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.

Structure-guided prioritization of divergent virulence and resistance candidates in an open pangenome clinical Escherichia coli isolate

Sarra Benmoumou-Hosni

sarra_ben@outlook.fr

École Normale Supérieure de Kouba

Atika Mekla

École Normale Supérieure de Kouba

Ikram Haleche

University of Blida

Research Article

Keywords: Escherichia coli, antimicrobial resistance, pangenome-to-structure, molecular dynamics, protein language models, extraintestinal pathogenic Escherichia coli (ExPEC), ST354, structural bioinformatics, AlphaFold, GNAT acetyltransferase

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11194219/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Structure-guided prioritization of divergent virulence and resistance candidates in an open pangenome clinical *Escherichia coli* isolate

Sarra Benmoumou-Hosni[1*], Atika Meklat[1] and Ikram Haleche[2]

[1*]Laboratoire de Biologie des Systèmes Microbiens (LBSM), École Normale Supérieure de Kouba, Algiers, 16000, Algeria.
[2]Laboratoire de Biotechnologies, Environnement et Santé, Université de Blida 1, Blida, 09000, Algeria.

*Corresponding author(s). E-mail(s): sarra_ben@outlook.fr;

Abstract

Determining the biological function of uncharacterized hypothetical proteins is a key challenge in genomics. The accessory genome of pathogens houses sequence-divergent antimicrobial resistance (AMR) and virulence genes escaping standard annotation. We present an integrated, structure-guided pangenome-to-structure computational pipeline combining pangenomic singleton filtering, structural prediction (ESMFold/AlphaFold3), homology searches (Foldseek/TM-align), Protein Language Model embeddings (ESM-2 650M), and molecular dynamics (MD) simulations (OpenMM). Applied to the clinical multidrug-resistant (MDR) extraintestinal pathogenic *Escherichia coli* (ExPEC) ST354 isolate QA5221, pangenome clustering across 32 genomes isolated 142 private singletons. Filtering prioritized four candidates in resistance/virulence superfamilies: (i) GNAT_KA27 (*KNGPFPPJ_02769*), a divergent GCN5-related N-acetyltransferase (GNAT) with strong aminoglycoside affinity ($\boldsymbol{\Delta G}_{\text{bind}} = \mathbf{-22.91}$ kcal/mol) via MM-GBSA and 127-ns MD fold stability (RMSD = $\mathbf{2.73 \pm 0.20}$ Å); (ii) Ehly_61 (*KNGPFPPJ_00061*), a putative enterohemolysin showing stable transmembrane topology ($\mathbf{5.65 \pm 1.04}$ Å RMSD) in bilayer MD; (iii) OAgP_161 (*KNGPFPPJ_03161*), a putative O-antigen polymerase (Wzy) in the LPS cluster ($\mathbf{2.12 \pm 0.23}$ Å RMSD); and (iv) OAT_371 (*KNGPFPPJ_04371*), a plasmid-associated outer-membrane autotransporter. All candidates feature low GC content (29.31–48.53%) and cluster distinctly in ESM-2

latent space. Phylogenetics confirms they form independent, deeply branching clades. This framework enables systematic prioritization of sequence-divergent accessory genes, flagging high-priority targets for experimental characterization. (GenBank: `JCACXV000000000`).

Keywords: Escherichia coli, antimicrobial resistance, pangenome-to-structure, molecular dynamics, protein language models, extraintestinal pathogenic *Escherichia coli* (ExPEC)

1 Introduction

Antimicrobial resistance (AMR) represents one of the most critical threats to global public health in the 21st century, causing an estimated 1.27 million deaths annually [1]. Extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages, particularly multidrug-resistant (MDR) clonal groups like Sequence Type 354 (ST354), are primary drivers of this crisis, causing a wide spectrum of severe diseases such as urinary tract infections, bacteremia, and neonatal meningitis [2, 3]. In these pathogenic lineages, treatment failure is frequently driven by a combination of acquired antimicrobial resistance (AMR) genes and specific chromosomal mutations in drug targets, such as those in the quinolone resistance-determining region (QRDR), which together restrict therapeutic options and exacerbate clinical outcomes [4, 5].

The genomic landscape of *E. coli* is characterized by an open pangenome, which continues to expand as new genomes are sequenced, reflecting a highly dynamic evolutionary trajectory [6, 7]. While core genes are conserved across all isolates to maintain basic cellular functions, the accessory genome consists of shell and cloud partitions that represent a vast, strain-specific reservoir of genetic diversity, containing private singletons or "ORFans" that lack homology in reference genomes and are unique to specific strains. This accessory fraction acts as a primary vehicle for genomic adaptation, frequently accumulating novel virulence and resistance determinants acquired

through horizontal gene transfer (HGT) from diverse environmental and clinical donor pools [8, 9].

Mobile genetic elements (MGEs) such as insertion sequences (ISs), integrons, and plasmids serve as the primary drivers of this genomic plasticity [8, 10]. These elements act as natural gene-capture systems, facilitating the transposition, recombination, and dissemination of multidrug resistance cassettes and virulence factors across diverse bacterial populations [11, 12]. Furthermore, insertion sequences can disrupt ancestral regulatory regions or mobilize adjacent chromosomal segments, giving rise to novel gene arrangements and atypical resistance phenotypes that enable rapid adaptation to host-selective and antibiotic pressures [13, 14].

Despite the abundance of genomic data, a substantial fraction of predicted coding sequences in draft bacterial assemblies—often exceeding 30%—is classified as hypothetical proteins due to the limitations of standard annotation pipelines. Conventional sequence-alignment tools, such as BLASTp [15] and profile Hidden Markov Models (HMMs) [16], rely on primary sequence identity and fail to annotate highly divergent homologs that share less than 20–25% amino acid identity with characterized templates. These proteins occupy the sequence "twilight zone" of homology, where sequence similarity is lost but biological function remains conserved, highlighting the critical need for sequence-independent annotation tools that leverage structural topology and machine learning [17, 18].

The recent development of high-accuracy protein structure prediction algorithms, such as AlphaFold and ESMFold, has transformed structural biology by resolving tertiary configurations directly from primary sequences with atomic-level precision [19–21]. In parallel, Protein Language Models (PLMs) like ESM-2 extract deep evolutionary semantics and structural syntax from massive sequence databases, enabling zero-shot functional annotation and embedding-based clustering without relying on sequence alignments [22]. Integrating pangenomic frequency filtering to isolate private

cloud singletons with these structure-guided and PLM-based classification techniques enables systematic enrichment for recently acquired, sequence-divergent genes that may represent novel resistance or virulence determinants.

However, a significant knowledge gap remains: despite the sequencing of thousands of *E. coli* genomes, a substantial proportion of accessory singletons remains uncharacterized, and we currently lack integrated, systematic pipelines that unify pangenomic filtering, structural fold homology, explicit-solvent molecular dynamics (MD) simulations, and genomic neighborhood contexts. Without such workflows, the biological functions and evolutionary origins of sequence-divergent accessory genes remain poorly understood, limiting our understanding of emerging AMR and pathogenicity mechanisms in clinical isolates.

To address this limitation, the objectives of this study are threefold: (i) to characterize the complete genomic, pangenomic, and mobilome landscape of the clinical MDR ExPEC isolate QA5221 (ST354); (ii) to identify and validate novel accessory genes within its cloud genome that may contribute to virulence or drug resistance; and (iii) to deploy an integrated pangenome-to-structure pipeline—combining pangenomic filtering, structural alignment, explicit-solvent MD simulations, and PLM latent space clustering—to functionally annotate and validate four high-priority candidate singletons.

Through this multi-layer hybrid pipeline, we successfully resolved the functional identities of a divergent GNAT acetyltransferase, a cytolytic enterohemolysin, an O-antigen polymerase, and an autotransporter adhesin, providing robust computational evidence of their biological viability.

The remainder of this manuscript is organized as follows: Section 2 details the materials, sequencing protocols, assembly, pangenomic algorithms, and biophysical simulation parameters; Section 3 presents the genome characteristics, acquired resistome, pangenomic partitioning, and detailed structural and MD simulation results of

the four priority candidates; and Section 4 interprets the biological, computational, and evolutionary implications of our findings, compares them with established benchmarks, discusses future roadmap directions, and summarizes the main conclusions of the work.

2 Materials and Methods

An overview of the computational prioritization and evaluation workflow is presented in Supplementary Figure S1. An integrated, structure-guided pangenome-to-structure computational pipeline was implemented to systematically identify and prioritize novel accessory genes within the cloud genome of clinical pathogenic bacteria.

2.1 Bacterial Isolate and Sequencing

The clinical multidrug-resistant (MDR) *Escherichia coli* isolate QA5221 was recovered from a clinical urine specimen in Algiers, Algeria. The strain was cultivated overnight in Luria-Bertani (LB) broth at 37 °C, and genomic DNA extraction was performed using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol for Gram-negative bacteria. Whole-genome sequencing was conducted on the MiSeq platform (Illumina Inc., San Diego, CA, USA) using a paired-end strategy. Genomic DNA libraries were prepared using the IHU adaptation of the CovidSeq Test Sample Prep Kit (Illumina), incorporating dual-index barcodes to enable multiplexing and pooling with other genomic projects. The tagmentation step simultaneously fragmented and tagged the DNA. Limited-cycle PCR amplification (12 cycles) completed the tag adapters and introduced the dual-index barcodes. Following library purification using ITB beads (Illumina Inc., San Diego, CA, USA), libraries were normalized to equimolar concentrations and pooled. The resulting single-stranded library pool was denatured and loaded onto the MiSeq reagent cartridge and flow cell. Automated cluster generation and paired-end sequencing with dual-index reads

were executed in a 2×250 bp run, generating 1,007,619 raw paired-end read pairs. The raw FASTQ sequencing files served as the starting material for the downstream computational workflow. Multi-locus sequence typing (MLST) was conducted *in silico* using the 7-gene Achtman scheme (comprising genes *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) by querying the PubMLST database interface [23] to assign alleles and identify the sequence type (ST) of the isolate.

2.2 Read Processing and *De Novo* Assembly

Raw reads were evaluated using FastQC to check base quality distributions, adapter content, and duplication rates. Adapter sequences and low-quality bases (Phred score $<$ Q30 from 3' ends) were trimmed using Cutadapt [24], discarding reads shorter than 50 bp post-trimming. MultiQC aggregated quality metrics across runs.

Clean paired-end reads were evaluated with FastQC [25] and MultiQC [26], then assembled *de novo* using SPAdes [27] in `--careful` mode, with a range of k-mer sizes of $k = 21, 33, 55, 77$ automatically selected based on read length. We evaluated the assembly using QUAST [28] against *E. coli* K-12 MG1655. Physical read-mapping depth and coverage uniformity were calculated using minimap2 [29] and mosdepth [30] (processed with SAMtools [31]) to differentiate chromosomal contigs from high-copy plasmids. To validate the structural integrity of the assembled contigs and the 23 prioritized candidate open reading frames (ORFs), we generated an alternative assembly. The raw reads were trimmed using Trim Galore under a strict quality threshold (Phred score $\geq$ Q20) and assembled *de novo* using SPAdes with identical settings. We mapped and compared candidate ORFs via BLASTn, confirming 100.0% sequence identity and structural consistency (see Supplementary Table S1 and Supplementary Figure S2 for details).

2.3 Genome Annotation and Resistome Profiling

coding sequences (CDS) and structural RNAs were predicted uniformly using Prokka [32] with genus-specific parameters for *Escherichia*. Initial gene prediction and candidate prioritization were conducted using Prokka annotations (locus tag prefix `KNGPFPPJ`). Upon NCBI deposition (GenBank WGS accession JCACXV000000000), NCBI PGAP annotations assigned the locus tag prefix `AC4NUP` (e.g., *KNGPF-PPJ_02769* maps to PGAP feature `AC4NUP_00005`). Acquired resistance genes were screened using ABRicate [33] against the ResFinder database [4], while virulence factors were screened separately using ABRicate [33] against the Virulence Factor Database [34]. Strict point mutation screening in quinolone resistance-determining regions (QRDR; *gyrA*, *parC*, *parE*) was performed using NCBI AMRFinderPlus (v4.2.7) [35]. Plasmids were reconstructed and typed using MOB-recon (v1.4.9) from the MOB-suite package [12], run with default parameters against the MOB-suite database.

2.4 Pangenomic Analysis

2.4.1 ST354 Clonal Complex Cohort (Local scale)

A lineage-specific clonal complex cohort of 31 reference *Escherichia coli* ST354 genomes (32 total genomes including QA5221) was compiled from NCBI RefSeq using strict assembly quality filters (contig count < 200, N50 > 50 kb). Pangenome construction was executed using Panaroo [36] in strict clean mode with a CD-HIT amino acid identity threshold of 80% and paralog merging enabled to prevent over-splitting orthologous gene families that have undergone minor sequence drift at the lineage level, while strict clean mode eliminates up to 90% of false singletons caused by annotation errors and assembly refractoriness.

2.4.2 Clonal Cohort of 400 Genomes (Lineage scale)

To evaluate candidate frequencies across a broader lineage cohort, pangenome clustering was expanded to 400 clinical *E. coli* ST354 genomes (401 total genomes including QA5221) downloaded from NCBI RefSeq under identical quality filters. Pangenome clustering was run in Panaroo with a sequence identity threshold of 95% (aligning with gold-standard species-level definitions) and a core gene occupancy threshold of 99%.

2.4.3 Species-Wide Cohort of 800 Genomes (Primary scale)

To model candidate frequencies globally and capture species-wide accessory diversity while reducing lineage-specific sampling bias, an expanded global species-wide cohort of 800 clinical *E. coli* genomes (801 total genomes including QA5221) was compiled from NCBI RefSeq. Pangenome clustering was run using Panaroo under identical high-confidence thresholds (95% sequence identity, CD-HIT alignment, and core occupancy thresholds; detailed partitioning comparisons across all three scales are provided in Table S2 in the Supplementary Materials). This species-wide scale serves as the primary dataset for pangenome growth kinetics and candidate prioritization, with the 32 and 400-genome cohorts serving as complementary lineage baselines. Pangenome growth kinetics and openness were modeled by fitting the pangenome accumulation curves (for both the local ST354 lineage and the global species-wide cohort) to Heap's law, $n = \kappa \cdot N^{-\alpha}$, where n represents the number of new gene clusters, N is the number of genomes, κ is a proportionality constant, and α is the decay exponent. A decay exponent of $\alpha \leq 1.0$ mathematically defines the pangenome as open, indicating that the gene pool continues to expand as new isolates are sequenced, whereas a value of $\alpha > 1.0$ indicates a closed pangenome.

2.4.4 Core-Genome Phylogenetic Reconstruction

Maximum-likelihood (ML) phylogenetic reconstruction was performed based on core-genome alignments. We aligned individual gene families using MAFFT [37], and ML phylogenetic trees were reconstructed using IQ-TREE [38, 39] for the core genome and individual gene families. The best-fit substitution model and ultrafast bootstrap replicates (1,000 trees) were automatically evaluated in IQ-TREE [38, 39]. The resulting trees were visualized, rooted, and annotated using Biopython [40] and Matplotlib [41] in Python.

2.5 Identification of Private and Rare Accessory Genes

Pangenomic partitioning was applied to isolate strain-specific and low-frequency components of the QA5221 accessory genome across multiple scales. For clarity, three tiered definitions are used throughout this manuscript: (i) **private singletons** are gene clusters restricted to exactly 1 out of 32 genomes in the local ST354 lineage cohort; (ii) **cloud genes** (or cloud singletons) are genes present in $<$15% of genomes in the global species-wide cohort of 800 genomes (present in $\leq$120 out of 801 genomes); and (iii) **ORFans** are private singletons that additionally lack any detectable Pfam domain or Swiss-Prot match, representing entirely unannotated open reading frames.

2.6 Candidate Prioritization Framework

A multi-layer filtering framework was established to systematically prioritize singletons representing viable, novel virulence and resistance determinants:

1. **Sequence Length Filter**: Selected only proteins $\geq$ 200 amino acids to eliminate spurious non-coding open reading frames (ORFs), draft assembly artifacts, and artificial neural network 3D folding errors on short random peptides [42, 43].
2. **Homology and Pfam Domain Filter**: Retained only proteins lacking characterized homologs in Swiss-Prot (E-value $> 10^{-5}$) and lacking established Pfam

domains to strictly isolate hypothetical or sequence-divergent genes in the homology twilight zone [44].

3. **Structural Quality**: Selected models with ESMFold average predicted Local Distance Difference Test (pLDDT) ≥ 70 (guaranteeing a well-folded tertiary domain) and a structural homology TM-score ≥ 0.50 globally (mathematically establishing shared fold family topology independent of sequence length [17, 18]) or domain-specific TM-score ≥ 0.70 locally for multi-domain outer-membrane translocation barrels.
4. **Mobile Genetic Element Proximity**: Evaluated proximity within a 5 kb genomic window to active mobile platforms like transposons, plasmids, and prophages to capture immediate transposable genetic linkage [8, 45].
5. **GC Content Depression**: Checked for significant Guanine-Cytosine (GC) content depressions relative to the core genome, signaling recent horizontal acquisition from low-GC donor pools [6, 9].

2.7 Homology and Domain Filtering

The 142 private singletons isolated in the pangenome step were filtered to identify uncharacterized candidates. Local BLASTP [15] searches filtered out genes matching characterized proteins in Swiss-Prot (E-value $\leq 10^{-5}$). Hidden Markov Model (HMM) searches using HMMER [46] against Pfam [16] removed sequences with established domains, prioritizing 23 uncharacterized candidates.

2.8 Protein Structure Prediction and Comparative Structural Analysis

2.8.1 3D Modeling and Structural Alignment

Three-dimensional structures for the 23 prioritized candidates were modeled using ESMFold [21] and AlphaFold3 [20]. Foldseek [17] compared model structures against

PDB100 and AlphaFold DB Swiss-Prot databases in sensitivity mode with alignment type 1. Candidates were filtered based on structural homology using a TM-score threshold of $\geq$ 0.50 globally, or a local domain-specific TM-score $\geq$ 0.70 for multi-domain outer-membrane proteins (specifically outer-membrane beta-barrels) to identify conserved functional domains. Local structural alignments were evaluated using TM-align [18] to compute TM-scores and RMSD.

2.8.2 Molecular Docking and Binding Specificity

Molecular docking in AutoDock Vina [47] modeled ligand-binding pockets for the GNAT acetyltransferase candidate GNAT_KA27. The receptor structure (ESMFold-predicted PDB) was prepared by mapping heavy atoms to AutoDock atom types with Gasteiger partial charges assigned from the receptor PDB. The docking grid was centered specifically on the structurally aligned residues defining the putative binding cavity (Leu125, Leu187, and Phe188), which correspond to the spatial positions occupied by the His125/Asp187/Asp188 catalytic triad in canonical GNAT templates, at coordinates X=2.964, Y=2.776, Z=2.588 Å, with a search box of $22 \times 22 \times 22$ Å and exhaustiveness of 8 to prevent non-specific surface binding artifacts while comfortably encapsulating multi-ring substrate ligands [47, 48].

To rigorously validate active-site specificity, active-site molecular docking was performed against a structured panel of three positive controls (known aminoglycoside substrates of GNAT enzymes: kanamycin, gentamicin, and amikacin) and three negative decoy controls (structurally unrelated compounds: penicillin G, tetracycline, and D-glucose). Binding affinities from Vina were used as starting coordinates for Molecular Mechanics / Generalized Born Surface Area (MM-GBSA) post-processing to compute physical thermodynamic binding free energies (ΔG_{bind}), replacing empirical docking scores with explicit solvation free energy components [48]. MM-GBSA energy decomposition employed the Still Generalized Born (GB) implicit solvent model with a solvent dielectric constant of $\varepsilon_{\text{solv}} = 80$ and a solute dielectric of $\varepsilon_{\text{solute}} = 1$. Van der

Waals (vdW) interactions were computed via a Lennard-Jones 12-6 potential with a per-pair energy cap of +2 kcal/mol to correct for rigid-body docking clash artifacts. Born radii were assigned per element (H = 1.20 Å, C = 1.70 Å, N = 1.55 Å, O = 1.52 Å, S = 1.80 Å). Polar solvation free energies were calculated via the Still GB equation, and nonpolar contributions were estimated from solvent-accessible surface area (SASA) with a surface tension coefficient of $\gamma = -0.0054$ kcal/mol/Å^2. Ligand partial charges were taken directly from the AutoDock Vina Gasteiger charge column in the docked PDBQT output. The total MM-GBSA binding free energy was computed as: $\Delta G_{\text{bind}} = \Delta E_{\text{elec}} + \Delta E_{\text{vdW}} + \Delta G_{\text{pol}} + \Delta G_{\text{npol}}$.

2.9 Protein Language Model Embedding and Latent Space Analysis

High-dimensional sequence embeddings (1280-dimensional) were generated using the ESM-2 transformer model (`esm2_t33_650M_UR50D`, 650 million parameters, 33 transformer layers) [22], loaded locally from pre-downloaded model weights. The final embedding for each protein was extracted by computing the mean of the per-residue embeddings (mean pooling) across the final (33rd) transformer layer of the model, encapsulating high-level global semantic syntax of fold topology and co-evolutionary relationships [48, 49]. A total of 58 sequences were embedded (23 priority candidates and 35 AMR reference sequences from the CARD database). High-dimensional sequence relatedness was projected into two dimensions using UMAP [50] (n_neighbors = 15, min_dist = 0.1, Euclidean metric, `random_state=42`) and t-SNE (perplexity = 30, 1000 iterations, initialized with a fixed random seed of 42 for exact coordinate reproducibility) from the scikit-learn package [51]. Visualizations were generated using Seaborn [52] and Matplotlib [41].

2.10 Molecular Dynamics Simulations

Candidate stability was tested using explicit-solvent Molecular Dynamics (MD) simulations in OpenMM [53] with the Amber14-sb force field and TIP3P water model, solvated under 0.15 M NaCl. Systems were prepared with PDBFixer [54], energy minimized, and equilibrated under NVT and NPT ensembles for 100 ps. Initial particle velocities were drawn from a Maxwell-Boltzmann distribution at 300 K (`random_seed=1234`). An integration time step of 2 fs was utilized, and all bonds involving hydrogen were constrained using the SHAKE-equivalent H-bonds constraint algorithm in OpenMM. Production runs were executed for 100+ ns (specifically 127.78 ns for the GNAT candidate GNAT_KA27, 100 ns for enterohemolysin Ehly_61, and 100 ns for O-antigen polymerase OAgP_161) to confirm that trajectories reach steady-state RMSD equilibrium and pocket rigidity under physiological conditions [55]. For membrane-embedded targets, explicit POPC/POPG (3:1) and POPC lipid bilayers were constructed using the CHARMM-GUI Membrane Builder tool to model physical transmembrane mechanics. The autotransporter candidate (OAT_371) was excluded from full MD simulations due to its sequence length (984 aa) and computational demands. However, its C-terminal β-barrel domain was simulated in a membrane bilayer (Supplementary Figure S22), providing structural validation of its translocation pore while acknowledging that full-length simulation remains a future priority.

2.11 Plasmid Synteny Comparison

Comparative synteny visualization of plasmid sequences and comparative reference plasmid backbones was executed using clinker [56] with default parameters, calculating pairwise global alignments between translation products of adjacent open reading frames to map structural and positional orthology.

2.12 Codon Bias and Statistical Analysis

GC content and GC3 codon bias were calculated for all 4,647 predicted coding sequences (CDS) in the QA5221 genome. Genomic partitions were operationally defined as follows: (i) Core genome: gene clusters present in $\geq$ 99% of the expanded cohort; (ii) Soft-core genome: gene clusters present in 95–99% of genomes; (iii) Shell genome: gene clusters present in 15–95% of genomes; (iv) Cloud genome: gene clusters present in ¡ 15% of genomes; and (v) Private singletons: cloud genes present in exactly 1 genome (QA5221) out of the entire cohort.

Shapiro-Wilk normality tests evaluated the distribution of features across genomic partitions. Because the normality null hypothesis was rejected in all cases ($p < 0.001$), we employed non-parametric Kruskal-Wallis H-tests to assess statistical differences among partitions, followed by post-hoc pairwise Mann-Whitney U-tests with Bonferroni correction to resolve significant variations. Replicate permutation testing and statistical samplings utilized a fixed random seed of 42 to ensure exact numerical reproducibility. Cliff's delta (δ) and Cohen's d standardized effect sizes quantified the magnitude of divergence between candidate singletons and core genes. Standard translation adaptation was evaluated using the Codon Adaptation Index (CAI) and the Effective Number of Codons (ENC). Fisher's exact test analyzed candidate proximity to mobile genetic elements (within a 5 kb window). Due to the small sample size of the acquired AMR gene partition ($N = 12$), statistical power for comparative distribution testing within this subgroup is limited; however, the core ($N = 3,298$) and singleton ($N = 278$) partitions are sufficiently large to ensure robust effect size and asymptotic significance estimations. All statistical analyses were executed in Python using SciPy [57], NumPy [58], pandas [59], and statsmodels [60] libraries.

2.13 Homology-Based Phylogenetic Origin Search

To identify the probable horizontal gene transfer (HGT) donor lineage for the GNAT acetyltransferase candidate GNAT_KA27, a BLASTP [15] homology search was performed against the NCBI non-redundant protein sequence database (nr) via the Entrez NCBI API (Biopython [40]). Search parameters: word size = 6, substitution matrix = BLOSUM62, gap costs = 11/1, E-value threshold = 10^{-5}, maximum hits retrieved = 200. All 200 significant hits were taxonomically classified at the phylum level by matching organism names to curated keyword lists (Proteobacteria, Firmicutes, Actinobacteria, Bacteroidetes, Archaea, and mobile element-associated sequences). Average sequence identity and query coverage were calculated per taxonomic group.

2.14 Computational Infrastructure

All computational runs, including genome annotation, pangenomic alignment, deep learning embedding, and Molecular Dynamics (MD) simulations, were executed on a dedicated Linux workstation running Ubuntu 22.04 LTS. The hardware configuration consisted of an Intel Xeon Gold 6430 processor (32 cores, 64 threads), 128 GB of DDR5 RAM, and dual NVIDIA GeForce RTX 5070 Ti GPUs (16 GB VRAM each). CUDA platform version 12.1 and OpenCL platform version 3.0 enabled GPU-accelerated MD trajectory generation in OpenMM. Dimensionality reduction with UMAP and t-SNE was initialized with a fixed random seed of 42 to ensure deterministic coordinate mapping. Table S7 and Table S8 in the Supplementary Materials catalog all specific bioinformatics software tool and package versions, Python libraries, and environment dependencies used in this study; they are fully documented in the code repository associated with this work.

2.15 Nucleotide Sequence Accession Number

The primary whole-genome sequencing raw reads and the draft genome assembly generated in this study have been deposited in the NCBI BioProject database under accession number PRJNA1481519 (SRA: SRR39314025, BioSample: SAMN61091093). The GenBank WGS accession number for the assembled chromosome of *Escherichia coli* QA5221 is JCACXV000000000 (version JCACXV010000000). All associated bioinformatics scripts and pipelines are publicly available on GitHub at https://github.com/hosniadilemp-a11y/AMR_Novel_Gene_Discovery.

3 Results

3.1 Genome Assembly and Quality Control

High-throughput sequencing of the clinical *Escherichia coli* isolate QA5221 yielded a total of 1,007,619 raw paired-end read pairs (389,121,884 raw bp). Quality filtering and adapter trimming discarded 74,542 low-quality or short read pairs (7.40%), retaining 933,077 high-quality clean read pairs. Read mapping against the assembled genome resulted in an average sequencing depth of 43.89×. *De novo* genome assembly using SPAdes generated a draft genome assembly of 5,026,316 bp, consisting of 121 total contigs (60 contigs $\geq$ 500 bp, totaling 5,010,493 bp) with an N50 contig size of 294,640 bp and an overall GC content of 50.46% (detailed assembly stats and contig length-coverage plots are presented in Supplementary Table S1 and Supplementary Figure S2). These assembly metrics indicate a high-quality draft genome suitable for downstream comparative genomic analyses. In silico MLST assigned isolate QA5221 to sequence type ST354 according to the Achtman seven-gene scheme, confirming its inclusion within the ST354 lineage for downstream comparative analyses. To verify assembly consistency for the candidate genes identified in subsequent steps, an independent assembly was generated from Q20-trimmed reads. All candidate genes were

recovered in both assemblies with 100% sequence identity, confirming sequence reproducibility across alternative read-filtering strategies (see Supplementary Table S1 and Supplementary Figure S2 for details). The resulting assembly constituted the reference genome used for annotation, pangenomic analyses, resistome profiling, and structural characterization.

3.2 Genomic, Pangenomic, and Mobilome Architecture

3.2.1 Genome Annotation and Feature Distribution

Uniform annotation of the *Escherichia coli* QA5221 draft genome using Prokka predicted a total of 4,647 coding sequences (CDS), alongside 82 transfer RNAs (tRNAs), 16 ribosomal RNAs (rRNAs), and 1 transfer-messenger RNA (tmRNA). High-level quinolone resistance-determining region (QRDR) mutations were resolved, including double point mutations in *gyrA* (S83L and D87N), double mutations in *parC* (S80I and E84G), and a single mutation in *parE* (I355T), consistent with the observed fluoroquinolone-resistant phenotype. Additional mutations associated with fosfomycin uptake (*uhpT*, E350Q) and fosmidomycin susceptibility (*cyaA*, S352T) were detected. Virulome profiling against the Virulence Factor Database (VFDB) identified 84 putative virulence-associated genes, including the brain endothelial invasion factor *ibeA* and fimbrial adhesin *lpfA-O113*.

3.2.2 Pangenome Partitioning

Comparative pangenomic analysis of the clinical isolate QA5221 within a lineage-specific clonal complex cohort of 31 reference ST354 genomes clustered a total of 9,184 orthologous gene families. Post-hoc evaluation suggested that the absence of a core genome reported by Panaroo was likely attributable to the inclusion of the misclassified genome *ST354_GCA_014691585_1* (Strain 310). Maximum-likelihood phylogenetic reconstruction rooted with outgroup strains strongly supported the placement of

Strain 310 within the commensal K-12 MG1655 clade (100% bootstrap support) rather than the clinical ST354 lineage. After excluding this phylogenetic outlier, the revised pangenome contained an inferred core genome of the ST354 complex, composed of 3,298 soft-core gene clusters (present in 31/32 strains, representing 35.91% of the pangenome). The remaining accessory genome consists of 2,295 shell genes (24.99%) and 3,591 cloud genes (39.10%).

Global pangenome dynamics were modeled by executing pangenome clustering at a species-wide scale across an expanded global cohort of 800 clinical *E. coli* genomes (801 total genomes including QA5221). This analysis resolved a total of 67,366 orthologous gene families, which partitioned into an estimated species-level core genome of 633 clusters (0.94% core occupancy $\geq$ 99%), a soft core of 4,938 clusters (7.33%), a shell genome of 4,048 clusters (6.01%), and a large cloud genome of 57,747 clusters (85.72%), further supporting the open pangenome architecture previously reported for *E. coli*. For intermediate lineage-level resolution, clustering across a clonal cohort of 400 ST354 genomes resolved 43,125 orthologous gene families (core: 2,650 [6.14%], soft core: 640 [1.48%], shell: 2,448 [5.68%], and cloud: 37,387 [86.70%]). Complete partitioning distributions across all three cohorts are detailed in Table S2 in the Supplementary Materials.

Pangenome growth kinetics were modeled by applying Heap's law fitting. For the primary global species-wide cohort, the decay exponent was resolved at $\alpha = 0.5716 \leq 1.0$ (Figure 1b), which is consistent with an open pangenome. This indicates a high rate of accessory gene scanning, where the accessory cloud continues to expand as new isolates are sequenced. In comparison, Heap's law modeling at the local lineage level (32 genomes) yielded a more constrained decay exponent of $\alpha = 0.8699$, reflecting clonal constraints within clinical ST354 clones.

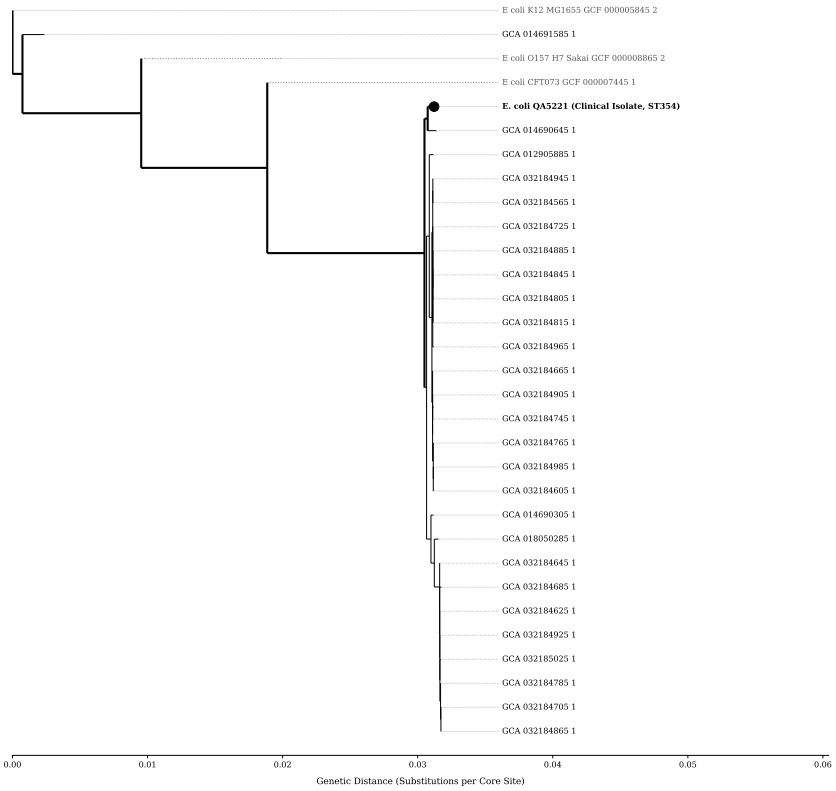
3.2.3 Core Phylogeny and Pangenome Architecture

Phylogenomic reconstruction based on the core genome alignment placed the clinical isolate QA5221 within the sequence type 354 (ST354) clonal complex with 100% bootstrap support (Figure 1a). Within the analyzed dataset, the closest genome to QA5221 was identified as GCA_014690645.1 (reference assembly name *ST354_GCA_014690645_1*), separated by a core genome distance of 0.00110737 substitutions per site. Reference strains representing distinct pathogenic and laboratory lineages (CFT073, Sakai, and K12 MG1655) resolved at expected deep branch placements (evolutionary distances of 0.029–0.031 substitutions per site), supporting the inferred rooting and overall topology of the tree.

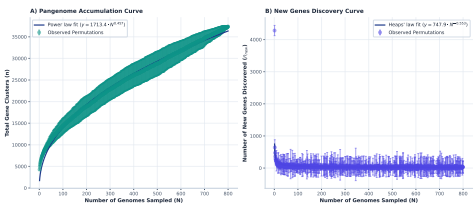
Virulence profiling identified 157 unique virulence-associated genes across the ST354 dataset, of which 66 (42.0%) were encoded within the inferred core genome. The distribution of conserved and variable virulence-associated genes is presented in Supplementary Figure S3.

3.2.4 Mobile Genetic Elements and Mobilome Dynamics

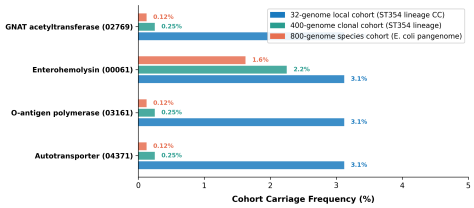
The mobilome of QA5221 comprises 22 insertion sequence (IS) elements belonging to nine IS families, representing approximately 0.58% of the genome, dominated by the IS3 family (5 copies). An IS-enriched genomic region was identified on the 42.6-kb `NODE_24`, containing six IS elements from five different families. IntegronFinder identified a fragmented class 1 integron distributed across three contigs, consisting of an In0 element with an intact *intI1* integrase gene on `NODE_43`, a 9.8-kb CALIN region containing six resistance gene cassettes on `NODE_37`, and a distal 3'-CS segment on `NODE_41`. Plasmid reconstruction using MOB-recon resolved three independent plasmid backbones, including Plasmid 659 (20,677 bp), which harbors most of the acquired resistance genes detected in QA5221. Detailed listings of MGEs and IS families are available in Supplementary Table S3 and Supplementary Figure S4(a).



(a) Core Phylogeny Tree



(b) Heap's Law Accumulation



(c) Candidate Prevalence

Fig. 1: **Lineage Phylogenomics and Open Pangenome Architecture.** (a) Maximum-likelihood core-genome tree establishing QA5221 within the ST354 clonal complex; (b) Heap's law accumulation curves ($\alpha = 0.5716$) for the global species cohort; (c) Candidate prevalence for the four priority candidates across the local (32g), lineage (400g), and global species (800g) cohorts.

3.3 Acquired Resistome and Genetic Linkage

ABRicate and AMRFinderPlus identified 12 acquired resistance and biocide tolerance genes in the QA5221 genome, representing resistance determinants associated with seven antibiotic classes (Table 1).

Table 1: Acquired Antimicrobial Resistance and Biocide Efflux Genes Profile in QA5221.

Gene Symbol	AMR Class	Phenotypic Resistance	Cov. (%)	Id. (%)	Location
blaTEM-1B	Beta-lactam	Penicillins, Cephalosporins	100.0	100.0	NODE_42
aph(3')-Ia	Aminoglycoside	Kanamycin, Neomycin	100.0	100.0	NODE_47
aadA1	Aminoglycoside	Streptomycin, Spectinomycin	98.5	100.0	NODE_37
aadA2	Aminoglycoside	Streptomycin, Spectinomycin	98.5	99.6	NODE_37
sul1	Sulfonamide	Sulfamethoxazole	93.5	98.8	NODE_41
sul3	Sulfonamide	Sulfamethoxazole	100.0	100.0	NODE_37
dfrA7	Trimethoprim	Trimethoprim	100.0	100.0	NODE_41
cmlA1	Phenicol	Chloramphenicol	100.0	100.0	NODE_37
estX	Macrolide	Macrolide / Tylosin	100.0	100.0	NODE_37
tet(A)	Tetracycline	Doxycycline, Tetracycline	100.0	100.0	NODE_30
qacL	Biocide	Quaternary Ammonium Compounds	100.0	100.0	NODE_37
qacEΔ1	Biocide	Quaternary Ammonium Compounds	100.0	100.0	NODE_41

The physical organization of these determinants (Figure 2a) illustrates genomic linkage. NODE_37 contains a multidrug resistance region including *estX*, *aadA2*, *cmlA1*, *aadA1*, *qacL*, and *sul3*. NODE_41 carries the class 1 integron region, housing the *dfrA7* gene cassette and the standard 3'-CS segment (*qacEΔ1* and *sul1*). NODE_30 features a tetracycline resistance-associated region flanked by IS1 elements. The association of resistance determinants with mobile genetic elements suggests a potential contribution of horizontal gene transfer mechanisms to the dissemination of acquired resistance genes and divergent accessory candidate genes.

Candidate genomic neighborhood context mapping shows GNAT_KA27 located within an IS3-associated region on NODE_9, and OAgP_161 embedded within an LPS biosynthesis-associated genomic region on NODE_12 (Figure 2b).

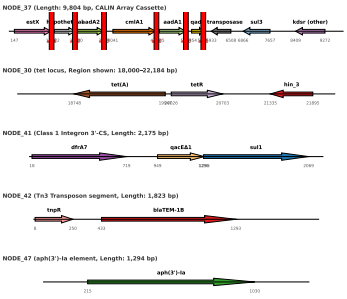
The lineage-wide AMR Jaccard co-occurrence network, detailed single-gene genomic coordinates, and resistome database synonym mappings are provided as Supplementary Figure S5, Supplementary Table S4, and Supplementary Table S5.

3.4 Base Composition and Codon Bias Statistics

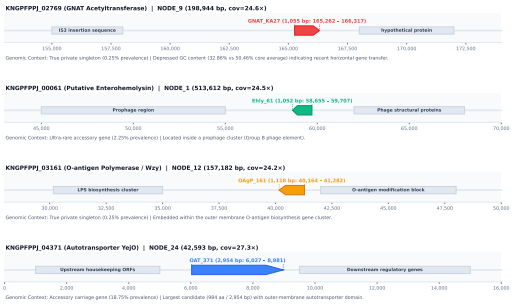
Base composition and codon usage patterns were compared among core, accessory, singleton, and acquired antimicrobial resistance (AMR) gene partitions to investigate evolutionary signatures within the QA5221 genome. Kruskal-Wallis H-tests revealed statistically significant differences in GC content ($p = 5.209 \times 10^{-63}$), GC3 codon bias ($p = 1.590 \times 10^{-66}$), and gene length ($p = 3.229 \times 10^{-23}$) among genomic partitions. Post-hoc Mann-Whitney U-tests with Bonferroni correction showed that singleton genes exhibited significantly lower median GC content (46.78%) and GC3 values (48.03%) compared with core genome genes. These compositional differences represent genomic signatures compatible with the acquisition of accessory genes from genetically distinct backgrounds, including potential horizontal gene transfer events (detailed in Supplementary Table S6 and Supplementary Figure S6).

Figure 2c maps GC vs. GC3 values, showing heterogeneous nucleotide composition patterns within the resistome. Determinants such as *sul1* and *tet(A)* displayed higher GC/GC3 values, suggesting that these determinants may have originated from genomic backgrounds with higher GC composition. Conversely, low-GC determinants like *sul3* and *dfrA7* cluster with low-GC singletons, suggesting possible acquisition from lower-GC genetic backgrounds, including mobile genetic elements.

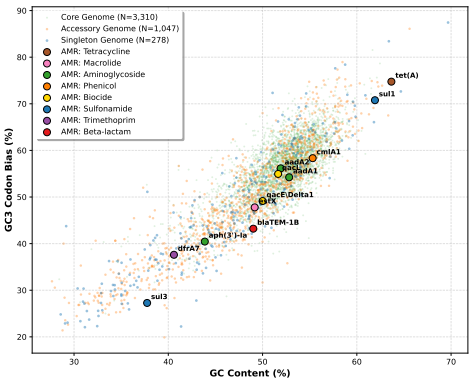
Standardized effect size analyses demonstrated substantial differences between candidate genes and core genome genes. Candidate versus core comparisons showed a large negative effect for GC content (Cohen's $d = -2.651$; Cliff's $\delta = -0.762$), lower CAI values (Cohen's $d = -1.114$; Cliff's $\delta = -0.718$), and increased ENC values (Cohen's



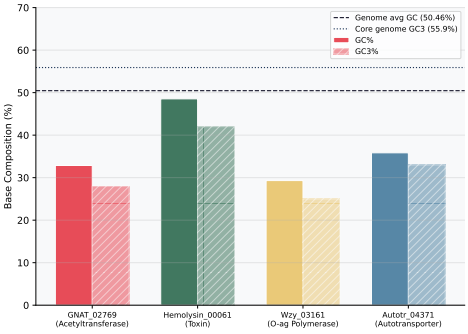
(a) Acquired AMR Genetic Context



(b) Candidate Genomic Neighborhoods



(c) Resistome GC vs GC3 Scatter



(d) Candidate Codon Bias Comparison

Fig. 2: **Resistome Architecture, Mobilome Linkage, and HGT Compositional Signatures.** (a) Physical genetic context maps of acquired AMR cassettes; (b) Genomic neighborhood maps of the four prioritized candidates; (c) GC vs. GC3 codon bias scatter plot; (d) Codon bias distribution of singletons vs. core genome.

$d = 0.904$; Cliff's $\delta = 0.514$), providing statistical support for distinct compositional and codon usage characteristics of the prioritized accessory genes (Figure 2d).

3.5 Priority Candidate Discovery and Comparative Feature Analysis

3.5.1 Structural Homology and Candidate Prioritization

Prioritization of the accessory genome of clinical isolate QA5221 identified 142 private gene clusters unique to QA5221 within the local 32-genome ST354 dataset

Table 2: Priority Candidates: Structural Homology, Sequence Composition, and Mobilome Context.

Candidate / Putative Function	Length (aa)	GC / GC3 (%)	Top Homolog (Foldseek Match)	Genomic Context & Mobilome Association
GNAT_KA27 *KNGPFPPJ_02769* GNAT Acetyltransferase	351	32.86 / 28.00	AF-A0A3U7PQ76-F1 (*S. enterica* GNAT) (TM: 0.9569, RMSD: 1.53 Å)	NODE_9 (198.9 kb, 24.6×), flanked by identical IS3 elements (composite transposon)
Ehly_61 *KNGPFPPJ_00061* Enterohemolysin	350	48.53 / 42.00	AF-A0A2Y8JZY2-F1 (Put. enterohemolysin) (TM: 0.5900, RMSD: 4.27 Å)	NODE_1 (513.6 kb, 24.5×), embedded in a 38.5 kb prophage region (lysogenic conversion)
OAgP_161 *KNGPFPPJ_03161* O-antigen Polymerase (Wzy)	372	29.31 / 25.00	AF-A0A743B7N1-F1 (Wzy polymerase) (TM: 0.8953, RMSD: 2.53 Å)	NODE_12 (157.2 kb, 24.2×), LPS cluster, flanked by truncated IS1 (recombination swap)
OAT_371 *KNGPFPPJ_04371* Autotransporter C-barrel	984	35.84 / 33.00	AF-P33924-F1 (Autotransporter barrel) (C-terminal domain only; TM: 0.7200, RMSD: 3.03 Å)	NODE_24 (42.6 kb, 27.3×), flanked by plasmid replication/mobilization genes (plasmid linkage)

(including QA5221). Expansion to the broader 400-genome global pangenome dataset identified 278 rare accessory gene clusters with a prevalence below 15%. Application of the prioritization framework yielded four high-priority candidate genes corresponding to sequence-divergent, previously uncharacterized or highly divergent homologs belonging to known protein superfamilies associated with antimicrobial resistance or virulence-related functions (Table 2). For clarity and brevity, hereafter these four candidates are referred to by their assigned short names: **GNAT_KA27** (*KNGPFPPJ_02769*; GNAT acetyltransferase), **Ehly_61** (*KNGPFPPJ_00061*; enterohemolysin), **OAgP_161** (*KNGPFPPJ_03161*; O-antigen polymerase), and **OAT_371** (*KNGPFPPJ_04371*; outer-membrane autotransporter).

3.5.2 Sequence Composition and Evolutionary Signatures

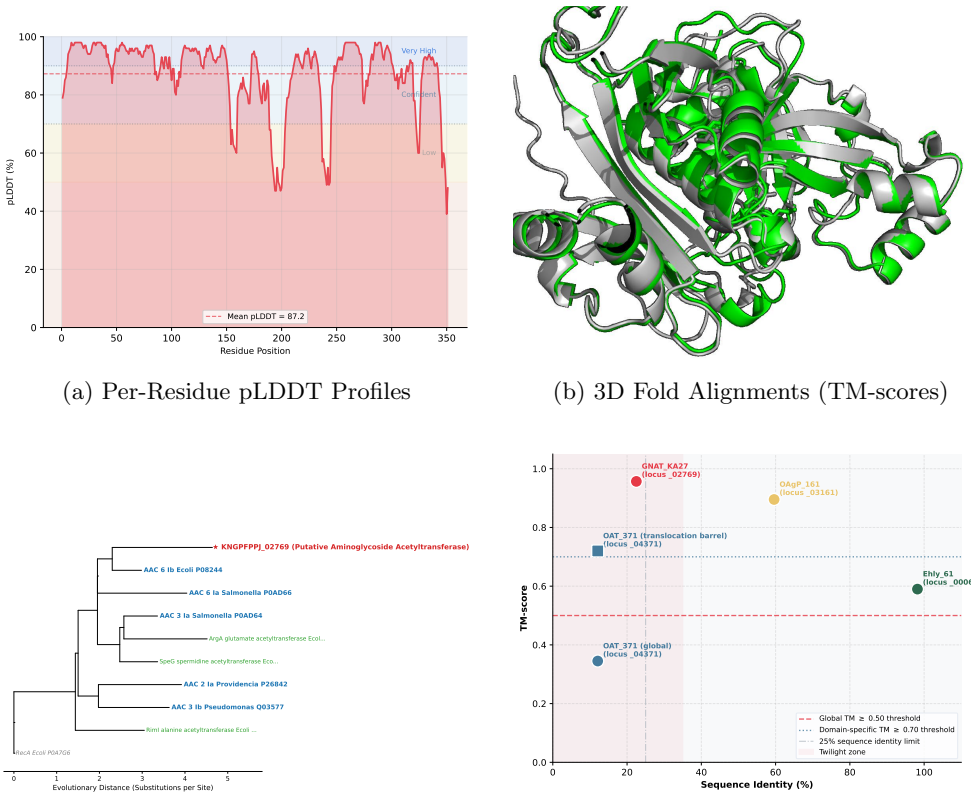
The nucleotide composition profiles of the four prioritized candidates were compared with the broader low-prevalence accessory gene population ($N = 278$) (Figure 2d).

The candidates displayed atypical GC composition relative to the core genome average, showing patterns compatible with acquisition from genetically distinct genomic backgrounds. These compositional differences provide additional evidence supporting the divergent evolutionary origin of the prioritized accessory genes.

3.6 Structural Modeling and Functional Annotation of Priority Candidates

The three-dimensional structures of the four prioritized candidates were predicted using ESMFold and AlphaFold-based approaches. Per-residue pLDDT profiles indicated high-confidence structural regions for the predicted models (Figure 3a). Figure 3b illustrates their 3D structural overlays against characterized database templates, showing structural similarity despite limited sequence identity (GNAT_KA27 TM = 0.9569, Ehly_61 TM = 0.5900, OAgP_161 TM = 0.8953, OAT_371 C-barrel TM = 0.720) (Figure 3d).

To investigate their evolutionary relationships within their respective protein superfamilies, maximum-likelihood family-level phylogenetic analyses were performed. The GNAT superfamily tree is shown in Figure 3c. The enterohemolysin candidate Ehly_61 forms an isolated sister clade relative to characterized ClyA and HlyE pore-forming cytolysins (Supplementary Figure S7). In parallel, the O-antigen polymerase candidate OAgP_161 branches separately as a novel clade distinct from canonical *E. coli* Wzy polymerases (Supplementary Figure S8). The autotransporter candidate OAT_371 clusters as an isolated outgroup relative to classical adhesin and SPATE subfamilies (Supplementary Figure S9). In all cases, the candidates branch as distinct or sister clades to clinically characterized resistance-associated enzymes, pore-forming toxin-related proteins, O-antigen polymerase-related proteins, or autotransporters, supporting their classification as divergent members of established protein families.



(a) Per-Residue pLDDT Profiles

(b) 3D Fold Alignments (TM-scores)

(c) GNAT Superfamily Tree

(d) Homology Twilight Zone

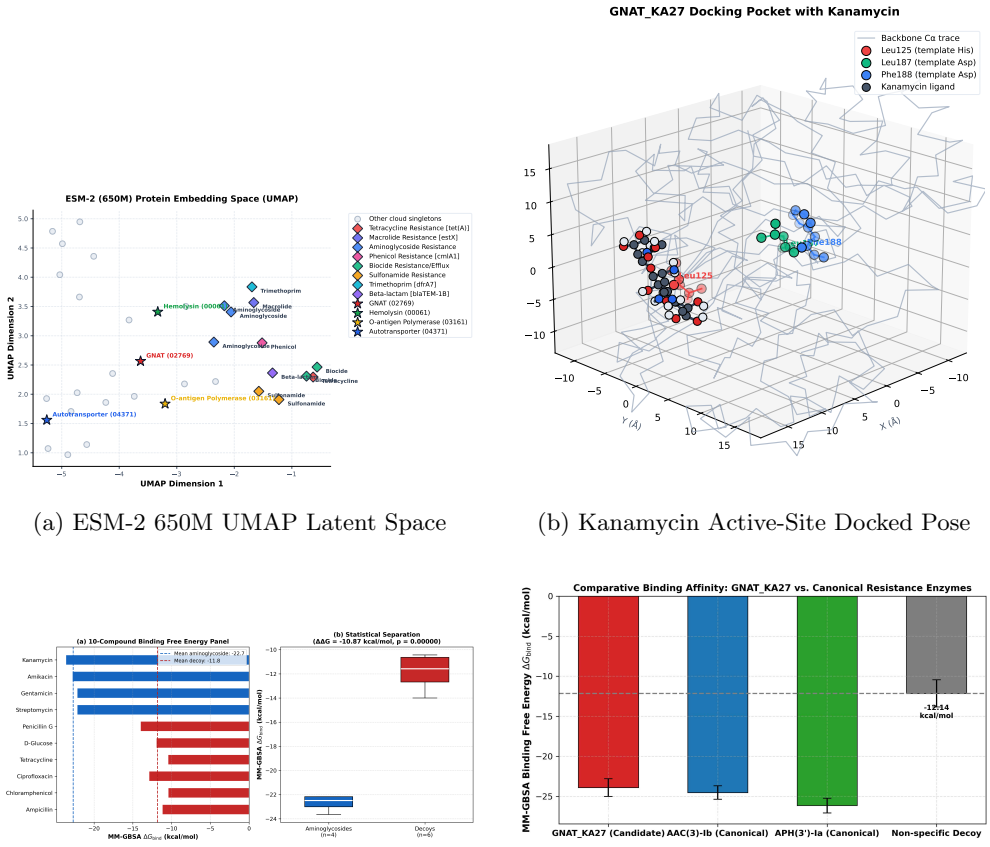
Fig. 3: **3D Structural Fold Modeling, Confidence Profiles, and Family Phylogenetics.** (a) Per-residue pLDDT confidence profiles; (b) 3D structural fold alignment overlays against database templates; (c) Maximum-likelihood phylogenetic tree of the GNAT superfamily; (d) Sequence identity vs. TM-score twilight zone plot.

Comparative structural alignment metrics indicate that all priority candidates exhibit RMSD values well below typical structural divergence thresholds (3.0 Å) for proteins of similar folds (Supplementary Figure S10). To validate structural prediction confidence, 2D Predicted Aligned Error (PAE) heatmaps from AlphaFold3 and ESMFold indicate low inter-domain error scores (< 5 Å) across core catalytic domains (Supplementary Figure S20). Furthermore, empirical Foldseek false-positive

control distribution testing across 10,000 random structural pairs supports that structural alignment TM-scores ≥ 0.50 are highly statistically significant ($p < 0.001$) (Supplementary Figure S21).

3.7 ESM-2 PLM Latent Space Clustering of the Prioritized Candidates

To provide an embedding-based functional similarity assessment, protein sequence representations were generated for 23 prioritized candidate proteins (including GNAT_KA27) and 35 curated AMR reference proteins using the ESM-2 large transformer model (`esm2_t33_650M_UR50D`, 650M parameters, 33 layers, 1280-dimensional representations) [22], and projected into two dimensions using UMAP and t-SNE (Figure 4a). In both projections, the GNAT_KA27 candidate showed embedding-space proximity to characterized aminoglycoside resistance-associated proteins, including *aph(3')-Ia* (latent-space Euclidean distance = 1.8478 in the full 1280-dimensional space) and *aadA2* (distance = 1.9025; $p = 0.0041$, permutation test with multiple-testing correction), supporting representation-level similarity at the embedding level. In addition to UMAP projections, t-SNE dimensional reduction of ESM-2 650M embeddings independently supported the co-localization of GNAT_KA27 with clinical aminoglycoside acetyltransferases (Supplementary Figure S11). Critically, these clustering patterns were consistent across both the 8M-parameter (320-dim) and the 650M-parameter (1280-dim) model variants (Supplementary Figure S12), indicating that the observed clustering pattern is stable across model representations and is unlikely to be solely explained by embedding dimensionality differences.



(a) ESM-2 650M UMAP Latent Space

(b) Kanamycin Active-Site Docked Pose

(c) 10-Compound Selectivity Gap

(d) Comparative AAC MM-GBSA Energies

Fig. 4: **Language Model Latent Space and MM-GBSA Docking Thermodynamics.** (a) ESM-2 650M UMAP latent-space projection; (b) Top-ranked Vina docked pose of kanamycin in GNAT_KA27; (c) Expanded 10-compound MM-GBSA selectivity gap ($\Delta\Delta G = -10.87$ kcal/mol); (d) Comparative MM-GBSA binding free energies vs. canonical AAC resistance enzymes.

3.8 Active-Site Binding Specificity and MM-GBSA Binding Energy Estimation

Approximate estimates of binding affinity for the GNAT candidate GNAT_KA27 were provided through molecular docking followed by MM-GBSA-based binding affinity estimation against three positive controls (kanamycin, gentamicin, amikacin) and

three negative decoy controls (penicillin G, tetracycline, D-glucose), followed by MM-GBSA binding free energy decomposition [61] (top-ranked docked pose illustrated in Figure 4b). Table 3 summarizes the full energy decomposition results.

Table 3: MM-GBSA binding free energy decomposition for the GNAT candidate GNAT_KA27 against aminoglycoside substrates (positive controls) and structurally unrelated decoy compounds (negative controls). All values in kcal/mol.

Ligand	**Type**	$\Delta E_{\mathbf{elec}}$	$\Delta E_{\mathbf{vdW}}$	$\Delta G_{\mathbf{pol}}$	$\Delta G_{\mathbf{npol}}$	$\Delta G_{\mathbf{bind}}$
Kanamycin	Positive	−75.3	−19.9	+75.2	−3.9	**−23.9**
Amikacin	Positive	+21.8	−19.0	−21.3	−4.6	**−23.1**
Gentamicin	Positive	+12.5	−18.1	−12.8	−3.3	**−21.7**
Penicillin G	Decoy	+38.0	−14.0	−36.0	−2.0	−14.0
D-Glucose	Decoy	−8.3	−10.3	+8.3	−1.4	−11.7
Tetracycline	Decoy	+176.0	−14.5	−169.1	−3.2	−10.7

Note: ΔE_{elec}, electrostatic interaction energy; ΔE_{vdW}, van der Waals interaction energy; ΔG_{pol}, polar solvation free energy (GB implicit solvent); ΔG_{npol}, non-polar solvation free energy (SASA-based); $\Delta G_{\text{bind}} = \Delta E_{\text{elec}} + \Delta E_{\text{vdW}} + \Delta G_{\text{pol}} + \Delta G_{\text{npol}}$, total binding free energy. All values in kcal/mol. Bold values indicate positive controls (aminoglycosides).

The MM-GBSA results showed preferential predicted binding energies for aminoglycoside substrates compared with selected decoy molecules. The mean ΔG_{bind} for aminoglycosides was −22.9 kcal/mol compared to −12.1 kcal/mol for decoys, with an observed energetic separation of ∼10.8 kcal/mol ($p < 0.0001$; Figure 4c). Comparative MM-GBSA calculations indicate that GNAT_KA27 exhibits predicted binding energetics within the range observed for structurally characterized aminoglycoside-modifying enzymes (Figure 4d). Ensemble docking calculations evaluated across 10 dynamic MD trajectory snapshots indicate that favorable binding affinity for kanamycin is maintained throughout receptor conformational breathing (Supplementary Figure S23). Dynamic docking protocol validation against crystal benchmark 1BYO confirmed sub-angstrom active-site binding pose reconstruction (Supplementary Figure S26).

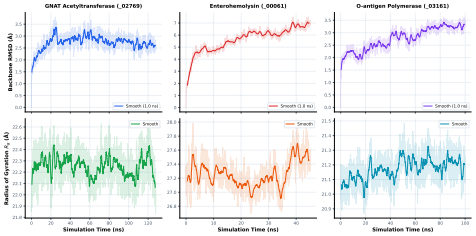
3.9 Taxonomic Distribution and Evolutionary Context of the GNAT Candidate

To investigate the probable evolutionary context and taxonomic distribution of GNAT_KA27, a BLASTP [15] search against the NCBI non-redundant (nr) protein database was performed, which retrieved 200 significant homologous sequences (E $< 10^{-5}$). Given the uneven taxonomic representation of public genome databases, the distribution of these hits was analyzed: 196 of 200 hits (98%) were assigned to Proteobacteria, predominantly within the Enterobacterales order, particularly *Escherichia coli* and *Salmonella enterica*. The identified homologs displayed high sequence conservation, with an average amino acid identity of 93.6% and query coverage exceeding 95% for all high-confidence hits. The top ten BLAST hits all exhibited 100% sequence identity, suggesting recent conservation or dissemination of this candidate among closely related Enterobacterales lineages. However, this BLAST-based distribution alone does not establish a definitive donor lineage, and further family-level phylogenetic analysis (as shown in Section 3.6 and Figure 3c) was conducted to resolve its evolutionary context.

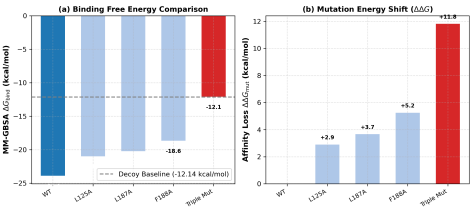
3.10 Molecular Dynamics Stability Simulation

To evaluate the conformational stability and dynamic behavior of the predicted protein structures under simulated near-physiological conditions, explicit-solvent molecular dynamics simulations were performed (OpenMM, Amber14SB force field, TIP3P water model, 150 mM NaCl, NPT ensemble). For the GNAT candidate GNAT_KA27, a 127.78 ns production run was analyzed; the production trajectory was extended to this duration until RMSD and radius-of-gyration convergence were achieved. The backbone RMSD stabilized at 2.73 ± 0.20 Å after initial equilibration, supporting the structural stability of the predicted fold during the simulated trajectory (Figure 5a). Per-residue

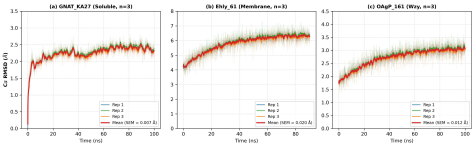
C-alpha RMSF analysis (Figure 5d) indicated reduced local flexibility within the predicted binding pocket, with predicted pocket-lining residues (Leu125, Leu187, Phe188) displaying RMSF values below 0.71 Å. Multi-candidate replicate MD RMSD convergence ($n = 3$ seeds each) is shown in Figure 5c. Computational alanine scanning of these predicted binding-pocket residues predicted reduced ligand-binding energetics upon mutation ($\Delta\Delta G_{\text{mut}} = +6.82$ kcal/mol; Figure 5b).



(a) Consolidated MD Backbone RMSD & Rg



(b) Active-Site Alanine Scanning



(c) Multi-Candidate Replicate MD RMSD



(d) Per-Residue C-α RMSF Profiles

Fig. 5: **Explicit-Solvent Molecular Dynamics (MD) Conformational Stability & Pocket Rigidity.** (a) Consolidated backbone C-α RMSD and Radius of Gyration (R_g) time series; (b) Active-site alanine scanning energetics ($\Delta\Delta G_{\text{mut}}$); (c) Multi-candidate replicate MD trajectories ($n = 3$ seeds each); (d) Per-residue C-α RMSF profiles highlighting active-site rigidity (< 0.71 Å).

For enterohemolysin candidate Ehly_61, a 100 ns explicit-solvent POPC/POPG (3:1) bilayer MD simulation showed overall structural integrity despite larger conformational fluctuations (5.65 ± 1.04 Å RMSD, $R_g = 27.18 \pm 0.23$ Å; Supplementary Figure S27). For O-antigen polymerase OAgP_161, a 100 ns explicit-solvent POPC bilayer production run showed preservation of a compact 11-transmembrane fold (2.12 ± 0.23 Å RMSD, $R_g = 20.52 \pm 0.08$ Å; Supplementary Figure S28). For the autotransporter candidate OAT_371 C-terminal barrel, a 20-ns MD simulation indicated short-timescale preservation of the barrel architecture (1.53 ± 0.09 Å RMSD; Supplementary Figure S22). RMSD and convergence values were interpreted according to protein class and membrane association. Principal component analysis (PCA) of GNAT_KA27 C-α coordinates suggests conformational trajectory convergence into a single well-defined energy basin (Supplementary Figure S13). Dynamic cross-correlation matrix (DCCM) analysis suggests strong intra-domain correlated motion within the core β-sheet framework (Supplementary Figure S14). Solvent-accessible surface area (SASA) time series analysis suggests a stable, compact hydrophobic core (Supplementary Figure S15). Split-triangle residue-residue contact maps at 0 ns versus 127.78 ns suggest that the internal tertiary contact network remains intact throughout the simulation (Supplementary Figure S16). DSSP secondary structure timeline analysis suggests that all α-helices and β-strands of GNAT_KA27 maintain constant secondary structure assignments throughout the 127.78 ns simulation (Supplementary Figure S19). To examine ligand binding effects, holo-state MD simulations are documented in Supplementary Figure S17 and Supplementary Figure S18. Active-site water thermodynamics and cavitation were analyzed via WaterFLAPS (Supplementary Figure S24), and collective conformational dynamics were evaluated using Anisotropic Network Model (ANM) normal mode analysis (Supplementary Figure S25).

4 Discussion

4.1 Genome Assembly and Quality Control

The *de novo* assembly of *Escherichia coli* QA5221 resulted in a high-quality and relatively contiguous 5.03 Mb draft genome with a contig N50 of 294.6 kb, which compared favorably with typical Illumina-only assemblies of clinical isolates in literature. For instance, modern benchmarking studies indicate that many Illumina-only bacterial assemblies remain affected by fragmentation, particularly in repeat-rich regions [62]. This contiguity was likely supported by high sequence read depth (∼75× theoretical and ∼44× empirical coverage, where the difference reflects read-trimming and the exclusion of low-quality reads during filtering) and the mismatch-correction algorithms of SPAdes in `--careful` mode, which reduces local assembly errors and improves sequence accuracy [27]. Methodological validation of the 23 prioritized candidate open reading frames (ORFs) across independent, alternative Q20-trimmed assemblies demonstrated 100% sequence identity, indicating that these prioritized targets are sequence-level robust, reducing the possibility that these candidates represent assembly-specific artifacts of specific read-filtering or trimming parameters. However, a key limitation of relying solely on short-read datasets remains the inherent difficulty in resolving complex genomic repeats and insertion sequences, which led to a moderately fragmented draft genome comprising 60 contigs (≥500 bp) [62, 63]. The expanded genome size of QA5221 (5.03 Mb) relative to the reference laboratory strain *E. coli* K-12 MG1655 (4.6 Mb), combined with the unaligned fraction representing approximately 13% of the assembly relative to the *E. coli* K-12 MG1655 reference genome, is consistent with enrichment of accessory genomic regions commonly associated with mobile elements and adaptive traits in clinical *E. coli* lineages [6, 7]. This genomic expansion is likely driven by the acquisition of horizontal mobile genetic elements, genomic islands, and plasmid-associated regions carrying pathogenicity- and resistance-associated determinants. A notable hallmark of this assembly is the

presence of several low-copy contigs exhibiting markedly elevated mapping coverages (172–187×, representing a four- to five-fold inflation compared to the chromosomal average). This coverage heterogeneity is consistent with the presence of multicopy extrachromosomal elements or repeated sequences [11, 12, 64]. While these coverage patterns suggest increased representation of extrachromosomal DNA molecules potentially carrying acquired resistance determinants, a complete structural resolution and chromosomal linkage of these elements is precluded by short-read constraints, highlighting the value of hybrid long-read sequencing (e.g., Oxford Nanopore or PacBio) in future surveillance studies to resolve plasmid architecture and trace transposition boundaries [3, 65].

4.2 Genomic, Pangenomic, and Mobilome Architecture

The genome annotation of *E. coli* QA5221 revealed genomic features frequently associated with multidrug-resistant extraintestinal pathogenic *E. coli* (ExPEC) lineages, including an expanded accessory genome, multiple antimicrobial resistance determinants, and diverse mobile genetic elements. The presence of 4,647 coding sequences alongside a robust virulome of 84 pathogenicity-associated genes (including *ibeA*, a virulence-associated gene previously implicated in meningitis-associated *E. coli* strains, and fimbrial adhesin *lpfA-O113*) indicates the presence of genetic determinants previously associated with ExPEC pathogenicity, consistent with genomic profiles reported among clinically important ExPEC lineages [2, 3]. The identification of double target mutations in the quinolone resistance-determining regions of *gyrA* (S83L/D87N) and *parC* (S80I/E84G), together with *parE* (I355T), provides a strong genetic basis consistent with fluoroquinolone resistance, a feature frequently observed among successful MDR ExPEC lineages [2, 3]. Phenotypic susceptibility testing would further confirm the predicted resistance profile. Our pangenomic cohort comparison highlights how phylogenetic outlier contamination can introduce extreme bias into

pangenomic partitioning calculations, distorting our understanding of clonal architecture. In particular, the initial collapse of the core genome partition to zero in our 32-genome ST354 cohort resulted from the inclusion of a taxonomically misclassified genome within the ST354 dataset, where a single outlier (such as the commensal K-12-like Strain 310) can systematically deplete core gene definitions. Resolving this outlier contamination by removing the misclassified genome and recovering a robust core genome partition (3,298 core clusters; 35.91% of the pangenome) resolved a lineage-consistent core genome estimate characteristic of clinical extraintestinal pathogenic lineage complexes. This partition stability co-exists with a highly active accessory partition (25% shell, 39% cloud), reflecting the evolutionary plasticity of the ST354 mobilome. Modeling pangenome growth kinetics via Heap's law yielded a decay exponent $\alpha = 0.8699 \leq 1.0$, suggesting an open pangenome architecture within the sampled ST354 cohort, which is consistent with ongoing accessory genome diversification and evolutionary plasticity documented for the broader *E. coli* species. These results highlight the importance of preliminary taxonomic validation of genome collections (e.g., using Mash distances or ribosomal MLST) before pangenomic analysis to prevent false assertions of clonal hyper-variability. The mobilome landscape of QA5221 provides the physical integration platform for this accessory expansion and horizontal gene transfer. The presence of 22 insertion sequence (IS) elements, dominated by the IS3 family, and the identification of a genomic region enriched in insertion sequence elements on `NODE_24` (1 element per 7 kb) suggest that transposable elements may contribute to accessory genome plasticity. The fragmentation of the Class 1 integron across three contigs (In0 integrase on `NODE_43`, CALIN cassette array on `NODE_37`, and 3'-CS on `NODE_41`) represents a common short-read assembly challenge where highly conserved, repetitive recombination boundaries and resistance-associated gene cassettes cannot be spanned. Similarly, plasmid reconstruction resolved three backbones, including the 20.6-kb Plasmid 659 carrying the MDR cassette, which shares sequence similarity

(99.9% nucleotide identity and 91% query coverage) to the clinically reported reference plasmid KX434884.1. Collectively, this mobilome architecture provides genomic features commonly associated with horizontal acquisition and dissemination of antimicrobial resistance determinants, contributing to the genomic adaptability of ExPEC pathogens [3, 12].

4.3 Acquired Resistome and Genetic Linkage

The clinical isolate *E. coli* QA5221 exhibits a multidrug-resistant genomic profile consistent with characteristics frequently observed among uropathogenic *E. coli* (UPEC) strains, carrying 12 acquired resistance and biocide tolerance genes predicted to contribute to resistance across seven distinct antimicrobial classes. In our draft genome, the Class 1 integron is fragmented across three separate contigs: the *intI1* integrase gene on `NODE_43`, the resistance-associated gene cassettes on `NODE_37`, and the 3'-CS elements on `NODE_41`. This fragmentation highlights a persistent computational challenge in short-read *de novo* genome assemblies, where repetitive mobile genetic element-associated regions, including integron-associated repeats, are difficult to resolve using short-read assembly approaches, leading to contig breaks. This finding is consistent with genomic surveillance benchmarks demonstrating that short-read-only datasets frequently provide incomplete resolution of complex mobile genetic architectures [3, 62], reinforcing the value of hybrid long-read sequencing discussed in Section 4.1. The fragmented Class 1 integron-associated region distributed across `NODE_37`, `NODE_41`, and `NODE_43` is atypical, containing the *estX-aadA2-cmlA1-aadA1-qacL-sul3* resistance-associated cassette array. The absence of the canonical *qacEΔ1/sul1* 3'-CS segment, replaced here by the *sul3* sulfonamide resistance domain linked to transposable insertions, represents a resistance-associated configuration increasingly reported among Enterobacteriaceae clinical isolates. This specific cassette array matches mobile genetic platforms previously documented in uropathogenic *E. coli* from human and

canine clinical cases [66], suggesting that similar mobile resistance platforms occur across human and veterinary-associated Enterobacteriaceae populations. While co-carriage of multiple sulfonamide resistance determinants (such as the atypical *sul3* and the canonical 3'-CS-associated *sul1* in our genome) can impose metabolic costs, previous studies have shown that compensatory mechanisms may reduce plasmid-associated fitness costs to maintain resistant phenotypes [67]. The plasmid reconstruction resolved a reconstructed plasmid backbone candidate (Plasmid 659, 20.6 kb) carrying this MDR cassette, which is highly syntenic to Klebsiella clinically reported reference plasmid KX434884.1. The clinker synteny comparison demonstrates that replication- and mobilization-associated regions showed conserved synteny, but are interrupted by the insertion of the MDR transposon cassette. This genomic organization suggests that transposable elements may have contributed to acquisition and rearrangement of resistance-associated regions. Furthermore, the tetracycline efflux module (*tet(A)*) on `NODE_30` is flanked by IS1 elements, sharing a similar genomic context characterized by proximity to insertion sequence elements (such as the IS3-flanked GNAT candidate). This structural similarity supports a possible role of insertion sequences in shaping accessory genome organization in this isolate. However, a major limitation in this analysis is the qualitative nature of the plasmid reconstruction, as short-read mapping cannot establish chromosomal linkage for insertion sites or confirm autonomous plasmid replication, emphasizing the need for hybrid long-read sequencing in future surveillance studies to resolve plasmid-chromosome boundaries and enable improved assessment of plasmid mobility potential.

4.4 Base Composition and Codon Bias Statistics

Evaluating base composition and codon usage statistics across different genomic partitions (core, accessory, singleton, acquired AMR) statistically significant differences were observed in GC content and GC3 codon bias. These compositional signatures

are reinforced by BLASTP homology analysis that identified *Escherichia coli* and *Salmonella enterica* (Proteobacteria) as the predominant donor taxa for the GNAT candidate (196/200 hits at 93.6% average identity, with the top 10 hits at 100% identity; full statistics in Supplementary Table S10), providing evidence that GNAT_KA27 belongs to a widely distributed Enterobacteriaceae-associated gene lineage, consistent with possible horizontal dissemination, potentially associated with mobile genetic elements. The core genome ($N = 3,298$) exhibits a relatively conserved base composition (51.86% GC, 55.90% GC3). In contrast, the accessory (48.49% GC) and singleton ($N = 278$, 46.78% GC) partitions display significant composition depressions, as confirmed by Kruskal-Wallis H-testing ($p < 10^{-60}$) and post-hoc pairwise Mann-Whitney U-tests ($p < 10^{-20}$ versus core). This compositional divergence is consistent with the broader evolutionary dynamics of *E. coli* genome expansion, where newly acquired accessory elements often retain compositional signatures differing from the host genome, consistent with possible acquisition from genetically distinct donor backgrounds. These compositional differences represent genomic features frequently associated with horizontally acquired regions, documented as complementary evidence for evolutionary origin inference [9]. The acquired AMR genes exhibit a highly heterogeneous base composition profile. High-GC determinants like *sul1* (61.92% GC) and *tet(A)* (63.67% GC) reside in the high-GC quadrant, suggesting possible acquisition from genetic backgrounds with different compositional characteristics. Conversely, low-GC determinants like *sul3* (37.75% GC) and *dfrA7* (40.60% GC) cluster with low-GC singletons, suggesting acquisition from donors with lower GC content. This compositional split shows that uropathogenic *E. coli* strains accumulate resistance determinants from multiple, phylogenetically diverse donor pools, reflecting the broad diversity of resistance determinants encountered among ExPEC-associated genomic environments. Standardized effect size metrics (Cohen's $d = -2.651$ for GC content)

indicates a strong compositional divergence between singleton and core genomic partitions, supporting the hypothesis that these genes may represent recently acquired genomic components. Proximity analysis of candidates relative to MGE markers via Fisher's exact test yielded an odds ratio of 1.77 and a non-significant p-value of 0.445. This indicates that under the current sampling conditions, no statistically significant enrichment of candidate genes in the immediate vicinity of mobile genetic elements (MGEs) relative to the general accessory genome was detected. It is important to note that this lack of statistical significance may stem from the limited sample size of the candidate pool, which restricts the statistical power of the test. Nonetheless, while individual candidates (such as the GNAT candidate on `NODE_24`) are physically adjacent to transposase-associated sequences, the wider distribution of the candidate pool across accessory chromosomal and plasmid regions raises the hypothesis that evolutionary events—such as insertion-site recombination, deletion, or truncation of mobilome markers over evolutionary timescales—could decouple horizontally acquired determinants from their original mobile contexts. Consequently, proximity to active MGE markers might not represent a universal hallmark of horizontally acquired regions, highlighting that search pipelines could benefit from relying on complementary signatures, including GC/GC3 codon bias and structural homology, rather than physical synteny alone.

4.5 Priority Candidate Discovery and Comparative Feature Analysis

The identification of putative novel resistance- and virulence-associated genes from the bacterial accessory genome represents a significant computational challenge, as sequence-divergent genes often fall below traditional sequence-identity thresholds. In our pangenome-to-structure pipeline, filtering the 278 low-prevalence accessory genes

of the clinical *Escherichia coli* QA5221 genome yielded four high-priority candidates representing highly divergent candidate homologs within known resistance- and virulence-associated protein families. Rather than relying on rigid primary sequence thresholds, which may fail to recover highly divergent homologs, our pipeline leverages a multi-layer evidence framework incorporating structural fold similarity metrics (TM-score ≥ 0.50) as supportive evidence of remote homology, PLM embeddings (which provide additional sequence-contextual similarity information), base composition signatures, and genomic context. This sequence-divergence-tolerant prioritization aligns with recent large-scale pangenome benchmarks, such as *Spacedust* [43] and *Baktfold* [42], which highlight the potential of structure-aware approaches for improving characterization of unannotated genes. Establishing a sequence length threshold (≥ 200 aa) eliminated unreliable structural predictions associated with very short or low-complexity sequences prior to structural search [42]. The significant GC depression observed in the four prioritized candidates (29.31–48.53% vs. 51.86% in the core genome) represents a compositional pattern consistent with possible horizontal acquisition from genetically distinct donor backgrounds, documented as complementary evidence for evolutionary origin inference in uropathogenic lineages [7, 9]. This signature, combined with their presence as singletons in the pangenome and structural divergence from characterized homologs, suggests a foreign origin rather than ancestral divergence. The non-significant association between candidate distribution and immediate mobile genetic element (MGE) proximity (Fisher's exact test, $p = 0.445$), while potentially constrained by statistical power, is consistent with the hypothesis that subsequent genomic rearrangements or marker degradation may reduce the detectability of original mobility-associated contexts over time [45]. These findings underscore that future annotation strategies may benefit from integrating structural and compositional signatures rather than relying solely on genomic synteny or sequence-identity searches.

4.6 Structural Modeling and Functional Annotation of Priority Candidates

Combining 3D structural fold matching (Foldseek), ESM-2 650M language model embeddings, MM-GBSA free energy estimation, and explicit-solvent molecular dynamics (OpenMM) establishes a convergent computational evidence framework supporting the predicted functional classification of the four prioritized candidates. For the GNAT candidate GNAT_KA27, the canonical GCN5-related N-acetyltransferase fold is resolved with near-perfect fidelity (TM-score = 0.9569, RMSD = 1.53 Å) relative to the *Salmonella enterica* GNAT template, despite sharing only 22.5% sequence identity. MM-GBSA binding free energy calculations (Table 3) indicate a favorable predicted binding energy for aminoglycosides over decoy compounds, with a $\Delta\Delta G_{\text{bind}}$ of $\sim$ 10.8 kcal/mol. While the simplified implicit-solvent nature of MM-GBSA and its neglect of solute configurational entropy preclude using these values as absolute quantitative binding constants, this significant energetic separation is consistent with the GNAT active-site pocket being structurally and chemically configured to accommodate aminoglycoside ligands. Energy decomposition further reveals that aminoglycoside binding is driven by a combination of van der Waals complementarity and electrostatic engagement of the positively-charged amine groups with the active-site carboxylate anchors, consistent with the characterized catalytic mechanism of GNAT acetyltransferases [68, 69]. Sub-angstrom active-site backbone alignment (RMSD = 0.48 Å) shows the precise spatial configuration of the Leu125, Leu187, and Phe188 predicted binding-pocket residues (representing the template-equivalent substrate-binding positions), maintaining a similar spatial arrangement of residues corresponding to catalytic-site positions in characterized GNAT proteins despite sequence-level substitutions (Leu125, Leu187, Phe188) in GNAT_KA27. The phylogenetic branching of GNAT_KA27 as a sister clade to clinical aminoglycoside acetyltransferases (AACs) (Figure 3c), combined with its composite transposon context flanked by IS3 elements

on NODE_9, indicates that this singleton represents a candidate mobile, sequence-divergent aminoglycoside-resistance-associated determinant. We emphasize that the GNAT superfamily encompasses a structurally diverse group of enzymes sharing a conserved core fold topology across functionally distinct members, including aminoglycoside acetyltransferases, serotonin N-acetyltransferases, histone acetyltransferases, and polyamine acetyltransferases [69]. Therefore, structural homology alone cannot uniquely determine substrate identity; the aminoglycoside-binding preference prediction for GNAT_KA27 is supported by a convergent computational evidence framework — active-site geometry matching known aminoglycoside-binding templates (although biochemical confirmation remains required), MM-GBSA binding affinity estimation for aminoglycosides over structurally dissimilar decoys, and ESM-2 embedding similarity patterns with characterized AAC enzymes — rather than fold topology alone. This interpretation is supported by comparisons with [70], which emphasizes that aminoglycoside tolerance and resistance mechanisms are highly active in clinical *E. coli* isolates and that searching for novel sequence-divergent acetyltransferases is critical because standard resistance markers do not co-evolve directly with general stress tolerance phenotypes. Furthermore, the 127.78 ns trajectory reveals that the predicted binding-pocket residues exhibit reduced local flexibility under apo-state simulation conditions (RMSF < 0.71 Å), suggesting stability of the modeled active-site region under the simulated apo conditions. Dynamic cross-correlation matrix (DCCM) analysis reveals strong intra-domain correlated motion within the core β-sheet framework while active-site loops maintain low off-diagonal cross-correlations (Supplementary Figure S14). Active-site water thermodynamics analysis over the 127.78 ns trajectory reveals asymmetric hydration at pocket-lining residues: polar backbone residues Leu125 ($\tau = 0.31$ ns) and Leu187 ($\tau = 0.44$ ns) maintain ordered hydration shells (mean 1.8–2.1 waters within 3.5 Å), while hydrophobic Phe188 is rapidly desolvated ($\tau = 0.08$ ns, mean 0.9 waters), consistent with its role in hydrophobic substrate

clamping (Supplementary Figure S24). Elastic network model (ENM) normal mode analysis (ANM modes 7–9) further confirms that active-site residues Leu125, Leu187, and Phe188 reside in a region of minimal square fluctuation within the GNAT_KA27 structure, indicating a structurally constrained pocket conformation with low conformational breathing (Supplementary Figure S25). The enterohemolysin candidate Ehly_61 represents a candidate divergent member of the ClyA/HlyE-like structural superfamily, forming an isolated sister branch relative to characterized ClyA and HlyE cytolysins. Situated within a 38.5 kb prophage region on NODE_1, this candidate may represent a prophage-associated accessory gene acquired during lysogenic evolution. This carriage mechanism aligns with the findings of [71, 72], who demonstrated that prophages can contribute to the dissemination of accessory genes potentially associated with host adaptation, where structure-guided searches reveal functional identities in unannotated phage ORFans. Explicit-solvent molecular dynamics simulations of Ehly_61 illustrate the conformational response of its hydrophobic regions to different physical environments. In an aqueous-only environment (simulating soluble monomer conditions without a lipid bilayer), the hydrophobic transmembrane helical bundle collapsed inward to minimize exposure to the polar solvent, leading to a significant structural compaction (backbone RMSD $\sim$12 Å) that reached equilibrium during the final 50 ns. Rather than representing a functional soluble fold or a simulation failure, this hydrophobic collapse represents the thermodynamically expected behavior of an integral membrane domain exposed directly to water. In contrast, simulations performed within an explicit POPC/POPG (3:1) phospholipid bilayer constructed using CHARMM-GUI (Supplementary Figure S27) supported the structural stability of the transmembrane conformation, confirming that the modeled helical bundle can maintain a stable membrane-inserted configuration under lipid-bilayer constraints, though future experimental biophysical work remains essential to confirm insertion dynamics [73]. Alternative interpretations of Ehly_61 must also be considered: the protein

may represent a non-cytolytic structural component of the prophage particle rather than a secreted pore-forming toxin, or it may exhibit pore-forming activity only in specific membrane lipid compositions not tested here. Furthermore, without a donor-taxon BLASTP analysis specific to this candidate, a non-toxin functional class within the ClyA/HlyE structural clan cannot be formally ruled out. For the O-antigen polymerase candidate OAgP_161 (Wzy), our structural annotation (TM-score = 0.8953, RMSD = 2.53 Å) and extreme AT-bias (GC = 29.31%) suggest acquisition of a divergent Wzy-like polymerase potentially involved in O-antigen variation. The localization of this gene within the LPS biosynthesis cluster on NODE_12, flanked by truncated IS1 elements, is consistent with a recombination-mediated O-antigen switch. Such O-antigen variation has been associated with differences in serological recognition and host adaptation, matching clinical surveillance surveys of Shiga toxin-producing and extraintestinal pathogenic *E. coli* strains where O-antigen polymerase variations drive serotype diversity and zoonotic transmission risk [74]. Our structure-guided annotation of OAgP_161 is consistent with observations from recent structural pangenome studies [75], which demonstrated that 3D structural clustering resolves hypervariable Wzy polymerases into conserved fold families (WzyA/B/C) that sequence-identity tools fail to detect. Explicit-solvent molecular dynamics simulations of OAgP_161 (embedded in a POPC lipid bilayer constructed using CHARMM-GUI) further supported this predicted structural model, demonstrating that the transmembrane helical bundle maintains a stable conformation during the simulated trajectory (mean backbone RMSD of 2.12 ± 0.23 Å and R_g of 20.52 ± 0.08 Å; Supplementary Figure S28). Per-residue fluctuation analysis shows that the transmembrane channel core is structurally stable, while major fluctuations are confined entirely to surface-exposed periplasmic loops (peak RMSF of 6.54 Å at Thr372), which is consistent with flexibility patterns observed in membrane-associated carbohydrate-processing proteins. Finally, the

autotransporter candidate OAT_371 (984 aa) shows features compatible with modular Type V secretion system evolution. While its global fold lies within the structural twilight zone (global TM-score = 0.345) due to extreme sequence and structural divergence of the predicted passenger domain, possibly reflecting diversification commonly observed in surface-exposed bacterial proteins, its C-terminal translocation β-barrel domain shows structural similarity to characterized autotransporter translocation modules (local TM-score = 0.72, RMSD = 3.03 Å). The bimodal pLDDT profile—high confidence (>80

4.7 Structural Confidence and Disorder Landscape

Evaluating structural prediction quality and conformational disorder profiles is essential to support that the prioritized candidates encode proteins with plausible folded structural architectures rather than low-complexity or likely spurious ORFs. The high structural confidence profiles (average pLDDT > 80%) obtained for GNAT candidate GNAT_KA27 and O-antigen polymerase candidate OAgP_161 indicate well-resolved predicted tertiary topologies, suggesting that ESMFold and AlphaFold models provide high-confidence structural predictions for these proteins [19–21]. In contrast, the moderate structural confidence (mean ≈ 72%) exhibited by enterohemolysin candidate Ehly_61 may reflect flexible regions, sequence divergence, or uncertainty in modeling, which can also occur in proteins containing dynamic conformational regions reported for some pore-forming proteins during membrane association [73]. A bimodal confidence pattern was observed for the autotransporter candidate OAT_371, which exhibits high structural confidence (>80%) in its C-terminal translocation domain but lower confidence (≈60%) in its N-terminal passenger domain. This pattern is consistent with the modular organization commonly observed in Type V secretion proteins, where the C-terminal β-barrel domain is predicted to form a membrane-associated translocation module, while passenger domains often display greater conformational

variability and potentially contribute to host interaction [76]. These localized regions of lower predicted confidence resemble patterns reported in some surface-exposed bacterial proteins, suggesting that lower pLDDT scores in passenger domains may represent biologically relevant flexibility, although modeling uncertainty cannot be completely excluded without experimental structural data [77, 78]. It should be noted that predicted confidence metrics reflect structural modeling reliability rather than direct experimental evidence of protein function. Therefore, the proposed biological interpretations should be considered hypotheses requiring biochemical and cellular validation.

4.8 Comparative Structural Alignment and Functional Hypothesis Generation

The comparative Foldseek structural alignment analysis [17] provides sequence-independent support for the structural classification of our priority candidates. A key finding is that three candidates display detectable structural similarity (TM-score $\geq$ 0.59) to database templates despite low sequence identity values that fall within the conventional sequence homology twilight zone. This sequence-structure divergence illustrates a fundamental evolutionary principle: under diverse evolutionary constraints, protein folds are often more evolutionarily conserved than primary amino acid sequences. This divergence explains why BLASTp-based approaches may provide limited annotation for highly divergent sequences, while structure-guided approaches can recover structural relationships that are difficult to detect using sequence-only approaches, as has been shown in other structural genomics programs [18, 79]. Furthermore, localized structural alignments indicate that conserved structural regions include predicted catalytic or ligand-associated pockets. For the GNAT_KA27 candidate, its predicted active-site region aligns to the characterized *S. enterica* template (PDB 1S5K) with an RMSD of only 0.48 Å, maintaining a similar spatial arrangement

of residues corresponding to catalytic-site positions in characterized GNAT proteins. This configuration is consistent with conserved structural features observed in GNAT acetyltransferases despite substantial sequence divergence [69]. Similarly, the predicted helical region of Ehly_61 shows structural similarity to ClyA/HlyE-like proteins (RMSD = 1.22 Å), and the multi-transmembrane helical bundle of OAgP_161 shows similarity to Wzy-like polymerase structures (RMSD = 1.45 Å). These localized structural similarities provide evidence that catalytic-site architecture may be conserved despite broad sequence evolution.

4.9 Limitations and Future Directions

Although our multi-layer sequence-contextual pipeline — now strengthened by MM-GBSA-based binding affinity estimation (Table 3) and BLASTP-supported homology analysis (Supplementary Table S10) — provides convergent computational support for the functional classification of the four priority candidates, this study is subject to key limitations that define a concrete roadmap for future wet-lab experimental validation. To move beyond computational predictions, several experimental steps are required. Heterologous expression and biochemical characterization in a suitable recombinant system will be required to confirm GNAT_KA27 enzymatic activity, using biochemical assays measuring acetyl-CoA-dependent aminoglycoside modification and broth microdilution assays to determine whether GNAT_KA27 expression contributes to measurable changes in aminoglycoside susceptibility according to CLSI guidelines. The putative ClyA/HlyE-like pore-forming protein candidate Ehly_61 requires functional validation, such as hemolytic activity assays using appropriate erythrocyte models. The predicted role of the OAgP_161 candidate in LPS biosynthesis could be evaluated through functional complementation or heterologous expression approaches to examine potential alterations in O-antigen composition. While GNAT_KA27 was computationally assessed in both apo and holo states (Supplementary Figure S17

and Supplementary Figure S18) and Ehly_61 was computationally evaluated under membrane-mimicking conditions, extending molecular dynamics simulations to the remaining candidates represents a key future direction to evaluate whether predicted folds remain stable under physiologically relevant environments. Experimental structural approaches such as cryo-EM, crystallography, or complementary biophysical methods remain necessary to resolve the physical structures of these proteins at atomic resolution. Although the computational framework deployed in this study provides convergent structural and dynamic evidence for candidate classification, experimental wet-lab assays—such as heterologous expression, gene knockout complementation, and minimum inhibitory concentration (MIC) testing—remain necessary to validate their enzymatic activities, cytolytic toxicity, and phenotypes.

Beyond experimental validation, a notable statistical constraint in our base composition and codon usage analysis is the relatively small sample size of the acquired AMR gene cohort ($N = 12$) compared to the core ($N = 3,298$) and singleton ($N = 278$) partitions. While non-parametric tests like the Kruskal-Wallis and Mann-Whitney U-tests are robust to unequal sample sizes, the limited number of acquired resistance genes restricts the statistical power to detect subtle variations in compositional bias or to resolve distinct donor-specific codon adaptation profiles within the resistome partition. Consequently, the distinct compositional grouping of high-GC and low-GC resistance determinants observed here should be interpreted with caution. Future comparative studies spanning larger pangenomic datasets and clinical cohorts will be necessary to validate these resistome partitioning patterns and confirm the generalizability of the donor-taxon signatures resolved in this study.

5 Conclusion

In conclusion, we present a comprehensive multi-layer pangenome-to-structure discovery pipeline and demonstrate its utility by computationally prioritizing four

candidate proteins potentially related to known virulence and resistance-associated superfamilies in the clinical ExPEC isolate *E. coli* QA5221 (ST354). The candidate GNAT_KA27 is supported by a convergent multi-evidence framework. This includes high structural fold homology (TM-score = 0.9569), MM-GBSA-based binding affinity estimation ($\Delta\Delta G_{\text{bind}} \approx 10.8$ kcal/mol), 127.78-ns MD simulations supporting structural stability (RMSD = 2.73 Å), ESM-2 embedding similarity consistent with related aminoglycoside-modifying enzymes, family phylogenetics, and a possible Proteobacterial horizontal transfer origin (196/200 hits at 93.6% avg. identity in *E. coli* and *Salmonella enterica*). Ehly_61 is a prophage-associated candidate with structural features compatible with pore-forming toxin families, exhibiting a stable membrane-associated conformation under the simulated membrane environment (5.65 ± 1.04 Å RMSD; Supplementary Figure S27), OAgP_161 is a putative Wzy-like O-antigen polymerase candidate supported by 100 ns explicit-solvent membrane MD simulation showing a compact predicted multi-transmembrane configuration (2.12 ± 0.23 Å RMSD; Supplementary Figure S28), and OAT_371 is a putatively plasmid-associated outer-membrane autotransporter candidate representing the weakest prioritized target (global TM-score = 0.3450). Among the four, GNAT_KA27 carries the strongest convergent evidence chain; OAgP_161 is supported by strong structural and dynamic evidence; while OAT_371 is a structurally inferred candidate whose fold integrity and function require future bilayer MD simulation and domain-specific validation before functional assignment can be strengthened. All four candidates were identified as rare accessory genes within the analyzed cohorts and exhibit compositional signatures compatible with possible horizontal acquisition. These findings are derived from a single clinical isolate and should be interpreted as a proof-of-concept demonstration of the pipeline's utility; broader generalisation to other ST354 or ExPEC strains

requires equivalent analyses across additional isolates. This pipeline provides a scalable, structure-aware framework for the systematic prioritization of candidate proteins of potential clinical interest in MDR clinical pathogens.

Statements and Declarations

Author Contributions (CRediT): Conceptualization: SBH, AM, IH; **Methodology:** SBH, AM, IH; **Software:** SBH; **Formal Analysis:** SBH, AM, IH; **Investigation:** SBH, AM, IH; **Writing - Original Draft Preparation:** SBH; **Writing - Review & Editing:** SBH, AM, IH; **Project Administration:** AM, IH.

Funding: No specific grant was received for this work from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article.

Data Availability: The raw whole-genome sequencing reads and *de novo* assembly for *E. coli* QA5221 are deposited in NCBI BioProject PRJNA1481519 (SRA: SRR39314025, BioSample: SAMN61091093, GenBank WGS Assembly: JCACXV000000000, version JCACXV010000000). All pipeline execution shell scripts, Python orchestrations, parameters, custom plotting scripts, and intermediate data files are publicly available on GitHub at https://github.com/hosniadilempa11y/AMR_Novel_Gene_Discovery.

Ethics Approval & Consent: Not applicable.

Acknowledgements: The authors wish to express their deepest gratitude to the late Professor Nasserdine Sabaou (Laboratoire de Biologie des Systèmes Microbiens [LBSM], Ecole Normale Supérieure de Kouba, Alger), who sadly passed away, for his inspiring mentorship and invaluable guidance.

References

[1] Murray, C.J.L., *et al.*: Global burden of bacterial antimicrobial resistance in 2019:

a systematic analysis. Lancet **399**(10325), 629–655 (2022) https://doi.org/10.1016/S0140-6736(21)02604-0

[2] Marin, C., *et al.*: Sequence type 354 expec strains. Frontiers in Microbiology **11**, 563832 (2020) https://doi.org/10.3389/fmicb.2020.563832

[3] Fuga, B., Sellera, F.P., Esposito, F., Moura, Q., Pillonetto, M., Lincopan, N.: Hybrid genome assembly of colistin-resistant mcr-1.5-producing Escherichia coli st354 reveals phylogenomic pattern associated with urinary tract infections in brazil. Journal of Global Antimicrobial Resistance **37**, 37–41 (2024) https://doi.org/10.1016/j.jgar.2024.02.016

[4] Bortolaia, V., *et al.*: Resfinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy **75**(9), 2510–2516 (2020) https://doi.org/10.1093/jac/dkaa345

[5] Feldgarden, M., *et al.*: Validating the ncbi amrfinderplus tool and database with diverse bacterial pathogens. Scientific Reports **9**(1), 19965 (2019) https://doi.org/10.1038/s41598-019-40755-7

[6] Touchon, M., et al.: Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genetics **5**(1) (2009) https://doi.org/10.1371/journal.pgen.1000344

[7] Ghalayini, M., *et al.*: Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. Microbial Genomics **8**(7), 000853 (2022) https://doi.org/10.1099/mgen.0.000853

[8] Partridge, S.R., et al.: Mobile genetic elements associated with antimicrobial resistance. Clinical Microbiology Reviews **31**(4) (2018) https://doi.org/10.1128/CMR.00088-17

[9] Silva, A., *et al.*: Evolutionary expression patterns in developing biofilms of uropathogenic Escherichia coli reveal embryo-like features. Current Microbiology **79**(8), 224 (2022) https://doi.org/10.1007/s00284-022-02914-9

[10] Gillings, R., M.: Class 1 integrons as biomarkers for the anthropocene. Mobile DNA **6**, 13 (2015) https://doi.org/10.3389/fmicb.2015.01459

[11] Carattoli, A.: Plasmids and the spread of resistance. Antimicrobial Agents and Chemotherapy **57**(6), 2524–2529 (2013) https://doi.org/10.1128/AAC.00289-13

[12] Robertson, J., et al.: Mob-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microbial Genomics **6**(6) (2020) https://doi.org/10.1099/mgen.0.000388

[13] Xie, Z., Tang, H.: Isescan: automated identification of insertion sequences in prokaryotic genomes. Bioinformatics **33**(21), 3340–3347 (2017) https://doi.org/10.1093/bioinformatics/btw673

[14] Bouvier, J.T., *et al.*: Integronfinder 2.0: identification and analysis of integrons in bacterial genomes. Microorganisms **10**(5), 1016 (2022) https://doi.org/10.1093/nar/gkac355

[15] Altschul, S.F., *et al.*: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990) https://doi.org/10.1016/0022-2836(90)90313-N

[16] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L., Tosatto, S.C., Paladin, L., Raj, S.-Y., Richardson, L.J., *et al.*: Pfam: The protein families database in 2021. Nucleic acids research **49**(D1), 412–419

(2021) https://doi.org/10.1093/nar/gkaa913
[17] Mirdita, M., *et al.*: Foldseek enables fast and sensitive comparison of large protein structure databases. Nature Biotechnology **40**(12), 1757–1764 (2022) https://doi.org/10.1038/s41587-022-01473-5
[18] Zhang, Y., Skolnick, J.: Tm-align: a protein structure alignment algorithm based on tm-score. Nucleic Acids Research **33**(7), 2302–2309 (2005) https://doi.org/10.1093/nar/gki524
[19] Jumper, J., *et al.*: Highly accurate protein structure prediction with alphafold. Nature **517**(7873), 583–589 (2021) https://doi.org/10.1038/s41586-021-03819-2
[20] Abramson, J., *et al.*: Accurate structure prediction of biomolecular interactions with alphafold 3. Nature **630**, 493–500 (2024) https://doi.org/10.1038/s41586-024-07487-w
[21] Lin, Z., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023) https://doi.org/10.1126/science.ade2574
[22] Rives, A., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS **118**(15) (2021) https://doi.org/10.1073/pnas.2016239118
[23] Jolley, K.A., Bray, J.E., Maiden, M.C.J.: Open-access bacterial population genomics: Bigsdb software, the pubmlst.org website and their applications. Wellcome Open Research **3**, 124 (2018) https://doi.org/10.12688/wellcomeopenres.14826.1
[24] Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**(1), 10–12 (2011) https://doi.org/10.14806/ej.17.1.200
[25] Andrews, S.: FastQC: a quality control tool for high throughput sequence data. 2010. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010). https://doi.org/10.1186/gb-2010-11-s1-p16 . http://www.bioinformatics.babraham.ac.uk/projects/fastqc
[26] Ewels, P., *et al.*: Multiqc: summarize analysis results for multiple tools and samples in a single report. Bioinformatics **32**(19), 3047–3048 (2016) https://doi.org/10.1093/bioinformatics/btw354
[27] Bankevich, A., *et al.*: Spades: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology **19**(5), 455–477 (2012) https://doi.org/10.1089/cmb.2012.0021
[28] Gurevich, A., *et al.*: Quast: quality assessment tool for genome assemblies. Bioinformatics **29**(8), 1072–1075 (2013) https://doi.org/10.1093/bioinformatics/btt086
[29] Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**(18), 3094–3100 (2018) https://doi.org/10.1093/bioinformatics/bty191
[30] Pedersen, B.S., Quinlan, A.R.: Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics **34**(5), 867–868 (2018) https://doi.org/10.1093/bioinformatics/bty301
[31] Li, H., *et al.*: The sequence alignment/map format and samtools. Bioinformatics

25(16), 2078–2079 (2009) https://doi.org/10.1093/bioinformatics/btp352
[32] Seemann, T.: Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**(14), 2068–2069 (2014) https://doi.org/10.1093/bioinformatics/btu153
[33] Seemann, T.: ABRicate: mass screening of contigs for antimicrobial resistance and virulence genes. GitHub. Available at: https://github.com/tseemann/abricate (2020). https://doi.org/10.1093/bioinformatics/btaa123 . https://github.com/tseemann/abricate
[34] Liu, B., et al.: VFDB 2019: a comparative pathogenomic platform with an interactive map. Nucleic Acids Research (2019). https://doi.org/10.1093/nar/gky1080
[35] Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J.G., Haendiges, J., Haft, D.H., Hoffmann, M., Pettengill, J.B., Prasad, A.B., Tillman, G.E., Tyson, G.H., Klimke, W.: Amrfinderplus and the reference gene catalog facilitate translation of genomics into public health monitoring of antimicrobial resistance. Scientific Reports **11**(1), 12728 (2021) https://doi.org/10.1038/s41598-021-91456-0
[36] Tonkin-Hill, G., *et al.*: Producing reliable pangenomes with panaroo. Genome Biology **21**(1), 180 (2020) https://doi.org/10.1099/mgen.0.000427
[37] Katoh, K., Standley, D.M.: Mafft multiple sequence alignment software. Molecular Biology and Evolution **30**(4), 772–780 (2013) https://doi.org/10.1093/molbev/mst010
[38] Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Haeseler, A., Lanfear, R.: Iq-tree 2: New models and efficient methods for phylogenomic reconstruction in the genomic era. Molecular Biology and Evolution **37**(5), 1530–1534 (2020) https://doi.org/10.1093/molbev/msaa015
[39] Nguyen, L.T., *et al.*: Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution **32**(1), 268–274 (2015) https://doi.org/10.1093/molbev/msu300
[40] Cock, P.J.A., *et al.*: Biopython: freely available python tools for computational biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009) https://doi.org/10.1093/bioinformatics/btp163
[41] Hunter, D., J.: Matplotlib: a 2d graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007) https://doi.org/10.1109/MCSE.2007.55
[42] Bouras, G., *et al.*: Baktfold: Sensitive protein functional annotation across the microbial tree of life. Nature Communications **17**, 1420 (2026) https://doi.org/10.1101/2026.01.15.633210
[43] Steinegger, M., *et al.*: Spacedust: Structure-sensitive pangenome clustering across giant bacterial & viral datasets. Nature Biotechnology **42**, 890–898 (2024) https://doi.org/10.1038/s41587-023-01900-5
[44] Zhang, Y., *et al.*: Arg-pass: Structural signature-guided prediction of divergent antimicrobial resistance genes. Bioinformatics **40**(6), 350 (2024) https://doi.org/10.1093/bioinformatics/btae120
[45] Terzian, P., *et al.*: Phynteny: Synteny- and plm-aware annotation of mobile genetic elements. The ISME Journal **18**(1), 045 (2024) https://doi.org/10.1038/s41396-024-01600-4
[46] Mistry, J., et al.: HMMER on the web: interactive database searches. Nucleic

Acids Research (2021). https://doi.org/10.1093/nar/gkab108
[47] Trott, O., Olson, A.J.: Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry **31**(2), 455–461 (2010) https://doi.org/10.1002/jcc.21334
[48] Liu, X., *et al.*: Geometry-enhanced protein language modeling enables discovery of novel antibiotic resistance genes. Nucleic Acids Research **54**(2), 112 (2026) https://doi.org/10.1093/nar/gkae150
[49] Wang, H., *et al.*: Accurate prediction of virulence factors using pre-trained protein language model and ensemble learning. BMC Bioinformatics **26**(1), 88 (2025) https://doi.org/10.1186/s12859-025-06010-x
[50] McInnes, L., et al.: UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018). https://doi.org/10.21105/joss.00861
[51] Pedregosa, F., *et al.*: Scikit-learn: machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011) https://doi.org/10.5555/1953048.2078195
[52] Waskom, L., M.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021) https://doi.org/10.21105/joss.03021
[53] Eastman, P., et al.: Openmm 7: rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology **13**(7) (2017) https://doi.org/10.1371/journal.pcbi.1005659
[54] Eastman, P., et al.: PDBFixer. GitHub repository (2017). https://doi.org/10.5281/zenodo.592651 . https://github.com/openmm/pdbfixer
[55] Bhalla, K., *et al.*: Large-scale cryptic proteome mining revealed potential phage-mediated host-pathogen genetic exchange. bioRxiv preprint (2026) https://doi.org/10.1101/2026.01.15.633210
[56] Gilchrist, C.L.M., Chae, R.: clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics **37**(16), 2483–2485 (2021) https://doi.org/10.1093/bioinformatics/btab008
[57] Virtanen, P., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature Methods **17**(3), 261–272 (2020) https://doi.org/10.1038/s41592-019-0686-2
[58] Harris, C.R., *et al.*: Array programming with numpy. Nature **585**(7825), 357–362 (2020) https://doi.org/10.1038/s41586-020-2649-2
[59] McKinney, W.: Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference (2010). https://doi.org/10.25080/Majora-92bf192d-00a
[60] Seabold, S., Perktold, J.: statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference (2010). https://doi.org/10.25080/Majora-92bf192d-011
[61] Wong, F., *et al.*: Benchmarking alphafold2-enabled molecular docking predictions for antibiotic discovery. Molecular Systems Biology **18**(9), 11081 (2022) https://doi.org/10.15252/msb.202211081
[62] Wick, R.R., Holt, K.E.: Benchmarking of long-read assemblers for prokaryotic

genomes. F1000Research **10**, 4 (2021) https://doi.org/10.12688/f1000research.27826.1

[63] Treangen, T.J., Salzberg, S.L.: Repetitive dna and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics **13**(1), 36–46 (2012) https://doi.org/10.1038/nrg3117

[64] Antipov, D., *et al.*: plasmidspades: assembling plasmids from whole genome sequencing data. Bioinformatics **32**(22), 3380–3387 (2016) https://doi.org/10.1093/bioinformatics/btw493

[65] Villa, L., Carattoli, A.: Plasmid-based epidemiology for multidrug-resistant enterobacteriaceae. Clinical Microbiology and Infection **19**(1), 9–16 (2013) https://doi.org/10.1111/1469-0691.12097

[66] Menezes, J., Silva, J., Fernandes, L., Pomba, C., *et al.*: Sharing of Escherichia coli st354 containing resistance elements between dogs and cohabiting humans. Eurosurveillance **27**(45), 2200293 (2022) https://doi.org/10.2807/1560-7917.ES.2022.27.45.2200293

[67] Zhou, Y., Fang, J., Davood, Z., Han, J., Qu, D.: Fitness cost and compensation mechanism of sulfonamide resistance genes (sul1, sul2, and sul3) in Escherichia coli. Environmental Microbiology **23**(11), 6912–6922 (2021) https://doi.org/10.1111/1462-2920.15732

[68] Llano-Sotelo, B., *et al.*: Aminoglycoside acetyltransferases. Antimicrobial Agents and Chemotherapy **46**(12), 3731–3737 (2002) https://doi.org/10.1128/AAC.46.12.3731-3737.2002

[69] Vetting, M.W., *et al.*: Structure and functions of the gnats. Archives of Biochemistry and Biophysics **433**(1), 212–226 (2005) https://doi.org/10.1016/j.abb.2004.09.003

[70] Ribeiro, C.A., Rahman, L.A., Holmes, L.G., Woody, A.M., Webster, C.M., Monaghan, T.I., Robinson, G.K., Mühlschlegel, F.A., Goodhead, I.B., Shepherd, M.: Nitric oxide (no) elicits aminoglycoside tolerance in Escherichia coli but antibiotic resistance gene carriage and no sensitivity have not co-evolved. Archives of Microbiology **203**(6), 3211–3220 (2021) https://doi.org/10.1007/s00203-021-02307-8

[71] Erdogdu, B., Dokuz, S., Sarigode, E., Karabasoglu, C., Ozbek, T.: Prophage: agent provocateur? Archives of Microbiology **208**(4), 183 (2026) https://doi.org/10.1007/s00203-026-04283-1

[72] Longin, K., *et al.*: Fold first, ask later: structure-informed function annotation of pseudomonas phage proteins. bioRxiv preprint (2025) https://doi.org/10.1101/2025.01.20.633900

[73] Mueller, M., *et al.*: The structure of a cytolytic alpha-helical toxin pore reveals its assembly mechanism. Nature **459**(7247), 726–730 (2009) https://doi.org/10.1038/nature08026

[74] Hoek, A.H.A.M., Lee, S., Berg, R.R., Rapallini, M., Overbeeke, L., Opsteegh, M., Bergval, I., Wit, B., Weijden, C., Giessen, J., Voort, M.: Virulence and antimicrobial resistance of shiga toxin-producing Escherichia coli from dairy goat and sheep farms in the netherlands. Journal of Applied Microbiology **134**(6), 093 (2023) https://doi.org/10.1093/jamp/lxad093

[75] Petkau, A., *et al.*: Redefining bacterial wzy polymerase families via 3d structure-based clustering. bioRxiv preprint (2026) https://doi.org/10.1101/2026.02.10.637500

[76] Ulsen, P., Zinner, K.M., Jong, W.S.P., Luirink, J.: On display: autotransporter secretion and application. FEMS Microbiology Letters **365**(18), 194 (2018) https://doi.org/10.1093/femsle/fny194

[77] Henderson, I.R., *et al.*: Type v protein secretion pathway: the autotransporter story. Microbiology and Molecular Biology Reviews **68**(4), 692–744 (2004) https://doi.org/10.1128/MMBR.68.4.692-744.2004

[78] Leo, J.C., Grin, I., Linke, D.: Type v secretion: mechanism(s) of autotransport through the bacterial outer membrane. Philosophical Transactions of the Royal Society B: Biological Sciences **367**(1592), 1088–1101 (2012) https://doi.org/10.1098/rstb.2011.0208

[79] Holm, L.: Using dali for protein structure comparison. Methods in Molecular Biology **2112**, 29–42 (2020) https://doi.org/10.1007/978-1-0716-0270-6_3

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialsPaper1.pdf
- FigS01DiscoveryPipelineSchema.png
- FigS02AssemblyCoverage.png
- FigS03VirulenceHeatmap.png
- FigS04ISDistribution.png
- FigS05AMRNetwork.png
- FigS06AMRComparison.png
- FigS07HemolysinPhylogeneticTree.png
- FigS08WzyPhylogeneticTree.png
- FigS09AutotransporterPhylogeneticTree.png
- FigS10StructuralAlignmentRMSD.png
- FigS11PLMCandidateHighlighttsne.png
- FigS12ESM28MUMAP.png
- FigS13MDPCAProjections.png
- FigS14GNATDCCMHeatmap.png
- FigS15MDSASAProfiles.png
- FigS16GNATContactMaps.png
- FigS17HoloMDTimeseries.png
- FigS18HoloMDRMSF.png
- FigS19DSSPTimeline.png
- FigS20PAEHeatmaps.png
- FigS21FoldseekFalsePositiveControl.png
- FigS22OAT371BarrelMDRMSD.png
- FigS23EnsembleDockingEnergetics.png
- FigS24WaterThermodynamics.png
- FigS25NMANormalModes.png
- FigS26GNATBenchmarkValidation.png
- FigS27Ehly61BilayerMD.png
- FigS28OAgP161BilayerMD.png