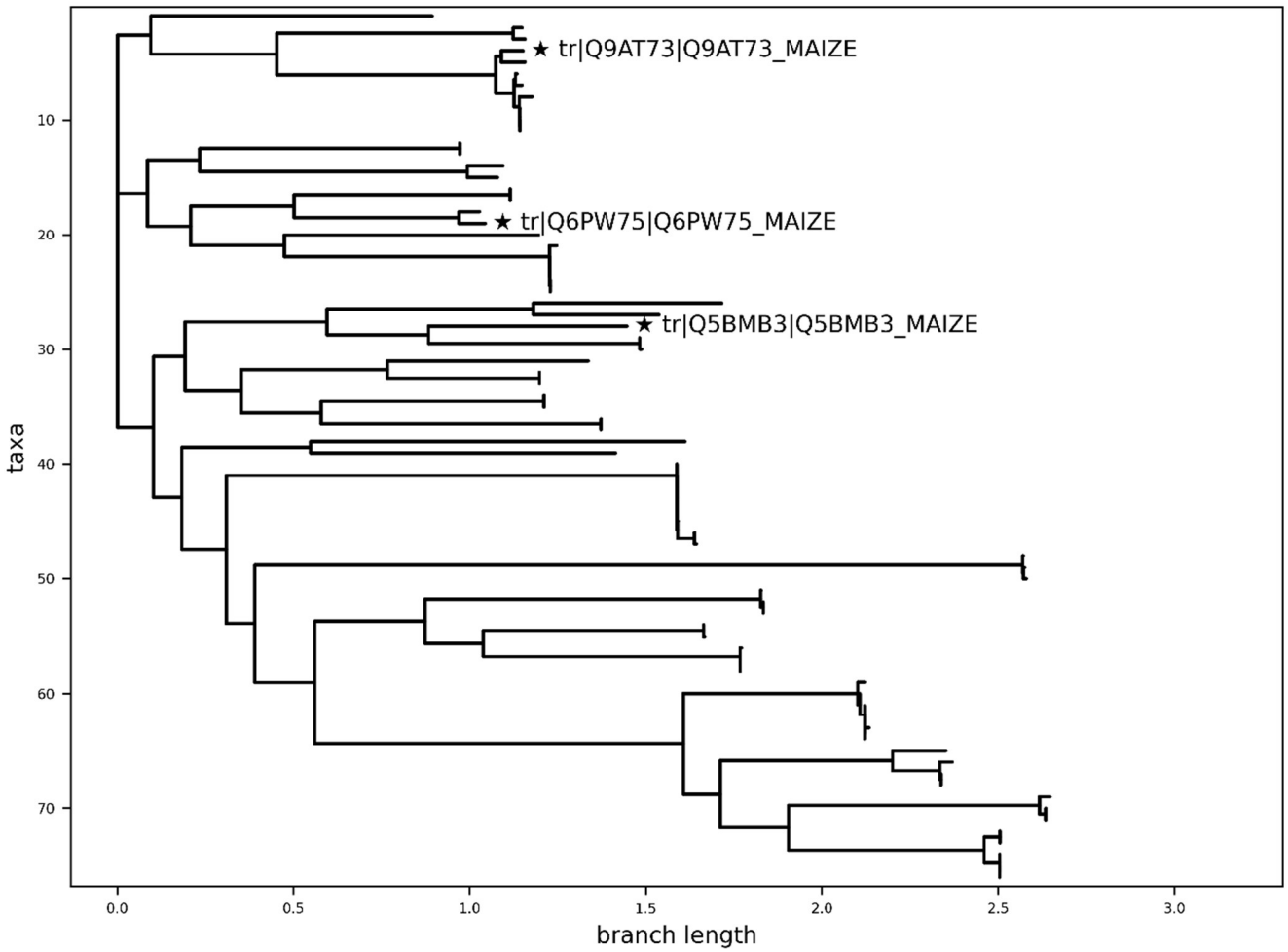


**Figure 3.** Maximum-likelihood phylogeny of maize NLR proteins inferred from MAFFT [21] alignments using FastTree [22]. Experimentally characterised receptors (Rxo1, Rp3-1, and Rp1-D) are highlighted.

### Structural Comparisons Reveal Partial Correspondence between Embeddings and Structural Conservation

Foldseek-based structural comparisons revealed measurable but moderate conservation across maize NLR proteins, with pairwise similarity matrices showing consistent structural relationships among receptors. A statistically significant but weak correlation was observed between embedding-derived similarity and structural similarity (Pearson $r = 0.256$, $p = 5.69 \times 10^{-44}$), suggesting that protein language model embeddings capture limited but detectable structural signals relevant to NLR organisation (Figure 4).

However, embedding similarity values were generally higher than structural similarity scores, suggesting that embeddings encode additional evolutionary and functional constraints beyond direct structural alignment.

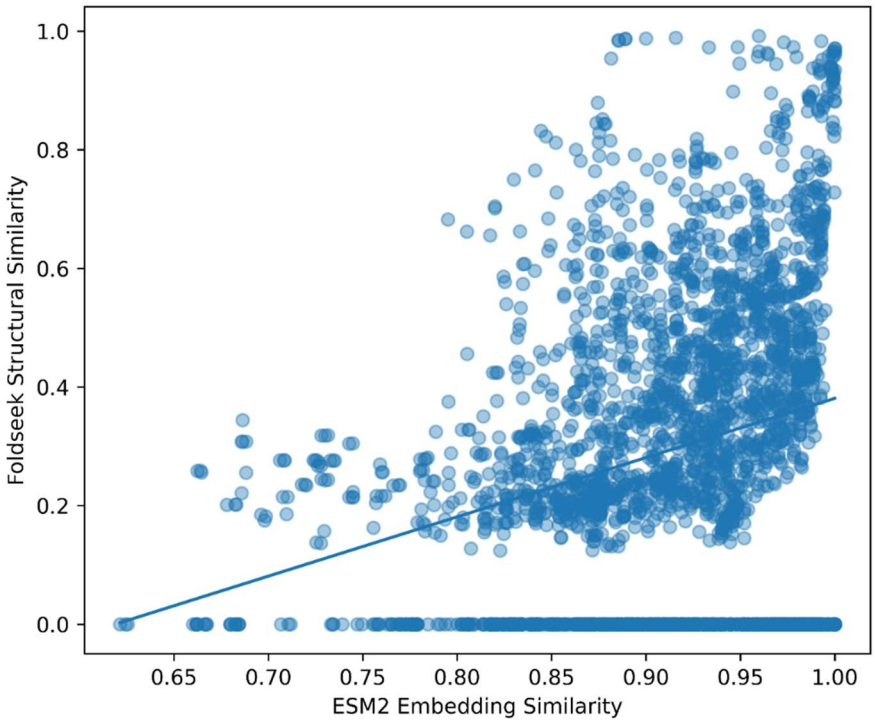


**Figure 4.** Relationship between embedding similarity and structural similarity.

### Negative Controls Support Biologically Meaningful Embedding Organisation

Random protein sequences matched to NLR length distributions showed reduced embedding similarity relative to authentic maize NLR proteins. Shuffled-sequence controls retained partial similarity but disrupted native sequence organisation, indicating sensitivity of embeddings to residue ordering. Embedding similarity values between validated NLR proteins and random controls ranged approximately from 0.58 to 0.70, significantly lower than similarities observed among true NLR proteins.

The biological negative control analysis showed separation between validated NLRs and LRR-containing non-NLR proteins in the embedding space. PCA projection followed by K-means clustering (k = 3) grouped all three validated NLR targets (Q5BMB3, Q6PW75, and Q9AT73) into the same cluster, whereas three of the four LRR-containing proteins (A0A1D6F5N5, Q940E8, and P0DL10) were assigned to a separate cluster. Q5GAP9 represented an exception, forming a third cluster despite being classified as an LRR-containing protein. This protein also showed the highest cosine similarity to NLRs (up to 0.85) and has not yet been reviewed by UniProt, suggesting that incomplete annotation or atypical sequence features may contribute to its distinct embedding position.

### Domain-Level Analyses Reveal Strong Signalling Conservation

Domain-resolved similarity analyses across target proteins (Rxo1, Rp3-1 and Rp1-D) revealed strong conservation across all major NLR-associated domains (Figure 5). CC domain similarities ranged from 0.962–0.978, NB-ARC similarities from 0.935–0.975, and LRR similarities from 0.964–0.982. The consistently high similarity values indicate extensive conservation throughout both signalling-associated and recognition-associated regions. While minor differences were detectable between receptors, no domain class exhibited markedly reduced conservation relative to the others.

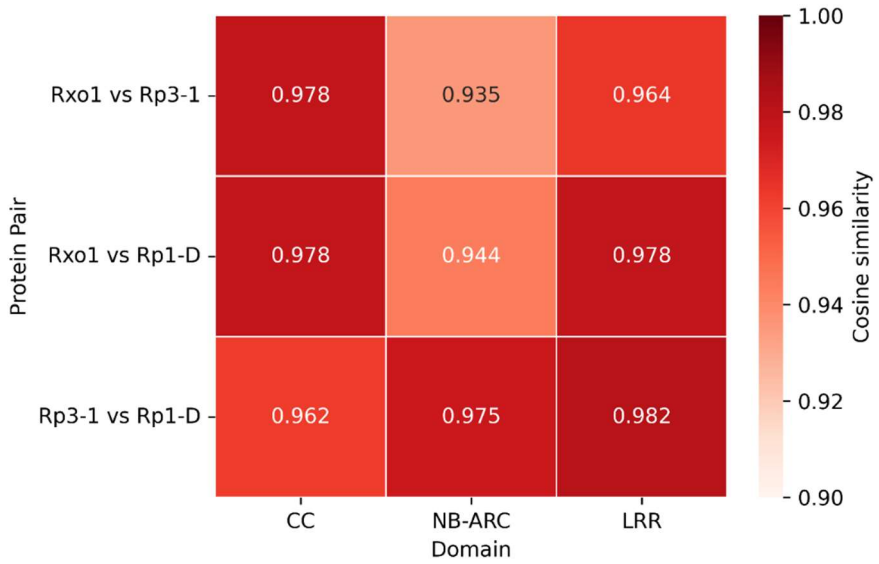


**Figure 5.** Domain-level conservation among target proteins. Heatmap showing: CC similarity, NB-ARC similarity, LRR similarity. Across: Rxo1, Rp3-1, Rp1-D.

### Engineering-Oriented Analyses Reveal Structured Compatibility and Perturbation Landscapes across Maize NLRs

Embedding-based compatibility scoring across maize NLR proteins identified a structured distribution of candidate receptor combinations, rather than uniformly interchangeable architectures. Experimentally characterised receptors, including Rxo1 (Q5BMB3), Rp3-1 (Q6PW75), and Rp1-D (Q9AT73), occupied distinct positions within this compatibility landscape, supporting their use as biologically grounded reference points for comparative analysis. Thought the Rxo1–Rp3-1 pair exhibited a compatibility score of approximately 0.54 and generated multiple high-scoring chimera configuration, other pairs achieve compatibilities > 0.6 (Figure 6A).

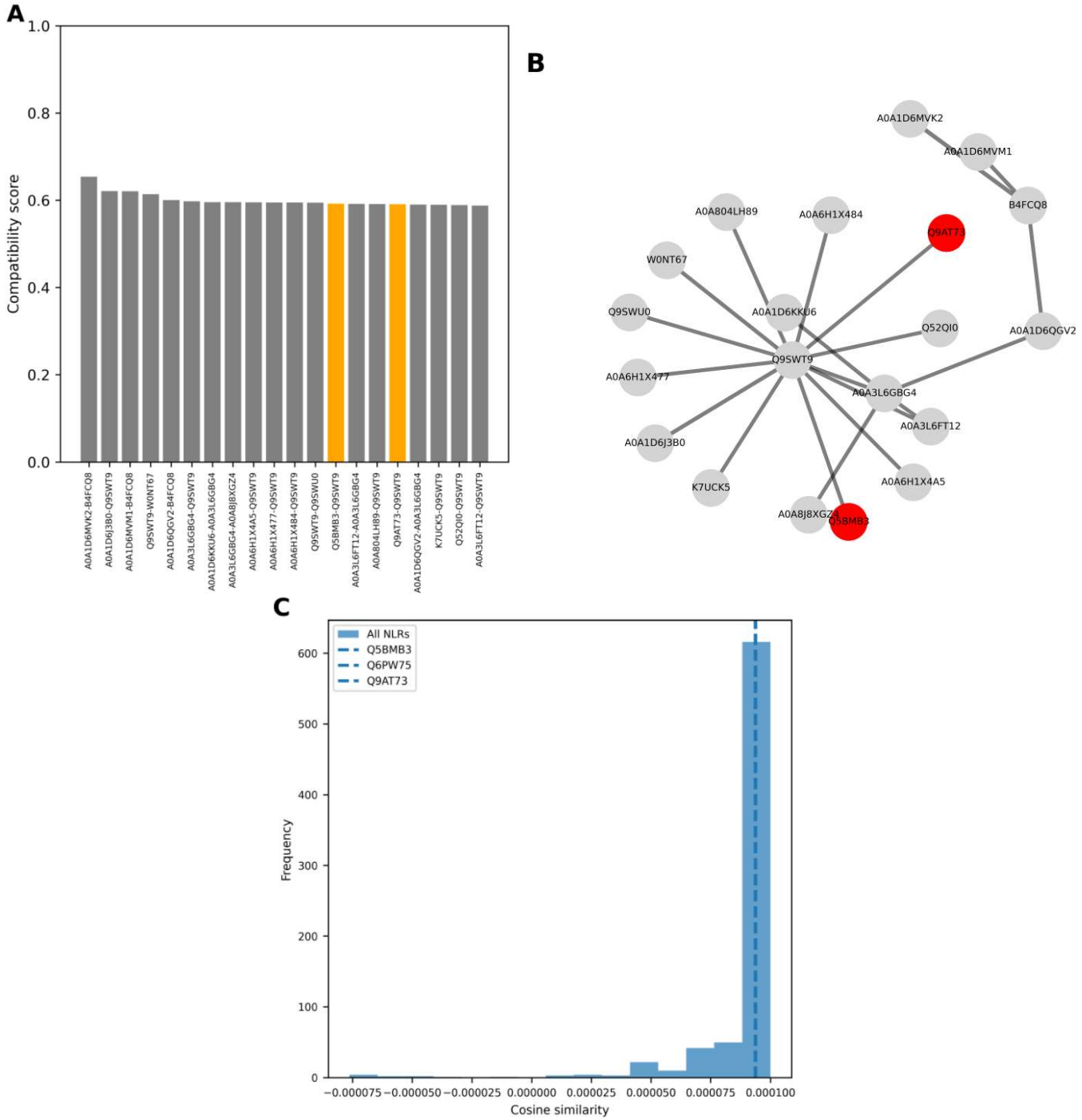


**Figure 6.** Computational prioritisation landscape across maize NLR proteins. (**A**) Ranked chimera compatibility scores across candidate NLR pairs. Experimentally characterised proteins are highlighted in yellow within the global compatibility landscape. (**B**) Chimera compatibility network analysis revealed a structured interaction landscape among maize NLR receptors. Target proteins are highlighted in red. (**C**) Distribution of LRR mutational perturbation effects across maize NLR proteins, showing high retention of embedding similarity following amino acid substitutions, with target receptors highlighted within the global perturbation landscape. Target proteins represented with dashed line, they overlapped due to their embedding similarity (Cosine similarity).

Chimera compatibility network analysis revealed a structured interaction landscape among maize NLR receptors (Figure 6B). Nodes represent individual NLR proteins, and weighted edges represent predicted compatibility scores for domain-respecting chimera construction. The resulting network exhibits non-uniform connectivity, with experimentally characterised receptors (Rxo1, Rp3-1, and Rp1-D) embedded within a broader compatibility space rather than forming isolated clusters. Edge-weight variation reflects heterogeneous recombination potential across receptor pairs, indicating structured constraints in NLR architectural compatibility.

Population-scale LRR mutational analyses revealed broadly high tolerance to amino acid substitutions, with most mutations preserving near-identical embedding similarity relative to wild-type sequences (Figure 6C). Mutation similarity distributions were tightly clustered near 1.0, and target proteins exhibited narrow perturbation profiles, suggesting locally robust regions within recognition-associated domains. These results highlight candidate regions for future experimental exploration of NLR robustness and evolvability.

**Experimental Consistency and R13 Perturbation Landscape**

The chimera prioritisation framework reproduced the experimentally observed compatibility difference in the Gpa2/Rx1 domain-swap system, assigning a higher score to the functional Gpa2CN/Rx1L construct than to the non-functional Rx1CN/Gpa2L construct (functional: 0.00161; non-functional: −0.00154; $\Delta$ = 0.00314). This demonstrates that the scoring function captures sequence and domain-level compatibility differences associated with experimentally characterised functional outcomes.

Application to the R13 recombination series revealed a structured compatibility landscape across altered LRR configurations. Relative embedding similarity scores varied within a bounded range (−0.00181 to 0.00121), demonstrating that different LRR recombination patterns produce distinct compatibility states rather than equivalent perturbations. The observed non-monotonic response is consistent with the context-dependent nature of LRR-mediated recognition, where functional compatibility depends on specific domain configurations rather than simple additive sequence changes. These results indicate that the framework can detect experimentally relevant differences in engineered NLR architectures and capture graded perturbational effects within recognition-associated regions (Figure 7).

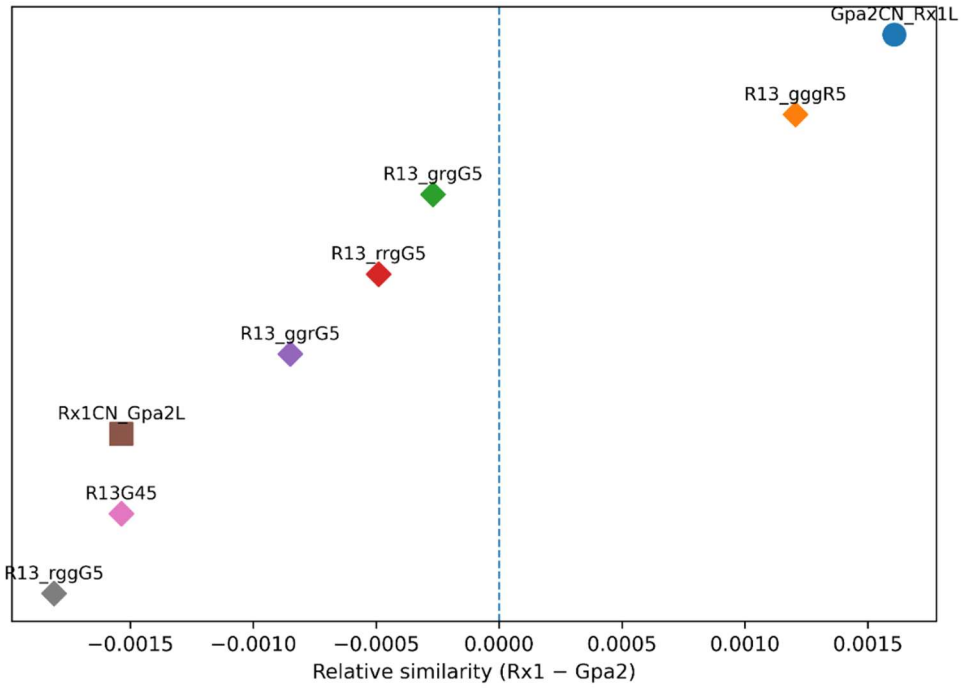


**Figure 7.** Validation against experimentally characterised Gpa2/Rx1 chimeras. Relative similarity correctly ranked the published functional chimera above the non-functional chimera, with the R13 recombination series occupying intermediate positions between the two reference constructs.

### Experimental Consistency S258E Mutation Preserves Global Architecture While Introducing Interface-Level Compatibility Changes

Protein language model embeddings indicated that the experimentally characterized S258E mutation caused negligible disruption to the overall Pikp-1 sequence representation. The WT Pikp-1 and S258E variant showed near-identical global similarity (0.99999), indicating that the single residue substitution introduced minimal perturbation within the conserved HMA framework (Figure 8A).

Structure-guided analysis of the Pikp-HMA^SNK-EKE/AVR-PikC complex (PDB: 7QPX) identified 11 HMA interface residues and 6 AVR-PikC interface residues. The S258E variant retained complete conservation of the experimentally defined interface region, producing an interface compatibility score of 1.0, equivalent to WT Pikp-1.

Although the overall interface architecture was maintained, residue-level contact analysis revealed that Glu258 introduced additional potential short-range interactions with AVR-PikC. In the S258E model, Glu258 formed multiple predicted heavy-atom contacts with AVR-PikC residues, predominantly involving residues 69–76, 204, and 217, with several contacts occurring below 5 Å. These interactions are consistent with a localized extension of the HMA–effector interaction network while preserving the native interface framework. Biochemical feature analysis showed that S258E produced only a minor change relative to WT Pikp-1

(Feature_Distance = 0.038), consistent with a conservative substitution that preserves overall HMA characteristics.

In contrast, OsHIPP19 HMA replacement produced a substantially larger biochemical shift (Feature_Distance = 3.139), consistent with broader perturbation of local HMA properties. Although global embedding similarity remained high (0.9992), the domain swap reduced Pikp-1 interface conservation (0.09) and altered predicted local compatibility, reflecting a larger sequence-level and interface-level perturbation than the single-site S258E modification (Figure 8B).

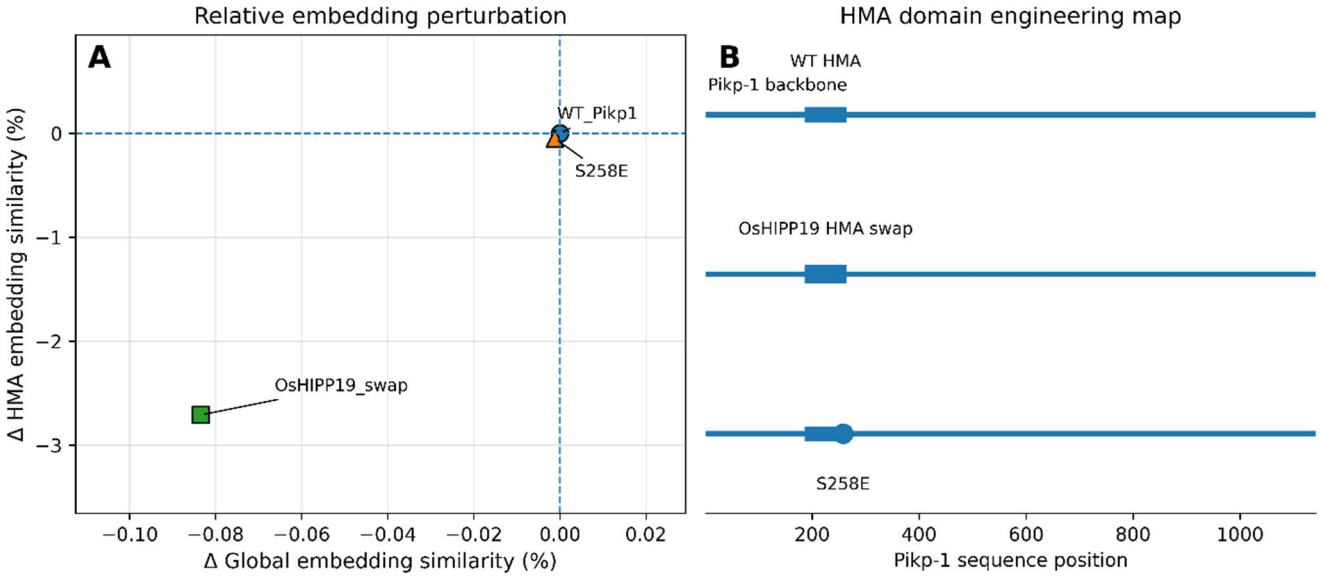


**Figure 8.** (**A**) Relative changes in protein language model embedding similarity compared with WT Pikp-1, showing the impact of HMA swap and S258E mutation. (**B**) Schematic representation of HMA domain engineering, highlighting the OsHIPP19 HMA replacement and the S258E point mutation.

## DISCUSSION

This study presents an integrated computational framework combining protein language model embeddings, structural analyses, domain annotation, phylogenetic inference, and engineering-oriented perturbation analyses to investigate conservation and diversification across maize NLR receptors.

Our findings support a model in which maize NLR proteins maintain highly conserved signalling-associated architectures while permitting limited diversification within recognition-associated regions. Across the analysed dataset, NB-ARC and CC domains exhibited particularly strong conservation, consistent with previous observations that core immune signalling machinery remains evolutionarily constrained in plant intracellular receptors [25].

In contrast, LRR-associated regions displayed increased variability, consistent with their role in pathogen recognition specificity and adaptive diversification. Although overall LRR similarity remained relatively high

among the selected maize receptors, mutational analyses identified positions predicted to tolerate amino acid substitutions without strongly perturbing embedding-derived sequence organisation.

A statistically significant relationship was observed between protein language model embedding similarity and structural conservation. Although the correlation between ESM2 embeddings and Foldseek (Alphafold) derived structural similarity was moderate rather than strong, this suggests that embedding space captures partial structural constraints while also reflecting divergence driven by rapid domain level evolution in NLR proteins [4]. In fact, protein language models encode a combination of structural, evolutionary, and sequence-derived features beyond merely reproducing structural similarity. These findings are consistent with previous work showing that protein language models learn biologically meaningful representations beyond sequence similarity alone [16]. To further evaluate whether the embedding space captured NLR-specific biological features, we incorporated biologically relevant negative controls consisting of maize LRR-containing proteins lacking canonical NB-ARC architecture. Unlike random or shuffled sequence controls, these proteins represent a stringent comparison because they share immune-related functions and LRR-associated features with NLRs while belonging to distinct receptor classes. PCA projection and K-means clustering grouped the validated NLRs separately from the majority of LRR controls, supporting the ability of the embeddings to distinguish canonical NLR architectures from related immune proteins. Interestingly, the LRR protein, Q5GAP9, showed a divergent embedding position and relatively high similarity to NLRs; however, this protein remains unreviewed by UniProt, highlighting the potential influence of incomplete annotation or atypical sequence features.

Validation using experimentally characterised NLR engineering systems demonstrated that the framework captures biologically relevant compatibility constraints. The Gpa2/Rx1 domain-swap constructs and R13 LRR recombination series were correctly ranked according to experimentally observed functional behaviour, indicating that the scoring framework can detect both large-scale domain incompatibility and graded perturbations within recognition-associated regions. Furthermore, the Pikp S258E system provided an independent test of fine-scale interface perturbation. The experimentally validated S258E mutation produced minimal changes in global embedding space while introducing predicted interface-level modifications consistent with enhanced AVR-PikC recognition. Together, these validation systems demonstrate that the framework can capture both domain-level compatibility and residue-level structural changes relevant to NLR engineering.

The computational framework reported here differs from conventional sequence or structure only analyses as it integrates complementary analytical layers into a unified workflow for ranking engineering candidates. Rather than predicting functional resistance directly, it

prioritises receptor variants and chimeric combinations that preserve conserved signalling architectures while enabling controlled exploration of recognition-associated variation.

Also, the results presented here are computational and should not be interpreted as experimentally validated predictions of receptor activity. Chimera compatibility scores and mutational perturbation analyses represent prioritisation tools rather than direct measures of immune function, and experimental validation is required to confirm signalling competence and resistance phenotypes in vivo. Case studies of Rxo1, Rp3-1, and Rp1-D illustrate the utility of the framework. These receptors exhibit strong conservation in signalling associated domains while maintaining detectable diversification in recognition regions, making them informative models for exploring engineering feasibility in maize resistance proteins.

Despite the limitations of this study, we demonstrate that integrated embedding, structure, and domain level analyses can provide a means of narrowing experimental search space by ranking and contextualising candidate receptor variants prior to experimental screening, rather than directly predicting resistance outcomes.

## CONCLUSIONS

This study integrated protein language model representations with sequence, phylogenetic, structural, domain-level, interface, and perturbation analyses to develop a computational framework for evaluating engineering-relevant properties of plant immune receptors. Analyses of Rxo1, Rp3-1, and Rp1-D showed that maize NLR receptors maintain highly conserved signalling-associated architectures across sequence, domain, and structural levels while permitting greater diversification within recognition-associated regions.

The framework was benchmarked against experimentally characterised NLR engineering systems. Evaluation of Gpa2/Rx1 domain swaps demonstrated the ability to distinguish compatible and incompatible engineered architectures, while analysis of the R13 LRR recombination series captured graded perturbation patterns associated with recognition-region variation. The Pikp-1 S258E system further showed that the framework can identify fine-scale residue-level modifications that preserve overall receptor architecture while altering predicted interface interactions.

The results of this work demonstrate that the framework does not replace experimental validation but provides a quantitative strategy for prioritising candidate NLR variants and engineered receptor configurations. By integrating multiple computational signals, this approach offers a scalable means to reduce the search space for future engineering efforts in plant immunity and crop disease resistance.

## SUPPLEMENTARY MATERIALS

The following supplementary materials are available online, Table S1: Uniprot protein information arranged by group, Table S2: Uniprot protein information arranged by cluster.

## DATA AVAILABILITY

All data generated in this study is available from the author upon reasonable request.

## CONFLICTS OF INTEREST

The author declares that he has no conflicts of interest.

## FUNDING

The author declared no financial support was received for this work.

## ACKNOWLEDGMENTS

During the preparation of this work, the author used ChatGPT to improve the clarity and readability of the text, and to help with literature review. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## REFERENCES

1. Jones JDG, Staskawicz BJ, Dangl JL. The plant immune system: From discovery to deployment. Cell. 2024;187(9):2095-116. doi: 10.1016/j.cell.2024.03.045
2. Cui H, Tsuda K, Parker JE. Effector-triggered immunity: from pathogen perception to robust defense. Annu Rev Plant Biol. 2015;66:487-511. doi: 10.1146/annurev-arplant-050213-040012
3. Kourelis J, van der Hoorn RAL. Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. Plant Cell. 2018;30(2):285-99. doi: 10.1105/tpc.17.00579
4. Teasdale LC, Murray KD, Collenberg M, Contreras-Garrido A, Schlegel T, van Ess L, et al. Pangenomic context reveals the extent of intraspecific plant NLR evolution. Cell Host Microbe. 2025;33(8):1291-1305.e9. doi: 10.1016/j.chom.2025.07.011
5. Cadiou L, Brunisholz F, Cesari S, Kroj T. Molecular engineering of plant immune receptors for tailored crop disease resistance. Curr Opin Plant Biol. 2023;74:102381. doi: 10.1016/j.pbi.2023.102381
6. Wang J, Hu M, Wang J, Qi J, Han Z, Wang G, et al. Reconstitution and structure of a plant NLR resistosome conferring immunity. Science. 2019;364(6435):eaav5870. doi: 10.1126/science.aav5870
7. DeYoung BJ, Innes RW. Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol. 2006;7(12):1243-9. doi: 10.1038/ni1410

8. Kourelis J, Sakai T, Adachi H, Kamoun S. RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. PLoS Biol. 2021;19(10):e3001124. doi: 10.1371/journal.pbio.3001124
9. Ramakrishna W, Emberton J, Ogden M, SanMiguel P, Bennetzen JL. Structural analysis of the maize rp1 complex reveals numerous sites and unexpected mechanisms of local rearrangement. Plant Cell. 2002;14(12):3213-23. doi: 10.1105/tpc.006338
10. Prigozhin DM, Sutherland CA, Rangavajjhala S, Krasileva KV. Majority of the Highly Variable NLRs in Maize Share Genomic Location and Contain Additional Target-Binding Domains. Mol Plant Microbe Interact. 2025;38:275-84. doi: 10.1094/MPMI-05-24-0047-FI
11. Zhao BY, Ardales E, Brasset E, Claflin LE, Leach JE, Hulbert SH. The Rxo1/Rba1 locus of maize controls resistance reactions to pathogenic and non-host bacteria. Theor Appl Genet. 2004;109(1):71-9. doi: 10.1007/s00122-004-1623-y
12. Jines MP, Balint-Kurti P, Robertson-Hoyt LA, Molnar T, Holland JB, Goodman MM. Mapping resistance to Southern rust in a tropical by temperate maize recombinant inbred topcross population. Theor Appl Genet. 2007;114(4):659-67. doi: 10.1007/s00122-006-0466-0
13. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. bioRxiv [Preprint]. 2020 Dec 15. doi: 10.1101/2020.12.15.422761
14. Weissenow K, Rost B. Are protein language models the new universal key? Curr Opin Struct Biol. 2025;91:102997. doi: 10.1016/j.sbi.2025.102997
15. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123-30. doi: 10.1126/science.ade2574
16. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA. 2021;118(15):e2016239118. doi: 10.1073/pnas.2016239118
17. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523-D531. doi: 10.1093/nar/gkac1052
18. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Nov 16-20; Online. pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6
19. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7(10):e1002195. doi: 10.1371/journal.pcbi.1002195
20. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412-9. doi: 10.1093/nar/gkaa913
21. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80. doi: 10.1093/molbev/mst010

22. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490
23. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42(2):243-6. doi: 10.1038/s41587-023-01773-0
24. Slootweg E, Koropacka K, Roosien J, Dees R, Overmars H, Lankhorst RK, et al. Sequence Exchange between Homologous NB-LRR Genes Converts Virus Resistance into Nematode Resistance, and Vice Versa. Plant Physiol. 2017;175(1):498-510. doi: 10.1104/pp.17.00485
25. Maidment JHR, Shimizu M, Bentham AR, Vera S, Franceschetti M, Longya A, et al. Effector target-guided engineering of an integrated domain expands the disease resistance profile of a rice NLR immune receptor. eLife. 2023;12:e81123. doi: 10.7554/eLife.81123
26. De la Concepcion JC, Franceschetti M, Maqbool A, Saitoh H, Terauchi R, Kamoun S, et al. Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. Nat Plants. 2018;4(8):576-85. doi: 10.1038/s41477-018-0194-x
27. Wang GF, Balint-Kurti PJ. Maize Homologs of CCoAOMT and HCT, Two Key Enzymes in Lignin Biosynthesis, Form Complexes with the NLR Rp1 Protein to Modulate the Defense Response. Plant Physiol. 2016;171(3):2166-77. doi: 10.1104/pp.16.00224

How to cite this article:
Camargo Rodriguez AV. Computational Mapping of Maize NLR Sequence Space Reveals Conserved Architectures Supporting Receptor Engineering Prioritisation. Crop Breed Genet Genom. 2026;8(3):e260020. https://doi.org/10.20900/cbgg20260020.