Conserved protein sequence-structure signatures identify emerging antibiotic resistance genes from the human microbiome

Bartrop L.[1,2], Beauchemin-Lauzon E.[2], Grenier F.[2], Rodrigue S.[2], Haraoui L.P.[1] 1 - Department of Microbiology and Infectious Diseases, Université de Sherbrooke, QC, Canada; 2 - Department of Biology, Université de Sherbrooke, QC, Canada

Abstract

Bacteria exhibiting antimicrobial resistance (AMR) is a problem that has grown to become a significant global public health challenge. Antibiotic resistance genes (ARGs), determinants of AMR, mostly emerge from non-clinical settings. Identifying novel ARGs from the human microbiome which confer resistance to clinical antibiotics is a crucial component of addressing AMR. We sought to leverage the increased accuracy of computational protein structure prediction by training a one-class support vector machine on the pairwise primary protein structure and tertiary protein structure distributions of conserved structural regions of various ARG classes. We applied the computational method to seven human microbiome project reference strains and functionally confirmed eight out of nine tested novel ARG predictions, belonging to ARG classes: DFR, class B and C β-lactamases and penicillin binding proteins. We also applied the method directly to protein structures within the AlphaFold database and functionally confirmed a novel metallo-β-lactamase with low homology to existing ARGs. In total, 70% of tested genes confer resistance to clinical antibiotics at CLSI resistant breakpoints, or are emerging from the 'pre-resistome' and may become clinically relevant in the future. This study represents a precise and computationally efficient method of identifying previously uncharacterized ARGs from DNA databases.

Introduction

The enthusiasm following the discovery of antibiotics was rapidly overshadowed by the rise of bacteria exhibiting antimicrobial resistance (AMR), defined as the ability of microbes including bacteria to survive treatment with antibiotics that once effectively killed them. AMR has since become one of the most pressing global public health challenges of our time (O'Neill, 2016). Although key drivers of AMR, antibiotic resistance genes (ARGs), predate the use of antibiotics (D'Costa et al., 2011). For the most part they emerge in environmental bacteria at increasing rates in response to anthropogenic antibiotic exposure and pose the greatest risk to human health once they are stably integrated into the human microbiome. This can occur through one of three main routes: (i) if they evolved in generalist or opportunist species which can also inhabit human microbiome niches; (ii) via horizontal gene transfer from environmental bacteria to human-associated bacteria by mobile genetic elements (MGEs); or (iii) by *de novo* evolution within the human microbiome (Crits-Christoph et al., 2022; Larsson & Flach, 2022). As such, identifying novel ARGs from environmental and human microbiomes that confer clinical levels of resistance to antibiotics is a crucial component of addressing AMR, and for accurate surveillance, diagnostics, and antibiotic stewardship.

By training a one-class support vector machine (SVM) on the pairwise primary protein structure and tertiary protein structure distributions of conserved structural regions of an ARG class, we developed a novel, precise, and computationally efficient method of identifying novel ARGs. We applied the approach to seven human microbiome project (HMP) reference strains and functionally confirmed eight out of nine predicted novel ARGs when expressed in *Escherichia coli*: four novel DFR genes conferring resistance to trimethoprim with less than 55% amino acid (aa) identity to existing DFRs; two class B β-lactamases from *Burkholderiales* sp., 1_1_47 and *Klebsiella* sp.; a cephalosporinase from *Acinetobacter junii* SH205; and a β-lactam-resistant penicillin binding protein (PBP) from *Acinetobacter radioresistens* SH164. Of those that were functional, six present clinical relevance reaching MICs at Clinical and Laboratory Standards Institute (CLSI) resistant breakpoints and the remainder are emerging from the 'pre-resistome', with risk of becoming clinically relevant in the future. While our primary dataset was HMP reference strains, the computational method also predicted a novel metallo-β-lactamase (MBL) directly from the AlphaFold database (AFDB), which has low homology to existing ARGs and confers a clinically relevant MIC to ampicillin when expressed in *E.coli.*

This research demonstrates the importance of combining recent advances in computational protein structure prediction and search with machine learning (ML), to identify emerging ARGs from DNA databases. Subsequent experimental verification confirms those with clinical relevance, particularly if they transfer to human pathogens.

Rationale

Classic approaches to identify novel ARGs rely on aa sequence alignment, generally adopting high (>60%) aa sequence identity (seqID) to databases of ARG sequences (Alcock et al., 2023; A.-N. Zhang et al., 2021). ML–based classifiers have also proved useful trained on the vast amounts of biological data generated since the rise of modern sequence technologies, reporting precisions which exceed 85% (Abbas et al., 2025, 2025; Arango-Argoty et al., 2018; Hamid & Friedberg, 2020; Naik et al., 2025) . However, most ML-based classifiers are trained on ARG sequences collated from various databases with inherent biases, for e.g., ARGs from clinical settings, with little sequence variation, or without experimental support (Hamid & Friedberg, 2020). Whilst performance is high on internal validation, their generalizability to more novel sequences from other DNA databases remains less clear. For example, ARG-Net only identified 3.99% of ARGs from ESKAPE sequences with high probability below 50% identity to known ARGs (Pei et al., 2024) and whilst high precision was achieved by training a low homology dataset, seqIDs of identified ARGs did not extend below 40% (Yagimoto et al., 2024). The amount of microbial proteins outside clinical settings with homology to ARGs increases with decreasing seqID and ARGs can share lower seqID, for example, at the time of discovery $bla_{\text{NDM-1}}$ and *sul4* shared the highest seqID of ~33% to bla_{VIM} and *sul2,* respectively (Razavi et al., 2017; Yong et al., 2009).

Efforts to address this limitation expand training beyond sequences (Ghosh et al., 2025; Rannon et al., 2025), such as combining models from recent computational protein structure tools like AlphaFold2 with sequence in graph and convolution neural networks (Qabel et al., 2023). As with all ML-based tools, experimental confirmation is essential to establish biological and clinical relevance. Pairwise comparative modelling (PCM) integrates protein structure and predicted novel ARGs from the intestinal resistome with an average seqID of 29.8% and was supported by experimental data (Ruppé et al., 2019). Non ML-based models such as profile hidden Markov models (HMMs) have also been successfully employed for ARG detection (Feldgarden et al., 2019; Gibson et al., 2015; Yin et al., 2018), including with experimental support for proteins with seqIDs below 70% (Berglund et al., 2019).

With rapid developments in computational protein structure prediction tools (Jumper et al., 2021; Wu et al., 2022), progress has been made in terms of accuracy (now in line with experimental solutions), and the amount of structures available (>200 million predicted by AlphaFold2) (Varadi et al., 2024). Antibiotic resistant protein (ARP) structures encoded by ARGs are often more conserved compared to their seqIDs, with protein structural similarity quantified using the template modelling (TM)-score, which ranges from 0 (no structural similarity) - 1 (a perfect structural match) (Y. Zhang & Skolnick, 2004). A pairwise analysis indicates ARP structures across well-known ARG classes have an average median seqID of 30.4% and TM-score of 0.77, so ARP structures can theoretically have low seqIDs whilst retaining higher structural similarity to perform the same function (Figure 1). Existing computational tools potentially miss a large number of potential ARGs which have less than 50% seqID to known ARGs. For example, although sharing low seqID, ARP structures encoded by the aforementioned *sul4* and

bla$_{NDM-1}$ share TM-scores of 0.9 to their respective closest homologs by seqID, encoded by *sul2* and *bla*$_{VIM}$, indicating their structures are highly conserved.

Nonetheless, ARP structures typically evolve from existing proteins (the proto-resistome, (Morar & Wright, 2010)) performing primary functions or are involved in the target pathway of an antibiotic (Baquero et al., 2021). Proteins in the proto-resistome often exhibit high TM-scores to ARP structures, however lack function. Small changes in critical aa residues can make the difference between a true ARG (belonging to the resistome) from genes that do not confer clinical resistance but could evolve there (the proto-resistome), so prioritizing proteins with high TM-scores to find novel ARP structures would likely result in the generation of a high false positive rate (FPR) (Kaderabkova et al., 2022; Venkatesan et al., 2023). Conversely, proteins with reduced TM-scores to ARP structures can still retain the same function, as depicted by a larger range of TM-scores shared between ARP structures, primarily for AAC's, class B β-lactamases, APH's and ANT's (Figure 1b). This could be explained by ARP structures performing functions by similar localized structural motifs encoded by important aa's (Koehler Leman et al., 2023). Proteins which contain these motifs without structural conservation in other regions would have reduced TM-scores to ARP structures so only prioritizing proteins with high TM-scores would also result in a high false negative rate (FNR).

Since experimental verification remains the primary bottleneck in the discovery pipeline, it is imperative that algorithms identify ARGs with high precision that either currently confer, or are likely to confer in the future, clinically relevant MICs. Therefore, we reasoned that integrating the seqID with the TM-score of structurally conserved regions of ARP structures would be a simple and highly predictive method of ARG function and aimed to validate the approach through experimental confirmation of predictions within standardized expression systems and a range of antimicrobial susceptibility testing procedures against CLSI breakpoints.

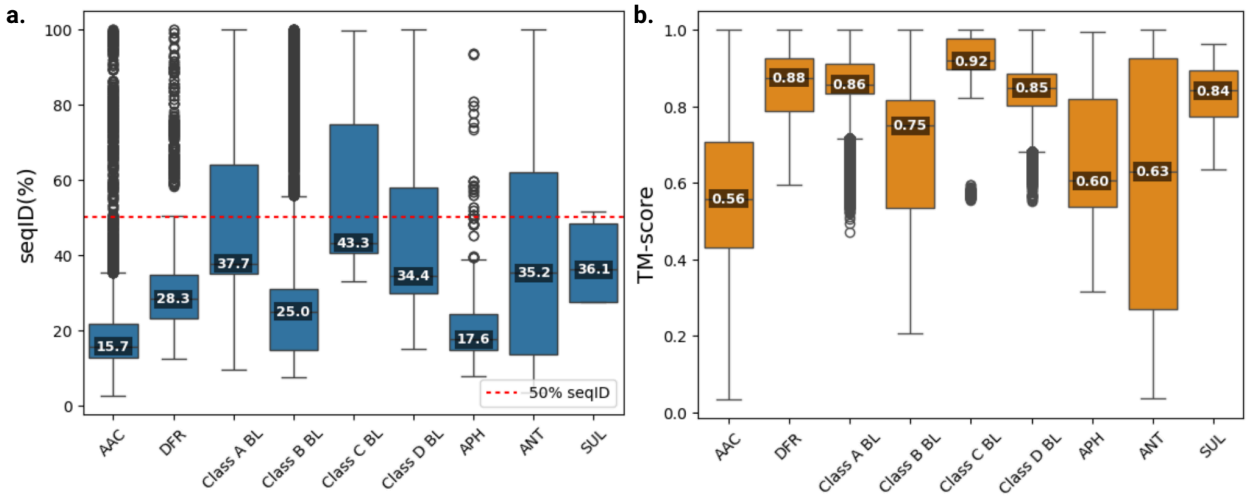


Figure 1. (a) Boxplots of pairwise sequence identities (seqID) and (b) TM-scores for protein structures from well-known ARG classes. The median value is shown on each boxplot. ARG classes include: aminoglycoside acetyl-transferases (AAC, number of pairwise comparisons (*n)* = 6,335), dihydrofolate reductases (DFR, *n* = 2,070), Class A β-lactamases (Class A BL, *n* = 661,780), Class B β-lactamases (Class B BL, *n* = 77,840), Class C β-lactamases (Class C BL, *n* = 77,840), Class D β-lactamases (Class D BL, *n* = 178,506), aminoglycoside phospho-transferases (APH, *n* = 1,122), aminoglycoside nucleotidyl-transferases (ANT, *n* = 2,950), and mobile dihydropteroate synthases (SUL, *n* = 12). Pairwise seqIDs and TM-scores were calculated using Foldseek under default settings, with TM-scores normalized by the subject length.

Results

The computational method calculates pairwise primary and tertiary protein structure identities of conserved protein structural regions (*high-local Distance Difference Test (lDDT)*) of ARPs to generate an x, y distribution of seqIDs and TM-scores. A one-class support vector machine (SVM)[8], commonly used for outlier detection, is trained to learn the decision boundary surrounding the distribution. A query ARP (qARP) structure is aligned to each *high-lDDT* ARP structure, and its derived seqID and TM-score are calculated. If any points lie within the decision boundary (DB) of the x, y distribution, it is predicted as functional; if any points lie outside the DB, it is flagged as an outlier, predicting it as non-functional (Figure 2).

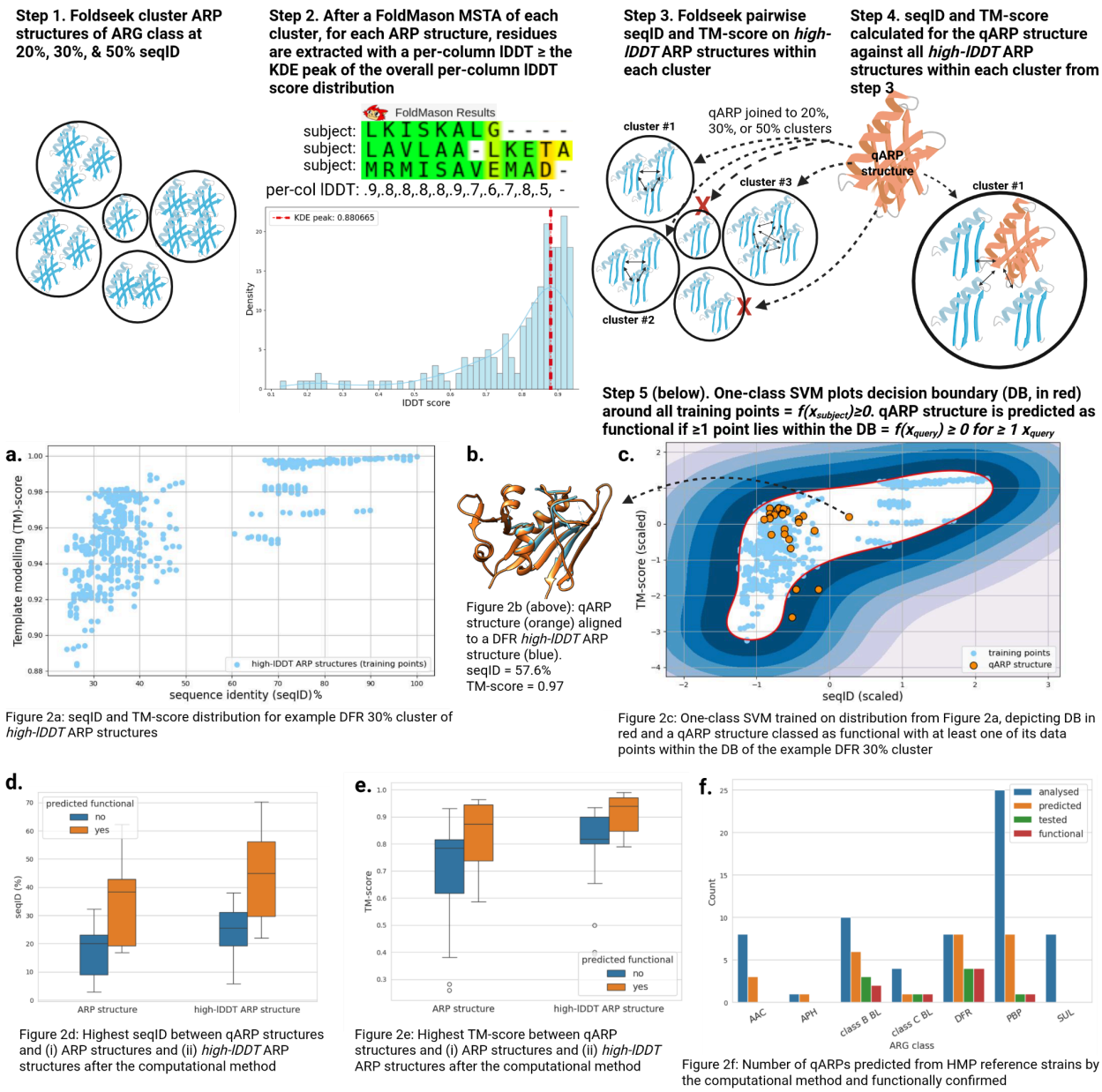


Figure 2b (above): qARP structure (orange) aligned to a DFR *high-lDDT* ARP structure (blue). seqID = 57.6% TM-score = 0.97

Figure 2a: seqID and TM-score distribution for example DFR 30% cluster of *high-lDDT* ARP structures

Figure 2c: One-class SVM trained on distribution from Figure 2a, depicting DB in red and a qARP structure classed as functional with at least one of its data points within the DB of the example DFR 30% cluster

Figure 2d: Highest seqID between qARP structures and (i) ARP structures and (ii) *high-lDDT* ARP structures after the computational method

Figure 2e: Highest TM-score between qARP structures and (i) ARP structures and (ii) *high-lDDT* ARP structures after the computational method

Figure 2f: Number of qARPs predicted from HMP reference strains by the computational method and functionally confirmed

Figure 2. Overview of the computational method. *Step 1*. ARP structures for a given ARG class are collected and clustered at 20%, 30%, and 50% seqID using Foldseek. *Step 2.* For each cluster, a multiple structure alignment (MSTA) is performed using FoldMason. Aligned residues with a per-column local Difference Distance Test (lDDT) score ≥ the Kernel Density Estimation (KDE) peak of the per-column lDDT distribution are extracted to create *high-lDDT* ARP structure .pdb files for each ARP structure. *Step 3.* The pairwise (all vs all) seqIDs and TM-scores are calculated for *high-lDDT* ARP structures using Foldseek TM-align. *Step 4.* qARP structures are assigned to ARP clusters from Step 1 based on their highest seqID to an ARP structure (determined by Foldseek, see Methods). SeqIDs and TM-scores of the qARP structure to *high-lDDT* ARP structures in clusters are calculated using Foldseek TM-align. Figure 2a. Example distribution of pairwise seqIDs (x) against TM-scores (y) for a *high-lDDT* ARP structure

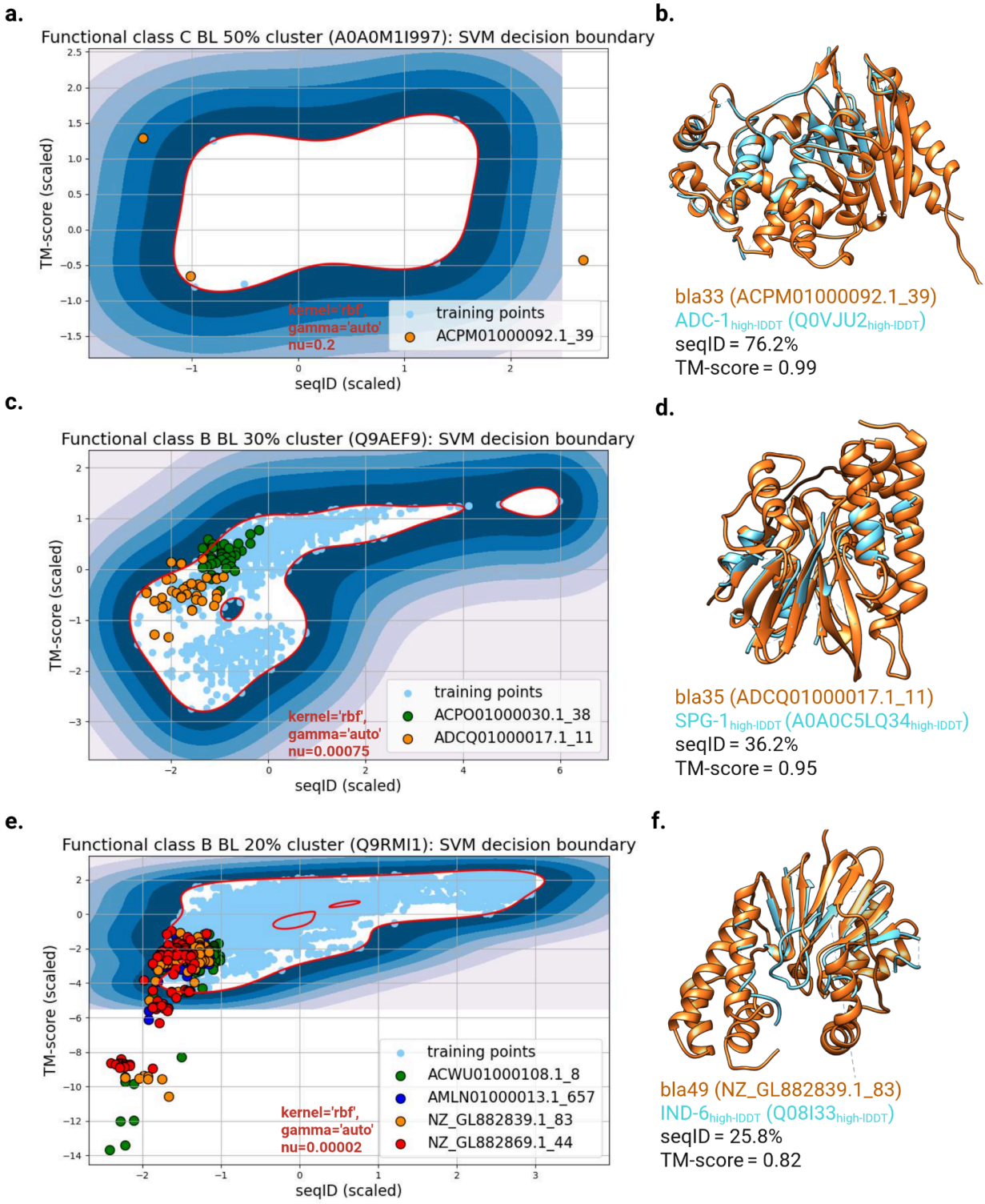
cluster (training points in blue). *Step 5.* A one-class support vector machine (SVM) is trained on the x, y distribution from Figure 2a (Figure 2c, now scaled). The decision boundary (DB, in red) surrounds all training points so that $f(x^{(i)}_{ARP\ structure}) \geq 0$. The seqIDs and TM-scores of the aligned region of a single qARP structure against each *high-lDDT* ARP structure from *Step 4* are plotted (data points in orange). If any data points lie on or within the DB, the qARP is classed as functional (i.e., $f(x^{(i)}_{qARP\ structure}) \geq 0\ for \geq x^{(i)}_{qARP\ structure}$) and otherwise the qARP is classed as non-functional. Figure created with BioRender.com.

The computational method identified emerging ARGs from Human Microbiome Project reference strains

Query antibiotic resistant genes (qARGs) were first predicted from coding sequences (CDS) originating from seven gram-negative Human Microbiome Project (HMP) reference strains sampled from the human gut (Supplementary Table 1) and downloaded from the project catalog (https://hmpdacc.org/hmp/catalog/, using an initial Diamond screening against a created database of ARG sequences from the Comprehensive Antibiotic Resistance Database (CARD, v4.0.1, see Methods, Supplementary Table 3) (Alcock et al., 2023). qARGs that were not considered known ARGs (>90% seqID), had their protein structures downloaded from the AFDB if available or were computed directly by Alphafold2, representing query antibiotic resistant protein (qARP) structures. qARP structures were then applied to seqID, TM-score distributions of clustered *high-lDDT* ARP structures (one-class SVM models) created for well-known ARG families (mobile sulfonamide-resistant dihydropteroate synthetases (SUL), class A, B, C, and D β-lactamases, class 1, 2, and 3 β-lactam-resistant penicillin binding proteins (PBP), trimethoprim (tmp)-resistant dihydrofolate reductases (DFR), class 2', 3', and 6' aminoglycoside acetyl-transferases (AAC), aminoglycoside phospho-transferases (APH), and aminoglycoside nucleotidyl-transferases (ANT)). *High-lDDT* ARP structures and Foldseek databases of *high-lDDT* ARP structures for created one-class SVM models will be available shortly.

In total, 27 qARP structures were classed as functional by the computational method, predicted to provide resistance to aminoglycosides, β-lactams, and trimethoprim (Figure 2f, Supplementary Table 2). The highest average seqID of qARP structures to *high-lDDT* ARP structures was 44.4%, and the average TM-score was 0.92, which increased from a seqID of 17.8% and a TM-score of 0.71 to full length ARP structures (Figure 2d and 2e, Supplementary Table 2). Nine qARPs belonging to four ARG classes (DFR, class B and C β-lactamases, and β-lactam-resistant PBP), as well as positive controls (known ARPs), were synthesized and cloned into the Antibiotic Resistance Platform and expressed in *Escherichia coli* BW25113 (ΔbamBΔtolC) for functional verification. Activity was detected in eight out of nine qARPs against their respective antibiotics (Supplementary Table 2) and further testing was performed to determine

minimum inhibitory concentrations (MICs) using broth microdilution (BM) and an adapted agar diffusion (AAD) method (Table 1).



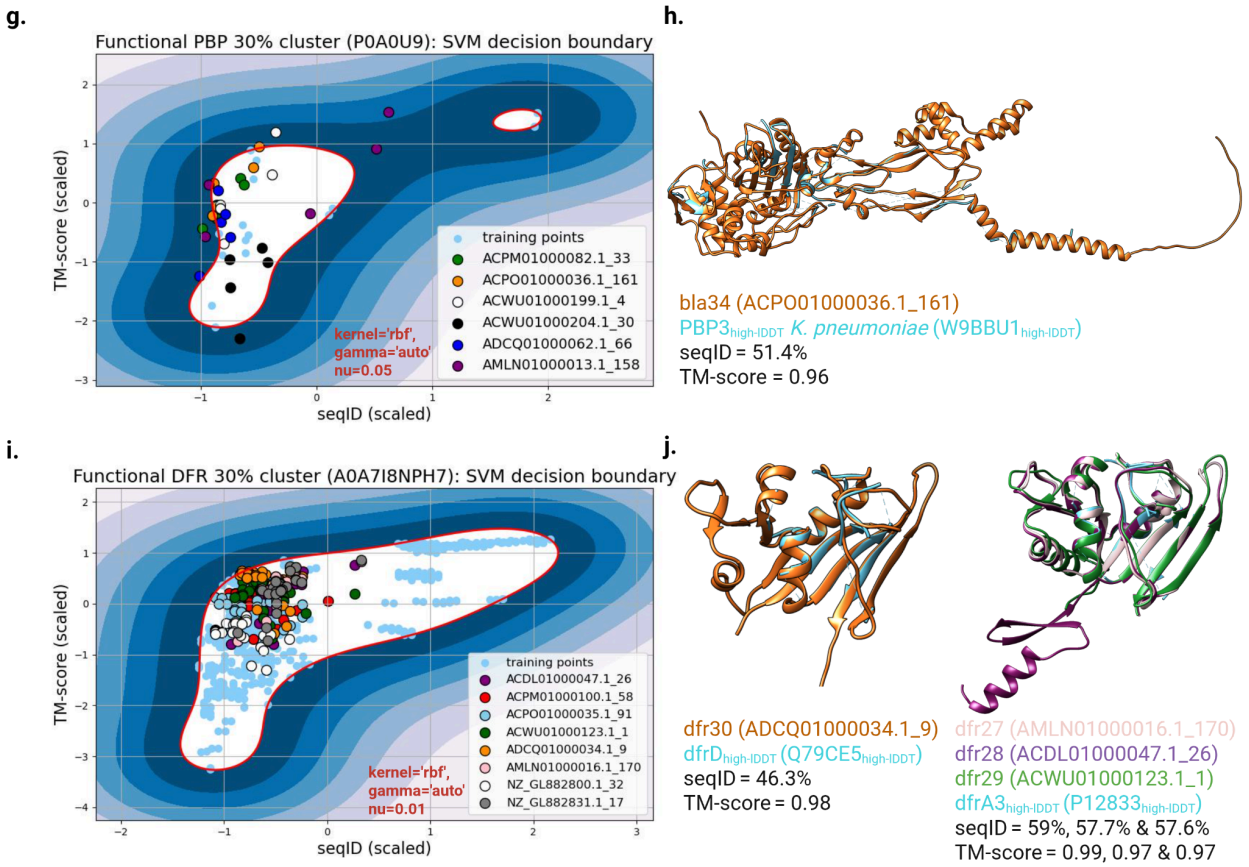


Figure 3. (a, c, e, h, i) One-class SVM models for qARP structures classed as functional by the computational method, depicted with at least one data point within the DB (in red) surrounding the training distribution. SVM hyperparameters used for DB calculation are also displayed in red on each plot. **(b, d, f, h, j)** qARP structures with activity detected following functional verification are overlaid with the *high-lDDT* ARP structure with the highest seqID and within the DB. Protein sequences of qARPs predicted functional by the computational method are available in the Supplementary data. One-class SVM plots were made using scikit-learn (Pedregosa et al., 2011) and AlphaFold2 protein structures were visualized with UCSF Chimera (Meng et al., 2006).

Six out of nine qARPs conferred MICs at CLSI resistant breakpoints

β-lactamases:

The pGDP2-bla33 construct, harbouring a potential cephalosporinase from *Acinetobacter junii* SH205 with 64.2% seqID to ADC-8 (Supplementary Table 3), conferred MICs above CLSI resistant breakpoints by both MIC determination methods for ampicillin (BM: MIC > 256 µg/mL; AAD: MIC > 64 µg/mL) and cefotaxime (BM: MIC 32 µg/mL; AAD: MIC 16 µg/mL). The qARP structure of bla33 was predicted as functional by the computational method by the data point within the DB of the functional SVM cluster with a seqID of 76.2% and TM-score of 0.99, which bla33 shares to ADC-8$_{\text{high-lDDT}}$ (Figure 3a and 3b). The qARP structure of bla33 also shares a seqID of

84.9% and TM-score of 0.99 to AVDC-1$_{high\text{-}lDDT}$, however this point is outside the DB (Figure 3a). AVDC-1 is an ampC cephalosporinase from *Acinetobacter venetianus,* however is not present in CARD as no MIC data has been reported. Based on low sequence homology to existing cephalosporinases and elevated MICs against β-lactams, we tentatively propose a gene name for the qARP bla33, of AJDC-1. Catalytic motifs known to be responsible for β-lactam hydrolysis are conserved in AJDC-1 (^{64}SXSK, ^{150}YXN, and ^{315}KTG) (Philippon et al., 2022) and were captured as conserved regions of *high-lDDT* ARP structures of known cephalosporinases by the computational method (Figure 3b). *Acinetobacter* sp. are common sources of ampC cephalosporinases, in particular ADC ('*Acinetobacter*-derived cephalosporinases) variants and often originate in *A. baumannii* (Hujer et al., 2005). AJDC-1 represents the first cephalosporinase functionally confirmed from HMP reference strain *A. junii* SH205, which may be the naturally occurring cephalosporinase from this species. Strains of *A. junii* are not commonly associated with disease but it has been reported as an emerging One-Health pathogen (Aguilar-Vera et al., 2024). Phylogenetic inference of AJDC-1 with ADCs from *Acinetobacter* sp. and other closest related cephalosporinases from *Acinetobacter* sp. (including those without MIC data), confirms AJDC-1 has phylogenetic proximity to ADC-8, functionally confirmed in 2007 from *Acinetobacter baylyi*, however is closest to the ampC cephalosporinase from *A. beijerinckii* (not present in CARD). Both AJDC-1 and ADC-8 are highly divergent from other functionally confirmed cephalosporinases (Zhao & Hu, 2012) (Supplementary Figure 1).

Construct pGDP2-bla49 harbouring a potential class B MBL, when determined by AAD, presented elevated MICs against ampicillin (MIC > 64 µg/mL), cefotaxime (MIC 8 µg/mL), and meropenem (MIC > 8 µg/mL) (Table 1), however not when measured by BM. This discrepancy is potentially explained by methodological differences between these experimental assays (discussed later). The qARP structure of bla49 (Alphafold ID: A0A1Y0Q343) is annotated as a 'MBL fold metallo-hydrolase' in the AFDB and it shares a highest seqID of 46.6% to the ARP structure encoded by VMH (AlphaFold ID: A0A347SHL3), identified in 2021 from the pathogen *Vibrio vulnificus*. VMH is not present in CARD likely because MIC data was not reported (Lu et al., 2021) and the ARP structure of VMH did not cluster with any known MBLs after the computational method. After alignment of all ARP structures encoded by class B MBLs, VMH does not contain the typical H-X-H-X-D zinc binding motif in MBLs, in its place there is a motif of G-G-D-P-D, most similar to the aligned S-G-D-P-D motif of bla49 in this position. As with VIM-4 from *P. aeruginosa* (Lassaux et al., 2011) and BcII from *Bacillus cereus* (Brem et al., 2015), conserved residues VI/VPGH in position 212-216 in VMH (Lu et al., 2021) may involve Zinc2 binding and these residues are also conserved in bla49 in position 217-221. Furthermore, the qARP bla36 was not functional after synthesis and it also shares a highest seqID of 45% to VMH, however it does not contain this motif. Bla49 shares a best hit to IMP-43 in CARD with 21.2% seqID (Supplementary Table 3), and after the computational method the qARP structure of bla49 shares a highest seqID of 25.8% and TM-score of 0.82 with the *high-lDDT* ARP structure, IND-6$_{high\text{-}lDDT}$ (Figure 3f). The IND-6 gene was reported from *Chryseobacterium indologenes* in 2009 from Burkina Faso (Zeba et al., 2009). As such, bla49 likely represents a novel sub-class B1 β-lactamase and we tentatively proposed its name of KMH-1 (*Klebsiella* sp.

metallo-hydrolase). A structural phylogeny of KMH-1 with closest known MBLs shows its shared evolutionary history with VMH-1 (Supplementary Figure 2).

Trimethoprim and trimethoprim-sulfamethoxazole resistance

The pGDP3-dfr27, pGDP3-dfr29, and pGDP3-dfr30 constructs all containing potential trimethoprim (tmp)-resistant DFR's from *K. pneumoniae* WGLW3, *P. aeruginosa* 2_1_26 and *Burkholderiales* sp. 1_1_47, respectively, conferred MICs above CLSI resistant breakpoints for tmp by BM (MIC 25 µg/mL for dfr27 and MIC > 256 µg/mL for dfr29 and dfr30) and by AAD (MIC > 16 µg/mL), similar to the the positive control, *dfrK*, and up to 127-fold higher than the negative control (MIC ≤ 2 µg/mL) (Table 1). The strain harbouring positive control *dfrK* also conferred resistance to sulfamethoxazole-trimethoprim (SXT) by AAD. As a structural analogue of p-aminobenzoic acid (PABA), sulfamethoxazole competitively inhibits dihydropteroate synthetase (dhps), which is an earlier step of where trimethoprim acts in the folate synthesis pathway (Ovung & Bhattacharyya, 2021). Clinical resistance to SXT is commonly reported by both a tmp-resistant DFR and one of the mobile SUL genes (sul1-sul4) and less likely by the presence of a single DFR or SUL resistance gene (Shin et al., 2015, Poey et al., 2024). Interestingly, a clinical MIC against SXT was not observed for the strain harbouring a known mobile SUL gene, *sul3*, as opposed to *dfrK* (Table 1).

The computational method determined a seqID of 57%-59% and TM-scores between 0.97-0.99 of the qARP structures of dfr27 and dfr29, to the *high-lDDT* ARP structure of dfrA3 ($dfrA3_{high\text{-}lDDT}$, Supplementary Table 2 and Figure 3j). Similarly, the qARP structure of dfr30 has a seqID of 47.6% and TM-score of 0.98 to the *high-lDDT* ARP structure $dfrD_{high\text{-}lDDT}$ (Supplementary Table 2, Figure 3j). Resistance to tmp has been reported by over-expression of dexcontextualized chromosomal folA genes on plasmids in *E.coli*, which 'swamps' tmp, allowing the chromosomal folA gene to function as normal (Palmer & Kishony, 2014; Soo et al., 2011). We posit that dfr29 and dfr30 represent true novel DFR variants (and not dfr27) based on multiple lines of evidence: (i) comparable tmp MICs to the positive control *dfrK* and lower tmp MICs observed for dfr27 when measured by BM; (ii) similar colony growth visually observed by the AAD method to the positive control at a concentration of 16 µg/mL of tmp, as opposed to smaller growth observed for dfr27 (Supplementary Figure 6a); (iii) dfr29 and dfr30 phylogenetically cluster with known DFR clades (Supplementary Figure 3); and (iv) after structural alignment with other ARP structures in the functional SVM cluster, dfr29 and dfr30 contain P21L and I94L substitutions that mediate tmp resistance in DFR proteins (Manna et al., 2021; Tamer et al., 2019). No known substitutions are present in dfr27. Dfr29 has a highest seqID of 54.7% to dfrA3 and dfr30 to dfrA26 of 40.6% (Supplementary Table 3). As such these likely represent novel dfrA genes and we tentatively propose their names dfrA50 and dfrA51 (Table 1).

Phylogenetic analysis with other known dfr's indicates dfrA50 and dfrA51 are distinct from existing dfr's (Supplementary Figure 3). DfrA50 from HMP reference strain *P. aeruginosa* 2_1_26 is on its own branch with the integron-encoded dhfr from *E.coli*, dfrA3 (Brolund et al., 2010). Strains of *P. aeruginosa* are often intrinsically resistant to tmp by low outer membrane permeability, efflux pumps, and the integron-encoded dfrB5

gene (Faltyn et al., 2019; Pang et al., 2019) and dfrA50 represents a new dfrA gene from this host. DfrA51 identified from HMP reference strain *Burkholderiales* sp. 1_1_47 is outgrouped from a clade containing other dfrA genes (dfrA8, dfrA18, dfrA42, and dfrA9) which are all also identified from gram-negative strains. *Burkholderiales* sp. may represent an environmental source of dfrB genes and to the best of our knowledge, dfrA51 from HMP reference strain *Burkholderiales* sp. 1_1_47 represents the first novel dfrA gene reported from this genus (Kneis et al., 2024).

Finally, it is worth noting dfr27 has a high seqID of 92% to dfr28 and no elevated MICs were observed in dfr28 measured by AAD. Residues present in dfr27 from dfr28 over the aligned length are V41I, H45L, I61V, L62I, Q65K,T68S, T73Q, K76S, D79E, P89E, E129D, D132E, and H149N, and dfr28 possesses an additional 37 residues constituting an α-helix at the N-terminus (Figure 3j). Further research is needed to deduce if these structural properties played a role in the over-expression to raise tmp MICs observed in dfr27 and not dfr28.

A membrane-associated PBP3 homolog was computationally predicted and functionally confirmed to provide resistance to β-lactams

The strain harbouring pGDP2-bla34, which encodes a predicted penicillin binding protein (PBP3) from *A. radioresistens* SH164, conferred MICs above CLSI resistant breakpoints against ampicillin by BM (MIC 256 µg/mL) and AAD, (MIC > 64 µg/mL) and cefotaxime by BM (MIC 16 µg/mL) and AAD (MIC 8 µg/mL, Table 1). Based on results of the Diamond search against the custom CARD database, bla34 (ACPO01000036.1_161) shares a seqID of 40.7% similar to the β-lactam resistant PBP3 mutant from *Klebsiella pneumoniae* (Supplementary Table 3). Computational prediction of membrane-associated proteins have lower relative confidence (Agarwal & McShan, 2024) and PBP3 proteins are often longer (i.e., >500 aa's). The computational approach was able to align distant residues of bla34 for a high structural match with a TM-score of 0.96 and seqID of 51.4% to the *high-lDDT* ARP structure of PBP3 from *K. pneumoniae* (Alphafold: W9BBU1) (Supplementary Table 2, Figure 3h). A Blastp search against the NCBI database, shows bla34 shares a highest seqID of 85% to a PBP3 from the clinical pathogen *A. baumannii* (NCBI: MFY0112749.1). To the best of our knowledge, the PBP3 identified in this study represents the first functionally verified report of a PBP3 from *A. radioresistens* SH164 providing β-lactam resistance when expressed in *E.coli*.

Reports linking alternate PBPs and important mutations for β-lactam resistance need further elucidation in *Acinetobacter* sp. (Sethuvel et al., 2023; Zapun et al., 2008). This explains their absence in the CARD and why they were not used in the training dataset- to create one-class SVM models for PBPs. Bla34 may have previously gone undetected as a responsible determinant for β-lactam resistance given the presence of OXA-905 in the same strain, determined by this study (Supplementary Table 3). Further study is needed to deduce potential substitutions responsible for resistance and it cannot be ruled out that overexpression of bla34 (which may have a reduced affinity for β-lactams (Zapun et al., 2008)), could have produced the observed phenotype, as described for other PBPs (Palmer & Kishony, 2014). Nonetheless, high homology of PBP3's shared

between *Acinetobacter* sp., increases the risk for the novel PBP3 from *A. radioresistens* SH164 to horizontally transfer to clinical pathogen *A. baumannii*. A phylogeny of the novel PBP3 from *A. radioresistens* SH164 in relation to other PBP's within the functional SVM cluster is presented in Supplementary Figure 4.

The *pre-resistome*: query antibiotic resistant proteins (qARP's) with activity but MICs below CLSI resistant breakpoints

E.coli BW25113 harbouring construct pGDP3-dfr28, containing a potential DFR (dfr28) from *E.coli* 53FAA conferred an increased tmp MIC compared to the negative control when measured by BM (MIC 12.5 µg/mL). The qARP structure of dfr28 has a seqID of 57.7% and TM-score of 0.97 to $dfrA3_{high\text{-}lDDT}$ (Supplementary Table 2, Figure 3j). *E.coli* BW25113 harbouring pGDP2-bla35 contains a potential MBL, bla35, from *Burkholderiales* sp. 1_1_47, which conferred an increased ampicillin MIC determined also by BM (MIC 4 µg/mL) against the negative control (MIC ≤2 µg/mL). The qARP structure of bla35 shares a seqID of 36.2% and TM-score of 0.95 to $SPG\text{-}1_{high\text{-}lDDT}$ (Supplementary Table 2, Figure 3d).

Whilst these constructs had activity against their respective antibiotics, they did not reach MICs at CLSI resistant breakpoints. Based on the computational method, the protein structures of both dfr28 and bla35 have at least one primary and tertiary structural feature that fits within the pairwise distribution of conserved primary and tertiary ARP structures within a clustered ARP class, so we posit these proteins are within the 'pre-resistome' – proteins with high potential to evolve into ARPs (i.e., with clinical relevance), however distinguished from the proto-resistome as they still hold some function. Moreover, the efflux-pump deficient *E.coli* BW25113 (ΔbamBΔtolC) was used to express qARPs to ensure all elevated MICs are attributed directly to the query enzyme. Although 'pre-resistome' proteins alone do not confer clinically relevant MICs against their respective antibiotics, CLSI resistant breakpoints may be reached when these proteins act in concert with, for example, increased efflux and decreased influx of carbapenems with bla35 (Pai et al., 2001; Tewawong et al., 2025)) and the presence of SUL genes (Hu et al., 2016), efflux pumps (Herrera-Heredia et al., 2017) and other folate transporters with dfr28 (Rodrigo et al., 2022), as well as increased expression to raise tmp MICs (Palmer & Kishony, 2014). The concept of the 'pre-resistome' in relation to the proto-resistome and resistome is depicted in Figure 4, exemplified with ampicillin MICs of three qARPs predicted as class B MBLs, experimentally tested in this study.

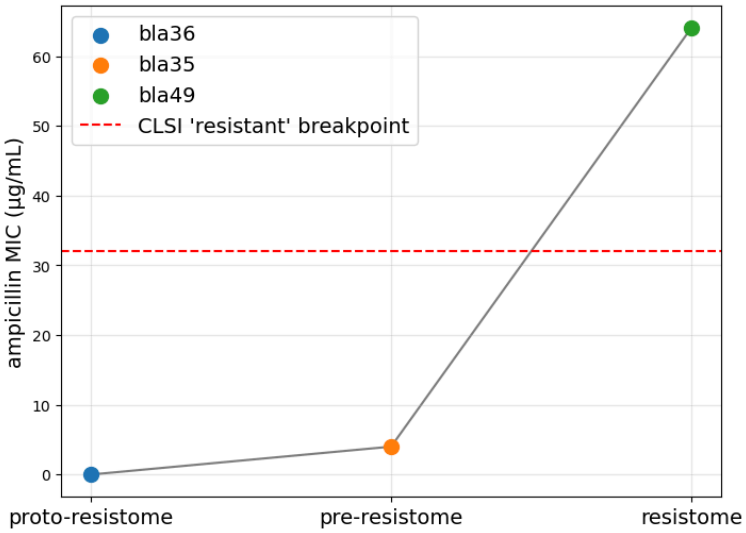


Figure 4. Varying ampicillin MICs of three experimentally tested qARPs, bla36, bla35, and bla49, which respectively belong to the proto-resistome, pre-resistome, and resistome.

Table 1. Functional confirmation results of qARPs and controls. All MICs were determined by broth microdilution (BM) unless marked and are measured in µg/mL.

MIC (µg/mL) for *E. coli* BW25113 (ΔbamBΔtolC) harboring	qARP	ARG	trimethoprim	sulfamethoxazole	SXT	ampicillin	cefotaxime	meropenem
pGDP2-bla33	bla33	*AJDC-1#*				**>256; >64***	**32; 16***	≤0.25; ≤2*
pGDP2-bla34	bla34	PBP3-*A. radioresistens#*				**256; >64***	**16; 8***	≤0.25; ≤2*
pGDP2-bla35	bla35					**4**; ≤16*	≤0.5; ≤2*	≤0.25; ≤2*
pGDP2-bla36	bla36					≤2; ≤16*	≤0.5; ≤2*	≤0.25; ≤2*
pGDP2-bla49	bla49	*KMH-1#*				≤2; **>64***	≤0.5; **8***	≤0.25; **>8***
pGDP1-bla50	bla50	*ECB-1#*				≤2; **32***	≤0.5; ≤28*	≤0.25; ≤2*
pGDP3-dfr27	dfr27		**25; >16***		≤2/38*			
pGDP3-dfr28	dfr28		**12.5;** ≤8*		≤2/38*			
pGDP3-dfr29	dfr29	*dfrA50#*	**>256; >16***		≤2/38*			
pGDP3-dfr30	dfr30	*dfrA51#*	**>256; >16***		≤2/38*			
pGDP2-*ACT-28*		*ACT-28*					**>32**	
pGDP2-*ccrA*		*ccrA*				**256**	**>32**	**>32**
pGDP2-*DIM-1*		*DIM-1*				**>64***	**>16***	**>8***
pGDP3-*dfrK*		*dfrK*	**>256; >16***		**>4/76***			
pGDP2/pGDP3-*sul3*		*sul3*		**>512***	≤2/38*			
pGDP1/pGDP2/pGDP3-empty vector			≤2; ≤8*	≤256*	≤2/38*	≤2; ≤16*	≤0.5; ≤2*	≤0.25; ≤2*
CLSI ‘resistant’ breakpoint			≥16	≥512	≥4/76	≥32	≥4	≥4

* = Measured by adapted agar diffusion (AAD); # = Proposed name of novel ARG; Bold text = MIC > MIC of negative control.

The computational method identified a novel class B metallo-β-lactamase with low homology to existing ARGs

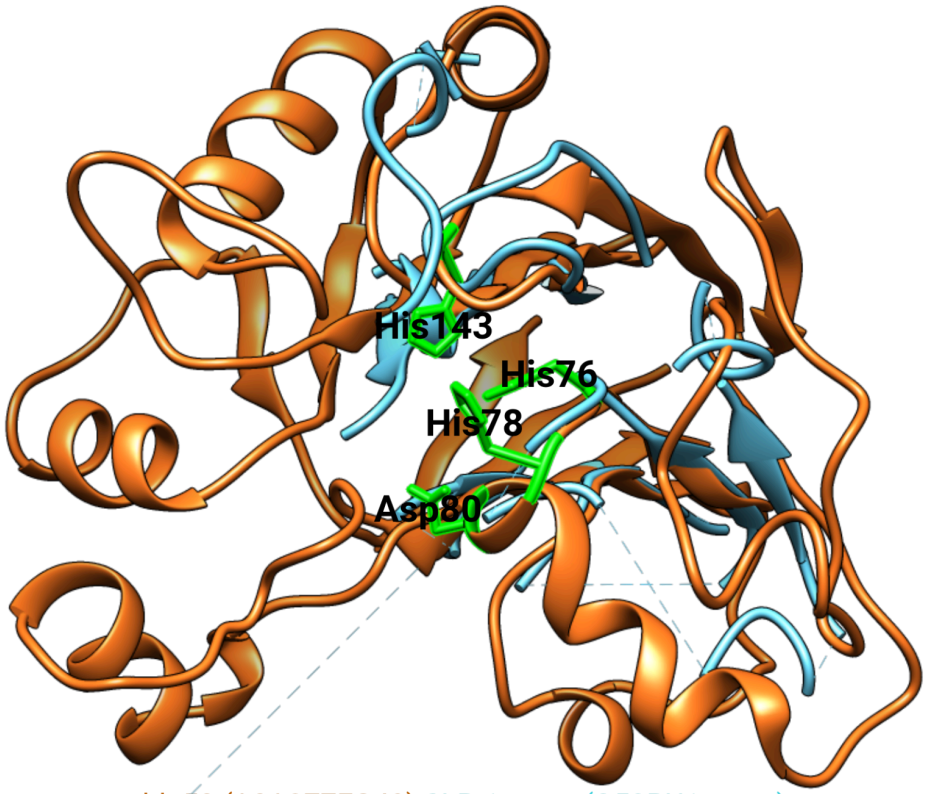
We next sought to further optimize functional predictions of qARPs by skipping the initial Diamond screening of a DNA database against the created custom Diamond database of ARG sequences obtained from the CARD, by downloading protein structures directly from the AFDB for functional prediction against created one-class SVM models. We were also motivated by the computational methods' ability to identify ARGs with low seqID to known ARGs so we focussed on proteins with remote homology to known ARGs. We prioritized MBLs given their importance in conferring resistance to last-resort carbapenems, as well as annotations originating from *E. coli*, to maximise chances of successful expression in the expression system *E.coli* BW25113, by searching the UniProtKB database with the following filter: (taxonomy_id:562) AND (protein_name: Metallo-beta-lactamase). We discarded qARP structures with >20% seqID to known MBL structures which resulted in the computational analysis of 172 qARP structures downloaded directly from the AFDB. Of these, 48 qARP structures were predicted as functional by the computational method which have an average seqID of 21.8% and TM-score of 0.81 to conserved structural regions (*high-lDDT*) of known MBLs (Supplementary Table 4).

We selected one qARP for functional verification and *E.coli* BW25113 harbouring construct pGDP1-bla50 and expressing qARP bla50 (AlphaFold/Uniprot ID: A0A377FG49), had an increased MIC against ampicillin when measured by AAD, (MIC 32 µg/mL), compared to the negative control (MIC ≤ 16 µg/mL), with growth observed at the intermediate CLSI breakpoint of 16 µg/mL. Increased activity against ampicillin was not observed compared to the negative control when measured by BM (Table 1), which is also likely due to differences of the methodological procedure.

A region of the qARP (covering only ~20% of its length) shares 30% seqID with POM-2 in CARD, with an e-value of 0.11967. Bla50 is annotated in Genbank as a phosphonate metabolism protein (PhnP, WP_000059951.1) and a carbon-phosphorus (C-P) lyase complex accessory protein (MBL superfamily) in UniProt from *E.coli* strains 412057 and NCTC9706 (non-HMP reference strains). The *phn* operon in *E. coli* breaks down phosphonates when phosphate is scarce (Yakovleva et al., 1998) and bla50 contains the conserved metal-binding active site for potential hydrolysis of penicillins. Based on the computational method, the qARP structure of bla50 shares a highest seqID with the *high-lDDT* ARP structure SLB-$1_{\text{high-lDDT.}}$ In MBL's the active site is encoded by residues H116, H118, D120, and H196 (Boyd et al., 2020) and were captured as a conserved region of the *high-lDDT* ARP structure, SLB-$1_{\text{high-lDDT}}$. These residues are structurally aligned to the same site present in bla50 by the computational method (encoded by residues H76, H78, D80, and H143, Figure 5). When the qARP structure of bla50 is aligned to the *full-length* protein structure encoded by SLB-1 using Foldseek, these proteins only share a seqID of 8.9% and TM-score of 0.54 (normalized by the subject length), reflecting the value of alignment against only conserved regions of ARP

structures. Due to elevated ampicillin MICs and low homology to existing MBLs we tentatively propose a name of this gene of ECB-1 (*E. coli* beta-lactamase). Following an MSTA, a phylogeny of ECB-1 relative to SLB-1 and PNGM-1 (the full length qARP structure of which ECB-1 shares the highest seqID (17.3%) and TM-score (0.71)) is presented in Supplementary Figure 2.

```
bla50 (A0A377FG49)          5 LTLTGTGGAQGVPAWGCECVACARARRSPQYRRQPCSGVVKFNEAITLIDAGLHDLADRWSPGSFQQFLLTHYHMDHVQG
                              + L                                VV       +ID                 + TH H D
SLB-1_high-lDDT (Q53BU1)    3 VYLH-----------------------------ANGLVV----QAFIIDTP------------PVASISTHSH-DRA--
bla50 (A0A377FG49)         85 LFPLRWGVGDTIPVYGPPDEQGCDDLFKHPGLLDFSHTVEPFVVFDLQGLQVTPLP-LNHSKLTFGYLLETAHSRVAWLS
                                L       I       +Q                                   H+K  +   L     ++ +
SLB-1_high-lDDT (Q53BU1)   34 Y--LN-R---GITTTV-TTQQI------------A-----------------VYDLGAGHTKDNLVVWL-----QILFGG
bla50 (A0A377FG49)        164 DTA-----G-LPEKTLKFLRNNQPQVMVIDCSHPPREDAPRNHCDLNTVLALNQVIRSPRVILTH
                                          T K        ++V                                  H
SLB-1_high-lDDT (Q53BU1)   73 CLISGEADLQQWL-TAK------V-IVVPG---------------------------------H
```



bla50 (A0A377FG49) SLB-1$_{\text{high-lDDT}}$ (Q53BU1$_{\text{high-lDDT}}$)
seqID = 19.9%; TM-score = 0.73

Figure 5. The protein sequence and Foldseek TM-align alignment of qARP bla50 with the *high-lDDT* ARP structure, SLB-1$_{\text{high-lDDT}}$. Active site residues in the alignment and corresponding structural atoms are highlighted in green.

Discussion

Recent studies have demonstrated the success of *in silico* prediction of novel ARGs with up to 80% accuracy of computational pipelines after experimental verification, even with low seqID to existing ARGs (Arango-Argoty et al., 2018; Ruppé et al., 2019).

Nonetheless, computationally predicted novel ARGs that are clinically important (i.e., confer MICs at CLSI resistant breakpoints), typically have seqIDs > 50% to known ARGs (Alcock et al., 2023; Berglund et al., 2017, 2019; Lund et al., 2023). Predictions of false positives increase at lower seqIDs, and high-throughput functional verification methods are still limited, owing to the primary bottleneck in the discovery pipeline. To address this, we leveraged recent advances in computational protein structure prediction and search to distinguish proteins from the proto-resistome that lack key amino acid residues or localized structural features associated with antibiotic resistance, focusing instead on conserved structural motifs (*high-lDDT* regions) in known ARGs and the amino acids that code for them. By measuring the similarity of distantly modeled residues and their pairwise tertiary structural distances, we were able to reduce noise from disordered or structurally divergent regions that can hamper full protein structure alignments and pairwise sequence structure comparisons (Illergård et al., 2009; Koehler Leman et al., 2023) and experimentally validated a novel, precise, and computationally efficient method to identify novel ARGs.

Eight out of nine (89%) proteins were functionally verified from HMP reference strains which had an average seqID of 46.2% to known ARGs and six out nine genes conferred MICs at CLSI resistant breakpoints. Furthermore, a protein with a MBL-fold in the AFDB with no significant homology over 70% subject length was functionally verified to provide clinically relevant MICs against ampicillin. Alignment of the qARP structure to *high-lDDT* ARP structures within each cluster and subsequent machine learning predictions are designed for computational efficiency and high accessibility and can be completed in under a few minutes on a standard desktop or laptop computer. As a result, the main required input to the computational method is a query protein structure. For proteins of ~300 amino acids in length, these predictions can be completed in under 5 minutes on a GPU and under 20 hours on 8 CPUs. Moreover, the AFDB can be queried for a vast amount of protein sequences already with computational solutions, reducing the need for computational protein structure predictions.

Clustering ARP structures at 20%, 30%, and 50% seqID prior to MSTA's is a critical component of the method as it limits pairwise seqID and TM-score distances of *high-lDDT* ARP structures, offering a balance of specificity and sensitivity. However, ARP structures with distant homology (i.e., <20% seqID) will be clustered as singletons or with <3 ARPs, restricting direct comparison of qARP primary and tertiary structural features to these ARPs, and highlights an important limitation of the algorithm. For example, the false positive qARP identified, bla36, shares a highest seqID to VMH, which clustered as a singleton. Bla36 was predicted functional within another 20% class B MBL cluster without comparison to conserved sequence and structural features important for function of VMH. This demonstrates the importance of wet-laboratory focussed methods (for e.g., functional metagenomics) to identify ARGs with less-common active sites and we anticipate that as our models are updated in the future with more ARP structures, clustering effects from more distant ARG homologs will be minimized.

We performed MIC testing by assembling qARPs into selected expression vectors (pGDP1, pGDP2, and pGDP3) of the Minimal Antibiotic Resistance Platform (Cox et al.,

2017), as well as negative and positive controls, in efflux-deficient *E.coli* BW25113 (Δ bamBΔtolC), so any increased activity could not be attributed to efflux pumps. Bla49 (from *K. pneumoniae* subsp. *pneumoniae* WGLW3) and bla50 (from *E. coli* 412057 and NCTC9706), were susceptible to ampicillin by BM, however reached CLSI resistant breakpoints when determined by AAD. Bla49 also reached clinical breakpoints for meropenem and cefotaxime when determined by AAD. Discrepancies between BM and agar disk diffusion methods have been reported with low agreement in *E. coli* for ampicillin (20%) (Palladini et al., 2023), as well as growth on ampicillin agar plates but not in Lysogeny Broth (LB) (Schoning, 2014). The pGDP series have a low-copy number and following gibson assembly low numbers of vectors which contain the qARP relative to those that don't, may have restricted the proliferation of cells expressing the qARP, because diffusion of ampicillin is expectedly higher in liquid as opposed to solid medium. An alternative explanation is the result of decontextualized qARPs expressed in another host with potential toxic effects that perturb plasmid maintenance. Plasmids may be maintained successfully with promoter regulation (Loew et al., 2010, (Flach et al., 2013; Razavi et al., 2017). These discordant results underscore the importance of verifying susceptible strains determined by BM with a methodology involving solid media, as well as determination of MICs alongside both positive and negative controls.

The computational prediction and functional confirmation of proteins identified from HMP reference strains: 4 novel DFR's, 1 novel class C and 2 class B β-lactamases, as well as a longer, membrane-associated PBP protein, and an ARG with little homology to existing ARGs, has provided a proof-of-concept of training a one-class SVM on pairwise primary and tertiary protein structure similarities in conserved regions of ARP structures to identify novel ARGs. High-precision algorithms are essential for identifying novel ARGs while minimizing false positives encountered in the wet-laboratory, and was achieved in this study through accurate functional verification of computational predictions using standardized experimental procedures and a range of susceptibility testing methods. A number of novel ARGs identified in this study already present clinically relevant MICs or are emerging from the 'pre-resistome', and their stable integration within the human gut microbiome increases their risk of showing up in clinical settings by transfer to human pathogens (Salyers et al., 2004). Finally, in an era of advancing computational structure prediction, we suggest conserved pairwise primary and tertiary protein structural feature distributions might be expanded beyond predicting functionality of ARGs alone and suggest researchers experiment with the approach on other datasets of proteins.

Data availability

Protein sequences of qARPs predicted functional by the computational method are available in the Supplementary data. Foldseek databases and *high-lDDT* ARP structure .pdb files of 20%, 30%, and 50% ARG class clusters and all code generated will be available as soon as possible.

Methods

Training set

Subject antibiotic resistance protein (ARP) structures

Computational and experimental (where available) ARP structures with an average pLDDT of at least 80 and encoded by protein sequences of well-known ARG classes obtained from CARD (Alcock et al., 2023) (AAC, APH, ANT, DFR, and SUL), the BLDB (class A, B, C and D β-lactamases) (Naas et al., 2017), and Berglund et al. (2017) (class B1 β-lactamases with detected activity), were downloaded from the AFDB or predicted by AlphaFold2 (Jumper et al., 2021) if unavailable, using full databases from 2023-05-30. The most recent version of AlphaFold, AlphaFold3, was not available at the time of analysis in 2023-2024. AlphaFold2 predictions were performed using Digital Research Alliance of Canada clusters or the COSMIC2 gateway web server (Cianfrocco et al., 2017). For the high number of β-lactamase variants with high seqID (i.e., >90%), at least one representative ARP structure of these variants was computationally predicted. ARP structures were then clustered by Foldseek (v.10-941cd33) (van Kempen et al., 2024) at 20%, 30%, and 50% seqID. Clusters with less than 3 ARP structures were omitted from subsequent analysis. AlphaFold ID's (for existing ARP structures) or own AlphaFold2 predictions of ARP structures predicted functional are listed in Supplementary Table 2.

Structurally conserved regions of ARP structures

To identify structurally conserved regions within clusters of ARP structures, we employed FoldMason (v.3-954d202) (Gilchrist et al., 2024), a tool designed for structural alignment and consensus modeling of protein structures at scale. FoldMason utilizes a per-column local distance difference test (lDDT) score (Mariani et al., 2013) to assess the consistency of atomic positions across aligned structures. The lDDT can estimate local structural conservation in multiple structural alignments (MSTAs) without relying on superposition, and enables identification of more conserved motifs even in protein structures with global conformational variability. There are limited established lDDT score thresholds in the literature that correspond to conserved structural regions of protein structures within an MSTA and they are often context dependent. In this case, we analyzed the distribution of lDDT scores (excluding values of -1, where only one protein structure is aligned in the column) by the application of a kernel density estimation (KDE) to smooth the distribution, minimize local noise, and reveal the global trend (Parzen, 1962; Rosenblatt, 1956). KDE is a non-parametric approach that avoids assumptions about the underlying distribution and estimates the probability density by centering a kernel (Gaussian) function at each data point. The peak of the KDE (argmax) was used as the representative lDDT score for structural conservation, as it is assumed to reflect the most probable structural similarity across the aligned residues and assumes conserved positions exhibit the highest consensus structural similarity. We also deemed this approach appropriate as it extracts conserved residues within the context of the particular MSTA; as opposed to applying a pre-defined lDDT threshold

across all MSTAs which both fluctuate in numbers of ARP structures and their structural variability. As such, we extracted residues of each ARP structure in FoldMason MSTAs with per-column lDDT values ≥ the KDE peak of the distribution of all lDDT scores in the MSTA to create *high-lDDT* ARP structures, represented as new .pdb structure files for each ARP structure. lDDT score distributions with KDE smoothing applied and associated KDE peak values for each MSTA within each cluster for class B β-lactamases as an example, are presented in Supplementary Figure 5. The Foldseek *createdb* command was used to create databases of *high-lDDT* ARP structures within each cluster.

Training

The training data represents pairwise primary and tertiary protein structure comparisons for *high-lDDT* ARP structures within 20%, 30% and 50% seqID ARP structure clusters, creating an x, y distribution of seqIDs (x) and TM-scores (y). This was computed using Foldseek TM-align with parameters --alignment-type 1 and --exhaustive-search specified and TM-scores were normalized by the subject length. seqIDs and TM-scores were scaled using scikit-learn's 'StandardScaler' (Pedregosa et al., 2011), which standardizes each feature independently by removing its mean and scaling to unit variance.

One-class SVM and hyperparameters

One-class support vector machines (one-class SVMs) have been previously implemented for outlier detection tasks, particularly when only positive (inlier) data are available (H. J. Shin et al., 2005). The algorithm learns a decision function that defines a boundary around the training data in a transformed feature space and called the decision boundary (DB). New data points are classified as inliers or outliers based on their relative position to the DB. For two-dimensional x, y distributions, the shape and tightness of the DB are primarily controlled by the parameters ν (nu) and γ (gamma) and values for these parameters were automated to avoid manual adjustments. The parameter ∈(0,1] sets an upper bound on the fraction of outliers and a lower bound on the number of support vectors. For the training data, representing a pairwise distribution of seqIDs (x) and TM-scores (y) of *high-lDDT* ARP structures, one-class SVM plots were created using scikit-learn (Pedregosa et al., 2011) and the default value for ν was selected as $1/N$, where N is the number of training points and assumes that all training data are inliers. The kernel parameter, γ, which defines the influence radius of a single data point in the radial basis function (RBF) kernel, was selected automatically using scikit-learn's 'scale' option, which sets γ=1/(number of features × data variance), deemed to provide a reasonable balance between overfitting and underfitting.

Validation using an existing protein structure based method of predicting ARGs

Validation of the computational method and the one-class SVM hyperparameters employed for training was performed on proteins previously experimentally verified by

an existing computational method also based on protein structure used to identify ARGs, called pairwise comparative modelling (PCM) (Ruppé et al., 2019). PCM predicted ARGs from a 3.9 million gene catalogue of the intestinal microbiota. Referring to Supplementary Table 1 of the PCM paper, we collected protein sequences for analysis from the project website (https://mgps.eu/Mustard/index.php?id=accueil), and only focussed on ARG classes which have both positive and negative functional synthesis results and for which one-class SVM models had been created in this study. Additionally, proteins with more than 80% seqID to a known ARP structure were excluded. Protein structures were downloaded directly from the AFDB or predicted by AlphaFold2 and previously confirmed to be absent from the training data. Computational predictions and functional synthesis results of our method compared to PCM for proteins assessed are presented in Supplementary Table 5. In summary, our computational method applied to this dataset has an accuracy of 0.76 and the precision increased from 0.647 of PCM to 0.79 (Table 2). Furthermore, precision remained high at seqIDs to a known ARP structure below 25% (0.75), with 18 and out of 24 true positive predictions being correct, compared to PCM (0.583, Table 2).

Table 2. Accuracy and precision of PCM compared to the computational method described in this study.

	PCM	This approach
Accuracy^	n.a.	0.76
Precision*	0.647	0.79
Precision* (<25% seqID)	0.583	0.75

^ = True positives +True negatives / True positives +True negatives + False positives + False negatives
* = True positives / True positives + False positives

Query antibiotic resistant proteins (qARPs) for functional prediction

qARP structures from HMP reference strains

Protein coding sequences (CDS) were predicted from nucleotide sequences of whole genome sequences (WGS) of selected HMP reference strains isolated from the human gut (Supplementary Table 1) by Prodigal (Hyatt et al., 2010) and underwent the initial Diamond screening by alignment to a custom database (Buchfink et al., 2015) of protein sequences obtained from CARD (v4.0.1) (Alcock et al., 2023) belonging to the same ARG families used to create one-class SVM models, with the following arguments specified: -e 10 --sensitive --query-cover 70 --subject-cover 70. CDS with >90% seqID to a known ARG were considered known ARGs. Protein structures of the remaining CDS were downloaded from the AFDB or predicted by AlphaFold2 if unavailable, representing qARP structures for functional prediction by the computational method.

qARP structures directly downloaded from the AFDB

To identify novel class MBLs with low homology to existing MBLs and skip the initial Diamond screening against the custom CARD database, we obtained UniProt accessions for proteins within the UniProtKB database (The UniProt Consortium, 2025) with the following filter: (taxonomy_id:562) AND (protein_name:Metallo-beta-lactamase). Using obtained UniProt accessions, their associated protein structures were downloaded directly from the AFDB. Protein structures with >20% seqID to known MBL structures, determined by Foldseek under default settings, were discarded with the remainder representing qARP structures for functional prediction by the computational method.

Cluster assignation

Prior to passing into one-class SVM models for functional prediction, qARP structures were aligned to a database of all ARP structures using Foldseek under default settings. This was performed so qARP structures could be assigned to 20%, 30% or 50% ARP structure clusters for a particular ARG class, based on their highest seqID to an ARP structure (Table 3).

Table 3. Assignation of qARP structures to ARP structure clusters.

Best seqID with an ARP structure (%)	Assigned to clusters
0 - 25	20%
> 25 - 50	30%
> 50	50%

Functional classification

Each qARP structure was aligned using Foldseek to *high-lDDT* ARP structure databases, to obtain an x, y distribution of qARP structure seqIDs (x) and TM-scores (y) against *high-lDDT* ARP structures (also scaled using scikit-learn's 'StandardScaler'). As with the training data, this was computed using Foldseek's TM-align mode, specified with --alignment-type 1, which uses the Needleman-Wunsch (NW) algorithm (Needleman & Wunsch, 1970) with affine gap costs. The NW algorithm was deemed appropriate to enable more accurate global alignments of shorter *high-lDDT* ARP structures to full-length qARP structures, in contrast to Foldseek's default alignment mode which uses the Smith-Waterman algorithm and is more suitable for local alignments (Smith & Waterman, 1981). Additionally, Foldseeks' --exhaustive-search was specified and the *high-lDDT* ARP structure (subject) length was used to normalize TM-scores and calculate seqIDs.

Let $f(x)$ be the decision function learned by the one-class SVM from the training data, where x represents a 2D feature vector (i.e., seqID and TM-score). For a given qARP

structure constituting multiple points $\{x^{(1)}_{qARP\ structure}, x^{(2)}_{qARP\ structure}, ..., x^{(n)}_{qARP\ structure}\}$, the qARP is classified as *functional* if *any* of its points lie on or within the DB:

$$\exists\ i\ \in \{1,...,n\}\ such\ that\ f(x^{(i)}_{qARP\ structure}) \geq 0$$

That is, the qARP is considered functional if at least one of its associated feature points has a decision score greater than zero, indicating it lies within the DB learned from the training data. Otherwise, the qARP is classified as non-functional. This criterion reflects the default decision threshold of the one-class SVM and assumes that any point outside the DB (i.e., $f(x^{(i)}_{qARP\ structure}) < 0$) is an outlier with respect to the training distribution.

Manual adjustment of Foldseek seqID calculations

Due to the alignment of shorter *high-lDDT* ARP structures against full-length qARP structures using Foldseek TM-align, underestimated seqID calculations were noticed due to alignment gaps not being properly accounted for in the aligned region of the qARP structure. Manual adjustment of seqIDs was performed by incorporating an alignment length equal to that of the length of the *high-lDDT* ARP structure.

Experimental verification

The Minimal Antibiotic Resistance Platform includes vectors grouped by a kanamycin (pGDP1 and pGDP2) or ampicillin (pGDP3 and pGDP4) selective marker, and by promoter strength, with the stronger β-lactamase pBla promoter (pGDP1 and pGDP3) and the weaker pLac promoter (pGDP2 and pGDP4) (Cox et al., 2017). After codon optimization for *E.coli*, DNA sequences of selected qARPs were synthesized by Twist Biosciences, Gibson assembled into appropriate vectors, and chemically transformed into *E.coli* MM294 chemically competent cells. It should be noted that successful assembly of the DNA sequence of qARP bla50, was only achieved in the pGDP1 series, which has the stronger *bla* promoter (as opposed to the weaker lac promoter in pGDP2 used to express remaining qARP β-lactamases). After 1h at 37°C, transformants were selected and grown overnight in Lysogeny Broth (LB). Plasmids were then extracted using a Plasmid DNA miniprep kit (New England Biolabs) and sequenced on a miniION Flongle Flowcell (Oxford Nanopore Technologies) to confirm a successful assembly. Plasmids were then re-transformed via electroporation into *E.coli* BW25113 (ΔbamBΔ tolC) and grown overnight on Mueller-Hinton (MH) II agar for 18h at 37°C containing the appropriate selective marker. Re-transformed colonies were then picked and stored at -80°C in glycerol. The Minimal Antibiotic Resistance Platform was a gift from Gerard Wright (Addgene kit #1000000143).

MIC measurement

Broth microdilution (BM)

For measurements of MICs by broth microdilution (BM), all strains were grown overnight in MHII broth for 18h at 37°C and applied in triplicate to a microdilution plate containing antibiotics (ampicillin, cefotaxime, and meropenem, for qARP β-lactamases and the PBP and trimethoprim and sulfamethoxazole-trimethoprim (SXT) for qARP DFRs), ranging around resistant CLSI breakpoints (CLSI M100, 34ed, 2024), as per CLSI M07 (CLSI, 2024). Following incubation for 18h at 37°C, visual inspection of wells (Wiegand et al., 2008) and optical density at 600nm (OD600) were recorded. All measurements were compared to previously transformed *E.coli* BW25113 (ΔbamBΔtolC) strains containing an empty pGDP vector (negative control), and a pGDP vector containing a known ARG in the respective class of the qARP (positive control). Transformed strains were considered resistant if visible growth and a higher OD600 was recorded for a higher antibiotic concentration for the strain expressing the qARP compared to the negative control, in all 3 replicates. For OD600 measurements, a *P* value of 0.05 was used to confirm significance.

Adapted agar diffusion (AAD) method

The adapted agar diffusion (AAD) method was developed in-house in an attempt to increase through-put and automation of MIC measurements. Briefly, 1 µL of culture was collected by a Rotor HDA (Singer Instruments) from each well of a 96-well plate containing a strain for MIC testing, previously grown for 18h at 37°C in 200 µL LB. The strain was then applied to MHII agar or MHII agarose (for SXT testing) plates with dissolved concentrations of relevant antibiotics at CLSI breakpoints, considered susceptible (s), intermediate (i), and resistant (r). Strains were applied to plates in the same relative position as they were collected from the 96-well plate. After 18h at 37°C, plates were visually inspected for colonies, with a strain considered resistant if any growth was present at a higher antibiotic concentration than the negative control, as well as all concentrations below the MIC of the strain harbouring the qARP (if applicable). Photos of plates and the associated layout of the 96-well plate corresponding to strain positions on plates are presented in Supplementary Figures 6a-6f and Supplementary Table 6, respectively.

Phylogenetics

To investigate evolutionary relationships between identified novel ARGs and closely related ARGs, FoldMason under default settings was used to create an MSTA. Maximum likelihood was applied as a search criterion in IQ-TREE v.3.0.0 using the best-fit model automatically selected by ModelFinder (Nguyen et al., 2015). An Ultrafast Bootstrap analysis was performed with 1000 ML replicates and an approximate likelihood ratio test search with 10,000 replicates. Phylogenetic trees were visualized and annotated with FigTree v.1.4.4 (available http://tree.bio.ed.ac.uk/software/figtree/).

References

Abbas, M. M., Ranjan, A., Hou, A., & Mukhopadhyay, S. (2025). Trans-ARG: Predicting Antibiotic Resistance Genes with a Transformer-Based Model and Pretrained Protein Language Model. *Proceedings of the 4th International Conference on AI-ML Systems*, 1–8. https://doi.org/10.1145/3703412.3703425

Agarwal, V., & McShan, A. C. (2024). The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nature Chemical Biology*, *20*(8), 950–959. https://doi.org/10.1038/s41589-024-01638-w

Aguilar-Vera, A., Bello-López, E., Pantoja-Nuñez, G. I., Rodríguez-López, G. M., Morales-Erasto, V., & Castillo-Ramírez, S. (2024). Acinetobacter junii: An emerging One Health pathogen. *mSphere*, *9*(5), e00162-24. https://doi.org/10.1128/msphere.00162-24

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M. C., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., … McArthur, A. G. (2023). CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, *51*(D1), D690–D699. https://doi.org/10.1093/nar/gkac920

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 23. https://doi.org/10.1186/s40168-018-0401-z

Baquero, F., Martínez, J. L., F. Lanza, V., Rodríguez-Beltrán, J., Galán, J. C., San Millán, A., Cantón, R., & Coque, T. M. (2021). Evolutionary Pathways and Trajectories in Antibiotic Resistance. *Clinical Microbiology Reviews*, *34*(4), e00050-19. https://doi.org/10.1128/CMR.00050-19

Berglund, F., Marathe, N. P., Österlund, T., Bengtsson-Palme, J., Kotsakis, S., Flach, C.-F., Larsson, D. G. J., & Kristiansson, E. (2017). Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, *5*(1), 134. https://doi.org/10.1186/s40168-017-0353-8

Berglund, F., Österlund, T., Boulund, F., Marathe, N. P., Larsson, D. G. J., & Kristiansson, E. (2019). Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*, *7*(1), 52. https://doi.org/10.1186/s40168-019-0670-1

Boyd, S. E., Livermore, D. M., Hooper, D. C., & Hope, W. W. (2020). Metallo-β-Lactamases: Structure, Function, Epidemiology, Treatment Options, and the Development Pipeline. *Antimicrobial Agents and Chemotherapy*, *64*(10), e00397-20. https://doi.org/10.1128/AAC.00397-20

Brem, J., van Berkel, S. S., Zollman, D., Lee, S. Y., Gileadi, O., McHugh, P. J., Walsh, T. R., McDonough, M. A., & Schofield, C. J. (2015). Structural Basis of Metallo-β-Lactamase Inhibition by Captopril Stereoisomers. *Antimicrobial Agents and Chemotherapy*, *60*(1), 142–150. https://doi.org/10.1128/aac.01335-15

Brolund, A., Sundqvist, M., Kahlmeter, G., & Grape, M. (2010). Molecular Characterisation of Trimethoprim Resistance in Escherichia coli and Klebsiella pneumoniae during a Two Year Intervention on Trimethoprim Use. *PLoS ONE*, *5*(2), e9233. https://doi.org/10.1371/journal.pone.0009233

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), Article 1. https://doi.org/10.1038/nmeth.3176

Cianfrocco, M. A., Wong-Barnum, M., Youn, C., Wagner, R., & Leschziner, A. (2017). COSMIC2: A Science Gateway for Cryo-Electron Microscopy Structure Determination. *Practice and*

Experience in Advanced Research Computing 2017: Sustainability, Success and Impact, 1–5. https://doi.org/10.1145/3093338.3093390

CLSI. (2024). *M07 | Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically* (Version 12).

Cox, G., Sieron, A., King, A. M., De Pascale, G., Pawlowski, A. C., Koteva, K., & Wright, G. D. (2017). A Common Platform for Antibiotic Dereplication and Adjuvant Discovery. *Cell Chemical Biology*, *24*(1), 98–109. https://doi.org/10.1016/j.chembiol.2016.11.011

Crits-Christoph, A., Hallowell, H. A., Koutouvalis, K., & Suez, J. (2022). Good microbes, bad genes? The dissemination of antimicrobial resistance in the human microbiome. *Gut Microbes*, *14*(1), 2055944. https://doi.org/10.1080/19490976.2022.2055944

D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., & Wright, G. D. (2011). Antibiotic resistance is ancient. *Nature*, *477*(7365), Article 7365. https://doi.org/10.1038/nature10388

Faltyn, M., Alcock, B., & McArthur, A. G. (2019). *Evolution and nomenclature of the trimethoprim resistant dihydrofolate (dfr) reductases*.

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., Tyson, G. H., Zhao, S., Hsu, C.-H., McDermott, P. F., Tadesse, D. A., Morales, C., Simmons, M., Tillman, G., Wasilenko, J., Folster, J. P., & Klimke, W. (2019). Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial Agents and Chemotherapy*, *63*(11), 10.1128/aac.00483-19. https://doi.org/10.1128/aac.00483-19

Flach, C.-F., Boulund, F., Kristiansson, E., & Larsson, D. J. (2013). Functional verification of computationally predicted qnr genes. *Annals of Clinical Microbiology and Antimicrobials*, *12*(1), 34. https://doi.org/10.1186/1476-0711-12-34

Ghosh, A., Vang, C. K., Brenner, E. P., & Ravi, J. (2025). Unlocking antimicrobial resistance with multiomics and machine learning. *Trends in Microbiology*, *0*(0). https://doi.org/10.1016/j.tim.2025.04.017

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, *9*(1), 207–216. https://doi.org/10.1038/ismej.2014.106

Gilchrist, C. L. M., Mirdita, M., & Steinegger, M. (2024). *Multiple Protein Structure Alignment at Scale with FoldMason* (p. 2024.08.01.606130). bioRxiv. https://doi.org/10.1101/2024.08.01.606130

Hamid, M.-N., & Friedberg, I. (2020). *Transfer learning improves antibiotic resistance class prediction* (p. 2020.04.17.047316). bioRxiv. https://doi.org/10.1101/2020.04.17.047316

Herrera-Heredia, S. A., Pezina-Cantú, C., Garza-González, E., Bocanegra-Ibarias, P., Mendoza-Olazarán, S., Morfín-Otero, R., Camacho-Ortiz, A., Villarreal-Treviño, L., Rodríguez-Noriega, E., Paláu-Davila, L., Maldonado-Garza, H. J., & Flores-Treviño, S. (2017). Risk factors and molecular mechanisms associated with trimethoprim–sulfamethoxazole resistance in Stenotrophomonas maltophilia in Mexico. *Journal of Medical Microbiology*, *66*(8), 1102–1109. https://doi.org/10.1099/jmm.0.000550

Hu, L.-F., Chen, G.-S., Kong, Q.-X., Gao, L.-P., Chen, X., Ye, Y., & Li, J.-B. (2016). Increase in the Prevalence of Resistance Determinants to Trimethoprim/Sulfamethoxazole in Clinical Stenotrophomonas maltophilia Isolates in China. *PLoS ONE*, *11*(6), e0157693. https://doi.org/10.1371/journal.pone.0157693

Hujer, K. M., Hamza, N. S., Hujer, A. M., Perez, F., Helfand, M. S., Bethel, C. R., Thomson, J. M., Anderson, V. E., Barlow, M., Rice, L. B., Tenover, F. C., & Bonomo, R. A. (2005). Identification of a New Allelic Variant of the Acinetobacter baumannii Cephalosporinase, ADC-7 β-Lactamase: Defining a Unique Family of Class C Enzymes. *Antimicrobial*

Agents and Chemotherapy, *49*(7), 2941–2948. https://doi.org/10.1128/aac.49.7.2941-2948.2005

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, *77*(3), 499–508. https://doi.org/10.1002/prot.22458

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

Kaderabkova, N., Bharathwaj, M., Furniss, R. C. D., Gonzalez, D., Palmer, T., & Mavridou, D. A. I. (2022). The biogenesis of β-lactamase enzymes. *Microbiology*, *168*(8), 001217. https://doi.org/10.1099/mic.0.001217

Kneis, D., Tskhay, F., de la Cruz Barron, M., & Berendonk, T. U. (2024). Bacteria of the order Burkholderiales are original environmental hosts of type II trimethoprim resistance genes (dfrB). *The ISME Journal*, *18*(1), wrae243. https://doi.org/10.1093/ismejo/wrae243

Koehler Leman, J., Szczerbiak, P., Renfrew, P. D., Gligorijevic, V., Berenberg, D., Vatanen, T., Taylor, B. C., Chandler, C., Janssen, S., Pataki, A., Carriero, N., Fisk, I., Xavier, R. J., Knight, R., Bonneau, R., & Kosciolek, T. (2023). Sequence-structure-function relationships in the microbial protein universe. *Nature Communications*, *14*(1), 2351. https://doi.org/10.1038/s41467-023-37896-w

Larsson, D. G. J., & Flach, C.-F. (2022). Antibiotic resistance in the environment. *Nature Reviews Microbiology*, *20*(5), 257–269. https://doi.org/10.1038/s41579-021-00649-x

Lassaux, P., Traoré, D. A. K., Loisel, E., Favier, A., Docquier, J.-D., Sohier, J. S., Laurent, C., Bebrone, C., Frère, J.-M., Ferrer, J.-L., & Galleni, M. (2011). Biochemical and Structural Characterization of the Subclass B1 Metallo-β-Lactamase VIM-4. *Antimicrobial Agents and Chemotherapy*, *55*(3), 1248–1255. https://doi.org/10.1128/aac.01486-09

Loew, R., Heinz, N., Hampf, M., Bujard, H., & Gossen, M. (2010). Improved Tet-responsive promoters with minimized background expression. *BMC Biotechnology*, *10*(1), 81. https://doi.org/10.1186/1472-6750-10-81

Lu, W.-J., Hsu, P.-H., & Lin, H.-T. V. (2021). A Novel Cooperative Metallo-β-Lactamase Fold Metallohydrolase from Pathogen Vibrio vulnificus Exhibits β-Lactam Antibiotic-Degrading Activities. *Antimicrobial Agents and Chemotherapy*, *65*(9), 10.1128/aac.00326-21. https://doi.org/10.1128/aac.00326-21

Lund, D., Coertze, R. D., Parras-Moltó, M., Berglund, F., Flach, C.-F., Johnning, A., Larsson, D. G. J., & Kristiansson, E. (2023). Extensive screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens. *Communications Biology*, *6*(1), 812. https://doi.org/10.1038/s42003-023-05174-6

Manna, M. S., Tamer, Y. T., Gaszek, I., Poulides, N., Ahmed, A., Wang, X., Toprak, F. C. R., Woodard, D. R., Koh, A. Y., Williams, N. S., Borek, D., Atilgan, A. R., Hulleman, J. D., Atilgan, C., Tambar, U., & Toprak, E. (2021). A trimethoprim derivative impedes antibiotic resistance evolution. *Nature Communications*, *12*(1), 2949. https://doi.org/10.1038/s41467-021-23191-z

Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: A local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, *29*(21), 2722–2728. https://doi.org/10.1093/bioinformatics/btt473

Meng, E., Pettersen, E., Couch, G., Huang, C., & Ferrin, T. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, *7*, 339. https://doi.org/10.1186/1471-2105-7-339

Morar, M., & Wright, G. D. (2010). The Genomic Enzymology of Antibiotic Resistance. *Annual Review of Genetics*, *44*(Volume 44, 2010), 25–51. https://doi.org/10.1146/annurev-genet-102209-163517

Naas, T., Oueslati, S., Bonnin, R. A., Dabos, M. L., Zavala, A., Dortet, L., Retailleau, P., & Iorga, B. I. (2017). Beta-lactamase database (BLDB) – structure and function. *Journal of Enzyme Inhibition and Medicinal Chemistry*. https://www.tandfonline.com/doi/abs/10.1080/14756366.2017.1344235

Naik, A., Naik, A., & Deshmukh, P. (2025). ProtGAT: Refining Antimicrobial Resistance Detection with Graph-Based Deep Learning*. *2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET)*, 1–6. https://doi.org/10.1109/ICAET63349.2025.10932144

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

O'Neill, J. (2016). *TACKLING DRUG-RESISTANT INFECTIONS GLOBALLY: FINAL REPORT AND RECOMMENDATIONS* (Review on Antimicrobial Resistance).

Ovung, A., & Bhattacharyya, J. (2021). Sulfonamide drugs: Structure, antibacterial property, toxicity, and biophysical interactions. *Biophysical Reviews*, *13*(2), 259–272. https://doi.org/10.1007/s12551-021-00795-9

Pai, H., Kim, J.-W., Kim, J., Lee, J. H., Choe, K. W., & Gotoh, N. (2001). Carbapenem Resistance Mechanisms in Pseudomonas aeruginosa Clinical Isolates. *Antimicrobial Agents and Chemotherapy*, *45*(2), 480–484. https://doi.org/10.1128/AAC.45.2.480-484.2001

Palladini, G., Garbarino, C., Luppi, A., Russo, S., Filippi, A., Arrigoni, N., Massella, E., & Ricchi, M. (2023). Comparison between broth microdilution and agar disk diffusion methods for antimicrobial susceptibility testing of bovine mastitis pathogens. *Journal of Microbiological Methods*, *212*, 106796. https://doi.org/10.1016/j.mimet.2023.106796

Palmer, A. C., & Kishony, R. (2014). Opposing effects of target overexpression reveal drug mechanisms. *Nature Communications*, *5*(1), 4296. https://doi.org/10.1038/ncomms5296

Pang, Z., Raudonis, R., Glick, B. R., Lin, T.-J., & Cheng, Z. (2019). Antibiotic resistance in *Pseudomonas aeruginosa*: Mechanisms and alternative therapeutic strategies. *Biotechnology Advances*, *37*(1), 177–192. https://doi.org/10.1016/j.biotechadv.2018.11.013

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pei, Y., Shum, M. H.-H., Liao, Y., Leung, V. W., Gong, Y.-N., Smith, D. K., Yin, X., Guan, Y., Luo, R., Zhang, T., & Lam, T. T.-Y. (2024). ARGNet: Using deep neural networks for robust identification and classification of antibiotic resistance genes from sequences. *Microbiome*, *12*(1), 84. https://doi.org/10.1186/s40168-024-01805-0

Philippon, A., Arlet, G., Labia, R., & Iorga, B. I. (2022). Class C β-Lactamases: Molecular Characteristics. *Clinical Microbiology Reviews*, *35*(3), e00150-21. https://doi.org/10.1128/cmr.00150-21

Poey, M. E., de los Santos, E., Aznarez, D., García-Laviña, C. X., & Laviña, M. (2024). Genetics of resistance to trimethoprim in cotrimoxazole resistant uropathogenic Escherichia coli: Integrons, transposons, and single gene cassettes. *Frontiers in Microbiology*, *15*. https://doi.org/10.3389/fmicb.2024.1395953

Qabel, A., Ennadir, S., Nikolentzos, G., Lutzeyer, J. F., Chatzianastasis, M., Boström, H., & Vazirgiannis, M. (2023). *Advancing Antibiotic Resistance Classification with Deep Learning Using Protein Sequence and Structure* (p. 2022.10.06.511103). bioRxiv. https://doi.org/10.1101/2022.10.06.511103

Rannon, E., Shaashua, S., & Burstein, D. (2025). DRAMMA: A multifaceted machine learning approach for novel antimicrobial resistance gene detection in metagenomic data. *Microbiome*, *13*(1), 67. https://doi.org/10.1186/s40168-025-02055-4

Razavi, M., Marathe, N. P., Gillings, M. R., Flach, C.-F., Kristiansson, E., & Joakim Larsson, D. G. (2017). Discovery of the fourth mobile sulfonamide resistance gene. *Microbiome*, *5*(1), 160. https://doi.org/10.1186/s40168-017-0379-y

Rodrigo, M. K. D., Saiganesh, A., Hayes, A. J., Wilson, A. M., Anstey, J., Pickering, J. L., Iwasaki, J., Hillas, J., Winslow, S., Woodman, T., Nitschke, P., Lacey, J. A., Breese, K. J., van der Linden, M. P. G., Giffard, P. M., Tong, S. Y. C., Gray, N., Stubbs, K. A., Carapetis, J. R., … Barnett, T. C. (2022). Host-dependent resistance of Group A Streptococcus to sulfamethoxazole mediated by a horizontally-acquired reduced folate transporter. *Nature Communications*, *13*(1), 6557. https://doi.org/10.1038/s41467-022-34243-3

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

Ruppé, E., Ghozlane, A., Tap, J., Pons, N., Alvarez, A.-S., Maziers, N., Cuesta, T., Hernando-Amado, S., Clares, I., Martínez, J. L., Coque, T. M., Baquero, F., Lanza, V. F., Máiz, L., Goulenok, T., de Lastours, V., Amor, N., Fantin, B., Wieder, I., … Ehrlich, S. D. (2019). Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature Microbiology*, *4*(1), Article 1. https://doi.org/10.1038/s41564-018-0292-6

Salyers, A. A., Gupta, A., & Wang, Y. (2004). Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends in Microbiology*, *12*(9), 412–416. https://doi.org/10.1016/j.tim.2004.07.004

Schoning, C. (2014, October 16). *Why do transformed TOP10 grow on amp plates and not in amp LB fluid?* [Online post]. https://www.researchgate.net/post/Why-do-transformed-TOP10-grow-on-amp-plates-and-not-in-amp-LB-fluid

Sethuvel, D. P. M., Bakthavatchalam, Y. D., Karthik, M., Irulappan, M., Shrivastava, R., Periasamy, H., & Veeraraghavan, B. (2023). β-Lactam Resistance in ESKAPE Pathogens Mediated Through Modifications in Penicillin-Binding Proteins: An Overview. *Infectious Diseases and Therapy*, *12*(3), 829–841. https://doi.org/10.1007/s40121-023-00771-8

Shin, H. J., Eom, D.-H., & Kim, S.-S. (2005). One-class support vector machines—An application in machine fault detection and classification. *Computers & Industrial Engineering*, *48*(2), 395–408. https://doi.org/10.1016/j.cie.2005.01.009

Shin, H. W., Lim, J., Kim, S., Kim, J., Kwon, G. C., & Koo, S. H. (2015). Characterization of trimethoprim-sulfamethoxazole resistance genes and their relatedness to class 1 integron and insertion sequence common region in gram-negative bacilli. *Journal of*

Microbiology and Biotechnology, *25*(1), 137–142. https://doi.org/10.4014/jmb.1409.09041

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Soo, V. W. C., Hanson-Manful, P., & Patrick, W. M. (2011). Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. *Proceedings of the National Academy of Sciences*, *108*(4), 1484–1489. https://doi.org/10.1073/pnas.1012108108

Tamer, Y. T., Gaszek, I. K., Abdizadeh, H., Batur, T. A., Reynolds, K. A., Atilgan, A. R., Atilgan, C., & Toprak, E. (2019). High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection. *Molecular Biology and Evolution*, *36*(7), 1533–1550. https://doi.org/10.1093/molbev/msz086

Tewawong, N., Kowaboot, S., Lektrakul, W., Supcharoengoon, U., Watanagul, N., & Pitaksajjakul, P. (2025). Mechanisms of fluoroquinolone resistance among Escherichia coli isolates from urinary tract infections in Thailand. *PLOS One*, *20*(5), e0325175. https://doi.org/10.1371/journal.pone.0325175

The UniProt Consortium. (2025). UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, *53*(D1), D609–D617. https://doi.org/10.1093/nar/gkae1010

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, *42*(2), 243–246. https://doi.org/10.1038/s41587-023-01773-0

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., … Velankar, S. (2024). AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, *52*(D1), D368–D375. https://doi.org/10.1093/nar/gkad1011

Venkatesan, M., Fruci, M., Verellen, L. A., Skarina, T., Mesa, N., Flick, R., Pham, C., Mahadevan, R., Stogios, P. J., & Savchenko, A. (2023). Molecular mechanism of plasmid-borne resistance to sulfonamide antibiotics. *Nature Communications*, *14*, 4031. https://doi.org/10.1038/s41467-023-39778-7

Wiegand, I., Hilpert, K., & Hancock, R. E. W. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nature Protocols*, *3*(2), 163–175. https://doi.org/10.1038/nprot.2007.521

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022). *High-resolution de novo structure prediction from primary sequence* (p. 2022.07.21.500999). bioRxiv. https://doi.org/10.1101/2022.07.21.500999

Yagimoto, K., Hosoda, S., Sato, M., & Hamada, M. (2024). Prediction of antibiotic resistance mechanisms using a protein language model. *Bioinformatics*, *40*(10), btae550. https://doi.org/10.1093/bioinformatics/btae550

Yakovleva, G. M., Kim, S.-K., & Wanner, B. L. (1998). Phosphate-independent expression of the carbon-phosphorus lyase activity of Escherichia coli. *Applied Microbiology and Biotechnology*, *49*(5), 573–578. https://doi.org/10.1007/s002530051215

Yin, X., Jiang, X.-T., Chai, B., Li, L., Yang, Y., Cole, J. R., Tiedje, J. M., & Zhang, T. (2018). ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*, *34*(13), 2263–2270. https://doi.org/10.1093/bioinformatics/bty053

Yong, D., Toleman, M. A., Giske, C. G., Cho, H. S., Sundman, K., Lee, K., & Walsh, T. R. (2009). Characterization of a New Metallo-β-Lactamase Gene, blaNDM-1, and a Novel Erythromycin Esterase Gene Carried on a Unique Genetic Structure in Klebsiella pneumoniae Sequence Type 14 from India. *Antimicrobial Agents and Chemotherapy*, *53*(12), 5046–5054. https://doi.org/10.1128/AAC.00774-09

Zapun, A., Contreras-Martel, C., & Vernet, T. (2008). Penicillin-binding proteins and β-lactam resistance. *FEMS Microbiology Reviews*, *32*(2), 361–385. https://doi.org/10.1111/j.1574-6976.2007.00095.x

Zeba, B., Luca, F. D., Dubus, A., Delmarcelle, M., Simporé, J., Nacoulma, O. G., Rossolini, G. M., Frère, J.-M., & Docquier, J.-D. (2009). IND-6, a Highly Divergent IND-Type Metallo-β-Lactamase from Chryseobacterium indologenes Strain 597 Isolated in Burkina Faso. *Antimicrobial Agents and Chemotherapy*. https://doi.org/10.1128/aac.01607-08

Zhang, A.-N., Gaston, J. M., Dai, C. L., Zhao, S., Poyet, M., Groussin, M., Yin, X., Li, L.-G., van Loosdrecht, M. C. M., Topp, E., Gillings, M. R., Hanage, W. P., Tiedje, J. M., Moniz, K., Alm, E. J., & Zhang, T. (2021). An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-25096-3

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

Zhao, W.-H., & Hu, Z.-Q. (2012). Acinetobacter: A potential reservoir and dispenser for β-lactamases. *Critical Reviews in Microbiology*, *38*(1), 30–51. https://doi.org/10.3109/1040841X.2011.621064