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.

CdMEC-A: An Automated Python Framework for Spatial Mobilome Profiling and Antimicrobial Resistance Risk Stratification in Clostridioides difficile

Reema Singh

reemasingh.gencompbio@gmail.com

LatchBio https://orcid.org/0000-0001-7049-5473

Method Article

Keywords: Bioinformatics, antimicrobial resistance, mobile genetic elements, horizontal gene transfer, software, genomic surveillance, Spatial genomics, Risk stratification, Clostridioides difficile

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11708297/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.

CdMEC-A: An Automated Python Framework for Spatial Mobilome Profiling and Antimicrobial Resistance Risk Stratification in *Clostridioides difficile*

Reema Singh[1,2*†]

[1] Independent Researcher, Saskatoon, SK, Canada

[2]LatchBio, San Francisco, CA, USA

[*]reemasingh.gencompbio@gmail.com

[†]**Author Note**: Present affiliation: LatchBio, San Francisco, CA, USA. This research was conducted independently by the author.

Abstract

Genomic surveillance of antimicrobial resistance (AMR) is frequently limited by computational workflows that catalog resistance genes in isolation while ignoring physical spatial context. Here, I present CdMEC-A, an open-source, highly scalable Python pipeline engineered to automate the detection of antimicrobial resistance gene (ARG) and mobile genetic element (MGE) associations, quantifying base-pair physical distances to inform horizontal gene transfer (HGT) risk. CdMEC-A systematically stratifies mobilization potential into actionable tiers: Stable, High-Risk, and Critical-Risk. Benchmarking against the *Clostridioides difficile* 630 reference genome demonstrated >99.5% feature mapping accuracy across complex genomic islands, conjugative transposons, and plasmid replicons. In a large-scale analysis of 495 C. difficile complete genomes, CdMEC-A identified 3,154 total associations, revealing that 10.9% of ARGs were embedded directly within MGE boundaries, while 83.2% resided in stable, unlinked chromosomal contexts. The pipeline exhibits near-perfect multi-threaded linear scaling across 16 CPU cores, achieving a high-throughput processing capacity exceeding 11,800 complete genomes per day. Designed with a modular architecture, CdMEC-A provides an immediate, standardized computational utility for high-throughput mobilome-aware AMR tracking, with extensible frameworks established for future integration into cross-reservoir One Health surveillance networks. CdMEC-A is implemented in Python and freely available on GitHub (https://github.com/ReemaSingh/CdMEC-A).

IMPORTANCE

AMR is increasingly propagated through MGEs that cross ecological boundaries. While current genomic pipelines catalog resistance genes, they lack the spatial resolution to determine how those genes are mobilized, failing to distinguish between stable chromosomal resistance and mobile cassettes poised for horizontal gene transfer. CdMEC-A addresses this gap by automating the quantification of physical base-pair distances between resistance genes and adjacent mobile elements. Processing over 11,800 genomes per day and standardizing risk into actionable tiers (Stable, High-Risk, and Critical-Risk), CdMEC-A provides a high-throughput tool to evaluate true resistance mobility. This capability shifts AMR surveillance from static inventory generation to predictive risk assessment across clinical and environmental microbiology.

Keywords: Bioinformatics, antimicrobial resistance, mobile genetic elements, horizontal gene transfer, software, genomic surveillance, Spatial genomics, Risk stratification, Clostridioides difficile

1. INTRODUCTION

Antimicrobial resistance (AMR) represents one of the most critical public health threats of the modern era, spanning complex interfaces across human, animal, and environmental domains. Advances in high-throughput whole-genome sequencing (WGS) have revolutionized AMR surveillance across these One Health sectors (1). However, existing computational and bioinformatic frameworks remain overwhelmingly centered on binary presence-or-absence profiling of antimicrobial resistance genes (ARGs) (2). While detecting an ARG confirms an organism's baseline genetic potential for resistance, simple gene identification fails to elucidate whether that gene is embedded within a mobilizable locus or how readily it may disseminate across ecological niches (3,4).

The evolution of AMR in pathogens across One Health reservoirs is largely driven by horizontal gene transfer (HGT), which allows bacteria to exchange genetic material (5). This process is facilitated by mobile genetic elements (MGEs), including plasmids, insertion sequences, and conjugative transposons (5). The potential for an ARG to be transferred is closely related to its genomic location, particularly its physical distance from nearby MGGs (6). ARGs located close to these elements are more likely to be mobilized and transferred between bacterial species and across different environmental settings than genes located within stable regions of the chromosome (6).

In multi-reservoir pathogens like *Clostridioides difficile*, distinguishing between intrinsic, vertically inherited resistance and actively mobilizable accessory cassettes is essential for accurate cross-sector risk assessment. As the primary cause of healthcare-associated diarrhea worldwide, *C. difficile* carries a 30-day mortality rate of 8.2% and an increasing

community-acquired burden (35%–41% of cases) (7,8). Mobile genetic elements comprise 11% to 30% of its genome, dominated by conjugative transposon families like Tn*916* and Tn*6194* carrying *tet(M)* and *erm(B)*, making mobilome architecture the central driver of its resistance dissemination (9, 10). However, manually inspecting assembly contigs to calculate spatial offsets, identify flanking insertion sequences, and evaluate structural co-localization across population-scale datasets creates a severe bioinformatic bottleneck, obstructing rapid One Health surveillance.

Existing computational tools often focus either broadly on generalized mobile element detection or strictly on resistance gene cataloging, leaving a significant gap in automated, high-throughput spatial risk stratification. While platforms like Comprehensive Antibiotic Resistance Database (CARD) (2) and AMRFinderPlus (11) catalog ARGs, and tools like MOB-suite (12) analyze plasmids, they operate in functional isolation and are not optimized for high-throughput distance calculations across integrative conjugative elements (ICEs). Genomic epidemiologists require a dedicated framework that bridges feature detection with precise coordinate mapping to translate raw genome assemblies into standardized, biologically meaningful risk tiers that reflect true mobility across host interfaces.

To address this critical gap, I developed CdMEC-A, an automated, open-source Python pipeline engineered to quantify base-pair distances between ARGs and MGEs at scale, systematically stratifying horizontal transfer risk into actionable tiers. Here, I detail the architecture of CdMEC-A, demonstrate its mapping precision against the *C. difficile* 630 reference genome, and benchmark its multi-threaded performance across 495 complete RefSeq assemblies. Finally, I illustrate its utility for mobilome-aware AMR surveillance, establishing a robust computational foundation for future expansion into automated, cross-reservoir One Health tracking networks.

2. MATERIALS AND METHODS

2.1 Data Selection and Validation: CdMEC-A was developed and validated using a multi-tiered genomic framework to resolve complex AMR architectures. The pipeline integrates the CARD (2) for protein-level resistance identification alongside a curated, species-specific database of *Clostridioides difficile* transposable elements and plasmids. To evaluate global scalability and spatial mapping precision, a benchmarking dataset comprising all complete NCBI RefSeq (13) *C. difficile* genomes (accessed 23 December 2025) was utilized. The selection of circularized assemblies over fragmented drafts was critical to ensuring gap-free mapping of high-risk structural architectures across diverse genome assemblies (Table S1). Mapping fidelity was verified by cross-referencing the *C. difficile* 630 reference genome (AM180355.1) (14) against expert annotations for multi-copy *erm(B)* loci within the Tn*5398* transposon to quantify positional precision. To benchmark cross-platform performance across an expanded panel of 25 distinct *C. difficile* isolates (Table S2), assemblies were parallel-processed through standalone RGI (v6.08; using 'Perfect and Strict' paradigms, with a ≥95% identity nudge for Loose hits, utilizing the 'High quality/Coverage' sequence selection criteria) and species-agnostic NCBI AMRFinderPlus (V3.12_2024-05-02.2; Galaxy wrapper; amrfinderplus database: V3.12-2024-05-02.2). To ensure an unbiased coordinate baseline, AMRFinderPlus (11) parameters were strictly constrained to≥80 nucleotide identity, ≥50% protein coverage, NCBI Translation Table 11, and a 5' flanking window size of 0 to capture exact local alignment boundaries. Failed predictions or database exclusions across baseline engines were explicitly logged as 'No Hit' to evaluate pipeline-specific rescue sensitivity.

2.1.1 Reference Database Compilation and Homology Filtering Thresholds

To ensure reproducible and high-fidelity identification of determinants, CdMEC-A couples curated genomic targets with rigid stringency criteria. For ARG identification, the comprehensive reference framework was established using the official Protein Homolog Model dataset extracted from the CARD (2). To definitively mitigate false-positive profiles stemming from domain-level alignments or non-functional structural homologs, the homology engine enforces a mandatory sequence identity constraint (≥80) alongside a minimum query sequence length coverage constraint (≥80), discarding non-significant matches (E-value ≥ 10^{-10}).

2.1.2 MGE Target Curation

To establish localized genomic context, a species-specific reference repository targeting *C. difficile* mobile genetic architectures was systematically compiled to yield the local database (i.e., combined_C_Diff_mge_nucl_db). This reference library comprises verified, high-confidence mobile elements categorized into two primary functional structural groups: conjugative/mobilizable transposons, and integrons, insertion sequences, or pathogenicity islands. Specifically, the database incorporates comprehensive sequence profiles for the Tn*916*-like superfamily, Tn*5397*, Tn*5398*, Tn*4453a-b*, Tn*6000*, Tn*6194*, Tn*6218*-like environments, and Tn*6944*. Additionally, core elements encompassing the *C. difficile* Insertion Site element 1 (CdISt1; GenBank: AJ579717.1), the Chloramphenicol-Florfenicol Resistance (CFR) localized region, and distinct tracking models covering the IS*30* and IS*L3* structural families were integrated into the database architecture. Definitive validation and alignment control baselines for genomic configuration testing were anchored via established standard reference coordinates including GenBank accessions AM180356.2, AY350745.1, and MG973074.1. All compiled nucleotide reference datasets were processed using standard NCBI "makeblastdb" protocols. To ensure computational interoperability and biological validity

during downstream spatial proximity mapping, match resolution thresholds for MGE profiles were strategically relaxed to a minimum of 70% sequence identity and 30% length coverage. This asymmetric filtering framework was intentionally established to accommodate the natural evolutionary mechanics of mobile elements, which frequently undergo truncation, fragmentation, or partial deletions over time, ensuring that degraded but epidemiologically active mobile genetic landscapes are captured without losing true resistance-linked genomic context.

2.1.3 Two-Tiered Validation of the MGE Reference Database

To guarantee the structural accuracy and biological specificity of the curated reference repository ,a two-tiered cross-validation benchmark was established prior to downstream analytical runs. First, structural alignment fidelity was evaluated via local command-line blastn (ncbi-blast-2.14.0+) (15) queries against the standard *Clostridioides difficile* 630 reference genome (AM180355.1), utilizing the pipeline's engineered filtering parameters (≥70% sequence identity and ≥30% length coverage). Second, to verify functional mobilization potential and eliminate stable core chromosomal background noise, an orthogonal profile Hidden Markov Model (pHMM) audit was executed. Sequences were subjected to a 6-frame translation and scanned using hmmscan (HMMER 3.4) (16) against the Pfam-A database (17) to confirm the presence of highly conserved, mobilization-specific catalytic domains (e-value < 10^{-10}).

2.2 Pipeline Architecture and Workflow

The overall architecture and computational workflow of the CdMEC-A framework are outlined in Figure 1. CdMEC-A is structured into four core functional modules:

1. **Homology Search and Feature Discovery:** Ingests FASTA assemblies and performs parallel homology searches. ARGs are identified via BLASTX (15) against CARD (2), while MGEs are detected via BLASTN (15) against the curated MGE database.
2. **Spatial Mapping and Distance Calculation:** The algorithm iterates through each identified ARG, scanning the surrounding genomic architecture on the same contig to compute the exact intergenic base-pair distance to upstream and downstream MGEs using a 1-based inclusive coordinate system.
3. **Proximity Thresholds and Categorization:** CdMEC-A applies a standardized 10,000 bp (10 kbp) scanning window to classify mobilization potential into mutually exclusive statuses:
    - **Embedded within MGE:** Characterized by an intergenic distance of 0 bp (overlapping or boundary-touching continuity) and categorized as Critical Risk due to maximum transfer potential.
    - **MGE-Associated (Upstream/Downstream):** Characterized by an MGE located within 10 kbp upstream or downstream of the ARG and categorized as High Risk due to elevated mobilization potential.
    - **Chromosomal (Unlinked):** Characterized by distances exceeding 10 kbp or an absence of MGE markers, categorized as a Stable State with low immediate transfer risk.
4. **Automated Reporting and Visualization:** Compiles structured JSON and TSV summary reports. An integrated visualization module generates global frequency distribution plots, tiered linear genomic maps of resistance clusters, and orientation analyses.

2.3 Computational Benchmarking Strategy

To find the fastest setup, computational performance was tested using two parallelization strategies. Task-level testing fixed the pipeline to process one genome at a time (-t 1) while increasing the CPU threads assigned to BLAST (15) searches (-bt 1, 4, 8), representing a typical desktop workstation setup. Process-level testing scaled the system to run multiple genomes at once (-t 4, 8, 16) using a single thread per search (-bt 1), simulating a high-throughput server. This approach tested both single-sample speed and large-scale processing capacity.

3. RESULTS

3.1 *In silico* Accuracy and Detection Sensitivity

To validate the spatial accuracy of CdMEC-A, I analyzed the *Clostridioides difficile* 630 reference genome (AM180355.1). The pipeline correctly identified both isogenic copies of the *erm(B)* gene, mapping them to the transposon Tn*5398* (AF109075.2) as 'Embedded within MGE' (Table 1), consistent with established genomic maps (18,19). Coordinates matched expert NCBI annotations with >99.5% accuracy (Supplementary Material). Routine $3\ bp$ shifts at start positions reflect standard start-codon (ATG) boundary calls, confirming pipeline precision when resolving duplicated ARGs within mobile elements.

Cross-platform benchmarking across 25 *C. difficile* isolates against RGI and NCBI AMRFinderPlus demonstrated high coordinate concordance (Supplementary Table S2). CdMEC-A matched AMRFinderPlus boundaries with a $0\ bp$ delta across shared loci. Minor $3\ bp$ shifts relative to RGI corresponded to open reading frame (ORF) translation starts, whereas localized truncations (such as a $-78\ bp$ shift in *ant(6)-Ia* and a $-909\ bp$ shift in *mef(H)*) reflected CdMEC-A's ability to terminate alignments at physical boundaries of degraded genes near MGEs.

CdMEC-A rescued four "Pipeline Unique" loci (*CDD-2*, *clbA*, *cfr(B)*, *AAC6_Ie_APH2_Ia*) where RGI returned no hits. AMRFinderPlus (11) exclusions highlighted database design differences rather than pipeline inaccuracies; for instance, native chromosomal markers (*cdeA*, *CdnimB*) excluded from NCBI clinical outbreak catalogs (11, 20) were successfully detected by CdMEC-A. Direct alignment architecture also resolved localized sequence dropouts (e.g., *cplR*), confirming diagnostic sensitivity across diverse assemblies.

3.2 Structural Validation of the MGE Reference Database

To confirm that the curated database accurately captures mobilome architecture, I mapped *C. difficile* 630 against the MGE repository using blastn (Supplementary Table S3). Under default filtering thresholds (>=70% identity, >= 30% coverage), the database resolved landmark mobile structures without chromosomal background noise.

The validation alignment captured an $81.8kb$ genomic island shared with *C. difficile* strain LSHTM-24 (CAAKAB010000006.1) at 98.61% identity ($E - value = 0.0$). Secondary High-Scoring Segment Pairs (HSPs) spanning $12.0kb$, $7.2kb$, $4.9kb$, and $4.2kb$ (>= 96.3% identity) down to sub-$500bp$ alignments resolved internal duplications and insertion sequences characteristic of modular pathogenicity islands.

The database also isolated classic integrated elements, including Tn*6194*-like (HG475346.1) and Tn*916*-like (KC414929.1) conjugative transposons ($9,448bp$ and $3,866bp$ alignments; $E - value = 0.0$), alongside the extra-chromosomal pCD630 plasmid replicon (AM180356.2).

Functional domain profiling via hmmscan against Pfam confirmed the purity of the MGE database (Supplementary Table S4). The repository resolved essential HGT machinery, including phage integrases (Phage_integrase), retroviral-like integrases (rve), site-specific transposases (Y1_Tnp), resolvases, and replication initiation factors (Rep_trans). Intercellular transfer systems (T2SSE, TraG-D_C) and mobile resistance markers (*ermC*, class A/D β-lactamases) were strongly represented. Non-mobile metabolic domains (e.g., glycosyl hydrolases) were localized exclusively within large modular islands where metabolic machinery co-evolves via hitchhiking. The absence of random chromosomal background confirms high target specificity.

3.3 Comparative Evaluation of CdMEC-A against Existing Frameworks

To evaluate functional positioning, CdMEC-A was benchmarked against established bioinformatic tools (Table 2). While tools like AMRFinderPlus (11) and RGI provide sensitivity for ARG detection, they lack automated spatial distance calculations to contextualize resistance within the mobilome. Conversely, MOB-suite (12) and VRprofile2 (21) offer insights into plasmid mobility and flanking regions but are not optimized for high-throughput, base-pair distance quantification across conjugative transposon (CTn) landscapes.

CdMEC-A bridges this gap by unifying ARG detection and MGE coordinate mapping into a single automated workflow. By enforcing a $10kb$ scanning window, the tool enables objective risk stratification ('Embedded', 'MGE-Associated', or 'Chromosomal'), moving beyond presence/absence matrices toward actionable horizontal transfer risk scoring.

3.4 Large-Scale Genomic Validation and Performance Benchmarking

3.4.1 High-Resolution Mapping of the *C. difficile* Mobilome

I deployed CdMEC-A across 495 complete *C. difficile* RefSeq genomes (Supplementary Table S1), identifying 3,154 ARG-MGE associations (Table 3). Most features resided in stable chromosomal contexts (2,625 ARGs; 83.2%). However, CdMEC-A resolved an active mobilome footprint: 344 ARGs (10.9%) were directly embedded within MGE boundaries (Critical Risk; Figure 2A). Flanking associations accounted for the remainder, with 111 ARGs (3.5%) positioned upstream and 74 (2.3%) downstream within $10kb$.

Distance calculations revealed a dense spatial architecture across the $10kb$ window, with a mean intergenic distance of $1,705.24bp$, a median of $0.00bp$, and a range of $0\ to 9,648bp$ (

$SD = 3,121.86bp$). The median value of $0.00bp$ reflects the high proportion of ARGs embedded directly within mobile vehicles. The clustering of non-embedded flanking features within hyper-proximal hotspots (Figure 2B) underscores a genomic architecture primed for rapid resistance exchange.

Although CdMEC-A was validated here using a species-wide dataset, its core parsing engine operates independently of host origin. The pipeline natively ingests FASTA assemblies from human clinical, agricultural, and environmental isolates, providing an immediate computational tool to calculate base-pair ARG–MGE distances across diverse ecological niches without requiring pipeline modifications.

3.4.2 Computational Benchmarking and Scalability

Evaluating performance across 495 complete genomes (Supplementary Table S1) revealed critical scaling properties along process-level (-t) and task-level (-bt) execution axes (Table 4). Single-core baseline execution required $954.29minutes$ ($\sim 15.9hours$; speedup $S = 1.0$, efficiency E = 100%).

CdMEC-A achieved maximum acceleration exclusively through process-level scaling (-t) (Figure 3). Task-level scaling (-bt) across 4 and 8 cores within a single process yielded flat-lined execution times of $971.77\ minutes$ ($S = 0.98$, E = 24.55%) and $962.22\ minutes$ ($S = 0.99$, E = 12.40%), respectively, due to internal threading overhead during database queries.

Conversely, process-level execution (-t) demonstrated linear scaling. Distributing assemblies across 4, 8, and 16 concurrent worker processes yielded speedups of 4.00, 7.88, and 15.89. The 16-core configuration compressed total execution time to $60.05minutes$ ($1hour3seconds$)

while maintaining 99.32% hardware efficiency, demonstrating a processing throughput exceeding 11,800 complete genomes per 24-hour period ($\sim 7.3 seconds/genome$).

3.5 Future Directions: Cross-Reservoir One Health Integration

While CdMEC-A currently provides high-throughput spatial distance calculations for individual genomic assemblies, future development will expand the framework into an automated One Health comparative platform. Planned feature updates include:

- **Metadata-Driven Cohort Comparisons:** Automated parsing of BioSample metadata (e.g., host species, isolation source, geographic location) to compute differential mobilization risk scores across human, animal, and environmental interfaces.
- **Conserved Mobilome Structural Tracking:** Algorithmic detection of shared, identical ARG–MGE junction signatures (e.g., conserved physical base-pair offsets across plasmids and transposons) to trace horizontal gene transfer events across species barriers.
- **Dynamic Database Updaters:** Automated syncing with expanding One Health resistance catalogs and emerging transposable element databases to maintain surveillance sensitivity across non-model species.

4. DISCUSSION

The exceptional spatial precision and detection sensitivity exhibited by CdMEC-A against the benchmark *Clostridioides difficile* 630 reference genome, alongside cross-platform comparisons with RGI and NCBI AMRFinderPlus (9), validate the structural and diagnostic fidelity of this framework for large-scale genomic surveillance. By resolving both isogenic copies of the *erm(B)* gene embedded within the Tn*5398* transposon, the pipeline demonstrates a robust algorithmic capacity to navigate complex duplications and historical fragmentation events without generating artifactual copy-number inflation (16,17). Furthermore, the systematic $3\,bp$ shifts observed relative to RGI align precisely with predictive open reading frame (ORF) translation starts, while localized truncation captures (such as those observed in *ant(6)-Ia* and *mef(H)*) underscore the tool's capacity to accurately map degraded yet epidemiologically relevant mobile genetic landscapes. Crucially, CdMEC-A's ability to rescue "Pipeline Unique" loci (*CDD-2*, *clbA*, and *cfr(B)*) where standalone platforms fell short, combined with its programmatic resilience against reference database curation gaps (such as NCBI's omission of core chromosomal elements like *cdeA* and *CdnimB*), highlights the clear operational advantage of its direct alignment architecture. Rather than indicating discrepancies, these variations reflect fundamental nomenclature and curation design choices between the CARD (8) and NCBI frameworks (11,13, 20). While CARD includes intrinsic chromosomal multidrug efflux components (*cdeA*) and basal enzymes requiring specific promoter mutations for phenotypic resistance (*CdnimB*), automated public pathogen databases like NCBI Pathogen Detection intentionally restrict outbreak catalogs to acquired determinants and validated point mutations to minimize surveillance redundancy. By integrating these distinct paradigms into a unified spatial resolution engine, CdMEC-A

minimizes false negatives and establishes a dependable, reproducible baseline for tracking resistance-linked genomic contexts across large-scale datasets.

From a computational standpoint, the development of CdMEC-A addresses a critical operational void in microbial genomics. While powerful platforms exist for independent ARG cataloging (2, 11) or plasmid typing (12), the manual cross-referencing of coordinate outputs across hundreds of large datasets introduces an unsustainable data-parsing bottleneck. CdMEC-A bypasses this limitation by integrating homology exploration and spatial distance parsing into a standardized runtime environment. Benchmarking experiments conducted across 495 complete *C. difficile* genomes demonstrated that task-level multi-threading yields negligible speedups due to internal parallelization overhead within individual alignment processes. In contrast, the decentralized process-level file sharding architecture scaled with linear efficiency. By compressing a 15.9-hour sequential workload down to just 60 minutes on a 16-core configuration while maintaining a 99.3% hardware utilization efficiency, CdMEC-A provides the computational scalability required for real-time surveillance. This performance transitions genomic surveillance from static gene cataloging to high-throughput, spatial mobilization-risk mapping across large-scale epidemiological cohorts.

While the pipeline achieves high spatial coordinate resolution across complete genomic assemblies, these computational insights establish a foundation for future functional and experimental validation. Although a $10kb$ scanning window serves as a robust genomic proxy for likely co-mobilization, future long-read sequencing applications across dynamic, actively replicating cultures will be required to capture intermediate circularized transposable forms and confirm physical co-transfer kinetics. Additionally, planned computational expansions of the pipeline's reference databases to encompass an even wider array of emerging non-clostridial

mobile genetic structures will continue to refine CdMEC-A's capacity to map rare, inter-species resistance flux across automated, cross-reservoir One Health tracking networks.

5. CONCLUSION

This study presents the development, validation, and deployment of CdMEC-A, an open-source bioinformatic pipeline that automates spatial coordinate analysis to quantify the mobilization potential of ARGs in *Clostridioides difficile*. Through a multi-tiered validation approach, I demonstrate that the pipeline achieves base-pair resolution structural precision when resolving complex genomic architectures, including multi-copy isogenic loci, while establishing a standardized, highly reproducible metric for mobilome-aware risk assessment. Combined with its verified process-level scalability (>11,800 complete genomes/day), CdMEC-A provides a robust, high-throughput foundation to shift AMR surveillance from static gene cataloging toward automated, mobilome-aware risk assessment across large-scale genomic cohorts.

DATA AVAILABILITY

The genomic sequences analyzed in this study were retrieved from the NCBI RefSeq database (Accession numbers provided in Supplementary Table S1). The CdMEC-A pipeline and source code are available at https://github.com/ReemaSingh/CdMEC-A.

COMPETING INTERESTS

The author declares no competing interests.

AUTHOR CONTRIBUTIONS

Reema Singh: Conceptualization, Methodology, Software (CdMEC-A), Validation, Formal Analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing. Reema Singh is the sole author and was responsible for all aspects of this study.

ACKNOWLEDGMENTS

The author utilized Gemini (Google) for language refinement, structural flow, and grammatical consistency during the preparation of the manuscript. All suggestions were critically reviewed and edited by the author, who maintains full accountability for the integrity of the final work.

REFERENCES

1. Djordjevic SP, Jarocki VM, Seemann T, Cummins ML, Watt AE, Drigo B, Wyrsch ER, Reid CJ, Donner E, Howden BP. 2024. Genomic surveillance for antimicrobial resistance—a One Health perspective.Nat Rev Genet 25:142-57.
2. Alcock BP, Raphenya AR, Lau TT, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen ALV, Cheng AA, Liu S, Min SY. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res*, *48*:D517-D525.
3. Partridge SR, Kwong SM, Firth N, Jensen SO. 2018. Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev* 31:10-1128.
4. Klümper U, Fang P, Li B, Xia Y, Frigon D, Hamilton KA, Quon H, Berendonk TU, de la Cruz Barron M. 2025. Towards the integration of antibiotic resistance gene mobility into environmental surveillance and risk assessment. npj Antimicrob Resist 3:81
5. Baquero F, Coque TM, Martínez JL, Aracil-Gisbert S, Lanza VF. 2019. Gene transmission in the one health microbiosphere and the channels of antimicrobial resistance. *Front Microbiol* 10:2892.
6. Grodner B, Shi H, Farchione O, Vill AC, Ntekas I, Diebold PJ, Wu DT, Chen CY, Kim DM, Zipfel WR, Brito IL. 2024. Spatial mapping of mobile genetic elements and their bacterial hosts in complex microbiomes. *Nat Microbiol* 9:2262–2277.
7. Di Bella S, Sanson G, Monticelli J, Zerbato V, Principe L, Giuffrè M, Pipitone G, Luzzati R. 2024. *Clostridioides difficile* infection: history, epidemiology, risk factors, prevention, clinical manifestations, treatment, and future options. *Clin Microbiol Rev* 37:e00135-23.

8. Canadian Nosocomial Infection Surveillance Program. 2025. Healthcare-associated infections and antimicrobial resistance in Canadian acute care hospitals, 2019–2023. *Can Commun Dis Rep* 51:249.
9. Mullany P, Allan E, Roberts AP. 2015. Mobile genetic elements in *Clostridium difficile* and their role in genome function. *Res Microbiol* 166:361–367.
10. Schüler MA, Riedel T, Overmann J, Daniel R, Poehlein A. 2024. Comparative genome analyses of clinical and non-clinical *Clostridioides difficile* strains. Front Microbiol 15:1404491.
11. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep* 11:12728.
12. Robertson J, Nash JH. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 4:e000206.
13. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:D61–D65.
14. Jose S, Mukherjee A, Abhyankar MM, Leng L, Bucala R, Sharma D, Madan R**.** 2018. Neutralization of macrophage migration inhibitory factor improves host survival after *Clostridium difficile* infection. *Anaerobe* 53:56–63.
15. McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32:W20–W25.
16. Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7:e1002195.

17. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL. 2014. Pfam: the protein families database. *Nucleic Acids Res* 42:D222–D230.
18. Farrow KA, Lyras D, Rood JI. 2000. The macrolide-lincosamide-streptogramin B resistance determinant from *Clostridium difficile* 630 contains two *erm*(B) genes. *Antimicrob Agents Chemother* 44:411–413.
19. Hussain HA, Roberts AP, Mullany P. 2005. Generation of an erythromycin-sensitive derivative of *Clostridium difficile* strain 630 (630Δ*erm*) and demonstration that the conjugative transposon Tn*916*ΔE enters the genome of this strain at multiple sites. *J Med Microbiol* 54:137–141.
20. Feldgarden M, Brover V, Fedorov B, Haft DH., Prasad AB, Klimke W. 2022. Curation of the AMRFinderPlus databases: applications, functionality and impact. *Microb Genom 8*:000832.
21. Wang M, Goh YX, Tai C, Wang H, Deng Z, Ou HY. 2022. VRprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens. *Nucleic Acids Res* 50:W768–W773.
22. Seemann T. 2026. Abricate: mass screening of bacterial genome assemblies against curated gene sequence databases. https://github.com/tseemann/abricate.

Figure Legends and Tables

Figure 1. Schematic representation of the CdMEC-A pipeline architecture and spatial risk stratification workflow.

Figure 2. ARG-MGE genomic proximity distribution across 495 *Clostridioides difficile* genomes.

Figure 3. Performance bottleneck analysis comparing task-level versus process-level parallelization strategies in CdMEC-A.

Table 1 : Validation of CdMEC-A positional accuracy against *C. difficile* 630 reference

Feature	**Strand**	**Reference Coordinates (NCBI)**	**CdMEC-A Coordinates**	**Base Pair (bp) Shift**	**Spatial Status**	**MGE Identity**
erm(B) (Copy 1)	Complement	2,316,967 – 2,317,704	2,316,970 – 2,317,704	3 bp*	Embedded within MGE	Tn5398 (AF109075.2)
erm(B) (Copy 2)	Complement	2,319,374 – 2,320,111	2,319,377 – 2,320,111	3 bp*	Embedded within MGE	Tn5398 (AF109075.2)

Note: *The 3-bp offset is systematic and corresponds to the identification of the translational start codon (ATG), whereas RefSeq annotations may include upstream leader sequences.* All distances and 'Embedded' status assignments were verified to have 100% computational fidelity, with the 'Embedded' status correctly identifying null intergenic spacing (0 bp distance) across all multi-copy *erm(B)* loci.

Table 2. Feature-based comparison of CdMEC-A with widely utilized antimicrobial resistance and mobile genetic element detection platforms. This table highlights the unique capacity of CdMEC-A to provide automated spatial linkage, risk-based classification, and integrated visualization for high-throughput bacterial genomic datasets.

Feature	**AMRFinderPlus (11)**	**RGI (CARD) (8)**	**VRprofile2 (21)**	**ABRicate (22)**	**CdMEC-A (Present study)**
ARG Detection	Yes	Yes	No	Yes	Yes
MGE Detection	No	No	Yes	(Partial)[1]	Yes
Automated ARG–MGE Linkage	No	No	(Partial)[2]	No	Yes
Integrated Spatial Visualization	No	Limited	Yes	No	Yes
High-Throughput Risk Profiling[3]	No	No	No	No	Yes

Optimized for *C. difficile* Context	Yes	Yes	Yes	Yes	Yes
Specialized CTn/ICE Detection	No	No	Partial	No	Yes

Note: MOB-suite identifies plasmid-bound genes but does not calculate intergenic distances for risk-stratification of chromosomal integrated elements. [1]ABRicate supports plasmid databases but does not identify integrated elements (ICEs/Transposons) autonomously. [2]VRprofile2 identifies flanking regions but lacks a standardized distance-based risk classification system for large datasets. [3]Refers to the ability to batch-process hundreds of genomes and generate standardized mobility reports.

Table 3: Summary Statistics for the CdMEC-A Pipeline Analysis

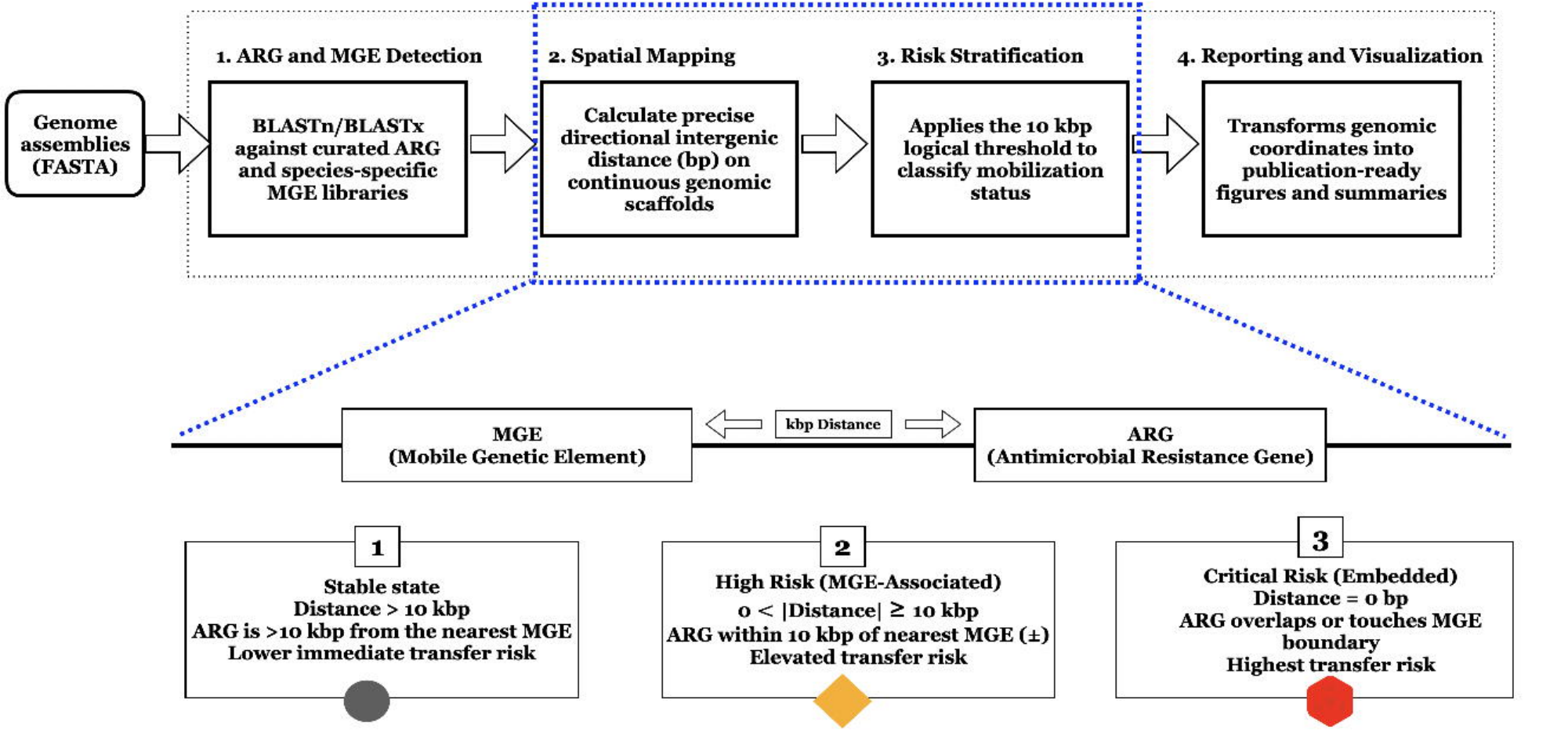
Total Samples Analyzed:	495	
Total Associated ARGs Found:	3154	
MGE Association Classifications	MGE-Associated (Upstream)	111 (3.5%)
	MGE-Associated (Downstream)	74 (2.3%)
	Embedded within MGE	344 (10.9%)
	Chromosomal (Unlinked)	2625 (83.2%)
Proximity Analysis (Distance in bp)	Mean Distance	1705.24 bp
	Median (50%)	0.00 bp
	Std Deviation	3121.86 bp
	Range	0 - 9648 bp

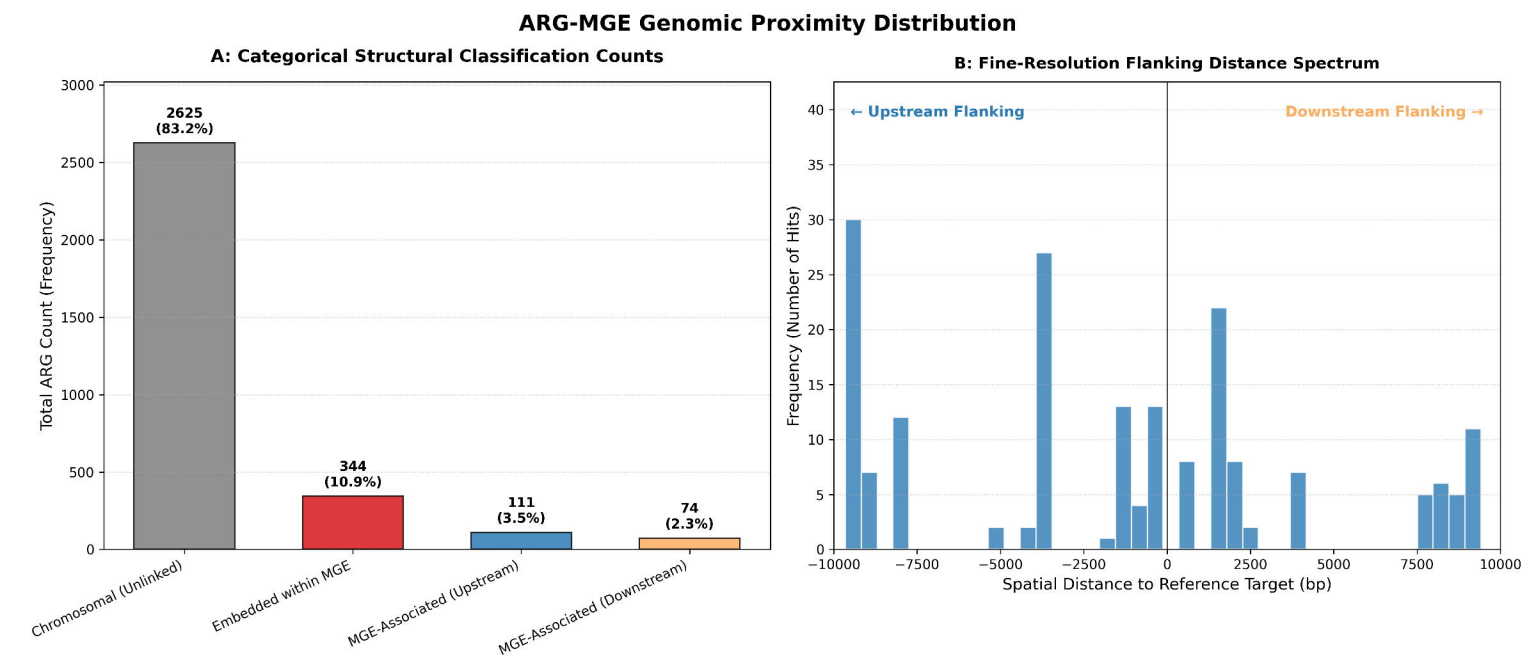
Table 4: High-Performance Computing Scaling Performance and Parallelization Efficiency Profiles of the CdMEC-A Pipeline.

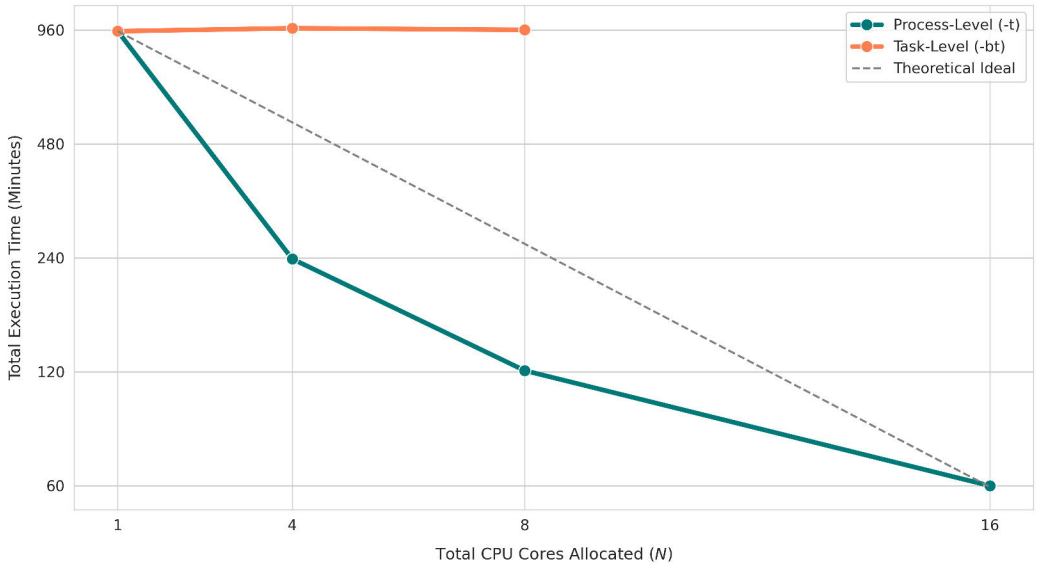
Scaling Mode (CLI Flag)	Allocated Cores (N)	Wall_Time_Actual_Min	Speedup	Efficiency(E, %)
Process-Level (-t)	1	954.294333	1.000000	100.000000
	4	238.844167	3.995468	99.886711
	8	121.027167	7.884960	98.561996
	16	60.053333	15.890780	99.317378
Task-Level (-bt)	1	954.294333	1.000000	100.000000
	4	971.771500	0.982015	24.550379
	8	962.216000	0.991767	12.397091

Note: Performance metrics were systematically evaluated using a standardized validation workload across a dataset of 495 isolates. Runtimes are presented directly in minutes (Wall_Time_Actual_Min). Speedup (S) represents a unitless ratio of computational acceleration calculated as $S = T_1/T_N$, where T_1 is the sequential baseline execution time on a single core (954.294333 min) and T_N is the observed execution runtime utilizing N total allocated CPU cores. Parallel Efficiency (E, %) evaluates real-world hardware optimization relative to ideal linear scaling behaviors and is calculated via the formula $E = (S/N) \times 100\%$. Under this formal metric, the single-core baseline configuration establishes an unsharded relative performance anchor profile (S = 1.000000, E = 100.000000 %). Extrapolating these metrics to high-throughput performance capacity, processing a batch of 495 isolates in 60.053333 min under maximum core allocation (N = 16) permits a daily operational throughput of 23.9787 discrete workflow cycles per 24-hour period (1,440 min / 60.053333 min), demonstrating a

scalable framework capable of analyzing 11,869.45 complete bacterial genomes per day (~ 11,800 genomes/24 h) on a single standard 16-core compute node.


1. ARG and MGE Detection
2. Spatial Mapping
3. Risk Stratification
4. Reporting and Visualization
Genome assemblies (FASTA)
BLASTn/BLASTx against curated ARG and species-specific MGE libraries
Calculate precise directional intergenic distance (bp) on continuous genomic scaffolds
Applies the 10 kbp logical threshold to classify mobilization status
Transforms genomic coordinates into publication-ready figures and summaries
MGE
(Mobile Genetic Element)
kbp Distance
ARG
(Antimicrobial Resistance Gene)
1
Stable state
Distance > 10 kbp
ARG is >10 kbp from the nearest MGE
Lower immediate transfer risk
2
High Risk (MGE-Associated)
0 < |Distance| ≥ 10 kbp
ARG within 10 kbp of nearest MGE (±)
Elevated transfer risk
3
Critical Risk (Embedded)
Distance = 0 bp
ARG overlaps or touches MGE boundary
Highest transfer risk


ARG-MGE Genomic Proximity Distribution
A: Categorical Structural Classification Counts
Total ARG Count (Frequency)
3000
2500
2000
1500
1000
500
0
2625
(83.2%)
344
(10.9%)
111
(3.5%)
74
(2.3%)
Chromosomal (Unlinked)
Embedded within MGE
MGE-Associated (Upstream)
MGE-Associated (Downstream)
B: Fine-Resolution Flanking Distance Spectrum
← Upstream Flanking
Downstream Flanking →
Frequency (Number of Hits)
40
35
30
25
20
15
10
5
0
−10000
−7500
−5000
−2500
0
2500
5000
7500
10000
Spatial Distance to Reference Target (bp)


Performance Bottleneck Analysis: Task vs. Process Parallelization
Process-Level (-t)
Task-Level (-bt)
Theoretical Ideal
Total Execution Time (Minutes)
960
480
240
120
60
1
4
8
16
Total CPU Cores Allocated (N)

Supplementary Files

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

- SupplementaryTables.xlsx
- SupplementaryMaterial.pdf