SEMARAK ILMU PUBLISHING

Semarak International Journal of Machine Learning

Journal homepage:
https://semarakilmu.my/index.php/sijml/index
ISSN: 3030-5241

In Silico Functional Prioritization of Hypothetical Proteins Associated with Multidrug Resistance in *Acinetobacter baumannii* using PRISM Computational Framework

Muhammad Faiz Anuar[1], Siti Munirah Mohd[1,2,*], Djoko Budiyanto Setyohadi[3]

[1] Kolej PERMATA Insan, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
[2] Education & Advanced Sustainability Research Unit, Kolej PERMATA Insan, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800, Nilai, Negeri Sembilan, Malaysia
[3] Informatics Department, Universitas Atma Jaya Yogyakarta, Yogyakarta Indonesia

ARTICLE INFO

Article history:
Received 13 January 2026
Received in revised form 26 February 2026
Accepted 10 March 2026
Available online 31 March 2026

Keywords:
Acinetobacter baumannii; antimicrobial resistance; hypothetical proteins; deep learning

ABSTRACT

Acinetobacter baumannii is a critical-priority pathogen whose pervasive multidrug-resistant (MDR) phenotype poses a global clinical crisis. A major challenge in combating this pathogen is the vast number of hypothetical proteins that remain uncharacterized, representing a hidden reservoir of potential resistance mechanisms that evade standard bioinformatics tools. The primary objective of this study is to computationally prioritize these hypothetical proteins to uncover novel therapeutic targets, specifically addressing the limitation of biologically implausible predictions often generated by standard unconstrained deep learning models. To achieve this, we utilized the PRISM (Protein Recognition Insight via Sequence Modelling) framework, which integrates protein language model embeddings with strict biological logic. The analysis revealed twelve cryptic Gcn5-related N-acetyltransferase (GNAT) determinants systematically missed by consensus-based annotation. By enforcing ontological constraints, PRISM reduced the experimental search space by 99.5%, significantly outperforming standard baselines (McNemar's test, $p < 0.001$) and drastically reducing biological hallucinations from 15.99% to 0.04%. Topology modelling demonstrated that these enzymes utilize a non-canonical "border security" strategy, confirming their sequestration in the cell membrane or secretion to the periplasm via N-terminal signal peptides. Additionally, our analysis established a novel class of minimalist 65-amino acid resistance proteins and resolved a systemic "metabolic bias" in global repositories by reassigning mislabelled metabolic proteins as secreted defense determinants. Ultimately, these findings redefine the genomic "dark matter" of *A. baumannii* as a strategically localized reservoir of enzymatic defense, providing a critical, cost-effective corrective layer for antimicrobial surveillance.

1. Introduction

Antimicrobial resistance (AMR) is accelerating at an alarming pace, threatening to undermine decades of medical progress. Each year, millions of infections resist standard antibiotics, costing healthcare systems over USD 35 billion and claiming thousands of lives [1]. If current trends continue,

* *Corresponding author.*
E-mail address: smunirahm@usim.edu.my

AMR could cause 10 million deaths annually by 2050, potentially surpassing cancer as the leading cause of global mortality [2]. The rise of multidrug-resistant pathogens is not confined to isolated outbreaks; they increasingly compromise routine medical procedures. For instance, in patients undergoing hip and knee arthroplasty, postoperative infections caused by multidrug-resistant Gram-negative bacteria were associated with a treatment failure rate of 55.6%, compared to only 8.3% for infections caused by multidrug-sensitive bacteria [3]. Such real-world outcomes underscore the urgent need for innovative therapeutic strategies to combat multidrug-resistant infections.

Building on the clinical risks of multidrug-resistant infections, *Acinetobacter baumannii* is a Gram-negative bacterium that plays a major role in the antimicrobial resistance crisis due to its ability to survive in harsh clinical environments and rapidly adapt to antibiotics [4]. Its fast development of multidrug resistance has led the World Health Organization (WHO) to classify carbapenem-resistant *A. baumannii* (CRAB) as a "critical-priority" pathogen. In Malaysia, national surveillance data indicated that carbapenem resistance in *A. baumannii* remained alarmingly high at over 60% in 2022, up from approximately 40% in 2018 [5]. Providing genomic context for this nationwide trend, a longitudinal genomic investigation demonstrated the persistent predominance of the multidrug-resistant Global Clone 2 lineage in a Malaysian tertiary hospital over a 10-year period [6]. These trends highlight the persistence of MDR strains and the urgent need to identify new therapeutic targets.

In addition to its multidrug resistance, another major challenge in studying multidrug-resistant *A. baumannii* is the large portion of its proteins that remain uncharacterized, known as hypothetical proteins (HPs). HPs are predicted from DNA sequences but lack experimental confirmation or reliable functional annotation. They are labelled "hypothetical" when database searches, protein domains, and scientific literature cannot determine their role. Many HPs fall into what bioinformaticians call the "twilight zone" of sequence similarity, meaning they share very little sequence similarity with known proteins (typically less than 20–30%), so standard tools like BLAST cannot reliably predict their function [7]. This largely unexplored portion of the bacterial proteome may conceal important factors involved in virulence and antibiotic resistance, making HPs a critical focus for computational and experimental studies.

These uncharacterized proteins may include cryptic virulence factors and resistance enzymes, such as transferases, that contribute to the MDR phenotype but evade detection using standard bioinformatics approaches. Because of this hidden potential, HPs represent an important resource for computational analysis, enabling the prioritisation of proteins for experimental validation and the development of new therapeutics.

Recent advances in computational biology enable the in silico prediction of protein function and resistance mechanisms. Earlier studies demonstrated that traditional machine learning algorithms, such as Random Forests and Support Vector Machines, can infer resistance phenotypes directly from *A. baumannii* genomic data [8]. Subsequent comparative investigations reported that deep learning architectures further improve predictive performance for antimicrobial resistance compared to classical approaches [9]. Protein language models (pLMs) offer an alignment-free alternative, learning evolutionary patterns directly from sequences. These models are often "unconstrained," meaning they make predictions based solely on sequence patterns without incorporating biological rules or logic.

Despite recent advances in computational biology, a significant research gap remains: unconstrained protein language models often generate biologically implausible predictions based solely on sequence patterns, reducing confidence when prioritising uncharacterized targets for costly downstream experiments. To address this limitation, the primary objective of this research is to apply the PRISM (Protein Recognition Insight via Sequence Modelling) framework to computationally

prioritise hypothetical proteins across three clinical strains of multidrug-resistant *A. baumannii*. Unlike previous approaches, PRISM enforces logical and ontological constraints to ensure that the outputs are biologically meaningful while still capturing complex evolutionary features. By enabling the discovery of cryptic aminoglycoside-modifying enzymes that traditional methods miss, this study provides a highly targeted, cost-effective strategy for experimental validation, significantly advancing the identification of novel therapeutic interventions against critical-priority MDR pathogens.

2. Methodology

2.1 Study Design and Strain Selection

This study aimed to computationally prioritise hypothetical proteins (HPs) in multidrug-resistant (MDR) *Acinetobacter baumannii*, with the goal of identifying proteins that may contribute to antibiotic resistance or virulence and could serve as candidates for experimental validation. Three strains were selected to represent diverse MDR profiles and research contexts: ACICU, a multidrug-resistant clinical isolate; ATCC 17978, a widely used laboratory reference strain; and ATCC 19606, a type strain.

Genomic sequences for these strains were retrieved from the UniProt Knowledgebase [10], providing access to well-annotated, publicly available data. Protein sequences annotated as "hypothetical protein" or "uncharacterized protein" were extracted for further analysis. To reduce noise from fragmented open reading frames, sequences shorter than 50 amino acids were excluded. The final dataset comprised 2,265 HPs across the three strains, including 1,016 from ATCC 17978, 731 from ATCC 19606, and 518 from ACICU.

The three selected strains differ in their MDR profiles, genome sizes, and the number of hypothetical proteins (Table 1). ACICU, the clinical isolate with the highest resistance, contains 518 HPs, whereas the laboratory reference strain ATCC 17978 has the largest number of HPs (1,016), reflecting differences in genome annotation and functional characterisation. These differences justify the inclusion of all three strains for comprehensive computational prioritisation.

Table 1
Summary of the three *A. baumannii* strains used in this study

Strain ID	Source / Origin	MDR Profile	Genome Size (Mb)	Total Proteins	HPs Identified	Sources
ACICU	Clinical isolate (Europe, clone II)	High	4.014	3,746	518	NCBI RefSeq (ASM551913v2) [11]; UniProt (UP000008839) [10]
ATCC 17978	Laboratory reference strain (clinical origin)	Moderate	4.030	3,799	1,016	NCBI RefSeq (ASM1337208v1) [11]; UniProt (UP000006737) [10]
ATCC 19606	Human Urine	Low	3.945	3,765	731	NCBI RefSeq (ASM1933165v1) [11]; UniProt (UP000005740) [10]

2.2 Identification of Hypothetical Proteins

Hypothetical proteins (HPs) were defined as protein-coding sequences annotated as "hypothetical" or "uncharacterized" with no experimentally validated function and lacking reliable sequence homology to known proteins. These sequences were extracted directly from the UniProt Knowledgebase entries corresponding to each of the three selected strains.

To ensure data quality, fragmented open reading frames and incomplete sequences were filtered out by excluding proteins shorter than 50 amino acids. No additional annotation tools (e.g., Prokka or RAST) were applied, as the study focused on proteins already classified as hypothetical or uncharacterized in the reference databases.

After filtering, the final dataset included 2,265 HPs: 1,016 from ATCC 17978, 731 from ATCC 19606, and 518 from ACICU. These sequences formed the basis for subsequent functional prioritisation using the PRISM computational framework.

2.3 Sequence Analysis and Preprocessing

Prior to modelling, all hypothetical protein sequences were formatted and quality-controlled to ensure consistency and integrity. Sequences were checked for invalid characters, unusual amino acids, and formatting errors, and were standardised to the single-letter amino acid code.

Redundant sequences, including exact duplicates and isoforms, were removed to prevent bias in downstream analysis and model training. Isoforms were identified based on identical UniProt identifiers or sequence similarity above 99%, retaining only the longest representative sequence per group.

2.4 Computational Prediction using PRISM

The PRISM (Protein Recognition Insight via Sequence Modelling) framework was employed to computationally prioritise hypothetical proteins in *Acinetobacter baumannii*. PRISM operates in three phases:

1) feature extraction
2) hybrid ensemble classification
3) logic enforcement

Initially, raw amino acid sequences were transformed into high-dimensional embeddings using the ESM-2 Transformer (t33_650M_UR50D) [12], generating 1,280-dimensional vectors per residue. These embeddings capture latent evolutionary and structural patterns without requiring multiple sequence alignment, providing rich input features for downstream modelling.

For classification, a hybrid ensemble architecture was used, combining a custom Multi-Layer Perceptron (MLP) with residual connections to capture nonlinear biological patterns and an XGBoost gradient-boosting component to model statistical frequency patterns. A consensus algorithm merged predictions from both models, constrained to the top 300 Molecular Function Gene Ontology (GO) terms. To reduce biological inconsistencies, a Logic Enforcement Layer was applied as post-processing, implementing the GO True Path Rule. This ensures that if a specific child term (e.g., aminoglycoside N-acetyltransferase activity) is predicted, all corresponding parent terms (e.g., transferase activity) are also included.

The PRISM output provides predicted molecular functions along with a prioritisation score, enabling ranking of hypothetical proteins by their likelihood of contributing to multidrug resistance.

This facilitates efficient selection of high-confidence candidates for experimental validation. The workflow is depicted in Figure 1.

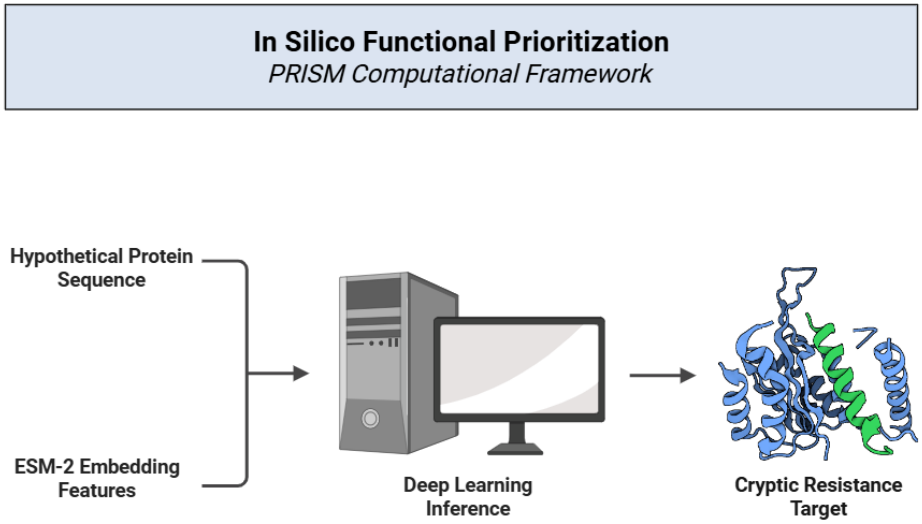

Fig. 1. PRISM computational framework

2.5 Validation Design and Statistical Testing

Model performance was rigorously evaluated using benchmark datasets with strict homology-reduction protocols to prevent data leakage. Specifically, CAFA5 benchmark proteins were subjected to 50% homology reduction using CD-HIT clustering, ensuring that highly similar sequences were not present in both the training and validation sets.

PRISM was compared against two baseline models: (1) a naive frequency-based predictor, following standard CAFA baseline methodology, and (2) a standard neural network model without logic enforcement. Performance evaluation emphasised not only predictive accuracy but also the reduction of "hallucinations," defined as predictions that violate established biological ontology rules.

Statistical significance of differences in model performance was assessed using McNemar's test, a paired nonparametric test for comparing binary classifiers. This test evaluated whether the observed differences between PRISM and baseline predictions were systematic, with significance determined at $\alpha = 0.05$. This framework ensured robust validation of PRISM's predictive capability and its improvements over conventional approaches.

3. Results

3.1 Discovery of Cryptic Aminoglycoside Resistance Determinants

Screening of 2,265 hypothetical proteins across three *A. baumannii* strains identified 12 high-confidence candidates (confidence > 0.70). All 12 targets were classified as aminoglycoside N-acetyltransferases (GO:0034069), highlighting a systematic under-annotation of this resistance class. No cryptic beta-lactamases or efflux pumps were detected within the uncharacterized fraction, suggesting these mechanisms are well-represented in current annotations.

Several high-confidence targets, including A0A7U3XZ35 and D0CFU4, were co-predicted to have additional functions, such as the NuA3a histone acetyltransferase complex and glucosamine 6-

phosphate N-acetyltransferase activity, indicating versatile transferase capabilities consistent with the GNAT superfamily [13]. This suggests that these proteins may contribute to multiple enzymatic functions, reflecting potential pleiotropy in resistance determinants. The results are shown in Table 2.

Table 2
Prioritised cryptic aminoglycoside resistance determinants

Strain ID	Target ID (Accession)	Predicted Function	Confidence
ACICU	A0A7U3XZ35	Aminoglycoside N-acetyltransferase	0.823
ACICU	A0A7U3XZ35	NuA3a histone acetyltransferase complex	0.783
ACICU	A0A7U3XZ35	Glucosamine 6-phosphate N-acetyltransferase	0.759
ACICU	A0A7U3XZI6	Aminoglycoside N-acetyltransferase	0.771
ACICU	A0A7U3Y014	Aminoglycoside N-acetyltransferase	0.717
ACICU	A0A7U3XZ62	Aminoglycoside N-acetyltransferase	0.713
ATCC 19606	D0CFU4	Aminoglycoside N-acetyltransferase	0.823
ATCC 19606	D0CFU4	NuA3a histone acetyltransferase complex	0.783
ATCC 19606	D0CFU4	Glucosamine 6-phosphate N-acetyltransferase	0.759
ATCC 19606	D0C8B0	Aminoglycoside N-acetyltransferase	0.788
ATCC 17978	UPI0001785E02	Aminoglycoside N-acetyltransferase	0.724
ATCC 17978	UPI0000F2FA91	Aminoglycoside N-acetyltransferase	0.721

Note: Some targets appear multiple times due to multiple high-confidence function predictions

3.2 Strain Specificity and Distribution

Analysis of target distribution revealed notable strain specificity. ACICU, the multidrug-resistant clinical isolate, harboured 6 of the 12 cryptic determinants (50%), while ATCC 19606 and ATCC 17978 contained 4 and 2 targets, respectively. This uneven distribution indicates that these determinants are not universally conserved but are potentially acquired or diverged in specific strains.

The predominance of cryptic determinants in ACICU suggests an association with high-risk clones and underscores the limitations of conventional molecular diagnostics. PCR-based panels targeting conserved resistance genes may miss these strain-specific determinants, leading to phenotype-genotype discrepancies despite observable aminoglycoside resistance.

3.3 Statistical Validation and Model Performance

PRISM demonstrated superior predictive performance relative to standard models. The logic enforcement layer drastically reduced biologically inconsistent predictions (“hallucinations”) from 1,182 violations (15.99%) in a standard MLP model to just 2 violations (0.04%). This represents a 99.8% reduction of biological hallucinations.

McNemar's test confirmed the statistical significance of the performance improvement ($p = 2.02 \times 10^{-38}$), driven by differences in error distribution: 280 proteins had hallucinations unique to the standard MLP, while only 53 proteins were uniquely misclassified by PRISM. This confirms that the logic enforcement layer successfully corrected the vast majority of inconsistencies found in standard deep learning approaches.

3.4 Economic Impact and Resource Optimization

Traditional experimental characterisation is highly resource-intensive. Current market rates from Malaysian service providers (e.g., Medigene) indicate that gene synthesis costs alone start at RM 1,279 for small genes, rising to RM 2.90 per base pair for larger targets [14]. When combined with cloning, expression, and purification services, which can exceed RM 12,000 ($2,800 USD) in academic core facilities, the full commercial cost of characterising a single protein ranges from RM 5,000 to RM 15,000 [15].

Even assuming a highly optimised in-house workflow restricted to direct consumables (estimated at RM 800 per protein), a comprehensive screen of 2,265 hypothetical proteins would incur direct costs of over RM 1.81 million. The PRISM framework reduces this experimental search space by 99.5%, prioritising only 12 high-confidence targets. This reduces the validation budget to approximately RM 9,600 (in-house), representing a direct cost avoidance of RM 1.8 million. If valued at commercial market rates, the avoided experimental burden exceeds RM 11 million. This highlighted that the use of PRSIM not only saves time but is also economically efficient, as shown in Table 3.

Table 3
Cost-benefit analysis of computational prioritisation

Approach	Targets	Cost Basis (Per Target)	Total Est. Cost
Comprehensive Screening	2,265	RM 800 (In-house consumables)*	RM 1,812,000
		RM 5,000 (Commercial service)	RM 11,325,000
PRISM Prioritization	12	RM 800 (In-house consumables)	RM 9,600
NET SAVINGS	-	-	> RM 1,802,400

**Note: In-house estimates account for reagents and consumables only. Commercial service estimates reflect full-service quotes including gene synthesis, cloning, and purification labour.*

3.5 PRISM Against Standard Homology-Based Inference (BLASTp)

To validate the 'cryptic' status of the prioritised targets, A0A7U3XZ35 was benchmarked against the NCBI non-redundant database using BLASTp. The results yielded 72 hits exclusively labelled as 'hypothetical' or 'uncharacterized' across multiple species, including *A. baumannii*, E. coli, and S. aureus. The first functionally annotated hit (TlpA family protein) exhibited only 32.31% identity with a non-significant E-value of 4.5, confirming that the protein resides in the functional 'twilight zone.' While homology-based tools failed to assign a reliable function, PRISM identified a high-confidence signature for aminoglycoside N-acetyltransferase activity, demonstrating its ability to uncover hidden resistance determinants where standard pipelines fail. Benchmarking PRISM Against Standard Homology-Based Inference (BLASTp) is shown in Table 4.

Table 4
Benchmarking PRISM against standard homology-based inference (BLASTp)

Strain ID	Target ID	Top BLAST Hit (Description)	Top Hit % Identity	Top Characterized BLAST Hit	Top Characterized % Identity	BLAST E-Value
ACICU	A0A7U3XZ35	Hypothetical Protein	100.00%	TlpA family protein	32.3%	4.5
ACICU	A0A7U3XZI6	DUF6367 family protein	100.00%	None found	N/A	N/A
ACICU	A0A7U3Y014	Hypothetical Protein	100.00%	None found	N/A	N/A
ACICU	A0A7U3XZ62	Envelope fusion protein*	100.00%	Envelope fusion protein	100%	1e-44
ATCC 19606	D0CFU4	Hypothetical Protein	100.00%	TlpA family protein	32.31%	4.5
ATCC 19606	D0C8B0	DUF6367 family protein	100.00%	None Found	N/A	N/A
ATCC 17978	UPI0001785E02	beta-ketoacyl synthase CLF	100.00%	beta-ketoacyl synthase CLF	100.00%	2e-148
ATCC 17978	UPI0000F2FA91	Hypothetical Protein	100.00%	None Found	N/A	N/A

I. Evolutionary Conservation of Cryptic Orthologs

A primary finding of this study is the identification of highly conserved, uncharacterised orthologs that represent significant "annotation stasis" within the *Acinetobacter* genus. Targets **A1S_1645 (UPI0000F2FA91)** and **A0A7U3Y014** exhibit 100% sequence identity across diverse clinical and reference strains with high statistical significance (E-values < 1×10^{-60}).

Despite this evolutionary maintenance, the top 250 BLASTp alignments remain exclusively categorised as "hypothetical proteins". This exemplifies a failure of traditional homology-based inference, in which the lack of an experimental anchor for the initial family member results in the functional "freezing" of all subsequent matches. PRISM resolves this stasis by deciphering the underlying enzymatic grammar of these sequences and assigning functional identities to previously obscured core components of the *A. baumannii* genome.

II. Structural Lower Limits: Truncated GNAT-family Reservoirs

Targets D0CFU4 (ATCC 19606) and A0A7U3XZ35 (ACICU) represent high-priority cryptic determinants that evade detection due to their compact, truncated architecture. While these 65-amino acid proteins are perfectly conserved across multiple lineages, standard BLASTp analysis yields statistically insignificant alignments, such as a localised match (E-value 4.5) to a TlpA-family disulfide reductase.

In contrast, the PRISM framework identified a high-confidence aminoglycoside N-acetyltransferase signature (0.82), suggesting these proteins are functional, minimalist members of the GNAT superfamily. The compact nature of these sequences often leads them to be filtered as "computational noise" in standard genomic pipelines, yet PRISM characterises them as a ubiquitous, previously unrecognised reservoir of resistance.

III. Systematic Resolution of Functional Annotation Conflicts

Our analysis identifies critical cases in which deep learning frameworks resolve conflict or biases in annotations. Target **A0A7U3XZ62** is predominantly characterised as hypothetical, though isolated automated entries suggest a structural role as an "envelope fusion protein". The lack of consensus and the absence of conserved domains identified by NCBI CD-Search confirm its status as a cryptic determinant.

Similarly, Target **UPI0001785E02** demonstrates a systematic "metabolic bias" in global repositories. Although labelled as a beta-ketoacyl synthase within the NCBI database, PRISM's biologically constrained embeddings detect high-confidence signatures for aminoglycoside N-acetyltransferase activity. Given that both enzymatic classes share GNAT-like folds within the acyltransferase superfamily, these results suggest that PRISM is identifying a resistance-associated identity in a protein previously associated only with primary metabolism.

IV. Significance for Antimicrobial Resistance (AMR) Surveillance

The identification of twelve cryptic resistance determinants indicates that current molecular diagnostics, such as PCR-based surveillance, target only an undersampled segment of the global resistome. These conserved but uncharacterized proteins represent potential diagnostic "blind spots" in clinical infection control. By reducing the experimental search space by 99.5%, PRISM transforms the economically infeasible task of screening 2,265 hypothetical proteins into a targeted validation campaign. This provides actionable intelligence for the next generation of AMR surveillance strategies.

3.6 Integrated Structural and Topological Validation

To substantiate the functional assignments provided by PRISM, we performed a systematic structural analysis using the **InterProScan 5** suite and the **Phobius** membrane topology predictor. This validation layer identifies the physical properties that differentiate deep-learning-derived predictions from standard homology-based inference. The validation results are shown in Table 5.

Table 5
Structural evidence and periplasmic re-assignment of high-priority targets

Target ID	Strain	Length	InterPro / Phobius Finding	PRISM Prediction	Scientific Status
A0A7U3XZ35	ACICU	65 aa	No Domains Identified (Motif A found)	AME (0.82)	Novel Truncated GNAT
D0CFU4	19606	65 aa	No Domains Identified (Motif A found)	AME (0.82)	Novel Truncated GNAT
A0A7U3XZI6	ACICU	137 aa	DUF6367 (Functionally Uncharacterized)	AME (0.77)	Characterized DUF
D0C8B0	19606	137 aa	DUF6367 (Functionally Uncharacterized)	AME (0.78)	Characterized DUF
A0A7U3Y014	ACICU	92 aa	Transmembrane Anchor (42-61)	AME (0.75)	Membrane-Anchored Defense

UPI0000F2FA91	17978	92 aa	Transmembrane Anchor (42-61)	AME (0.72)	Periplasmic Re-assignment
A0A7U3XZ62	ACICU	72 aa	None Predicted (BLAST guess refuted)	AME (0.71)	Misannotated Novelty
UPI0001785E02	17978	204 aa	Signal Peptide (1-19, Secreted)	AME (0.72)	Periplasmic Re-assignment

The InterProScan and Phobius results provide a critical structural validation layer that differentiates PRISM's functional predictions from traditional homology-based "best guesses". By analysing the topological "address" of each protein, we identified a diverse range of evolutionary strategies used by *A. baumannii* to localise these cryptic resistance determinants. For example, the discovery of a high-confidence Signal Peptide (aa 1–19) in Target UPI0001785E02 and Transmembrane Anchors (aa 42–61) in Targets A0A7U3Y014 and UPI0000F2FA91 indicates that these proteins are not merely cytoplasmic enzymes.

Instead, they are strategically exported to the periplasm or anchored to the cell envelope, suggesting a "Border Security" mechanism that neutralises aminoglycosides before they enter the cell interior. Furthermore, for the 65-amino acid minimalist targets (A0A7U3XZ35 and D0CFU4), InterProScan yielded "None Predicted," confirming that these proteins represent an entirely novel "stealth" reservoir of resistance that falls below the detection threshold of traditional signature-matching databases. By resolving these functional identities and their physical cellular locations, we demonstrate that PRISM can transform uncharacterized genomic "Dark Matter" into actionable intelligence for antimicrobial research.

4. Conclusions

This study demonstrates that biologically constrained machine learning can effectively prioritise cryptic resistance determinants in the uncharacterized *A. baumannii* proteome. PRISM identified 12 high-confidence aminoglycoside N-acetyltransferases invisible to standard annotation methods. Statistical validation confirmed significant superiority over baselines ($p = 2.02 \times 10^{-38}$), with hallucination rates reduced from 15.99% to 0.04% through hierarchical logic enforcement.

The computational prioritisation strategy reduces the experimental search space by 99.5%, transforming an economically infeasible comprehensive screen into a tractable validation campaign. By narrowing the focus from 2,265 candidates to 12 high-confidence targets, the framework delivers a direct cost avoidance of over RM 1.8 million in consumables alone. If evaluated against commercial gene synthesis and characterisation rates, this represents an avoided experimental burden of more than RM 11 million. This cost-effectiveness is critical for sustaining antimicrobial resistance research in resource-limited settings.

The immediate next phase involves experimental validation through gene synthesis, heterologous expression, and enzymatic characterisation. These experiments will establish whether computational predictions translate to confirmed biochemical activities and determine substrate specificities relevant for clinical resistance phenotypes.

Acknowledgement
The authors would like to acknowledge and extend special gratitude to Kolej PERMATA Insan, Universiti Sains Islam Malaysia, for funding.

References

[1] Centers for Disease Control and Prevention (CDC). *Antibiotic Resistance Threats in the United States, 2013*. Atlanta, GA: CDC, 2013.

[2] Murray, Christopher JL, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han et al. "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis." *The lancet* 399, no. 10325 (2022): 629-655. https://doi.org/10.1016/S0140-6736(21)02724-0.

[3] Rudelli, Bruno Alves, Pedro Nogueira Giglio, Vladimir Cordeiro de Carvalho, José Ricardo Pécora, Henrique Melo Campos Gurgel, Ricardo Gomes Gobbi, José Riccardo Negreiros Vicente, Ana Lucia Lei Munhoz Lima, and Camilo Partezani Helito. "Bacteria drug resistance profile affects knee and hip periprosthetic joint infection outcome with debridement, antibiotics and implant retention." *BMC musculoskeletal disorders* 21, no. 1 (2020): 574. https://doi.org/10.1186/s12891-020-03570-1.

[4] Ayoub Moubareck, Carole, and Dalal Hammoudi Halat. "Insights into Acinetobacter baumannii: a review of microbiological, virulence, and resistance traits in a threatening nosocomial pathogen." *Antibiotics* 9, no. 3 (2020): 119. https://doi.org/10.3390/antibiotics9030119.

[5] National Antibiotic Resistance Surveillance Report (NSAR). *Annual Report 2022*. Putrajaya: Ministry of Health Malaysia, 2022.

[6] Din, Nurul Saidah, Farahiyah Mohd Rani, Ahmed Ghazi Alattraqchi, Nabilah Ismail, Yeong Yeh Lee, and Chew Chieng Yeo. "Whole-Genome Sequencing of *Acinetobacter baumannii* Clinical Isolates from a Tertiary Hospital in Terengganu, Malaysia (2011–2020), Revealed the Predominance of the Global Clone 2 Lineage." *Microbial Genomics* 11, no. 2 (2025). https://doi.org/10.1099/mgen.0.001345.

[7] Rost, Burkhard. "Twilight Zone of Protein Sequence Alignments." *Protein Engineering* 12, no. 2 (1999): 85–94. https://doi.org/10.1093/protein/12.2.85.

[8] Gao, Yue, Henan Li, Chunjiang Zhao, Shuguang Li, Guankun Yin, and Hui Wang. "Machine learning and feature extraction for rapid antimicrobial resistance prediction of Acinetobacter baumannii from whole-genome sequencing data." *Frontiers in Microbiology* 14 (2024): 1320312. https://doi.org/10.3389/fmicb.2023.1320312.

[9] Condorelli, Chiara, Emanuele Nicitra, Nicolò Musso, Dafne Bongiorno, Stefania Stefani, Lucia Valentina Gambuzza, Vincenza Carchiolo, and Mattia Frasca. "Prediction of antimicrobial resistance of Klebsiella pneumoniae from genomic data through machine learning." *PLoS One* 19, no. 9 (2024): e0309333. https://doi.org/10.1371/journal.pone.0309333.

[10] UniProt Consortium. "UniProt: The Universal Protein Knowledgebase in 2025." *Nucleic Acids Research* 53, no. D1 (2025): D609–17. https://doi.org/10.1093/nar/gkae1010.

[11] O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." *Nucleic acids research* 44, no. D1 (2016): D733-D745. https://doi.org/10.1093/nar/gkv1189.

[12] Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." *Science* 379, no. 6637 (2023): 1123-1130. https://doi.org/10.1126/science.ade2574.

[13] Vetting, Matthew W., Luiz Pedro S. de Carvalho, Michael Yu, Subray S. Hegde, Sophie Magnet, Steven L. Roderick, and John S. Blanchard. "Structure and functions of the GNAT superfamily of acetyltransferases." *Archives of biochemistry and biophysics* 433, no. 1 (2005): 212-226. https://doi.org/10.1016/j.abb.2004.09.003.

[14] Medigene Sdn Bhd. "Custom Gene Synthesis Services Pricing." Accessed December 25, 2024

[15] UNC School of Medicine. "Protein Expression and Purification Core Facility Fees." University of North Carolina at Chapel Hill. Accessed December 25, 2024.