Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Computational Characterization of a Hypothetical Protein from the Jumbo Phage vB_Pae3705-KEN49 Targeting Multidrug-Resistant Pseudomonas aeruginosa

Meshack Tweya Omwega

meshacktweya2@gmail.com

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Moses Gachoya

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Justin Nyasinga

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Martin Georges

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

James Wachira

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Collins Kigen

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Eric Omori Omwenga

Kisii University

Geoffrey G Ratemo

Kisii University

Kevin Kamanyi

Kisii University

Erick Odoyo

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Lillian Musila

Kenya Medical Research Institute, Walter Reed Army Institute of Research–Africa

Research Article

Keywords: Pseudomonas aeruginosa, Phikzvirus, jumbo phage, hypothetical protein, AlphaFold3, DNABind, APBS electrostatics, HADDOCK protein–DNA docking, non-canonical DNA binding, regulatory

hypothesis, sequence-to-structure analysis

Posted Date: July 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10216254/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

Pseudomonas aeruginosa remains a WHO 2024 critical-priority pathogen, and jumbo bacteriophages of the *Phikzvirus* genus are emerging as viable therapeutic alternatives. Yet the majority of their encoded proteins remain functionally uncharacterised. We present a fully computational characterisation of WWV93726.1, a 721-residue conserved hypothetical protein from Pseudomonas jumbo phage vB_Pae3705-KEN49 (GenBank: PP456877). Genomic context places WWV93726.1 between phage-encoded RNA polymerase and major structural genes, a conserved syntenic position implicating a regulatory function. Physicochemical analysis identifies a globally acidic (pI 5.12), hydrophilic, cytoplasmic protein inconsistent with a structural virion role. Domain annotation reveals a bimodular architecture: an ordered N-terminal domain (residues 1–166); a disordered flexible linker (167–305); and an ordered C-terminal domain (306–721) with no significant sequence homology to any characterised protein. DNABind classified Domain 2 as a confident DNA binder (probability 0.804) versus non-binding Domain 1 (probability 0.006). Canonical DNA-binding motif screening was uniformly negative across four independent tools, suggesting a non-canonical binding mechanism. AlphaFold3 structure prediction confirmed the bimodular architecture. APBS electrostatic analysis revealed a focused electropositive patch on Domain 2, anchored by an RKKR cluster (+ 4 to + 5 kT/e; EP volume 12,951 Å³), absent from the neutral Domain 1. HADDOCK 2.4 docking against *Phikzvirus* promoter sequences confirmed energetically favourable interaction (HADDOCK score − 106.0 ± 6.4 kcal/mol; Z-score − 2.2), driven by electrostatic complementarity. Full-length protein docked 18.9 kcal/mol less favourably than isolated Domain 2, implicating Domain 1 in modulating DNA accessibility. We propose that WWV93726.1 is a non-canonical DNA-binding regulatory protein operating at the phage transcription-to-assembly transition.

1. Introduction

Pseudomonas aeruginosa is a Gram-negative opportunistic pathogen responsible for severe infections in immunocompromised patients, individuals with cystic fibrosis, and burn wound victims (1, 2). Its capacity for multidrug resistance, mediated by efflux pumps, beta-lactamases, carbapenemases, aminoglycoside-modifying enzymes, and robust biofilm formation, has rendered most available antibiotics ineffective against clinical isolates (3). The WHO 2024 Bacterial Priority Pathogens List designates carbapenem-resistant *P. aeruginosa* as a high-priority pathogen, underscoring the urgent need for alternative therapeutic strategies (4).

Jumbo bacteriophages (genomes > 200 kb) are of increasing therapeutic interest. Phages of the *Phikzvirus* genus, including phiKZ, 201phi2-1, and phiPA3, are particularly promising because they assemble proteinaceous nucleus-like compartments that physically sequester replicating phage DNA from host CRISPR-Cas systems and restriction-modification enzymes, enabling infection of otherwise highly resistant *P. aeruginosa* strains (5–7). Despite their therapeutic potential and mechanistic sophistication, *Phikzvirus* genomes encode large numbers of proteins with no functional annotation.

These constitute viral dark matter: sequences conserved by purifying selection across geographically distributed isolates, yet entirely uncharacterised (8, 9).

Functional annotation of such proteins is a prerequisite for understanding phage biology and for rationally engineering phage-based therapeutics. Computational characterisation, while unable to substitute for experimental validation, can generate specific, falsifiable hypotheses that prioritise experimental effort. A rigorous computational approach demands a clear analytical hierarchy: sequence-based evidence should be evaluated first to generate a hypothesis, and structure-based analyses should then test and mechanistically refine that hypothesis, not generate it independently..

This study applies this principle to WWV93726.1, a 721-residue conserved hypothetical protein from *Phikzvirus* vB_Pae3705-KEN49, isolated from hospital wastewater in Nairobi, Kenya (10). The protein was selected on three criteria: complete absence of functional annotation in any database; strict conservation (> 98% amino acid identity) across *Phikzvirus* isolates and a genomic position, between the phage-encoded RNA polymerase and major virion structural genes, suggestive of a regulatory rather than structural role. The goal of this study is not to prove a function, but to build a logically grounded, internally consistent computational hypothesis about protein function that can be directly tested experimentally.

2. Materials and Methods

2.1 Sequence Retrieval, Selection, and Genomic Context

WWV93726.1 (GenBank: WWV93726.1; genome PP456877) was retrieved from NCBI in October 2025. Selection criteria: (i) no functional annotation in NCBI or UniProt; (ii) BLASTp 2.14.0 E-value > 0.001 against all characterised proteins (identity ≥ 30%, coverage ≥ 70%); (iii) conservation across geographically distinct *Phikzvirus* isolates verified by BLASTp; (iv) genomic adjacency to functionally annotated genes. Genomic context was established by examining the syntenic arrangement of flanking genes across all available *Phikzvirus* genome sequences in GenBank.

2.2 Physicochemical Characterisation

Physicochemical properties were computed from the amino acid sequence using ExPASy ProtParam (11); molecular weight, isoelectric point, amino acid composition, net charge at pH 7.0, instability index (threshold > 40 predicts *in vitro* instability), aliphatic index, and GRAVY score. These sequence-derived properties test whether the physicochemical character of the protein is consistent with a regulatory (cytoplasmic, soluble) versus structural (hydrophobic, stable) function.

2.3 Domain Architecture and Disorder Prediction

Functional domain annotation used InterProScan 5.61-93.0 (12), integrating Pfam v35.0, PROSITE 2023_01, SMART v9.0, and NCBI CDD v3.20 (E-value ≤ 0.01). Remote sequence homology was searched using HHpred (13) against PDB70, SCOP70, Pfam-A, and CDD (confident: probability ≥ 90%; suggestive: ≥

50%). Intrinsic disorder was predicted using IUPRED3 (14) with a score ≥ 0.5 predicting disorder. Provisional domain boundaries were established from the convergence of these sequence-based tools.

2.4 Evolutionary Conservation Analysis

Evolutionary conservation of the protein was assessed through phylogenetic analysis using IQ-TREE version 2.2.5 (15). Homologous protein sequences obtained from BLASTp searches were aligned and used to construct a maximum-likelihood phylogenetic tree under the LG + G substitution model with 1,000 ultrafast bootstrap replicates.

2.5 Global DNA-Binding Classification - DNABind

Having identified a bimodular architecture and established that Domain 2 is the more evolutionarily conserved module, DNABind (16) was used to test whether the protein, or either of its domains, is predicted to bind DNA. DNABind employs a support vector machine trained on known DNA–protein interface structures and was applied to the full-length sequence, the Domain 1 sequence, and the Domain 2 sequence independently. Settings: FPR = 15.0%; score threshold = 0.125; probability threshold = 0.5313; expected sensitivity = 86.2%; MCC = 0.72. This analysis was completed before any structural prediction was performed.

2.6 Canonical DNA-Binding Motif Screening

Given the DNABind classification, the mechanism of DNA binding was investigated at the sequence level using four canonical DNA-binding motif searches: (i) HELIXTURNHELIX (EMBOSS 6.6.0) for canonical HTH motifs (score threshold 3.0; $p < 0.05$) (17); (ii) Pfam PF00196 and related winged-HTH families via HMMER 3.3 (18); (iii) PROSITE PS00028 for C2H2-type zinc fingers; and (iv) structure-based HTH prediction server applied to the AlphaFold3 model.

2.7 Tertiary Structure Prediction and Quality Validation

AlphaFold3 (19) predicted the tertiary structure without template restrictions. Per-residue pLDDT scores and PAE plots were exported. The provisional domain boundaries determined from sequence analysis were cross-validated against the pLDDT profile and PAE map. Model quality was independently validated using VADAR (20) (Ramachandran; threshold > 90% core), Verify3D (21) (threshold > 80% residues ≥ 0.1), QMEAN (22) (acceptable: Z within ± 2.0), and ProSA-web (23), before any structural inferences were drawn.

2.8 Structural Homology Analysis

FoldSeek (24) searched each domain structure against PDB, AlphaFold Database, and the Big Fantastic Virus Database. DALI (25) performed pairwise comparison for top hits. The taxonomic distribution of structural homologs was assessed separately for Domain 1 and Domain 2.

2.9 Binding Pocket Prediction

P2Rank 2.4 (26) and FPocket 4.0 (27) predicted functional binding pockets on the validated AlphaFold3 model. Pocket locations were recorded for correlation with electrostatic and per-residue interface analyses in subsequent sections. Drug scores below 0.3 indicate macromolecular interaction surfaces.

2.10 Electrostatic Surface Potential Analysis - APBS

To investigate the structural basis for non-canonical DNA binding, APBS 3.0 (28) calculated electrostatic surface potential at pH 7.0 after PDB2PQR 3.5 protonation with the CHARMM36 force field. Analysis was performed separately for Domain 1, Domain 2, and the full-length protein. Electropositive volume (EP vol) exceeding + 3 kT/e was quantified from the volumetric output for each construct. The percentage EP vol difference between isolated Domain 2 and the full-length protein was calculated to assess whether Domain 1 modulates the Domain 2 electropositive surface.

2.11 Electropositive Patch Residue Identification - BindUP

BindUP (https://bindup.technion.ac.il/) identified the largest electropositive surface patch on Domain 2 and its constituent residues. The positively charged residues (Arg, Lys, His) within the identified patch were designated as active interface residues (AIRs) for HADDOCK docking, providing a structural basis for the restraints.

2.12 Protein–DNA Docking - HADDOCK 2.4

HADDOCK 2.4 (29) was used to assess whether Domain 2 can form energetically favourable interactions with *Phikzvirus* promoter DNA sequences, testing the functional hypothesis. HADDOCK was selected for its established validation on protein–nucleic acid complexes and its support for AIR-guided docking with side-chain flexibility. Active residues from BindUP analysis (ARG617, ARG637, ARG689, HIS691, LYS692, ARG707) were used as protein AIRs for Domain 2 and full-length docking jobs; passive residues were defined automatically within 6.5 Å. For Domain 1 docking, ARG197, the sole positively charged residue identified in the Domain 1 largest surface patch by BindUP, was used as the single active residue, providing Domain 1 its best possible opportunity to dock and ensuring a fair internal comparison. Two *Phikzvirus* promoter sequences (25 bp B-form duplexes; Promoter-1: late-gene consensus from vB_PA32_GUMS; Promoter-2: early-gene consensus from published transcriptomic data (30), were docked against Domain 2, full-length protein, and Domain 1. DNA central bases (positions 12–31) were used as active residues. Sampling: 1,000 rigid-body; 200 semi-flexible; 200 water-refined structures. Domain 1 results serve as the internal negative control for the dataset.

3. Results

3.1 Genomic Context Suggests a Regulatory Role

WWV93726.1 is encoded in the genome of vB_Pae3705-KEN49 (PP456877), a *Phikzvirus* jumbo phage isolated from hospital wastewater in Nairobi, Kenya. BLASTp searches identified high-identity homologs (99–100%; full query coverage; E-value < 1×10^{-200}) exclusively in *Phikzvirus* phages: vB_PA32_GUMS

(Canada), OMK01 (USA), vb_Paem_HUST_1 (China), Sp.Brmt (Japan), and vB_PaePA01-KEN19 (Kenya). Greater than 98% pairwise amino acid identity across these geographically and temporally distributed isolates indicates that WWV93726.1 is under strong purifying selection, that is, its sequence is being actively maintained because mutations are harmful to phage fitness.

Syntenic genomic context analysis revealed that WWV93726.1 is consistently positioned between the phage-encoded non-virion RNA polymerase subunit genes and the major virion structural protein genes across all queried *Phikzvirus* genomes (Fig. 1). This boundary position, at the junction of the transcriptional apparatus and the structural assembly module, is a genomic signature that raises the first testable question: could this protein play a regulatory role at the transcription-to-assembly transition?

3.2 Physicochemical Properties Are Consistent with a Regulatory, Non-Structural Protein

To determine whether the basic physicochemical character of WWV93726.1 is consistent with a regulatory role, properties were computed from the primary sequence (Table 1). The protein is 721 residues (80,137.71 Da), acidic (pI 5.12; net charge − 24 at pH 7.0), and hydrophilic (GRAVY = − 0.346). Negatively charged residues (82 Asp + Glu) substantially outnumber positively charged residues (58 Arg + Lys). The GRAVY score indicates the hydrophilic character expected of cytoplasmic proteins. The instability index (42.26) exceeds the *in vitro* instability threshold (> 40).

Taken together, these properties are less characteristic of a canonical virion structural protein, which typically requires hydrophobic residues for capsid assembly and exhibits high *in vitro* stability, and are consistent with a cytoplasmic regulatory function. This narrows the hypothesis space and motivates the next question: what specific regulatory activity might this protein perform?

Table 1
Physicochemical properties of WWV93726.1 computed by ExPASy ProtParam.

Properties	**Value**
Molecular Weight	80137.71 Da
Isoelectric Point (pI)	5.12
Negative charged residues (Asp + Glu)	82
Positively Charged residue (Arg + Lys)	58
Molecular formular	$C_{3556}H_{5525}N_{963}O_{1084}S_{32}$
Instability Index	42.26
Aliphatic index	82.00
Grand average hydropathicity (GRAVY)	-0.346
Estimated half-life	30 hours (mammalian reticulocytes, in vitro)
	> 20 hours (yeast, in vivo).
	> 10 hours (*Escherichia coli*, in vivo)

3.3 Domain Architecture Reveals a Bimodular Protein with a Novel Domain 2

To determine whether WWV93726.1 contains recognisable functional modules, InterProScan (12), HHpred (13), and IUPRED3 (14) were applied to the sequence. The analyses converged on a tripartite architecture: a structured N-terminal Domain 1 (residues 1–166) with a borderline Med15-like signal (HHpred; $E = 3 \times 10^{-4}$) and a low-confidence zinc finger signal (PROSITE PS00028; E = 0.021); a central linker (residues 167–305) predicted to be intrinsically disordered throughout (IUPRED3 scores > 0.5); and a structured C-terminal Domain 2 (residues 306–721) that returned no significant sequence matches in Pfam, SCOP, CDD, SMART, or any HHpred database. (Fig. 2, Table 2).

The complete absence of sequence-based annotation for Domain 2 is a significant finding. It means Domain 2 cannot be assigned a function by sequence homology transfer and requires independent functional evidence. The disordered linker is also notable: intrinsically disordered regions in regulatory proteins often serve as flexible tethers that allow modular domains to adopt different relative orientations, enabling conformational switching. This architecture raises the key question: which of the two structured domains performs the core regulatory function?

Table 2
Functional domains and motifs in protein WWV93726.1

Motif/Domain	Database	Position	E-value
Med15 (ARC105)	NCBI-CDD (pfam09606)	110–274	3e-04
SP1-4_arthropods_N	NCBI-CDD (cd22553)	113–226	0.010
SP4_N	NCBI-CDD (cd22536)	65–226	0.058
Zap1_zf1 (C2H2)	Pfam (PF21816)	130–145	0.021
Formamidopyrimidine-DNA glycosylase (Fpg)	NCBI-CDD (PRK14811)	169–199	0.090

3.4 Evolutionary Conservation Analysis

Multiple sequence alignment of the homologous proteins revealed several highly conserved regions distributed across the sequence, indicating evolutionary retention of functionally important residues. Phylogenetic analysis using IQ-TREE grouped the homologous proteins into closely related clusters with strong bootstrap support, suggesting evolutionary relatedness and conservation among members of the protein family (Fig. 3). Short branch lengths observed among several homologs indicated limited sequence divergence, further supporting conservation of the protein across related taxa.

3.5 DNABind Independently Classifies Domain 2 as a DNA Binder Before Any Structure Is Consulted

Having identified Domain 2 as the core functional module, DNABind (16) was used to test whether the protein or any of its domains is predicted to bind DNA, entirely from sequence, before the AlphaFold3 structure was examined. The results provide a clean functional dissection (Table 3). Domain 2 alone is classified as a confident DNA binder (score = + 1.413; probability = 0.804). Domain 1 alone has a near-zero probability of DNA binding (score = − 5.179; probability = 0.006). The full-length protein receives the highest confidence (score = + 2.758; probability = 0.940), suggesting that the structural context of Domain 2 within the complete protein enhances the features that DNABind detects, but the activity is entirely attributable to the Domain 2 module.

These three results, obtained from sequence alone without reference to any structural prediction, constitute the primary evidence for the DNA-binding function hypothesis. The combination of genomic positioning, physicochemical consistency with a regulatory protein, identification of a novel conserved Domain 2, and its strong evolutionary conservation now converges with this direct functional classification to form a coherent hypothesis: WWV93726.1 is a DNA-binding regulatory protein, with Domain 2 as the DNA-binding module. The next question is: what mechanism does Domain 2 use to bind DNA?

Table 3

DNABind classification results for the full-length protein and isolated domains.

Construct	DNABind score	Probability	Classification
Domain 2 (residues 306–721)	+ 1.413	0.8043	DNA-binding
Full-length protein (721 aa)	+ 2.758	0.9403	DNA-binding
Domain 1 (residues 1–166)	−5.179	0.0056	Non-DNA-binding

*DNABind is a global (protein-level) classifier. It identifies which construct is predicted to bind DNA, but does not identify the specific interface residues. Per-residue surface analysis was deferred to the structure-based tier (*Section 3.10*).*

3.6 Canonical DNA-Binding Motif Screening: All Four Tools Return Negative Results

To determine whether Domain 2 uses a canonical HTH, winged-HTH, or zinc finger DNA-reading mechanism, four independent tools were applied to the sequence (Table 4). EMBOSS HELIXTURNHELIX (17) returned no hits above the score threshold of 3.0. Pfam PF00196 (winged-HTH) searched via HMMER 3.3 (18) returned no hits at any E-value. PROSITE PS00028 (C2H2 zinc finger) returned no hits. The structure-based HTH prediction server returned no hits on the AlphaFold3 model.

These uniformly negative results are an important scientific finding. They rule out all three classical sequence-specific DNA recognition mechanisms and establish that Domain 2, if it binds DNA, must do so through a non-canonical structural mechanism. Non-canonical DNA binding in phage regulatory proteins is typically mediated by electropositive surface patches that contact the DNA phosphate backbone through charge complementarity rather than inserting a reading helix into the major groove. This raises a specific structural question: does Domain 2 carry a focused electropositive surface patch consistent with this mechanism?

Table 4
Canonical DNA-binding motif screening results.

Tool	Search target	Result	Significance
HELIXTURNHELIX (EMBOSS 6.6.0)	Full-length sequence	No hits (score < 3.0)	No canonical three-helix HTH motif; protein does not use HTH-mediated major groove recognition
Winged-HTH (Pfam PF00196; HMMER 3.3)	Full-length sequence	No hits (any E-value)	Not a member of the winged-HTH superfamily
C2H2 zinc finger (PROSITE PS00028)	Full-length sequence	No hits	No zinc finger DNA-reading module
Structure-based HTH server	AlphaFold3 full-length PDB	No hits	No HTH arrangement in predicted structure; corroborates sequence-based searches

3.7 AlphaFold3 Structure Prediction Validates the Bimodular Architecture

Having established the sequence-level hypothesis, AlphaFold3 structure prediction (19) was performed, and the model was validated before any structural functional inference was drawn. The pLDDT profile (Fig. 4) independently confirms the tripartite architecture identified by sequence analysis: high confidence in Domain 1 (median pLDDT > 80) and Domain 2 (median pLDDT > 80) with markedly reduced confidence in the linker (median pLDDT < 50), consistent with intrinsic disorder. The PAE plot shows low intra-domain positional error and elevated inter-domain PAE reflecting genuine positional uncertainty due to linker disorder, not a failure of domain-level prediction.

All four quality metrics meet accepted thresholds (Table 5), validating the model for structural hypothesis generation. The three-dimensional architecture confirms the two-domain organisation: Domain 1 and Domain 2 are structurally independent folded units connected by a flexible disordered tether, consistent with a protein in which each domain can adopt different orientations relative to the other.

Table 5
AlphaFold3 model quality validation metrics.

Tool	Value	Threshold
VADAR - Ramachandran	Core 90.4%; Allowed 8.3%; Disallowed 0.5%	> 90% core required
Verify3D	81.28% ≥ 0.1	> 80% required
QMEAN Z-score	−1.8	± 2.0 of experimental structures
ProSA-web Z-score	−7.2	Experimental ~ 700 aa range: −8 to − 5

3.8 Structural Homology Confirms Differential Domain Histories

FoldSeek (24) and DALI (25) structural homology searches tested whether the contrasting sequence conservation profiles of Domain 1 and Domain 2 correspond to different structural evolutionary histories. The results confirmed this asymmetry. Domain 1 structural homologs are phylogenetically restricted to Pseudomonas-infecting *Phikzviruses* (top FoldSeek hit: PhiPA3 protein F8SK13; E = 1.19×10^{-11}; RMSD = 0.729 Å), suggesting a host-genus-specific function (Fig. 5). In contrast, Domain 2 yielded 23 structural homologs spanning phages that infect phylogenetically diverse bacterial hosts, *Pseudomonas, Ralstonia, Xanthomonas*, and *Serratia* (top DALI hit: 9rp3-A from vB_PA32_GUMS; Z-score = 32.0; RMSD = 1.2 Å; near-identical fold). Critically, every Domain 2 structural homolog is itself functionally uncharacterised. The broad taxonomic distribution confirms that Domain 2 performs a core function conserved across diverse large phages.

3.9 Binding Pocket Prediction Localises the Primary Functional Site to Domain 2

P2Rank (26) and Fpocket (27) identified three predicted binding pockets on the AlphaFold3 model (Fig. 6). The Domain 2 primary pocket is by far the most confident (P2Rank score = 53.07; probability = 0.981; 47 residues). Its low drug score (0.018) and high apolar interior (70.6%) are characteristic of a macromolecular, rather than small-molecule, interaction surface, consistent with a protein–DNA binding interface. The Domain 1 pocket is considerably less confident (score = 15.60; 17 residues) and, given the DNABind non-binding classification for Domain 1 (probability 0.006), is unlikely to mediate DNA interaction.

3.10 APBS Electrostatics: Domain 2 Carries a Focused Electropositive Surface Patch

To investigate whether the three-dimensional structure of Domain 2 carries a surface feature consistent with DNA interaction, as suggested by the sequence-level DNABind classification, APBS electrostatic analysis (28) was performed at pH 7.0 for all three constructs (Fig. 7).

Domain 1 is electrostatically neutral (mean = + 0.05 kT/e; SD = 0.82; EP vol > + 3 kT/e = 9,107 Å³). Its surface potential distribution is narrow and centred near zero, with only 1.2% of surface exceeding + 3 kT/e, consistent with the DNABind non-binding classification for Domain 1.

Domain 2 shows a bimodal surface potential distribution (mean = − 1.64 kT/e; SD = 1.90; EP vol > + 3 kT/e = 12,951 Å³). Despite a net-negative overall mean, Domain 2 has a distinct positive-tail population reaching + 4 to + 5 kT/e, corresponding to a spatially focused electropositive patch in the region of the

RKKR cluster (ARG342, ARG344, LYS346, LYS348). This surface feature is absent from Domain 1 and is consistent with the DNABind prediction of domain-specific DNA-binding activity.

The full-length protein is the most negative overall (mean = − 2.50 kT/e; SD = 2.14). The electropositive patch volume in the full-length protein is 3.6% smaller than in isolated Domain 2, indicating the patch is largely preserved in the complete protein context. Whether this difference is functionally significant cannot be determined from electrostatic analysis alone and requires experimental investigation.

3.11 BindUP Identifies Positively Charged Residues on the Domain 2 Surface Patch

BindUP was used to identify the specific residues composing the largest electropositive surface patch on Domain 2 and to define the residue set for docking. BindUP identified a 26-residue largest positive patch. The positively charged residues within this patch, ARG617, ARG637, ARG689, HIS691, LYS692, ARG707, form a spatially contiguous cluster on the Domain 2 surface, consistent with the + 4 to + 5 kT/e APBS signal and overlapping with the dominant P2Rank pocket. These six residues were used as active residues in HADDOCK docking. Whether these residues constitute the actual DNA-contact interface requires experimental validation.

The BindUP binary classification (non-DNA-binding; 44% confidence) conflicts with the DNABind classification (DNA-binding; 80.4% probability). Both results are reported. This disagreement reflects genuine uncertainty that only experimental data can resolve.

3.12 HADDOCK Docking: A Comparative Dataset Across All Constructs

HADDOCK 2.4 docking (29) was performed to assess whether Domain 2 produces energetically favourable predicted interactions with *Phikzvirus* promoter DNA sequences, and to compare this against the full-length protein and Domain 1. Six experiments were run: Domain 2 and full-length protein each against both promoters, and Domain 1 against both promoters as an internal comparison. Results are presented in Table 6.

Domain 2 against both promoters

Domain 2 docked against Promoter-1 produced a top cluster HADDOCK score of − 92.5 ± 7.4 kcal/mol (Z-score = − 1.6; cluster size = 4; electrostatic energy = − 464.2 ± 31.3 kcal/mol; vdW = − 28.9 ± 9.1 kcal/mol; BSA = 1,735 Å²). Against Promoter-2, the top cluster scored − 106.0 ± 6.4 kcal/mol (Z-score = − 2.2; cluster size = 24; intra-cluster RMSD = 3.9 ± 2.3 Å; electrostatic = − 422.0 ± 8.9 kcal/mol; vdW = − 48.1 ± 9.3 kcal/mol). The Promoter-2 result shows 24 independent docking runs converging to the same binding geometry (RMSD 3.9 Å) with a Z-score of − 2.2, indicating a well-defined top solution relative to all other clusters. Across both promoters, the electrostatic energy term is substantially larger than the van der Waals term, consistent with the surface electropositive patch identified by APBS and BindUP.

Full-length protein against both promoters

The full-length protein docked against Promoter-1 at − 73.6 ± 5.0 kcal/mol (Z-score = − 2.1; cluster size = 22; electrostatic = − 267.9 ± 39.6 kcal/mol; RMSD = 8.6 ± 0.3 Å). Against Promoter-2, the full-length protein scored − 76.4 ± 8.8 kcal/mol (Z-score = − 1.7; cluster size = 16; electrostatic = − 420.1 ± 40.6 kcal/mol; RMSD = 20.6 ± 0.2 Å). The full-length protein scores 18.9 kcal/mol (Promoter-1) and 29.6 kcal/mol (Promoter-2) lower than isolated Domain 2 against the respective promoters. This difference is consistent across both promoter sequences. The APBS analysis showed the Domain 2 electropositive patch is largely preserved in the full-length protein (3.6% EP vol reduction). The lower docking scores for the full-length protein compared to Domain 2 alone, alongside the increased intra-cluster RMSD and restraints violation energy, raise the hypothesis that Domain 1 influences DNA interaction — but the nature of this influence cannot be established from docking scores alone.

Domain 1 against both promoters

Domain 1 scored − 56.0 ± 8.1 kcal/mol against Promoter-1 (Z-score = − 2.7; cluster size = 14; electrostatic = − 140.9 ± 28.1 kcal/mol; RMSD = 31.2 ± 0.6 Å) and − 56.0 ± 1.8 kcal/mol against Promoter-2 (Z-score = − 2.0; cluster size = 29; electrostatic = − 151.0 ± 6.0 kcal/mol; RMSD = 35.0 ± 0.9 Å). The absolute HADDOCK scores are the lowest of all constructs. The electrostatic energy (approximately − 146 kcal/mol average) is substantially lower than Domain 2 (approximately − 443 kcal/mol average). The RMSD values (31–35 Å) are the highest of any run, indicating no convergent binding geometry was found. The identical score against two different promoter sequences is consistent with the DNABind non-binding classification for Domain 1 (probability 0.006). Domain 1 results serve as the internal comparison baseline for this dataset.

Table 6

HADDOCK 2.4 protein–DNA docking results for all six experiments. Domain 1 serves as the internal negative control.

Construct / DNA	HADDOCK score (kcal/mol)	Z-score	Cluster size	Electrostatic (kcal/mol)	vdW (kcal/mol)	BSA (Å²)
Domain 2 + Promoter-1	−92.5 ± 7.4	−1.6	4	−464.2 ± 31.3	−28.9 ± 9.1	1,735
Domain 2 + Promoter-2	−106.0 ± 6.4	−2.2	24	−422.0 ± 8.9	−48.1 ± 9.3	1,732
Full-length + Promoter-1	−73.6 ± 5.0	−2.1	22	−267.9 ± 39.6	−50.7 ± 1.9	1,436
Full-length + Promoter-2	−76.4 ± 8.8	−1.7	16	−420.1 ± 40.6	−31.1 ± 8.5	1,609
Domain 1 + Promoter-1	−56.0 ± 8.1	−2.7	14	−140.9 ± 28.1	−39.5 ± 2.7	1,178
Domain 1 + Promoter-2	−56.0 ± 1.8	−2.0	29	−151.0 ± 6.0	−39.2 ± 3.3	1,122

Domain 1 serves as the internal negative control: no coherent electropositive patch (APBS), non-DNA-binding by DNABind (probability 0.006), and a single active residue, ARG197, the only positively charged residue identified in the Domain 1 largest surface patch by BindUP.

4. Discussion

The primary objective of this study was to generate a specific, internally consistent, and experimentally testable hypothesis about the function of WWV93726.1. No single computational tool can establish protein function, but when multiple independent analyses based on different principles converge on the same prediction, confidence in the hypothesis increases. Across twelve analyses, the results point consistently toward Domain 2 (residues 306–721) as a predicted DNA-binding module.

At the sequence level, genomic context places WWV93726.1 at the boundary between the phage RNA polymerase cluster and the structural gene cluster, a position suggestive of a regulatory rather than structural role. Physicochemical properties are inconsistent with a virion structural protein. Domain annotation identifies Domain 2 as entirely novel at the sequence level, with no functional annotation in any database. DNABind (16) classifies Domain 2 as a probable DNA binder (0.804) and Domain 1 as a non-binder (0.006) from sequence alone. All four canonical DNA-binding motif screens return negative results, indicating the predicted activity, if real, does not proceed through an HTH, winged-HTH, or zinc finger mechanism.

At the structural level, AlphaFold3 (19) independently confirms the bimodular architecture defined by sequence analysis. Structural homology searches show Domain 2 to be broadly conserved across

phages of diverse bacterial hosts, consistent with a core function. The dominant predicted binding pocket localises to Domain 2. APBS analysis (28) identifies a focused electropositive surface patch on Domain 2 absent from the neutral Domain 1 surface, providing a plausible structural correlate for the DNABind prediction. HADDOCK docking (29) produces lower-energy, more geometrically convergent predicted interactions for Domain 2 than for Domain 1 against both *Phikzvirus* promoter sequences. None of these observations individually proves DNA binding, together they constitute a coherent and mutually reinforcing foundation for the hypothesis.

The uniformly negative canonical motif screening might appear to conflict with the DNABind classification. It does not. DNABind was not trained exclusively on HTH-mediated binders; it captures a broader set of sequence-level features associated with DNA interaction. The negative motif results establish only that WWV93726.1 does not use one of the three most common sequence-specific DNA recognition architectures. Many DNA-binding proteins, particularly regulatory proteins in bacteriophages, interact with DNA through non-canonical surfaces that are not captured by HTH or zinc finger consensus searches. The BindUP disagreement with DNABind (44% vs 80.4%) similarly reflects tool uncertainty for proteins that may fall outside the training distribution of either classifier. These discrepancies are reported transparently and are precisely the reason experimental validation is required.

The docking dataset reveals a consistent difference between isolated Domain 2 and the full-length protein: the full-length protein scores 18.9–29.6 kcal/mol lower than Domain 2 alone across both promoter sequences. This is an observation, not a mechanistic conclusion. Whether Domain 1 physically influences the Domain 2 surface, whether the difference reflects conformational effects of the linker, or whether it is an artefact of the larger molecular system in HADDOCK cannot be determined from these data. The APBS analysis shows the Domain 2 electropositive patch is largely preserved in the full-length protein (3.6% EP vol reduction), which is one observation that may be relevant to interpreting this difference. The role of Domain 1 in the biology of this protein is a question for experimental investigation, not a conclusion that can be drawn from computational docking scores alone.

Based on the totality of evidence presented, we propose the following hypothesis for experimental testing: WWV93726.1 is a DNA-binding protein in which the C-terminal Domain 2 (residues 306–721) is the functional module responsible for DNA interaction, while the role of Domain 1 (residues 1–166) remains to be determined. The conserved genomic position between the phage RNA polymerase and structural gene clusters, combined with the predicted DNA-binding activity of Domain 2, raises the possibility that WWV93726.1 participates in the regulation of phage gene expression, though this remains speculative.

Four experiments follow directly from this computational analysis and would together confirm or refute the core hypothesis: (i) electrophoretic mobility shift assay (EMSA) with purified Domain 2 and full-length protein against both promoter sequences to determine whether binding occurs and whether the two constructs differ in affinity; (ii) deletion mapping to confirm that Domain 2 is the minimal DNA-binding unit; (iii) site-directed mutagenesis of the six predicted interface residues (ARG617, ARG637, ARG689,

HIS691, LYS692, ARG707) to test whether they are required for binding; and (iv) chromatin immunoprecipitation or equivalent approaches in phage-infected cells to identify the *in vivo* genomic targets of WWV93726.1. These experiments are tractable with standard molecular biology methods and would provide the experimental foundation needed to move from hypothesis to function.

4.5 Limitations

All findings in this study are computational predictions. None constitutes experimental evidence of DNA binding, protein–DNA interaction, or regulatory activity. The AlphaFold3 model, despite meeting accepted quality thresholds, may contain local inaccuracies that affect the electrostatic and pocket analyses. HADDOCK docking results are sensitive to active residue selection; the six residues used were chosen based on BindUP analysis, but have not been experimentally validated as interface residues. The promoter sequences used in docking are consensus sequences derived from published transcriptomic data (30) and are not confirmed binding targets of this protein. DNABind and BindUP give conflicting binary classifications, and neither is definitive for a protein of this kind. All conclusions are framed as hypotheses requiring experimental validation before any functional claim can be made.

Conclusion

We present a sequence-to-structure computational analysis of WWV93726.1, a conserved hypothetical protein from *Phikzvirus* vB_Pae3705-KEN49. Twelve independent analyses converge on the hypothesis that Domain 2 (residues 306–721) is a predicted DNA-binding module. This protein has no sequence homology to any characterised protein, no canonical DNA-binding motif, and no prior functional annotation. The computational evidence presented here provides the foundation for a targeted experimental programme - EMSA, deletion mapping, mutagenesis, and *in vivo* chromatin association, that will determine whether WWV93726.1 is a genuine DNA-binding protein and, if so, what role it plays in the *Phikzvirus* infection cycle.

Declarations

Funding Declaration

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

Consent to Participate declaration: Not applicable

Consent to Publish declaration: Not applicable.

Availability of data and materials

The genome sequence of bacteriophage vB_Pae3705-KEN49 is publicly available in GenBank under accession number PP456877. The hypothetical protein analyzed in this study is available under

accession number WWV93726.1. All data generated or analyzed during this study are included in this published article.

Authors contributions

Conceptualization, M.T.O., L.M., and K.K; methodology, M.T.O., Mo.G., Ma.G., J.N., and L.M.; formal analysis, M.T.O., K.K., and E.O.O.; investigation, M.T.O., Mo.G., J.N., J.W., C.K., E.O.O., K.K., and E.O.; resources, L.M. and E.O.; data curation, M.T.O., E.O.O., K.K., and L.M.; writing--original draft, M.T.O.; writing--review and editing, L.M., Ma.G., J.N., and K.K., G.G.R.; visualization, M.T.O.; supervision, L.M. and K.K.; project administration, L.M.

Ethics declaration

Not Applicable

Competing Interest declaration

The authors declare that they have no competing interests.

Acknowledgments
The authors sincerely acknowledge the Phage Team and the Genomics and Bioinformatics Team at the Microbiology Hub, WRAIR–Kericho, for providing access to the bioinformatics laboratory and computational resources that enabled this work.

References

1. Al-Orphaly M, Hadi HA, Eltayeb FK, Al-Hail H, Samuel BG, Sultan AA et al. Epidemiology of Multidrug-Resistant Pseudomonas aeruginosa in the Middle East and North Africa Region. Gales AC, editor. mSphere. 2021;6(3):e00202-21. 10.1128/mSphere.00202-21
2. Gellatly SL, Hancock REW. *Pseudomonas aeruginosa*: new insights into pathogenesis and host defenses. Pathog Dis. 2013;67(3):159–73. 10.1111/2049-632X.12033.
3. Pang Z, Raudonis R, Glick BR, Lin TJ, Cheng Z. Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. Biotechnol Adv. 2019;37(1):177–92. 10.1016/j.biotechadv.2018.11.013.
4. WHO Bacterial Priority Pathogens List. 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control Antimicrobial Resistance. 1st ed. Geneva: World Health Organization; 2024. 1 p.
5. Chaikeeratisak V, Nguyen K, Khanna K, Brilot AF, Erb ML, Coker JKC, et al. Assembly of a nucleus-like structure during viral replication in bacteria. Science. 2017;355(6321):194–7. 10.1126/science.aal2130.
6. Mendoza SD, Nieweglowska ES, Govindarajan S, Leon LM, Berry JD, Tiwari A, et al. A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. Nature. 2020;577(7789):244–8.

10.1038/s41586-019-1786-y.

7. Gao Z, Feng Y. Bacteriophage strategies for overcoming host antiviral immunity. Front Microbiol. 2023;14:1211793. 10.3389/fmicb.2023.1211793.
8. Santiago-Rodriguez TM, Hollister EB. Unraveling the viral dark matter through viral metagenomics. Front Immunol. 2022;13:1005107. 10.3389/fimmu. 2022.1005107 PubMed PMID: 36189246; PubMed Central PMCID: PMC9523745.
9. Strathdee SA, Hatfull GF, Mutalik VK, Schooley RT. Phage therapy: From biological mechanisms to future directions. Cell. 2023;186(1):17–31. 10.1016/j.cell.2022.11.017.
10. Musila L, Bird JT, Margulieux KR, Kigen C, Mzhavia N, Filippov AA et al. Complete genome sequences of three *Pseudomonas aeruginosa* jumbo bacteriophages discovered in Kenya. Putonti C, editor. Microbiol Resour Announc. 2024;13(11):e00684-24. 10.1128/mra.00684-24
11. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. The Proteomics Protocols Handbook [Internet]. Totowa, NJ: Humana Press; 2005 [cited 2025 Jun 4]. pp. 571–607. Available from: http://link.springer.com/10.1385/1-59259-890-0:571 doi:10.1385/1-59259-890-0:571.
12. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinforma Oxf Engl. 2014;30(9):1236–40. 10.1093/bioinformatics/btu031 . PubMed PMID: 24451626; PubMed Central PMCID: PMC3998142.
13. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 2018;430(15):2237–43. 10.1016/j.jmb.2017.12.007.
14. Erdős G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. 2021;49(W1):W297–303. 10.1093/nar/gkab408.
15. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74. 10.1093/molbev/msu. 300 PubMed PMID: 25371430; PubMed Central PMCID: PMC4271533.
16. Liu R, Hu J, DNABind. A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches: DNA-Binding Residue Prediction. Proteins Struct Funct Bioinforma. 2013;81(11):1885–99. 10.1002/prot.24330.
17. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–7. 10.1016/S0168-9525(00)02024-2.
18. Eddy SR, Accelerated Profile HMM, Searches. Pearson WR, editors. PLoS Comput Biol. 2011;7(10):e1002195. 10.1371/journal.pcbi.1002195
19. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500. 10.1038/s41586-024-07487-w.

20. Willard L. VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res. 2003;31(13):3316–9. 10.1093/nar/gkg565.
21. Eisenberg D, Lüthy R, Bowie JU, [20]. VERIFY3D: Assessment of protein models with three-dimensional profiles. In: Methods in Enzymology [Internet]. Elsevier; 1997 [cited 2025 Aug 4]. pp. 396–404. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0076687997770228 10.1016/S0076-6879(97)77022-8
22. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009;37(Web Server issue):W510-514. 10.1093/nar/gkp322 PubMed PMID: 19429685; PubMed Central PMCID: PMC2703985.
23. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35(Web Server issue):W407-410. 10.1093/nar/gkm290 PubMed PMID: 17517781; PubMed Central PMCID: PMC1933241.
24. 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. 10.1038/s41587-023-01773-0.
25. Holm L. DALI and the persistence of protein shape. Protein Sci. 2020;29(1):128–40. 10.1002/pro.3749.
26. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminformatics. 2018;10(1):39. 10.1186/s13321-018-0285-8.
27. Kochnev Y, Durrant JD. FPocketWeb: protein pocket hunting in a web browser. J Cheminformatics. 2022;14(1):58. 10.1186/s13321-. 022-00637-0 PubMed PMID: 36008829; PubMed Central PMCID: PMC9414105.
28. Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, et al. Improvements to the APBS biomolecular solvation software suite. Protein Sci. 2018;27(1):112–28. 10.1002/pro.3280.
29. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc. 2003;125(7):1731–7. 10.1021/ja026939x.
30. Yakunina M, Artamonova T, Borukhov S, Makarova KS, Severinov K, Minakhin L. A non-canonical multisubunit RNA polymerase encoded by a giant bacteriophage. Nucleic Acids Res. 2015;gkv1095. 10.1093/nar/gkv1095.

Figures

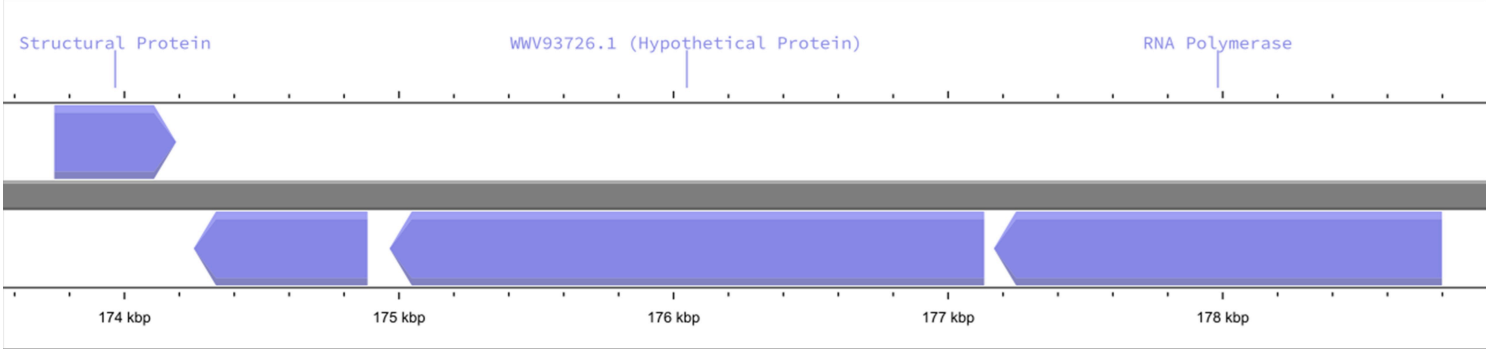


Figure 1

Genomic context of WWV93726.1 in phage vB_Pae3705-KEN49. The gene is positioned between phage-encoded RNA polymerase subunit genes and major virion structural protein genes. This syntenic arrangement is conserved across all queried *Phikzvirus* genomes and motivates the regulatory function hypothesis.

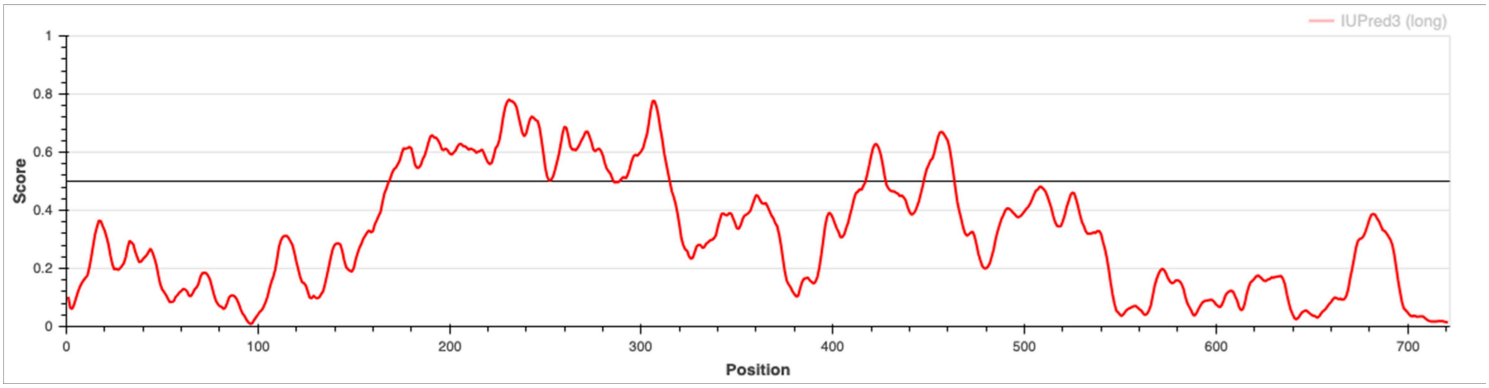


Figure 2

IUPRED3 per-residue disorder prediction profile for WWV93726.1. The linker region (residues 167–305) is consistently disordered (score > 0.5 threshold line). Domain 1 (1–166) and Domain 2 (306–721) are predominantly ordered.

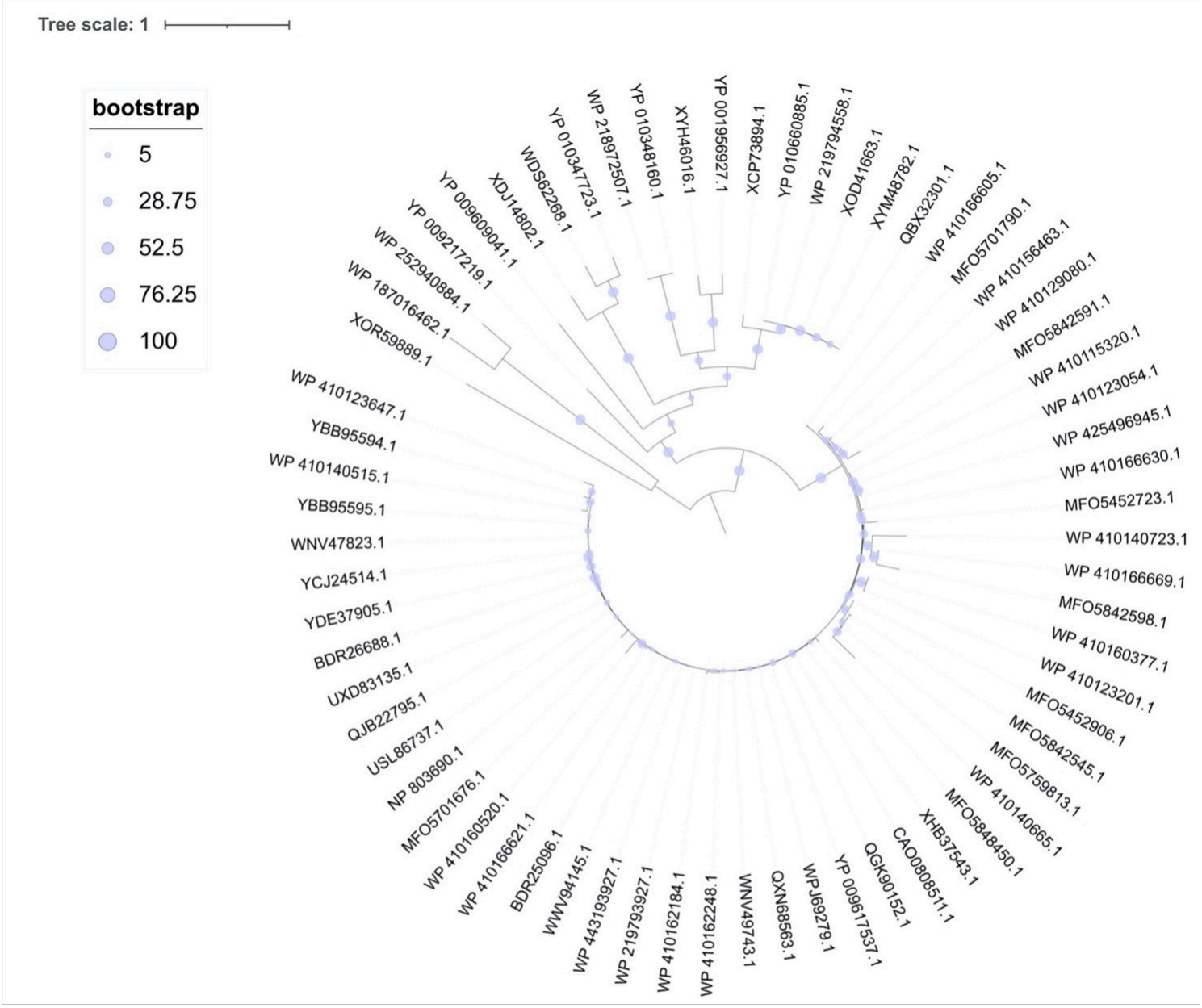


Figure 3

Maximum-likelihood phylogenetic tree of homologous proteins constructed using IQ-TREE under the LG+G model with 1,000 ultrafast bootstrap replicates. Conserved regions identified from the multiple sequence alignment are associated with closely related sequence clusters.

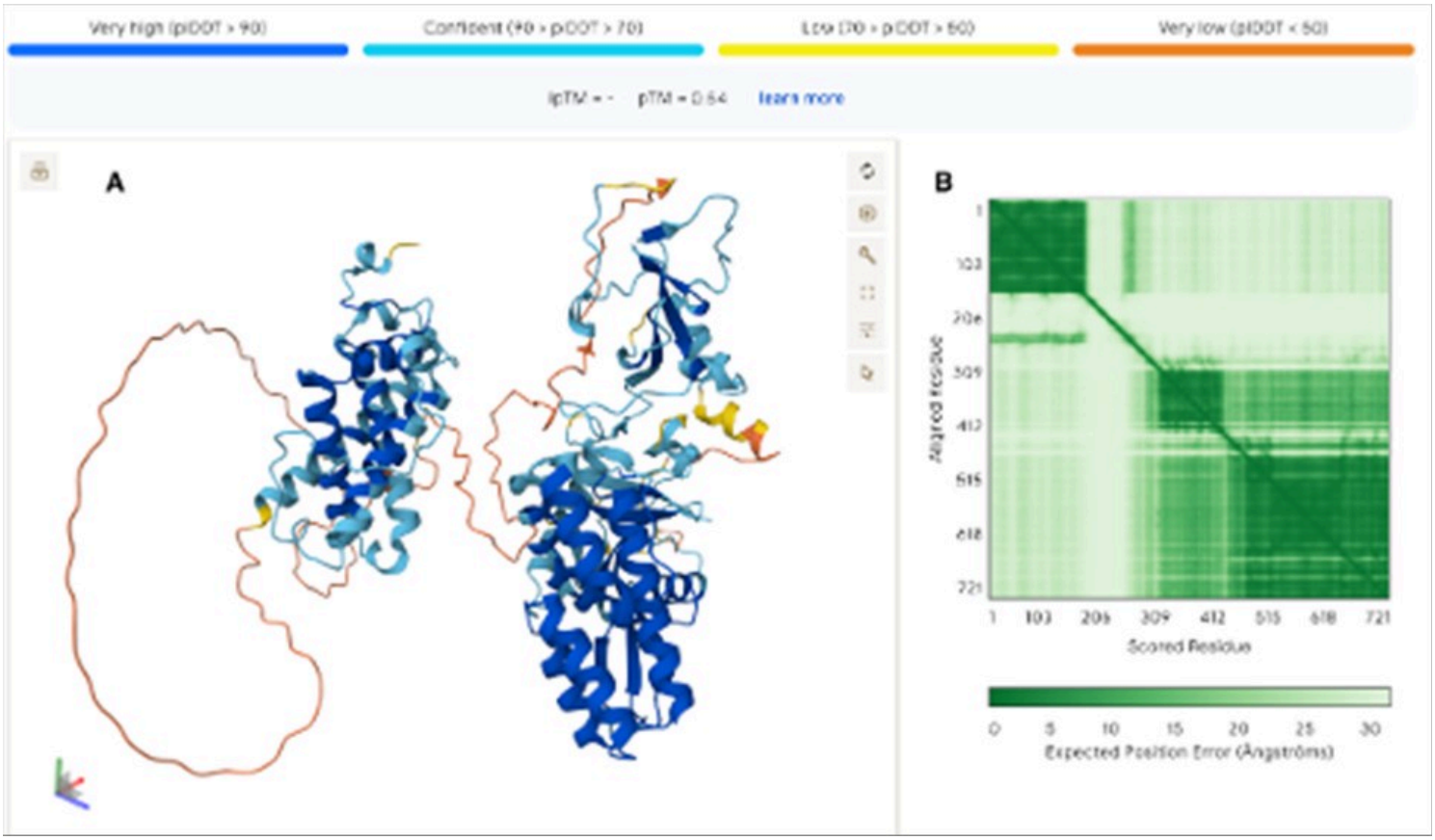


Figure 4

AlphaFold3 per-residue pLDDT confidence profile for WWV93726.1. Dark blue (> 70): high confidence; cyan (50–70): confident; yellow-orange (< 50): disordered. Domain 1 (1–166) and Domain 2 (306–721) are confidently modelled. The linker (167–305) shows collapsed confidence, independently validating the IUPRED3 disorder prediction.

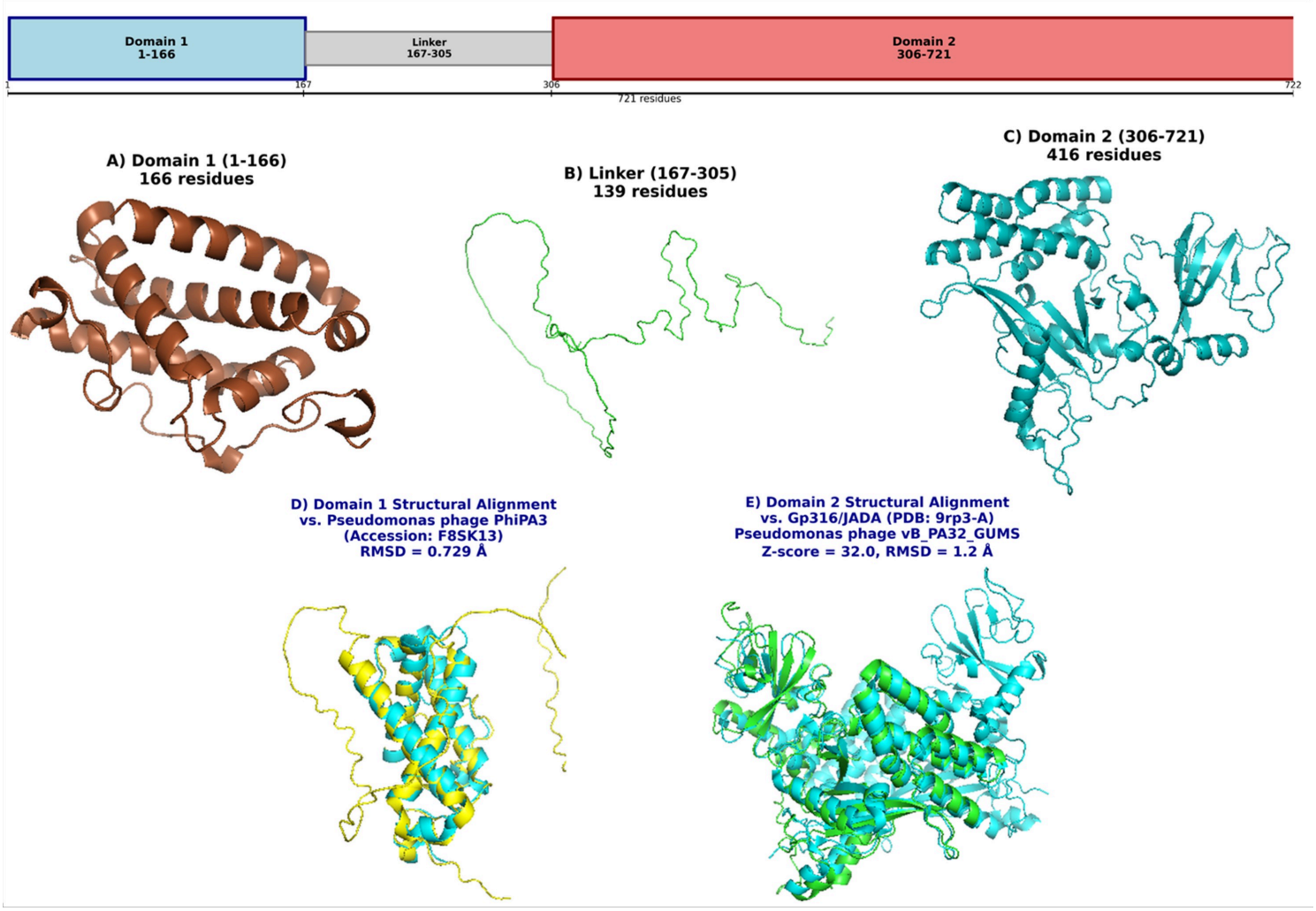


Figure 5

Domain architecture of WWV93726.1. Top panel: Linear schematic showing domain boundaries. **(A)** Domain 1 structure. **(B)** Flexibe linker region. **(C)** Domain 2 Structure. **(D)** Domain 1 has structural homologs restricted to Pseudomonas-infecting *Phikzviruses*, suggesting host-genus specificity**. (E)** Domain 2 has broadly conserved structural homologs across phages of diverse bacterial hosts, consistent with a core conserved function.

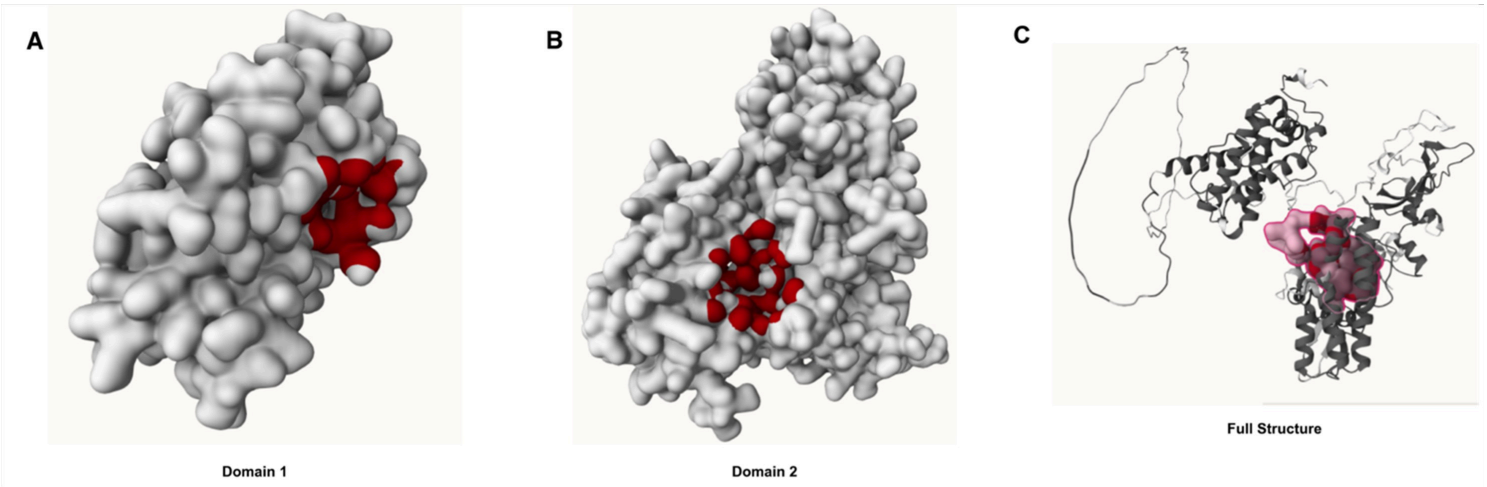


Figure 6

Predicted binding pockets. (A) Domain 1 showing primary binding pocket (red). (B) Domain 2 displays the most prominent binding cavity (red). (C) Complete protein structure with interdomain binding pocket highlighted (pink/red)

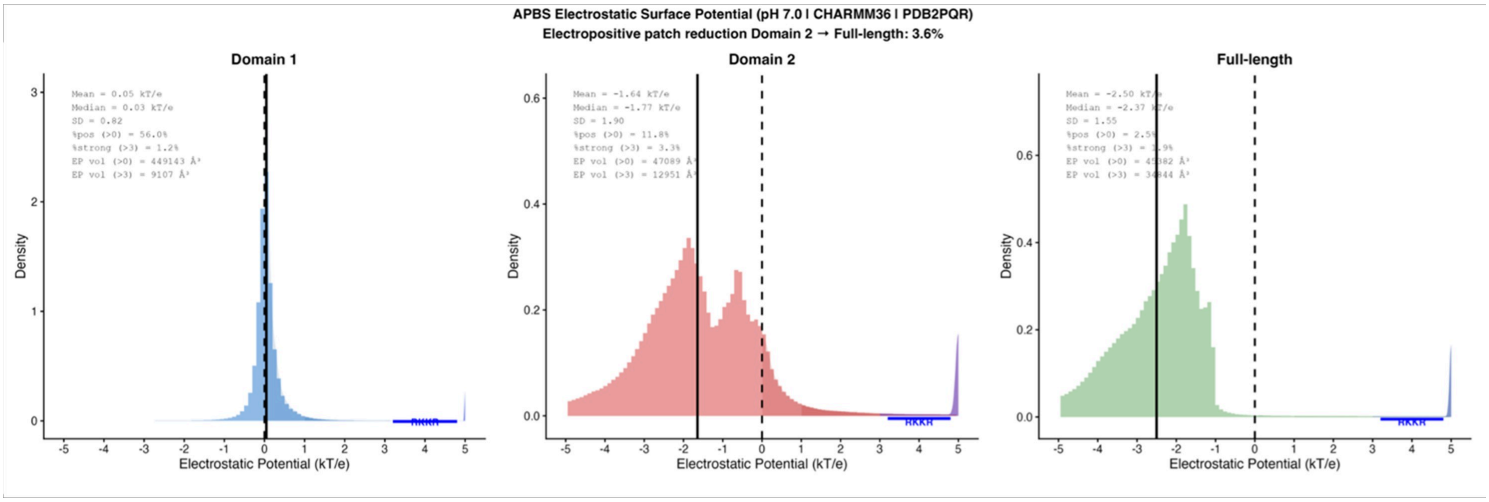


Figure 7

APBS electrostatic surface potential distributions for Domain 1 (blue), Domain 2 (red), and the full-length protein (green). Domain 1 is electrostatically neutral (mean +0.05 kT/e; SD 0.82). Domain 2 shows a bimodal distribution with a focused electropositive patch at +4 to +5 kT/e (RKKR cluster) representing the predicted DNA-contact surface. The full-length protein is the most negative overall; its 3.6% reduction in EP patch relative to isolated Domain 2 indicates the binding surface is electrostatically preserved in the complete protein.