DeepSEA: an alignment-free deep learning tool for functional annotation of antimicrobial resistance proteins

Tiago Cabral Borelli[a,b,c], Alexandre Rossi Paschoal[d,e], Ricardo Roberto da Silva[a,b]

[a] Computational Chemical Biology Laboratory, Department of BioMolecular Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto 14040-900, Brazil

[b] NPPNS, Department of BioMolecular Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, 14040-900, Brazil

[c] Cellular and Molecular Biology Program, Department of Cellular and Molecular Biology of Ribeirão Preto, School of Medicine, University of São Paulo, Ribeirão Preto, 14049-900, Brazil

[d] Bioinformatics and Pattern Recognition Group (Bioinfo-CP), Department of Computer Science, The Federal University of Technology – Paraná (UTFPR), Cornélio Procópio, Brazil

[e] Biological Data Science Group, Artificial Intelligence and Informatics Team, The Rosalind Franklin Institute, Harwell Science and Innovation Campus, Didcot OX11 0QS, UK

Abstract

Surveying antimicrobial resistance (AMR) is essential to track its evolution and spread. Alignment-based annotation tools use strict identity (>80%) cutoffs to distinguish between non-resistant (NRP) and resistant proteins (ARP) only annotating proteins similar to those in their databases. Deep learning and Hidden Markov Models (HMM) based tools also depend on protein alignment at some level. DeepARG filters input data to select the um SNP ARG-like proteins and HMMs are built on multi-sequence alignment (MSA) specific for the protein in a given family or group. Therefore, there is a need to remove the alignment dependency of AMR annotation tools to identify proteins with remote homology Here we present DeepSEA, an alignment-free tool fitted on antimicrobial-resistant sets of aligned and unaligned ARPs and NRP. DeepSEA outperforms the current multi-class AMR classifiers DeepARG, RGI and AMRfinder. Furthermore, DeepSEA trained weights cluster AMR by resistant mechanisms, indicating that the model's latent variables successfully captured distinguishing features of antibiotic resistance. Our tool annotated functionally validated tetracycline destructases (TDases) and confirmed the identification of a novel TDase found by HMM.

Availability and Implementation: DeepSEA is an open-source command line tool implemented in Python and TensorFlow and it is available at GitHub: https://github.com/computational-chemical-biology/DeepSEA-project.

Contact: Tiago Cabral Borelli, tiago.borelli@usp.br
Keywords: Deep learning, antimicrobial resistance, proteins alignment, genomic

Introduction

Penicillin discovery (1) enabled the treatment of harsh bacterial infections such as syphilis (2) and ensured the safety of patients through invasive or immunosuppressive procedures, which promoted a substantial increase in human life expectancy (2). However, producing and resisting antimicrobials are competitive mechanisms in microorganisms' evolution (3). Therefore, the overuse and misuse of these drugs have been selecting bacterial variants with genotypes that confer resistance and initiated a *Red Queen* (4) race named *The Antibiotic Crisis* (5–7).

In 2019, AMR directly caused 1.27 million deaths, and it was related to 4.95 million in total, causing more health burdens than HIV and Malaria (8) and has become one of the most dangerous threats in modern medicine (9). Although projecting deaths by AMR is an open debate guided by statistical inconsistencies (10,11), there is a need for improving surveillance strategies, drug discovery strategies, and infection treatment protocols. Also, to promote AMR and antimicrobial usage awareness, AMR needs to be approached from a One (12)Health perspective (13).

Together, whole (meta)genome sequencing (WGS) and sequence alignment enabled the screening of resistant determinants by antibiotic sustainability test (AST) (14). As a consequence, databases such as the Comprehensive Antibiotic Resistance Database (CARD) (15) were built upon experimentally validated antimicrobial resistance genes (ARG) or proteins, and alignment-based tools such as Resistance Genes Identifier (RGI), AMRFinderPlus (16), and Resfinder (17) were developed. Nevertheless, inferring homology and function from the excess of identity requires a cut-off since two homologous sequences are more similar than would be expected by chance (12). To annotate AMR proteins/genes alignment-based tools use strict identity values from 80% (17) to 90% (18), considering coverage at 100%, which limits the discovery of novel and less similar resistance determinants (19).

Deep learning tools emerged to reduce the alignment dependency due to ANN generalization capability. The pioneer DeepARG (19) was fitted on dissimilarity matrices of ARPs from 33 AMR classes and showed a lower false negative rate (recall = 0.9). ARG-SHINE (20), is a three-element ensemble tool composed of a Convolutional Neural Network (CNN), a K-Nearest Neighbor (KNN), and InterproScan (21) that combines features from protein sequences, BLASTP bit-score, and protein family domains to annotate ARPs. HMD-ARG (Hierarchical Multi-task Deep learning framework) (22) used multi-task learning CNNs to fit a three-step model that can predict if a protein is an APR, classify the positive ones into one of its 15 APR classes in its database, and predict if the APR came from horizontal gene transfer (HGT).

Large language models (LLM), such as ProteinBert (23), were adapted to proteomics to embed protein semantics in a high-dimensional space, since protein language models (pLM) learn text context. However, their main disadvantage is the need for huge amounts of data to produce the embedding matrices. Despite these advances, AMR annotation tools depend on protein alignment at some point in their pipelines. For instance, since DeepARG was fitted on dissimilarity matrices, querying data must be encoded as a vector containing the highest identity values to all DeepARG resistance protein classes. Therefore, the model is classifying alignment data. Also, since DeepARG was not fitted on proteins with functions

other than antimicrobial resistance, it uses Diamond to pre-filter ARP-like sequences before annotation, putting the tool under the same limitations as sequence alignment-based tools.

HMD-ARG used ARPs and non-resistant proteins (NRP) to train their models. However, the tool is available only on the webserver and with closed code, making it difficult to benchmark and limiting its use for screening large datasets. Also, HMD-ARG did not test the effect identity between ARPs and NRPs on the neural networks.

In this work, we present DeepSEA, an alignment-free AMR classifier for proteins. To construct DeepSEA, we fitted CNN and Recurrent Neural Networks (RNNs) to extract information directly from protein sequences. Our approach annotated functional proteins from less representative and broader AMR classes, such as tetracycline and multidrug (MDR), respectively. Fitting ANNs on NRP and ARPs refrain from assistant aligners to perform classification and enable screening of whole coding sequences from genomes. We demonstrated that our models outperform the "State of Art" (SoA) tools in an independent dataset composed of annotated functionally validated remote homologous APRs found through Hidden Markov Models (HMM).

Algorithm

Our neural network architectures were constructed using TensorFlow 2 (24). They consist of (i) two preprocessing layers: a text vectorization layer that encodes proteins (text) into integer-containing vectors with tokens = {1,20} for the canonical amino acids, and 0 for the padding value. An embedding layer that converts the integer vector into embedding matrices, (ii) feature extraction layers that are used in a mutually exclusive way: two Gated Recurrent Unit (GRU) layers or four 1D Convolutional (Conv1D) layers with ReLU as activation function. For Conv1D, we also tested the effect of gradually reducing the number of kernels used for convolution. And (iii) a Global Average Pooling layer followed by a Fully Connected layer with Softmax as an activation function. One hot encoded vector represented the antimicrobial-resistant protein classes to avoid ordinal biases that could result from a simple integer encoding. We used Adam as a stochastic gradient descent method with a learning rate of 0.001, Categorical Crossentropy as a loss function, and Categorical Accuracy as metrics to evaluate during the training process and hyperparameter selection. All models were evaluated using precision, recall, and F1-Score as metrics on the test set.

Implementation

We chose two databases that concatenated data from well-curated AMR sources using different strategies and retrieved ARP classes with at least a few hundred sequences. The HMD-ARG-DB (22) is a union of a series of AMR databases (See Methods) that were only filtered to eliminate duplicates (sequences with 100% amino acid identity). It originally contains data from 15 ARP families and totalizes 17828 examples. The Non-Redundant Comprehensive Database (NCRD) (25) also gathered proteins from AMR databases. However, the authors added an extra step to increase the database size: by aligning NCRD with the Protein Data Bank (PDB) and Non-Redundant Database (NR) they found AMR homologous sequences. They obtained 34,004 proteins from 29 AMR families for the version with 95% sequence identity at most (NCRD95).

To test the effect of NRPs we retrieved all 334,020 bacterial proteins non-related to AMR from the UniProt/Swiss-Prot end segmented into four different ways (See System and methods). Aligned: NRPs are similar to ARPs at some level, unaligned are NRPs with no alignment to ARPs but in similar NRP/ARP proportion, unaligned (50%) and unaligned 100% that contain respectively half and all the UniProt/Swiss-Pro.

Model evaluation on hold-out test set

We used three different ANN architectures to annotate APRs. They are (i) a CNN with a descending number of convolutional filters among its convolutional layers, (ii) CNN-S with a constant number of convolutional filters, and (iii) a gated recurrent unit (GRU) neural network. To set the hyperparameter values (SuppTable 1)**,** our group of models was fitted only on the most abundant classes of APRs from HMD (Fig. 2A) and NCRD95 (Fig 3A). Then, the NRPs from UniProt/Swiss-Prot were introduced in various amounts and identity degrees (SuppTables 2 and 3).

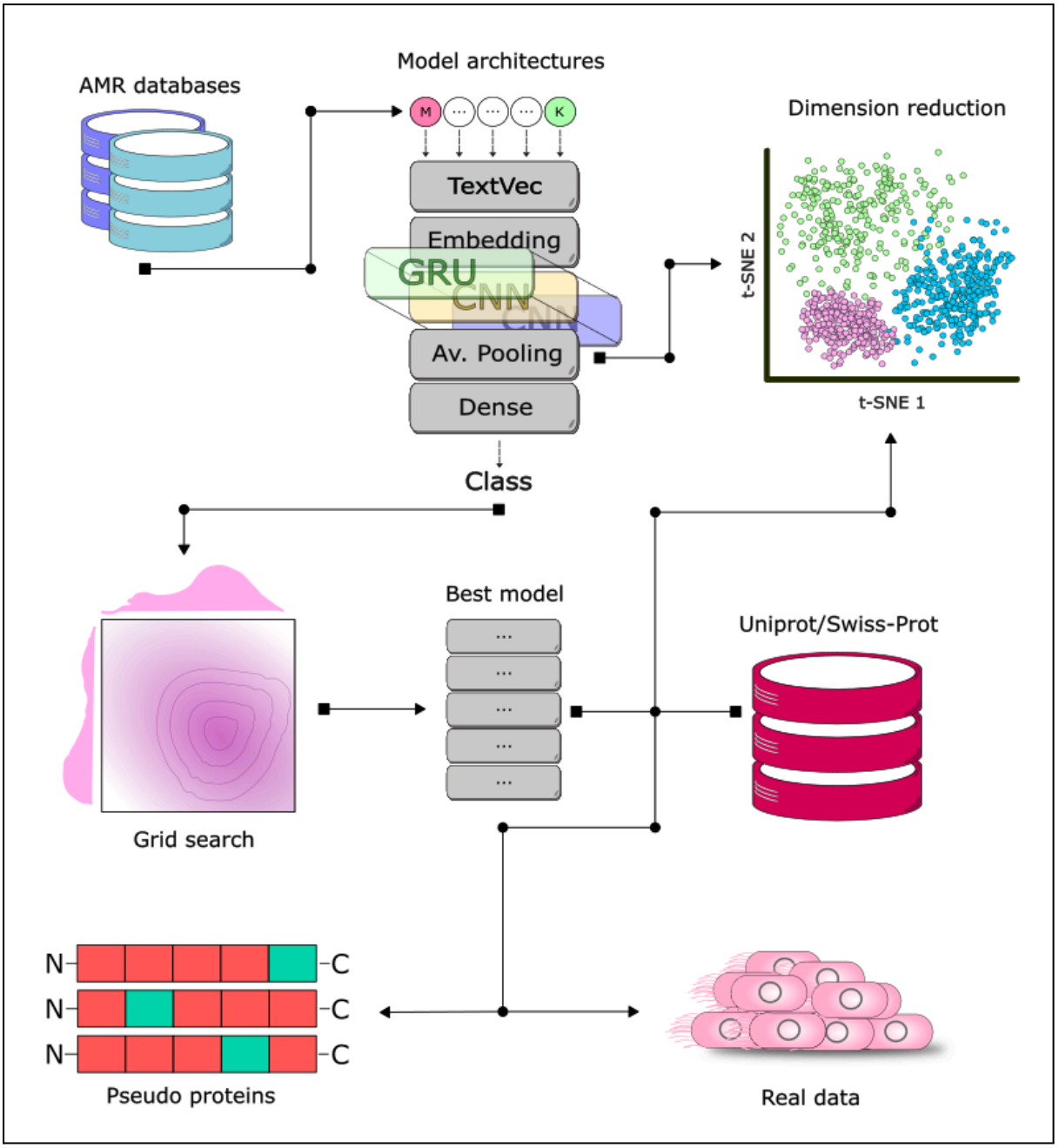


Figure 1. Overall DeepsSEA construction process. We used resistant proteins from HMD and NCRD95 databases to train and test three base architectures which vary only in the type of the third layer. The fourth layer (average pooling) summarizes the upstream matrix information. It passes to a fully connected layer that outputs a probability distribution where the highest value corresponds to the most probable AMR class. ARP set was used to hypeparams' definition, while the introduction of the NRP set aimed to force the model to classify AMR protein classes correctly. To evaluate model performance, the best models were applied to real and pseudo data.

When fitted on both databases, DeepSEA models showed high accuracy (>0.9). However, we measured precision, recall, and f1-score for each AMR class of ARP to avoid accuracy bias toward the majority class (26,27). For only-APRs from the HMD dataset, our ANNs presented values above 0.8 for precision, recall, and f1-score for all AMR classes (Fig 2 B), and the addition of aligned NPR impaired those metrics for the MDR class. However, unaligned NPRs (SuppTables 2) recovered performance for the MDR protein class (Fig 2C), although the increasing NPR set size harmed GRU architecture. ANNs on only ARPs (Fig

3B) from NCRD95 also yielded high metric values except for the CNN recall. GRU had variations on recall when fitted on ARP-NPR aligned (Fig 3C) and gradually increasing the NRP set size (Figs 3D, 3E, and 3F) affected the GRU performance for ARPs. The presence of 100% of unaligned NRPs exploded the gradients during model convergence (Fig 3F), precluding the calculations of f1-score, recall, and precision from ARPs.

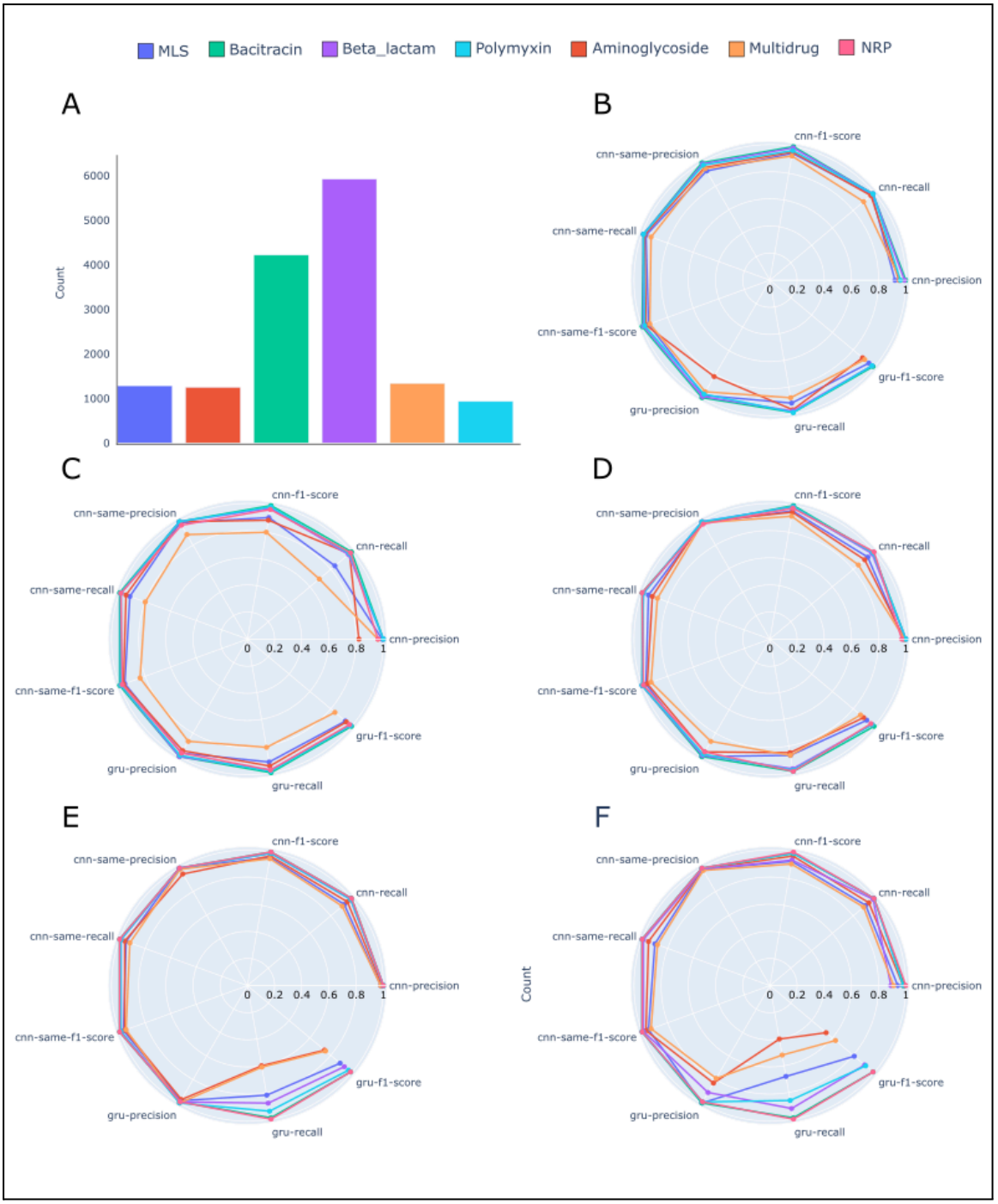


Figure 2. HMD database distribution and model metrics. (A) Number APR for each class in the training set. Metrics for ANN fitted on: (B) only APRs, (C)ARP-NPR-aligned proteins, (D) ARP-NPR-unaligned, (E) ARP and 50% of unaligned NPR proteins and (F) APR and 100% of unaligned NPR proteins.

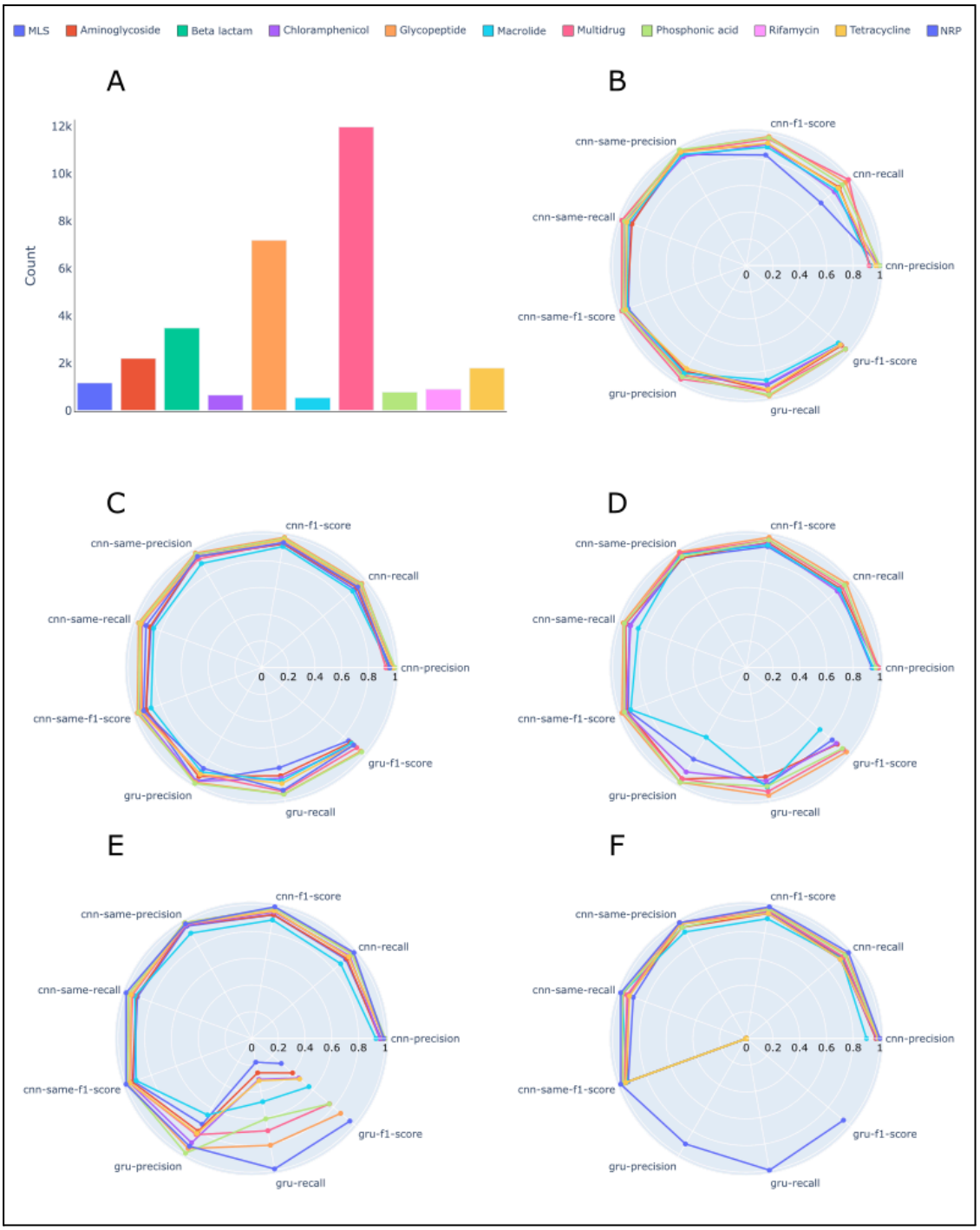


Figure 3. NCRD95 database distribution and model metrics. Metrics for ANN fitted on: (B) only APRs, (C)ARP-NPR-aligned proteins, (D)ARP-NPR-unaligned, (E) ARP and 50% of unaligned NPR proteins and (F) APR and 100% of unaligned NPR proteins

Investigating feature extraction in a limited-data scenario

After comparing model architectures and hyperparameter optimization (Supp Table 1), we investigated the patterns obtained by extracting the embedding matrices from the average pooling layer and applying t-distributed Stochastic Neighbor Embedding (t-SNE) for dimension reduction before and after training. Often pLMs embeddings are evaluated in low dimensional space on their ability to cluster amino acid biochemical features, protein subcellular localization, and protein families (28,29). However, the AMR data available is relatively limited and highly redundant since databases are often built by concatenating pre-existing ones (15,17,30,31). Thus we decided to insert the global average pooling layer after the Conv1D and GRU layers of all ANN to average all the feature extraction done by the convolutions layers or GRU units.

The three architectures clusterized ARPs (Fig 4A) and separated ARPs from NRP (Figs 4B and 4C), indicating the extraction of mutually exclusive features to annotate AMR. The CNN-Same model fitted on only ARPs (Fig 4A) better clusterized MDR proteins. When aligned NRP were introduced (Fig 4B) in training, CNN and GRU did not clearly define borders between MDR and NRP, and yet CNN-Same presented a better cluster definition, although misclassified NRP as MDRs due to recall decrease (Fig 2C). On the other hand, unaligned NRPs improved CNN separation between NRP and MLS, MRD, and aminoglycosides and CNN-Same completely split MLS, MRD, and aminoglycosides from each other and from NRP. t-SNEs of models fitted on datasets containing 50% (Supp. Fig 2A) and 100% (Supp. Fig 2B) agrees with data plotted in Figures 2E and 2F on the performance impairment of GRU architectures.

For NCRD95 data, the unfitted t-SNEs are more shuffled than the HMD's and contain more AMR classes (n = 10) than the counterpart (n=6) even though DeepSEA models could establish clear borders for them despite the lack of balance. Without NRPs (Fig 5A), all architectures formed well-defined clusters for all AMR classes. A cluster (red-dashed) formed by macrolide and MLS classes that share resistances to macrolide was formed by CNNs and GRU.

Fitting on NPR-APR-aligned (Fig 5B) resulted in misclassification mainly between NRP and MDR while the remaining classes were kept together. Unaligned NRP (Fig 5C) reduced the number of NRP misclassified as MDR for CNNs and kept macrolides near MLS proteins, nevertheless leading to poor GRU performance (Fig 3D). The same was observed in models fitted on half and a complete amount of NPR (Supp Fig 3A e 3B).

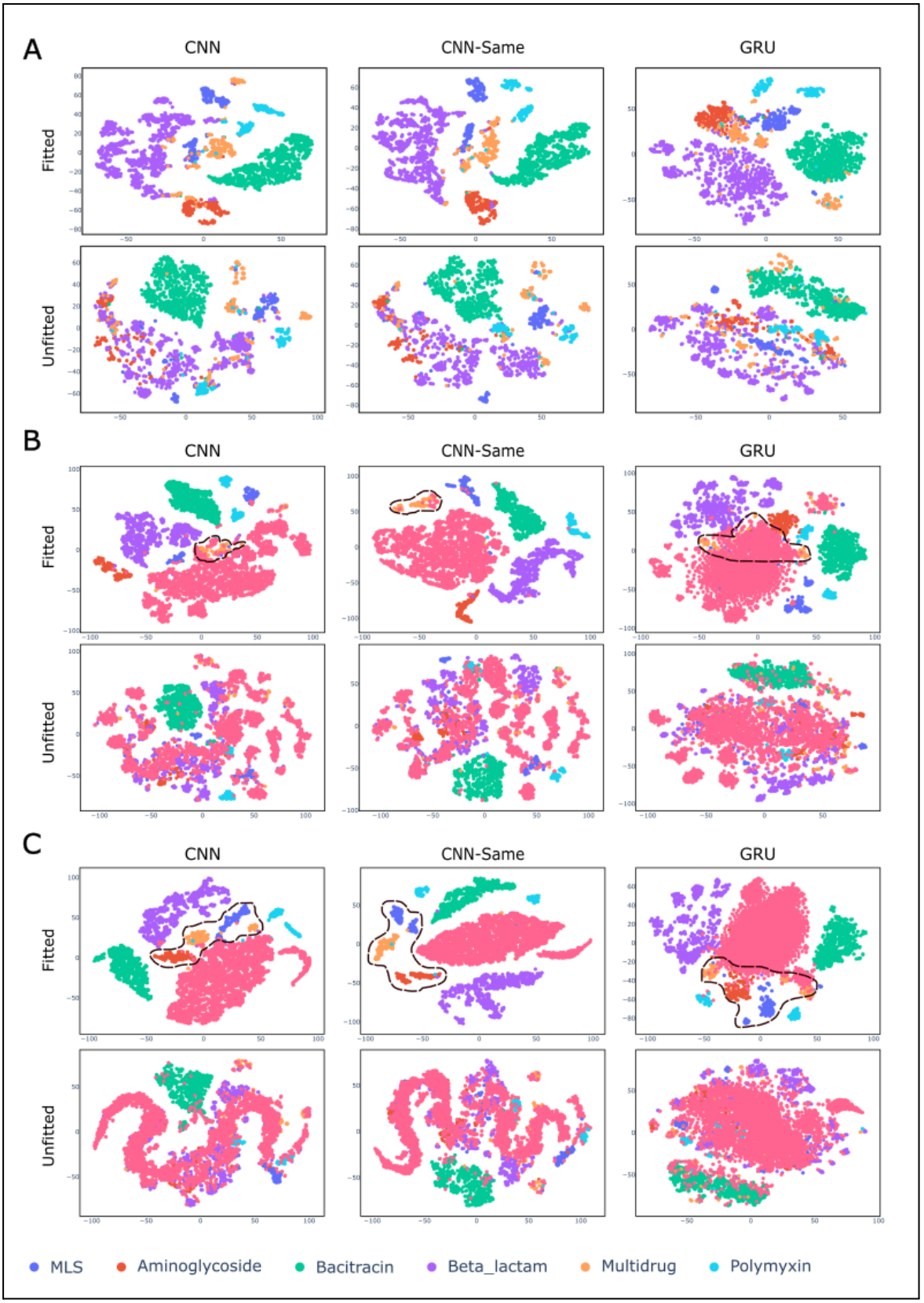


Figure 4. Visual comparison between classification boundaries before and after fitting DeepSEA models on HMD database. The average pooling layer outputs passed through t-SNEs dimension reduction with 1000 iterations and perplexity = 30. The figure contains results from: (A) fitting only on APRS, (B) ARP-NPR-aligned proteins, and (C) ARP-NPR-unaligned. The dashed area highlights the AMR classes that were better separated between CNN and CNN-Same.

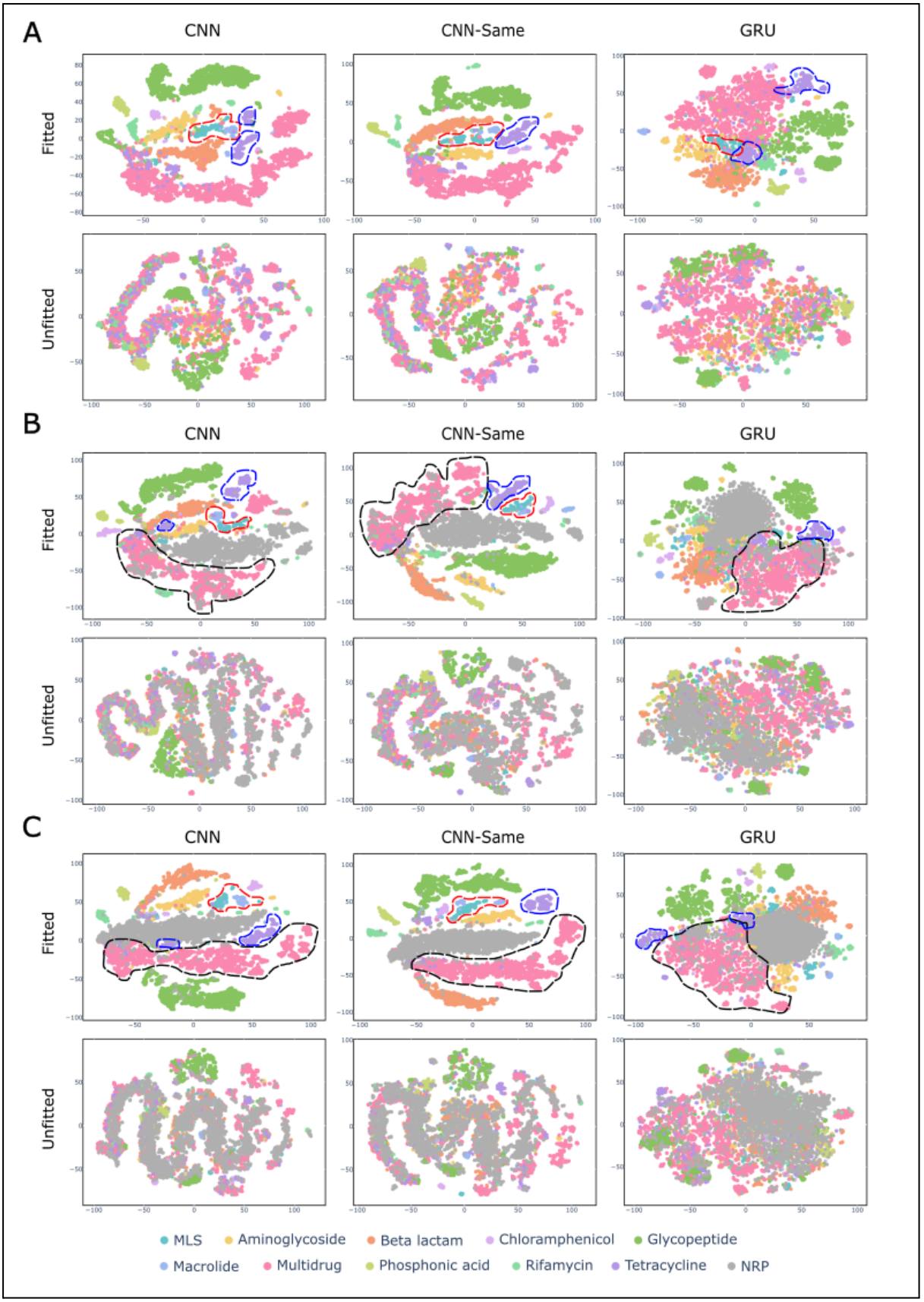


Figure 5. Visual comparison between classification boundaries before and after fitting DeepSEA models on the NCRD database. The average pooling layer outputs passed through t-SNEs dimension reduction with 1000 iterations and perplexity = 30. The figure contains results from: (A) fitting only on APRS, (B)

ARP-NPR-aligned proteins, and (C) ARP-NPR-unaligned. The dashed area highlights the AMR classes that were better separated between CNN and CNN-Same. The dark, red, and blue dashed areas highlighted clusters of MDRs with NPRs, MLS and macrolides and tetracycline, respectively.

Since macrolide and MLS are grouped, we hypothesized that ANNs learned features other than AMR classes. t-SNEs colored according to resistance mechanisms (Fig 6) revealed larger groups of a single mechanism as MDR-efflux pumps or glycopeptides-antibiotic target alteration. Interestingly, there is a cluster of APRs that share the "antibiotic (antimicrobial) inactivation" mechanism and may contain chloramphenicol, beta-lactam, rifamycin, phosphonic acid (fosfomycin), and tetracycline classes even though its composition depends on the ANN architectures and NRP-ARP relationship.

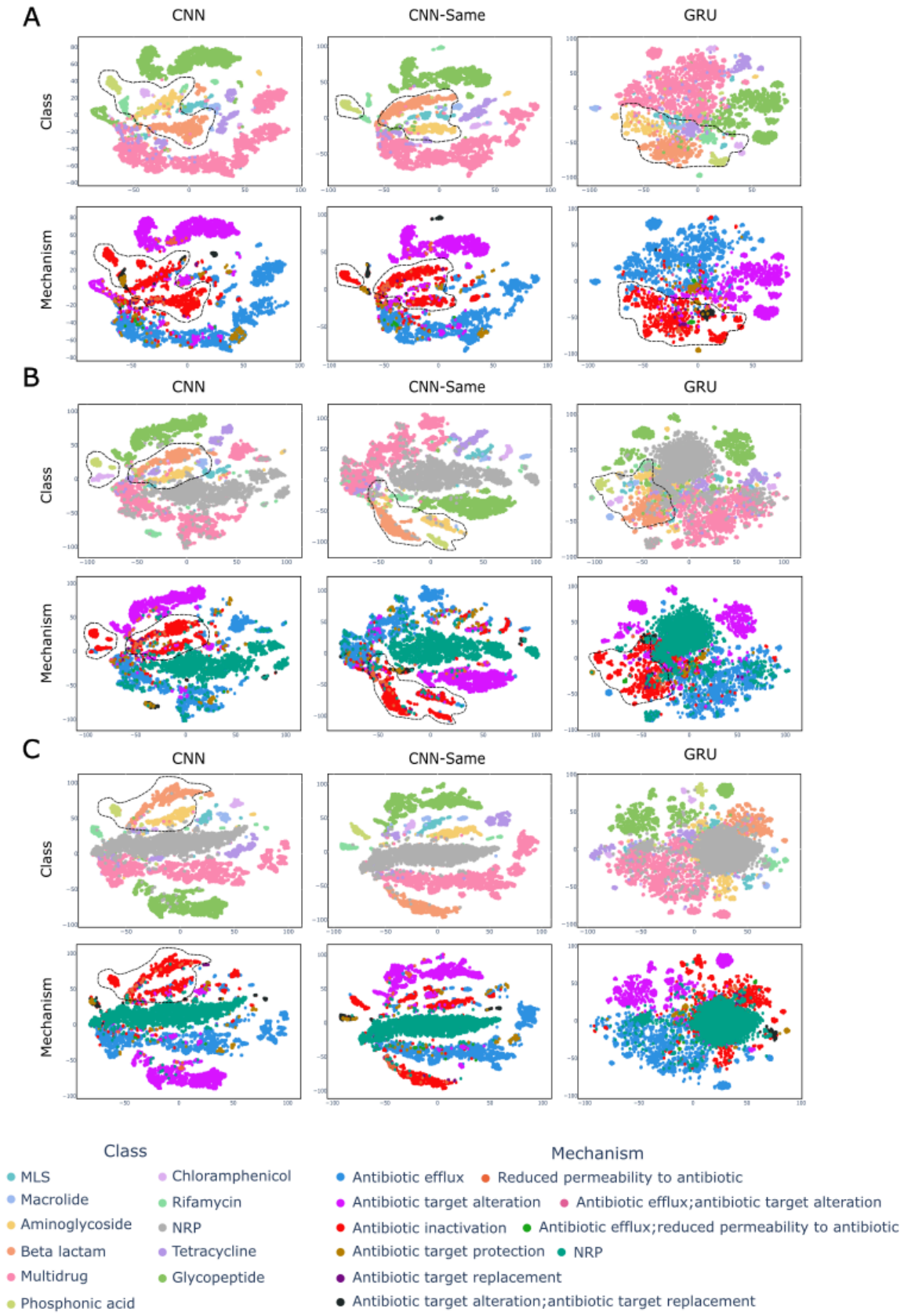


Figure 6. Comparison between cluster colored by AMR class and mechanism of ANN fitted on NCRD95 dataset. The figure contains results from: (A) fitting only on APRS, (B) ARP-NPR-aligned proteins, and (C)

ARP-NPR-unaligned. The dark-dashed area highlights a cluster of AMR proteins that hydrolyzate antimicrobial molecules.

Testing in simulated data and additional real data

To demonstrate that ANN extracted global and sequential patterns from our proteins, a set of 1000 PseudoARPs was generated, where each one is composed of ten shuffled 50-residue long kmers randomly selected, nine of them from NRPs and one from ARPs, similar to the DeepARG's method (19). The main difference is the use of NRPs to construct the PseudoARPs dataset. A comparison of the performance between ANNs fitted with and without NRPs (Fig 7A and Fig 7B) showed the false positive rates dropped from 98.6% to 1.1% and from 95.4% to 2.7% for CNNs fitted with the NCRD95 and HMD datasets, respectively (Table 1).

Table 4. False positive rates from pseudo-ARP dataset

	Only ARPs*		**Aligned.**		**Unaligned**		**Unalig. 50%**		**Unilig. 100%**	
	HMD	**NCRD**	**HMD**	**NCDR**	**HMD**	**NCRD**	**HMD**	**NCRD**	**HMD**	**NCRD**
CNN	91.1%	98.6%	4.8%	2.9%	5.9%	**1.1%**	17.8%	11.1%	30.2%	9.9%
CNN-SAME	95.4%	96.6%	18.9%	12.4%	16%	4.9%	25%	13.2%	20%	21.2%
GRU	82.1%	**69.5%**	2.4%	**1.4%**	5%	3.1%	**5.7%**	54.2%	**8%**	-

*Since the models fitted with Only ARPs data do not have non-resistant protein, we used a cut-off of 0.5 in the highest value from softmax output to distinguish between real and false results.

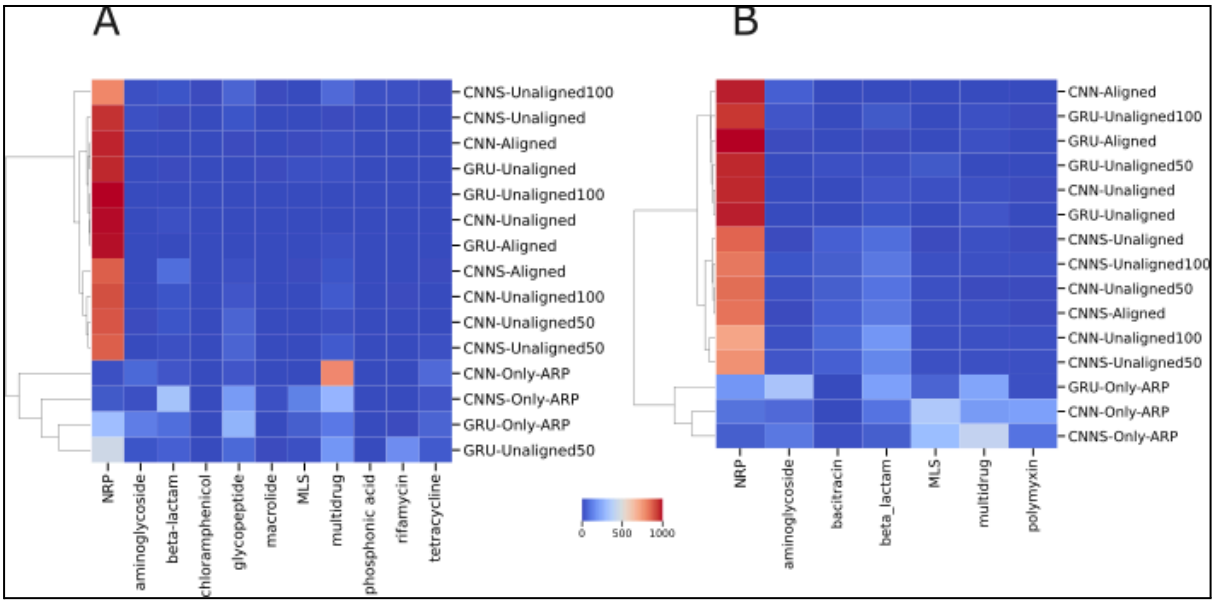


Figure 7. Comparison of performance on pseudo data among ANN architectures and dataset compositions. The row clusters were calculated using the Euclidian Distance of the number of respective AMC classes assigned by each ANN-dataset combination. NRP class of ANNs fitted on only APRs are sequences with probability < 0.5. The color scale represents the number of simulated synthetic proteins.

The bacterial genome does not encode only? ARPs, and, as the evaluation on Pseudo ARPs showed that fitting only on APRs results in high false positive rates, we hypothesized that ANN fitted only with ARPs data would fail to correctly classify a whole coding sequence data. To test it, our models were evaluated on proteins from reference genomes of the SKAPE-pathogen group: *Escherichia coli* (ASM584v2 and ASM886v2), *Klebsiella pneumoniae* (ASM24018v2), *Pseudomonas aeruginosa* (ASM676v1) *Acinetobacter Baumannii* (ASM863263v1), and the BRA006, a genome from our previous work closely related to *Micromonospora aurantiaca*. The results in Fig. 8A confirmed the positive effect of NRPs on CNNs training as they can correctly classify the majority of the proteome as NRP and the training on the NRP-ARP dataset increased the number of NRPs for all regardless of species (Fig. 8B). Also, DeepSEA demonstrated weak sequence length bias regardless of the species (Fig 8C).

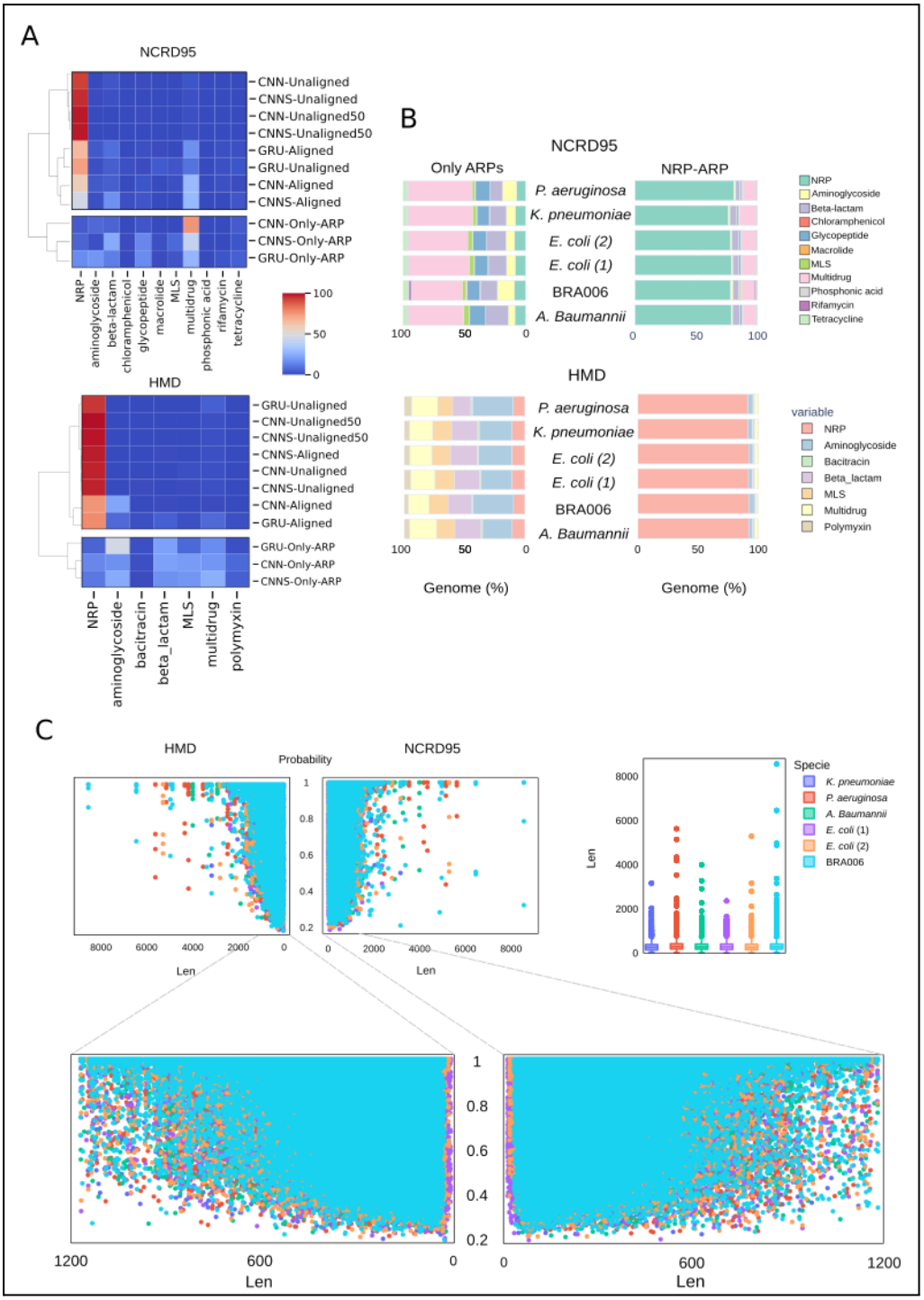


Figure 8. Application of DeepSEA models on proteins from SKAPE pathogens. A) The row clusters were calculated using the Euclidian Distance of the number of respective AMC classes assigned by each ANN-dataset combination. NRP class of ANNs fitted on only APRs are sequences with probability < 0.5. The color scale represents the number of proteins expressed in percentage. The models GRU-Unaligend50 and GRU-Unaliged100 were removed due to underfitting (See Supp. Fig. 1) B) Sum of the results of all models plotted per SKAPE pathogen. C) Correlation between probability and protein length. We highlight the length interval where the measurements are concentrated according to the distributions on boxplots.

Benchmarking and evaluation on independent data

To put DeepSEA in perspective we compared our model's performance against other "State of Art" (SoA) AMR annotation tools: the alignment-based AMRFinderPlus and RGI and the deep-learning-based DeepARG. DeepSEA area under the curve (AUC) scores from our test sets' ROC curves (Fig 9) outperformed DeepARG, AMRFinder, and RGI regardless of the ARP-NRP alignment. Also, The concurrent tools presented the lowest values of AUC scores when applied on ARP-NRP aligned, especially for the NCRD95 dataset, due to higher false positive rates.

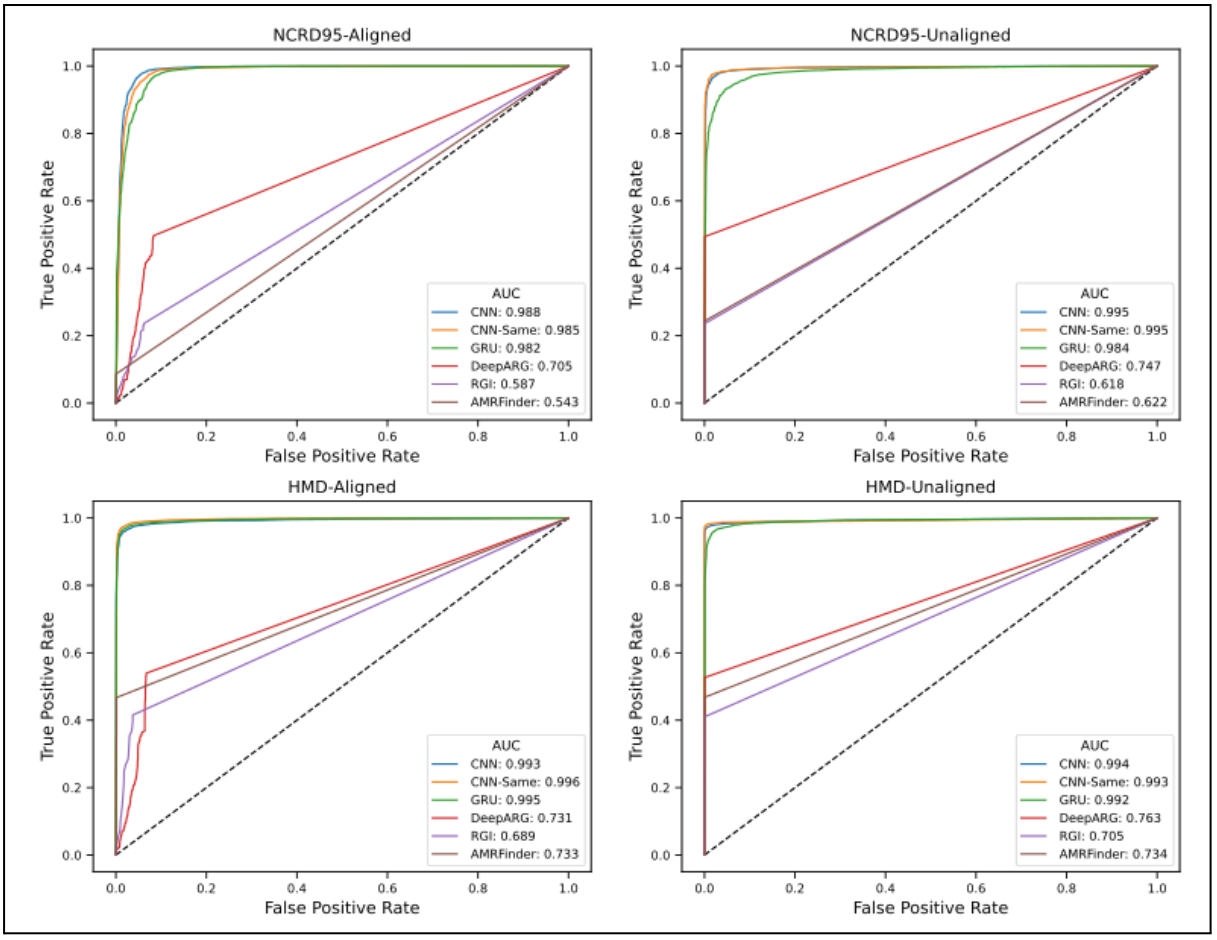


Figure 9. Benchmarking between DeepSEA and other AMR annotation tools. To compare the output scores all tools were tested with their respective default parameters. The result data was scaled as follows to obtain values between 0 and 1: DeepARG probabilities were kept from 0 to 1, AMRFinder's identity percentages were divided by 100. For RGI, we used a min-max normalization of the ratio of Past_Bitscore to Best_Hit_Bitscore. Since DeepSEA models have an NRP class, we apply the Complement Rule from the probability theory: $P(\overline{A}) = 1 - P(A)$, where a denotes the event of being assigned to any NRP and ac denotes the event of not being assigned to any NRP.

To compare DeepSea with the SoA tools in an external validation dataset, we retrieved protein sequences from the work of Blake et al., 2024 (26). Through a two-step HMM approach it was found 50 tetracycline destructase (TDase) enzymes of which 28% (14/50) increased the minimal inhibitory concentration (MIC) when cloned into a susceptible

E. coli host. DeepSEA models outperformed DeepARG in all measured metrics (Fig 10B). CNN-Unaligned achieved higher values of accuracy than RGI and AMRFinder. For precision, DeepARG models outperformed AMRFinder and CNN-Unaligned had a similar value to RGI. CNN-Unaligned also outperformed all other tools at F1-Score. Finally CNNS-Aligned achieved the same performance as AMRFinder and outperformed RGI for recall.

Considering functional TDases, DeepARG and RGI annotated 28% (4/14) and 64% (9/14) of the enzymes, respectively. AMRFinder and CNNS-Aligned annotated 92% (13/14), CNN-Aligned and CNN-Unaligned annotated 85% (12/14), and CNNS-Aligned, GRU-Aligned and GRU-Unaligned 78% (11/14). Only CNNS-Aligned and GRU-Aligned confirmed the identification of a novel functional TDase (TetHMM60) from Blake *et al.* (32) data despite the low identity (23.7%, with e-value = $2x10^{-14}$,bit-score = 72.8) between TetHMM60 and one TetX (SuppTable 3).

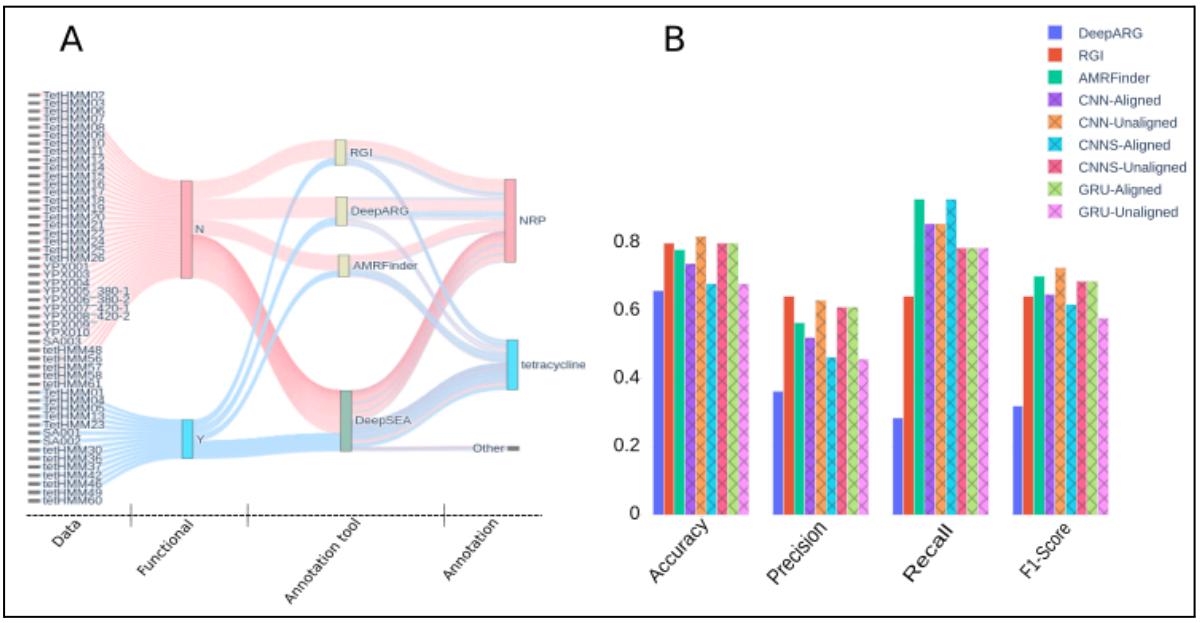


Figure 10. Evaluation of DeepSEA and other AMR annotation tools on TDases. Comparison between DeepSEA models and other tools on independent dataset with functional validation. A) AMR annotation tools' ability to identify functionally validated TDases. B) Performance metrics of DeepSEA and competitors. To establish labels we considered TDases with functional activity as positives (ARP) and TDases with no activity as negatives (NRP). Wrong AMR class annotations or lack of annotation of any tool were considered NRP. The hatched bars represent DeepSEA models and their respective ARP-NRP training set relation. Complete data are available in SuppTables 4, 5, and 6.

Discussion

Protein language models fitted using unsupervised learning for broad classification demand huge amounts of data. ProteinBert (23) was pre-trained on ~106 million proteins from UniProtKB/UniRef90 and ProtTrans (33) took 393 billion amino acids and 5616 GPUs to be trained. These pLMs can be fine-tuned to AMR annotation. However, large embedding matrix outcomes from pLMS demand high computation power. For instance, ProtTrans models encode proteins into 1024 dimension vectors which is a bottleneck to analyze whole

proteomes in usual hardware. DeepSEA neural networks with a global average pooling layer after Conv1D and GRU layers extract abstracted relations from limited and unbalanced AMR protein with up to 70 dimensions in the case of CNN and CNNS.

DeepARG (19) achieved higher recall depending on the ARP class than the bit-score method and evidenced deep learning benefits to AMR novelty annotation. However, DeepARG still needs Diamond to pre-filter likely ARPs before prediction per se. We demonstrate that fitting ANNs on only ARPs yielded a higher number of false positives on both pseudo and real data and that this issue can be bypassed by introducing a suitable NRP set on training. And that the weak dependence on alignment was enough to increase the DeepARG false positive rate (Fig. 9).

RGI takes advantage of CARD rigorous data curation and cut-offs to annotated AMR nevertheless, for being totally alignment dependent, it fails to identify remote homologous ARPs. This high dependence is demonstrated on AUC values (Fig 9) for RGI applied on aligned NRP-ARP. As expected, RGI performed worse than DeepARG due to higher alignment dependence. DeepSEA models were more independent of NRP-ARP similarity and differentiated MDR and NRP proteins. On the TDases test, CNN-Unaligned and RGI had similar precision values and DeepSEA CNN-Unaligned was superior at other metrics.

For AMRFinder, a noteworthy HMM drawback is the need for multi-sequence alignment to construct them for each ARP group aimed. DeepSEA models are multiclass classifiers fitted on NCRD95, an unbalanced dataset, and had equal or superior performance to AMRFinder on the TDases set and better AUC values on a large test set. Neither RGI, AMRFinder nor DeepARG could annotate the novel TDase discovered by Blake et al., 2024.

Even though HMD-ARG uses NRP in training, the authors did not test size and alignment effects. They state to use the most similar APR-NRP set to force their model to learn AMR class features. However, we demonstrated that aligned NRP-ARP sets caused performance impairment for the MDR class, which is one of the most concerning kinds of resistance and has been targeted in drug development research (34)

DeepSEA models did not need protein alignment for decision-making and had high recall values even for the less represented AMR protein classes in the test set. In this work, diamond was applied only to establish the NRP sets that helped our models to learn features related only to AMR protein. Still, t-SNEs from the global average pooling layer showed that unaligned NPR-ARP improved multidrug-resistant protein clustering. This AMR class is composed of a majority of efflux pumps that remove plenty of harmful substances from the bacterial cell, and therefore *in silico* distinguishing between general and antimicrobial efflux pumps is a difficult task. Thus, DeepSEA can be used to survey multi-drug resistant *Acinetobacter baumannii*, which has drawn attention from health entities worldwide (35).

DeepSEA models learned features related to resistance mechanisms even when fitted on NRP-ARP datasets. The antibiotic inactivation cluster contains classes of ARP whose proteins modify antimicrobial molecules and prevent their action. For instance, resistance to beta-lactam antimicrobials, such as penicillin, happens mostly due to the expression of beta-lactamases which cleave the lactam ring and avoid inhibiting penicillin-binding proteins (36) and aminoglycosides-modifying enzymes are the most widespread resistance mechanisms to aminoglycosides antimicrobials (37).

A central issue in AMR-bioinformatics is the discrepancy between wet lab and *in silico* observations (16). Thus, predicting correctly the genotype-phenotype map would point AST toward specific AMR targets. The results on the TDase set provided a "proof of concept" of our ANNs ability to identify functional and novel (remotely homologous) enzymes that RGI, AMRFinder, and DeepARG failed due to low sequence identity. Therefore, DeepSEA models can be conveniently applied for a more efficient genomic screening without assistant tools to find functional ARPs and provide class-driven AST.

System and methods

Data collection

We used two databases that gathered antimicrobial resistance proteins from specific sources through different methods. The HMD-ARG (22) database is a collection of 17,282 sequences spread into 15 protein resistance classes from 6 databases (39,17,29,19,30,40). The non-redundant comprehensive database (NCRD) (25) used DIAMOND (40) to align data from the Antibiotic Resistance Genes Database (ARDB) (41), the Structured Antibiotic Resistance Gene database (SARG) (42), and CARD against Protein Data Bank (PDB) (43) and Non-Redundant Protein Database (NR) to find homologous sequences and increase the number of sequences within NCRD, resulting in a total of ~34 thousands of sequences from 29 classes. Both works applied CD-HIT (44) to wipe out redundant data. To construct the negative portion of our dataset, we downloaded all 334 thousand bacterial protein sequences from UniProt/Swiss-Prot and removed those tagged with the antibiotic resistance label (KW-0046). Search parameters were: (taxonomy_id: 2) NOT (keyword:KW-0046)

Hyperparameters

A grid-search algorithm was used on our model candidates to determine suitable dimension values for embedding matrices produced by the embedding layer and the number of filters for the Conv1D layer or GRU units for the GRU layer. The models were fitted with the resistance proteins from HMD and NCRD datasets with 32, 64, 128, or 512 filter/unit values and generated embedding with 40, 50, 60, or 70 dimensions. No grid-searching was performed for the learning rate since TensorFlow adapts it along the epochs of training. The best combination of values was selected according to model accuracy on test sets.

Dataset composition and splitting

The neural networks were fitted with five dataset compositions for each resistance database collected. The ARPs portion was the same among them, and the NRP part, however, varied according to whether they aligned to resistant proteins in number. Finally, our datasets contain (i) only ARPs (ii) ARPs and aligned NRP that passed the e-value > 0.001 and coverage > 80% cut-offs by the Diamond tool, (iii) ARPs unaligned NRPs (here we randomly selected the same number of aligns NRPs with keeping proportions), (iv) ARPs and 50% of the unaligned NRPs and (v) ARPs and 100% of the unaligned NRPs. All the cited combinations were split into train and test at a 70/30 ratio, and 20% of the training set was randomly separated for validation during the training process. The unaligned NRPs for each dataset were picked randomly with SeqKit v2 tool (45).

Pseudo Data

We based our pseudodata construction method on DeepARG work (19), however instead of using only sequences from positive data, we added negative sequences to the pseudo database. A set of 1000 pseudo-protein sequences, all with 500 residues long, was constructed by randomly selecting nine 50-residue length kmers from negative sequences from Uniprot/Swiss-Prot plus one 50-residue length kmer from one resistance sequence randomly picked. Finally, the chosen kmers were shuffled to avoid order bias.

Data and software availability

DeepSEA code and software are available on https://github.com/computational-chemical-biology/DeepSEA-project. The complete data dataset, models, and code for analysis are available on https://github.com/computational-chemical-biology/DeepSEA-code

Funding

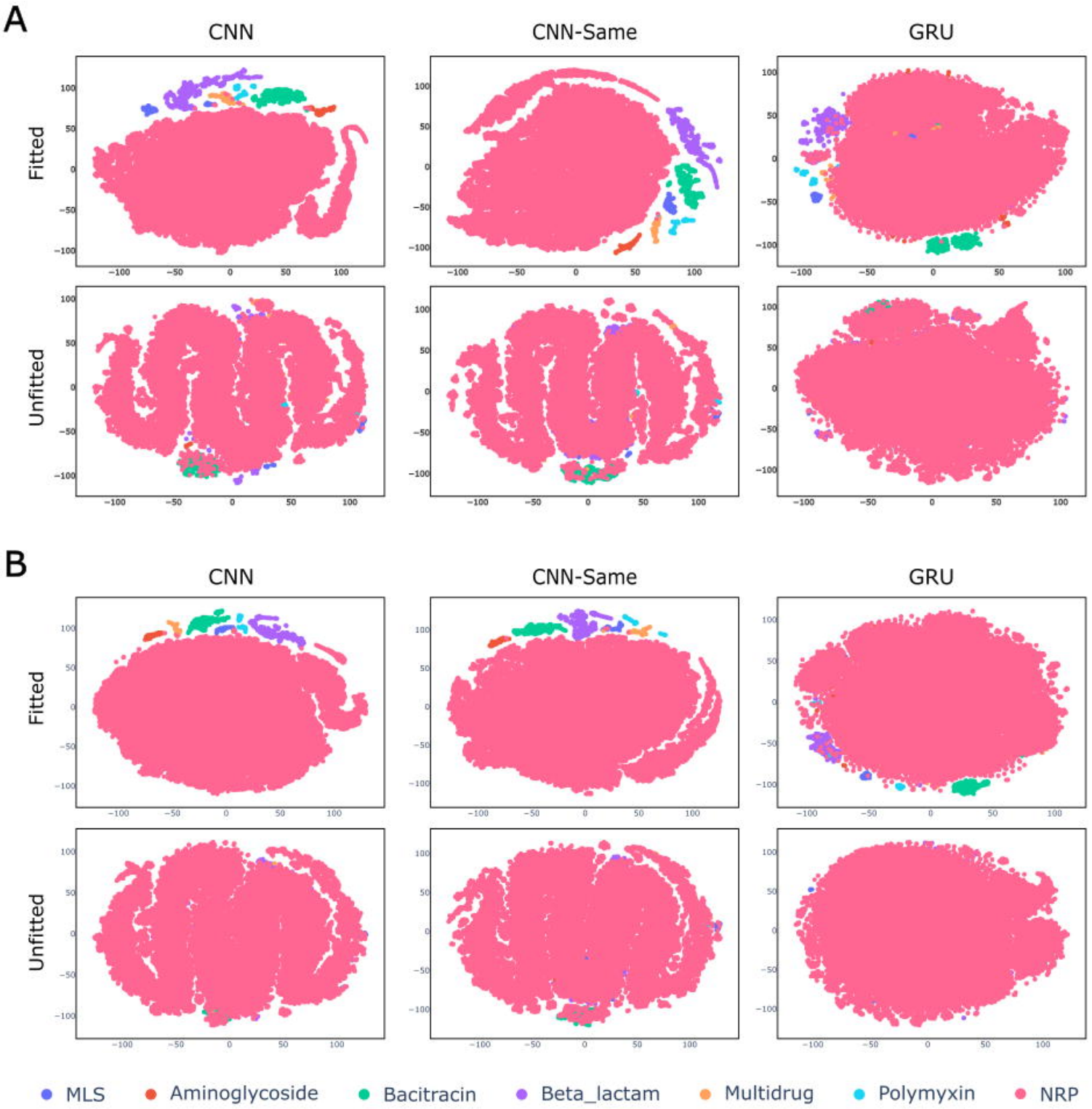
The authors were supported by the São Paulo Research Foundation [grant number 2021/08235-3 to T.C.B., and grant numbers 2017/18922-2, 2019/05026-4 and 2020/02207-5 to R.D.S.]. ARP was supported by Fundação Araucária project NAPI Bioinformática (# 66.2021).

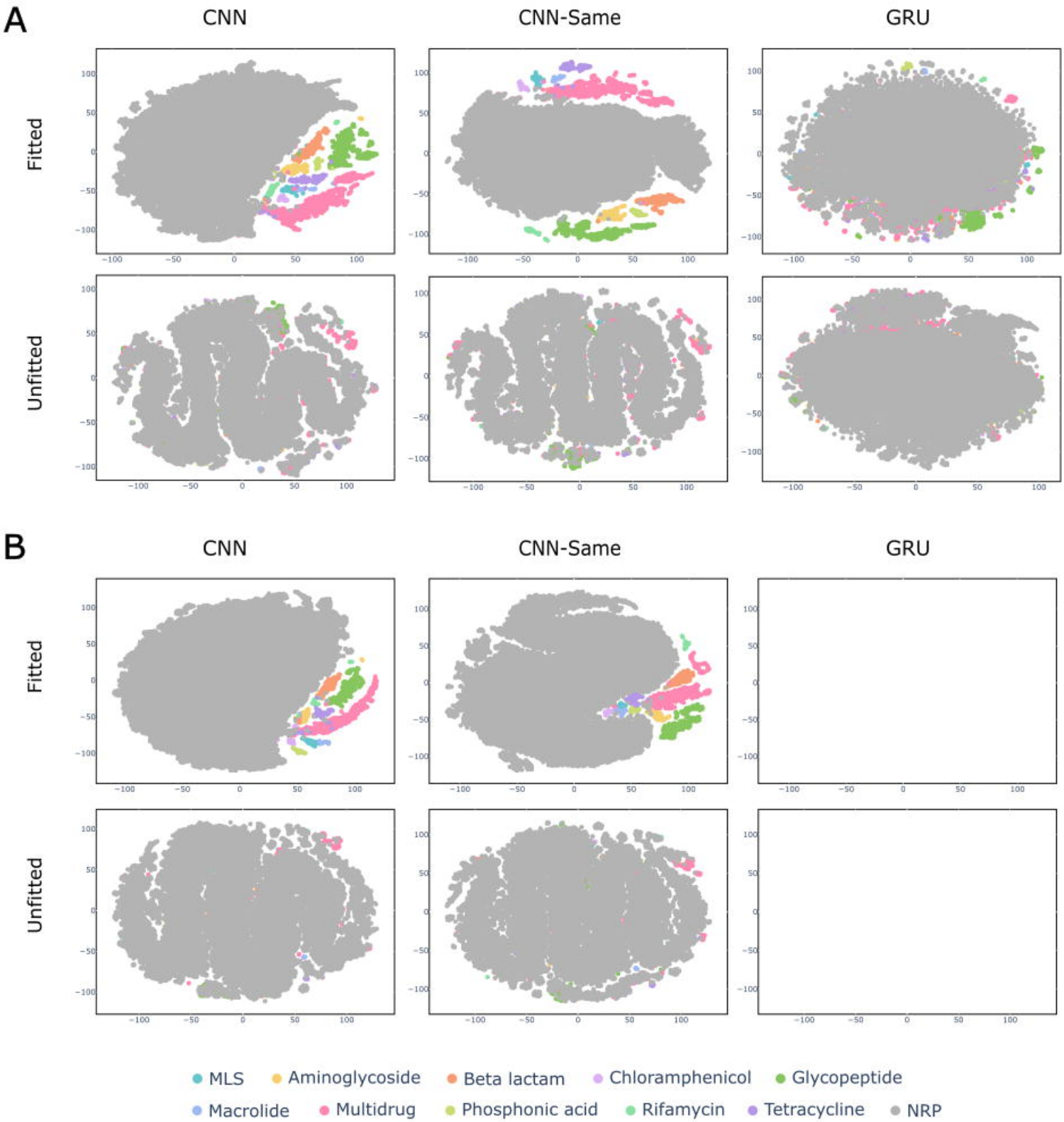
References

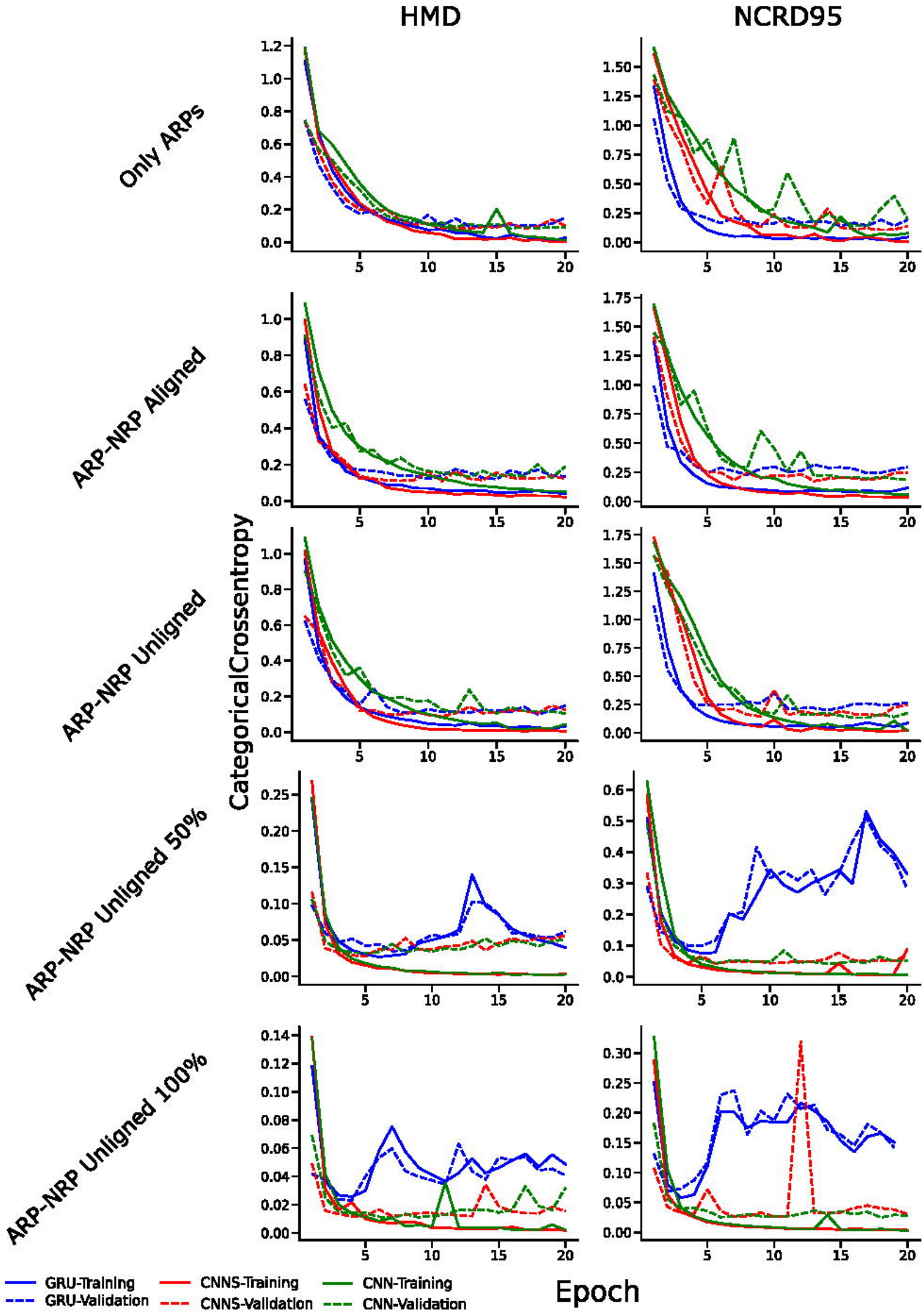
1. Fleming A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. Br J Exp Pathol. junho de 1929;10(3):226–36.
2. Yoshikawa TT. Antimicrobial Resistance and Aging: Beginning of the End of the Antibiotic Era? J Am Geriatr Soc. 2002;50(s7):226–9.
3. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 27 de novembro de 2020;11(1):6058.
4. Leigh Van Valen. A New Evolutionary Law. NEW Evol LAW. 1973;
5. Neu HC. The Crisis in Antibiotic Resistance. Science. 21 de agosto de 1992;257(5073):1064–73.
6. Rossolini GM, Arena F, Pecile P, Pollini S. Update on the antibiotic resistance crisis. Curr Opin Pharmacol. 1º de outubro de 2014;18:56–60.
7. Wright GD. Solving the Antibiotic Crisis. ACS Infect Dis. 13 de fevereiro de 2015;1(2):80–4.
8. Murray CJL, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet. 12 de fevereiro de 2022;399(10325):629–55.
9. Patel J, Harant A, Fernandes G, Mwamelo AJ, Hein W, Dekker D, et al. Measuring the global response to antimicrobial resistance, 2020–21: a systematic governance analysis of 114 countries. Lancet Infect Dis [Internet]. 16 de janeiro de 2023 [citado 18 de janeiro de 2023];0(0). Disponível em: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(22)00796-4/fulltext#supplementaryMaterial
10. Lord Jim O'Neill. AMR Review Paper - Tackling a crisis for the health and wealth of nations. 2014;
11. de Kraker MEA, Stewardson AJ, Harbarth S. Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050? PLOS Med. 29 de novembro de 2016;13(11):e1002184.
12. Pearson WR. An Introduction to Sequence Similarity ("Homology") Searching. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. junho de 2013;0 3:10.1002/0471250953.bi0301s42.
13. Djordjevic SP, Jarocki VM, Seemann T, Cummins ML, Watt AE, Drigo B, et al. Genomic surveillance for antimicrobial resistance — a One Health perspective. Nat Rev Genet. fevereiro de 2024;25(2):142–57.
14. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, et al. Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. J Clin Microbiol. janeiro de 2014;52(1):139–46.
15. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 8 de janeiro de 2020;48(D1):D517–25.
16. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. novembro de 2019;63(11):e00483-19.
17. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in

next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genomics. 24 de janeiro de 2022;8(1):000748.
18. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. The structure and diversity of human, animal and environmental resistomes. Microbiome. 7 de outubro de 2016;4(1):54.
19. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 1º de fevereiro de 2018;6(1):23.
20. Wang Z, Li S, You R, Zhu S, Zhou XJ, Sun F. ARG-SHINE: improve antibiotic resistance class prediction by integrating sequence homology, functional information and deep convolutional neural network. NAR Genomics Bioinforma. 1º de setembro de 2021;3(3):lqab066.
21. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinforma Oxf Engl. 1º de maio de 2014;30(9):1236–40.
22. Li Y, Xu Z, Han W, Cao H, Umarov R, Yan A, et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome. dezembro de 2021;9(1):40.
23. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 12 de abril de 2022;38(8):2102–10.
24. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.
25. Mao Y, Liu X, Zhang N, Wang Z, Han M. NCRD: A non-redundant comprehensive database for detecting antibiotic resistance genes. iScience. 17 de novembro de 2023;26(11):108141.
26. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning [Internet]. arXiv; 2016 [citado 17 de outubro de 2022]. Disponível em: http://arxiv.org/abs/1609.06570
27. Wang B, Mei C, Wang Y, Zhou Y, Cheng MT, Zheng CH, et al. Imbalance Data Processing Strategy for Protein Interaction Sites Prediction. IEEE/ACM Trans Comput Biol Bioinform. maio de 2021;18(3):985–94.
28. Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics. 1º de agosto de 2018;34(15):2642–8.
29. Pokharel S, Pratyush P, Heinzinger M, Newman RH, Kc DB. Improving protein succinylation sites prediction using embeddings from protein language model. Sci Rep. 8 de outubro de 2022;12(1):16933.
30. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. Antimicrob Agents Chemother. janeiro de 2014;58(1):212–20.
31. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res. 4 de janeiro de 2017;45(Database issue):D574–80.
32. Blake KS, Kumar H, Loganathan A, Williford EE, Diorio-Toth L, Xue YP, et al. Sequence-structure-function characterization of the emerging tetracycline destructase family of antibiotic resistance enzymes. Commun Biol. 16 de março de 2024;7(1):1–16.
33. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. outubro de 2022;44(10):7112–27.
34. Kyriakidis I, Vasileiou E, Pana ZD, Tragiannidis A. Acinetobacter baumannii Antibiotic

Resistance Mechanisms. Pathogens. março de 2021;10(3):373.
35. Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, Mohammed JC, et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. Nat Chem Biol. 25 de maio de 2023;1–9.
36. Ambler RP, Baddiley J, Abraham EP. The structure of β-lactamases. Philos Trans R Soc Lond B Biol Sci. janeiro de 1997;289(1036):321–31.
37. Garneau-Tsodikova S, J. Labby K. Mechanisms of resistance to aminoglycoside antibiotics: overview and perspectives. MedChemComm. 2016;7(1):11–27.
38. Smith KW, Alcock BP, French S, Farha MA, Raphenya AR, Brown ED, et al. A standardized nomenclature for resistance-modifying agents in the Comprehensive Antibiotic Resistance Database. Microbiol Spectr. 12 de dezembro de 2023;11(6):e0274423.
39. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. janeiro de 2015;9(1):207–16.
40. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. janeiro de 2015;12(1):59–60.
41. Liu B, Pop M. ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res. 1º de janeiro de 2009;37(suppl_1):D443–7.
42. Yin X, Jiang XT, Chai B, Li L, Yang Y, Cole JR, et al. 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. 1º de julho de 2018;34(13):2263–70.
43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 1º de janeiro de 2000;28(1):235–42.
44. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma Oxf Engl. 1º de julho de 2006;22(13):1658–9.
45. Shen W, Sipos B, Zhao L. SeqKit2: A Swiss army knife for sequence and alignment processing. iMeta. n/a(n/a):e191.


A
CNN
CNN-Same
GRU
Fitted
Unfitted
B
CNN
CNN-Same
GRU
Fitted
Unfitted
MLS
Aminoglycoside
Bacitracin
Beta_lactam
Multidrug
Polymyxin
NRP


A
CNN
CNN-Same
GRU
Fitted
Unfitted
B
CNN
CNN-Same
GRU
Fitted
Unfitted
MLS
Aminoglycoside
Beta lactam
Chloramphenicol
Glycopeptide
Macrolide
Multidrug
Phosphonic acid
Rifamycin
Tetracycline
NRP


HMD
NCRD95
Only ARPs
ARP-NRP Aligned
ARP-NRP Unligned
ARP-NRP Unligned 50%
ARP-NRP Unligned 100%
CategoricalCrossentropy
Epoch
GRU-Training
GRU-Validation
CNNS-Training
CNNS-Validation
CNN-Training
CNN-Validation