International Journal of
Molecular Sciences

MDPI

Article

Genome-Wide Mutation Scoring for Machine-Learning-Based Antimicrobial Resistance Prediction

Peter Májek [1], Lukas Lüftinger [1,2], Stephan Beisken [1], Thomas Rattei [3] and Arne Materna [1,*]

[1] Ares Genetics GmbH, Vienna 1030, Austria; peter.majek@ares-genetics.com (P.M.); lukas.lueftinger@ares-genetics.com (L.L.); stephan.beisken@ares-genetics.com (S.B.)
[2] Department of Computational Systems Biology, University of Vienna, Vienna 1030, Austria
[3] Centre for Microbiology and Environmental Systems Science, Division of Computational Systems Biology, University of Vienna, Vienna 1030, Austria; thomas.rattei@univie.ac.at
* Correspondence: arne.materna@ares-genetics.com

Abstract: The prediction of antimicrobial resistance (AMR) based on genomic information can improve patient outcomes. Genetic mechanisms have been shown to explain AMR with accuracies in line with standard microbiology laboratory testing. To translate genetic mechanisms into phenotypic AMR, machine learning has been successfully applied. AMR machine learning models typically use nucleotide k-mer counts to represent genomic sequences. While k-mer representation efficiently captures sequence variation, it also results in high-dimensional and sparse data. With limited training data available, achieving acceptable model performance or model interpretability is challenging. In this study, we explore the utility of feature engineering with several biologically relevant signals. We propose to predict the functional impact of observed mutations with PROVEAN to use the predicted impact as a new feature for each protein in an organism's proteome. The addition of the new features was tested on a total of 19,521 isolates across nine clinically relevant pathogens and 30 different antibiotics. The new features significantly improved the predictive performance of trained AMR models for *Pseudomonas aeruginosa*, *Citrobacter freundii*, and *Escherichia coli*. The balanced accuracy of the respective models of those three pathogens improved by 6.0% on average.

Keywords: machine learning; genomics; antimicrobial resistance; antibiotics; WGS; genome-wide mutation scoring

Citation: Májek, P.; Lüftinger, L.; Beisken, S.; Rattei, T.; Materna, A. Genome-Wide Mutation Scoring for Machine-Learning-Based Antimicrobial Resistance Prediction. *Int. J. Mol. Sci.* **2021**, *22*, 13049. https://doi.org/10.3390/ijms222313049

Academic Editors: Hiromi Nishida and Hiroshi Toda

Received: 15 October 2021
Accepted: 29 November 2021
Published: 2 December 2021

1. Introduction

Antibiotics are small molecules with bacteriostatic or bactericidal activity that are used to treat bacterial infections. The overuse of antibiotics has over time accelerated the emergence and spread of antimicrobial resistance (AMR) in bacteria [1]. Accurate antimicrobial susceptibility testing (AST) results are crucial to guide patient treatment as well as to inform antibiotic stewardship and outbreak monitoring. In this context, resistance predictions from whole pathogen genomes have the potential to complement or even replace phenotypic AST [2–5]. In recent years, various in silico methods have been applied to predict resistance phenotypes from whole genome sequencing data (WGS-AST) [6,7].

In principle, approaches for WGS-AST are either rules-based or machine learning (ML) based. The rules-based approach relies on the collection and curation of expert knowledge about AMR markers, e.g., genes and sequence mutations [8–13]. AMR prediction on a sequenced isolate is then performed by searching the isolate's genomic sequence for the presence of AMR markers and interpreting the findings. This approach offers high clinical interpretability [6], but requires continuous biocuration by domain experts to keep the underlying rule set up-to-date. Rules-based WGS-AST is expected to miss novel resistance mechanisms not yet described in the literature. In contrast, machine-learning-based approaches avoid manual collection of domain knowledge; instead, an ML algorithm

identifies mechanisms of AMR directly from a large dataset of sequenced isolates with known resistance phenotypes. The main prerequisite is a sizeable dataset to enable the ML algorithm to train accurate models. A major advantage of ML-based WGS-AST is that little expert biocuration is required (at the cost of clinical interpretability); the algorithms automatically pick new emerging modes of AMR resistance from new WGS-AST data.

ML-based WGS-AST typically uses nucleotide k-mer representations of either input genome assemblies or raw sequencing reads [14–18]. K-mer sets have been successfully used for various bioinformatics analyses, ranging from species identification [19] to genome assembly [20], as they offer advantages in computing efficiency and speed. To train a predictive AST model, an ML algorithm searches for correlations between a biological phenotype of interest and the presence of individual DNA k-mers. The dimensionality of DNA k-mer based feature spaces is typically orders of magnitude larger than the number of training samples. Training predictive models on such datasets requires careful regularization of the training process to avoid training set overfitting and poor model generalization. ML algorithms might not be able to detect biologically relevant signals in high dimensional training data or might incorrectly pick spurious correlations. The engineering of biologically relevant features pertaining to raw genomic sequences might thus be advantageous. For example, gene annotations translated into amino acid (AA) k-mers might more efficiently capture missense mutations and ignore silent mutations [21].

A key challenge for WGS-AST is that the presence of different point mutations can contribute to resistance. Different loss-of-function mutations in a single gene can decrease susceptibility, ultimately leading to resistance. In classical DNA or AA k-mer representations, resistance-contributing mutations result in the presence of different k-mers represented as different features, even though all cause a loss of function of the same causative gene. In a small training dataset, many such relevant k-mers are either missing or present in a few samples only and thus difficult to detect. Combining different mutations of a single gene or protein into a single aggregated score for each protein of interest addresses this challenge. Any disrupting mutation of a given gene would ideally generate a similar disruption score, despite being located on different positions. There are several bioinformatics tools that predict the structural effects of protein mutations and their usage in the context of AMR has been recently reviewed by Tunstall et al. [22]. Many of these tools have been used to evaluate mutations on limited sets of proteins, but to the best of our knowledge, this is the first systematic proteome-wide application of mutation scoring for WGS-AST. An efficient and robust mutation-scoring tool is needed to score millions of mutations across thousands of reference proteins within a reasonable run time and costs. The tool PROVEAN [23,24], evaluated in this study, generates predictions for approximately 10 mutations per second per the CPU core.

The performance of ML models can be improved in several ways. While the addition of training data is the most straightforward strategy, it is not always feasible. Previously, we explored the utility of different ML algorithms and their possible combinations [17]. In this study, we investigated improvements of WGS-AST ML models via the feature-space engineering approach. We evaluated whether extending a naïve DNA k-mer feature space with additional biologically relevant features, in particular automated mutation scoring, improves the predictive power of WGS-AST ML models. It is of note that a complementary strategy to improve the feature space of WGS-AST ML models is to focus on biologically relevant signals for AMR resistance phenotype determination, for example, by restricting the feature space to genes known to be associated with a given phenotype prior to model training. Such biologically informed WGS-AST models, including rule-based systems, were not considered in this study in order to focus on an exhaustive and purely statistical ML approach.

2. Results

2.1. Overview of the Data

Whole genome sequencing data together with resistant/susceptible phenotypes for nine clinically relevant pathogens and 30 antimicrobial compounds were used in this study. Altogether 110 datasets were prepared according to the protocol described in Materials and Methods section. Between 280 and 3681 isolates were used for training with a median training set size of 1368 isolates. Supplementary Table S1 contains a detailed description of each dataset.

2.2. WGS-AST Predictive Models on DNA K-Mers

Extreme gradient boosting (XGB) [25] models were trained on the 110 genomic k-mer datasets. The models achieved an average categorical agreement (accuracy) of 85.8%, an average major error (ME) of 11.0%, and an average very major error (VME) of 21.8%. Individual results of the DNA k-mer XGB models on the corresponding test sets are provided in Supplementary Table S2.

2.3. Feature Engineering

To improve WGS-AST models, we explored the utility of biologically relevant information, extending the original DNA k-mer feature space with AA k-mers, protein counts, multilocus sequence types (MLST), the presence of AMR markers, and DNA assembly quality control (QC) metrics. After the addition of the new feature types, the original DNA k-mers still made up 85% of the generated features.

On the set of 110 datasets studied, the models trained on the extended feature set achieved balanced accuracy (bACC) improved by 1.5% on average; 62 models improved, 10 models performed exactly the same, and 38 models worsened. The largest improvements in bACC were seen for models of *C. freundii* and *P. aeruginosa* with an average of 6.4% and 5.2%, respectively.

Encouraged by these improvements, the datasets were further extended with features scoring the functional alteration of protein mutations as provided by PROVEAN. Averaged over the 110 evaluated datasets, the proportion of feature types that passed the feature filtering step and were used as input for model training were split into DNA k-mers (82%), AA k-mers (12%), mutation scores (5%), protein counts (0.9%), AMR markers (0.012%), MLST types (0.0043%), and assembly QC metrics (0.0014%). Even though the mutation scores made up only about 5% of the feature space, their creation took about twice as long as the creation of all other features combined.

2.4. Models Trained with PROVEAN Features

Averaged over the 110 models trained, the first feature space extension improved the bACC by 1.5%. Adding PROVEAN features on top improved the bACC by an additional 1.6% on average (see Figure 1 and Supplementary Table S2). Twelve AMR models improved their bACC by more than 10% (Figure 2). Improvements in VME and ME are shown in Supplementary Figures S1 and S2.

Training models on fully extended feature spaces most improved the models for *C. freundii*, *P. aeruginosa*, and *E. coli*; their bACC improved on average by 9.3%, 7.5%, and 3.5%, respectively. The total number of errors on the test sets of these three pathogens was reduced from 2170 to 1704. Twelve of these models improved significantly, with a p-value < 0.05 as tested by McNemar's test. The difference on the aggregated confusion matrix of all models of these three pathogens was highly significant (p-value = 5.6×10^{-8}, McNemar's test). The improvements on *C. freundii*, *P. aeruginosa*, and *E. coli* models were mostly driven by sensitivity improvements, reducing VME by 16.8%, 11.9%, and 7.2%, respectively. The models of the other six pathogens seem to be less affected by the addition of the new features; their VME and ME decreased on average by only 0.2% and 2.3%, respectively. The details of particular models performance are provided in Supplementary Table S2. Out of all 110 models, 19 improved significantly and 5 downgraded significantly,

at significance 0.05 according to McNemar's test. The five downgrading models achieved a bACC lowered by 1.8% on average, while the 19 significantly improved models had on average a bACC higher by 10.8%. The two models that dropped most in performance after the addition of PROVEAN features were models of cefepime and gentamicin resistance on *E. cloacae*. For both of these models, training on the fully extended feature spaces slightly decreased the number of false positives but at the same time the number of false negatives increased comparably. As the test sets of these two models contained only 32 and 36 resistant samples, the drop of the models' sensitivities was higher than the small improvements in specificities, thus bACC dropped by about 8%.

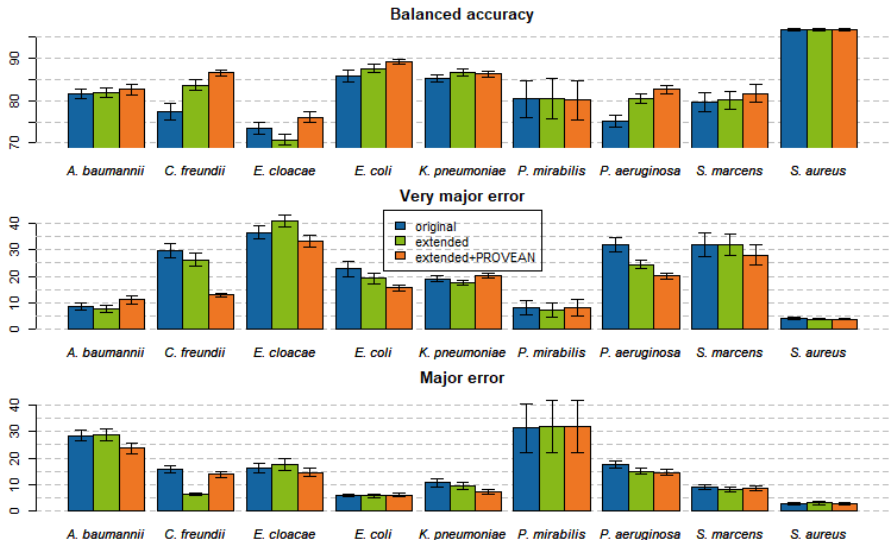

Figure 1. Balanced accuracy, very major error (1-sensitivity), and major error (1-specificity) of trained models as averaged across the compounds of the given pathogens. Error bars are the standard error of means.

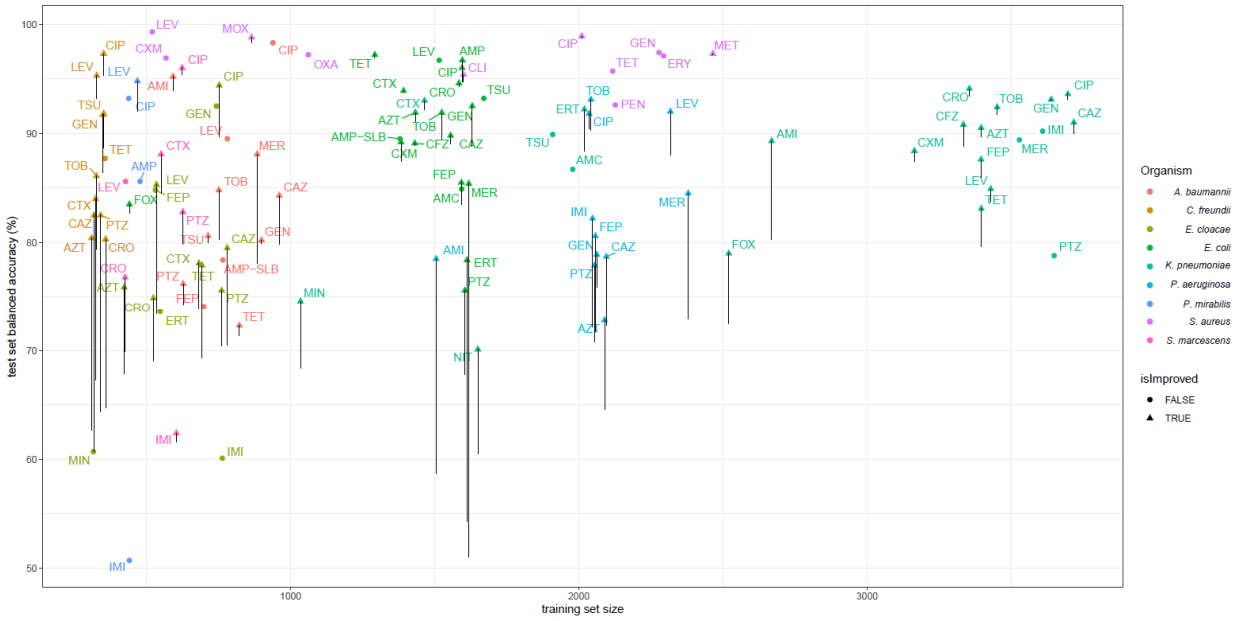

Figure 2. Balanced accuracy of 110 models evaluated on the 20% test set splits. The training set size is shown on the *x*-axis. For each dataset, the balanced accuracy of the best performing model of the three models considered, trained on different feature spaces (DNA k-mers, extended dataset, extended + PROVEAN features), is shown. Vertical lines indicate the increase in balanced accuracy on individual datasets from the original feature space. Section 4.2 lists all compound names and their abbreviations as shown here. A small scatter is added on the x-axis to avoid overlapping vertical lines.

2.5. Feature Importance

Even though PROVEAN features made up only about 5% of all features, for the models of *P. aeruginosa*, *C. freundii*, and *E. cloacae,* the mutation scores had about 50% contribution to the outcome of the new models. QC assembly metrics, MLST types, and AMR markers contributed little to the predictions (Figure 3).

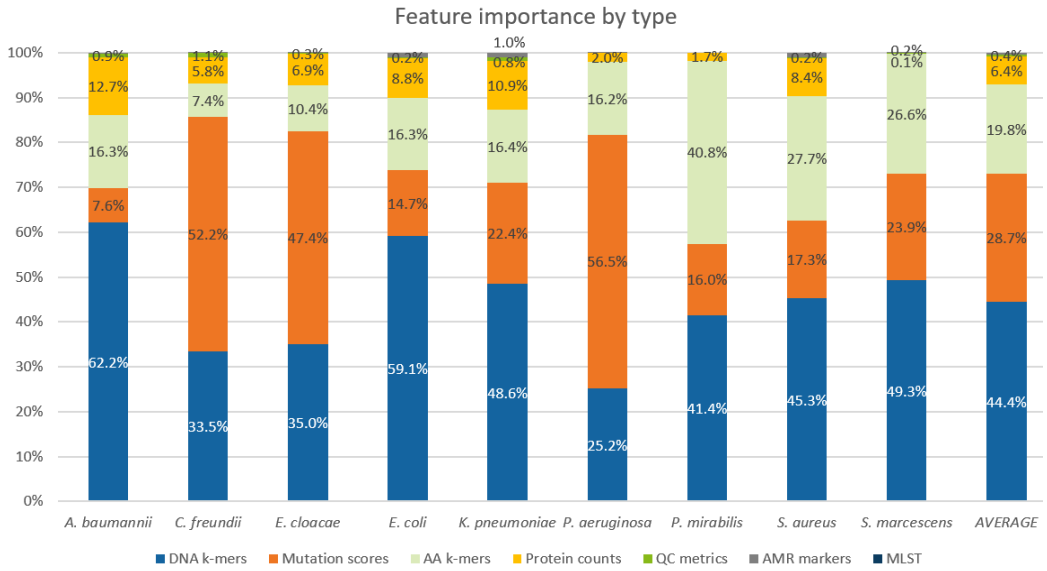

Figure 3. Relative importance of different feature types for individual pathogens as measured by SHAP contributions of features to model output. Values were summed up across all tested compounds.

Figure 4 demonstrates the predictive power of protein-wise aggregated PROVEAN scores on three XGB models that achieved a high improvement of balanced accuracy. The figure shows the distribution of mutations of *gyrA* and *ampD,* known markers for antimicrobial resistance in these pathogens [26,27]. PROVEAN scores on these genes were picked by XGB as the most important features of the respective models. Both genes had more than a hundred unique mutations, many of which occurred only several times in the provided datasets. As shown on the figure, the isolates with low PROVEAN scores for the corresponding gene were mostly resistant with strong positive predictive values of 96%, 98%, and 93%, respectively. These splits not only had a strong PPV but also were biologically meaningful and made the underlying models more interpretable.

The model for meropenem in *E. coli* improved most with regard to bACC after the addition of PROVEAN features. The original model detected only a single resistant isolate in the test set. Inclusion of the first feature space extension lowered VME a little, but the inclusion of PROVEAN scores lowered it further to 28.8%. The most important features of this model were the gene length of *ompC*, the presence of *aacA1*, and the length of *ompF*. Both membrane proteins OmpC and OmpF as well as aminoglycoside acetyltransferase *aacA1* are implied in antimicrobial resistance in *E. coli* [28–30]. In the test set, 58% of resistant isolates had OmpC shorter than 374 AA, while only 6% of susceptible isolates had such a short OmpC. The length of OmpC seems to be an informative signal for its truncation that might be harder to pick by DNA or AA k-mer features as the truncation can possibly happen on different sites.

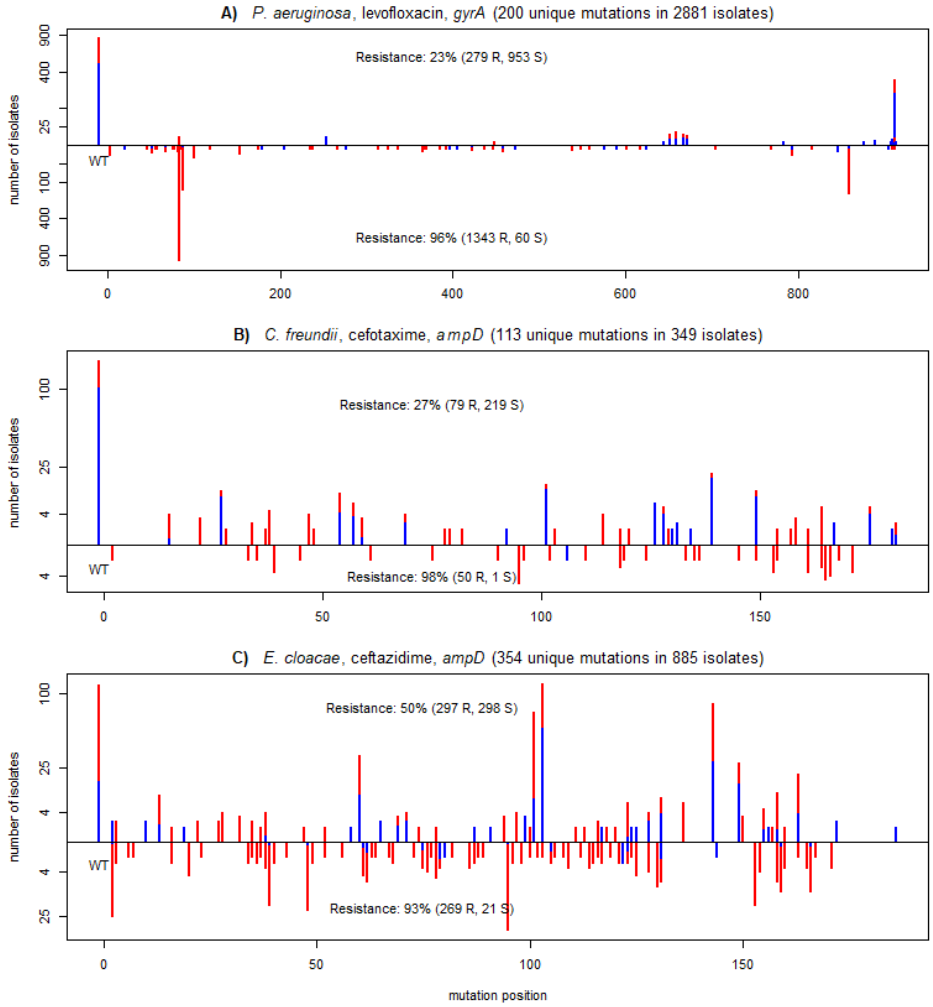

Figure 4. Mutational profiles across the three most important genes of the models for the following combinations: (**A**) *P. aeruginosa*, ciprofloxacin; (**B**) *C. freundii*, cefotaxime; and (**C**) *E. cloacae*, ceftazidime. Each plot shows the positional distribution of mutations of the most important gene for the model prediction. The bars show the number of assemblies having a given mutation at a given position, but only considering the most damaging mutation per assembly. Damaging mutations (PROVEAN scores lower than a gene specific threshold used by XGB models) are shown on the negative y-axis section; neutral mutations (score larger than or equal to the threshold) are shown in the positive y-axis section. The y-axis is shown in a square root scale. Red and blue colors correspond to assemblies resistant and susceptible to the given compound. The number of isolates identified as damaging or neutral along with the percentage of resistant isolates are shown for each group. Isolates with wild-type sequences are positioned at position -10 on the x-axis and labelled WT.

3. Discussion

Eighty AMR models distributed across all nine pathogens benefitted from the feature space extensions attempted in this study. In principle, if all possible mutations are present in a training dataset in statistically large enough quantities, DNA k-mer representation is powerful enough to pick the phenotype causative mutations. However, current WGS-AST datasets are limited in size; thus, engineered features that aggregate information across all possible mutations, e.g., from PROVEAN mutation scoring, provide advantages. Our findings confirm that for many pathogens, insufficient data are available for decision tree-like algorithms to confidently determine the relevant phenotype causing mutations from DNA k-mer feature space alone. DNA k-mers, mutation scoring features, AA k-mers, and protein counts had the highest importance in the trained models. Other considered feature types, among which were also indicators of the presence of AMR markers, had under 1%

contribution in all pathogens. This suggests that k-mer based features have enough power to accurately represent presence of AMR markers and implicitly annotated presence of markers does not contribute strongly to the trained ML models, at least not without expert insight about marker-phenotype correlations as is done in rules-based models.

We hypothesize that mutation scoring can effectively combine signals from different mutations of a given protein to provide features that have a strong predictive value even if individual mutations are present in low numbers. The mutagenesis landscape of individual genes can be complex with hundreds of different genotype-altering mutations. It is a non-trivial task to determine the causality of a specific mutation if it only occurs a few times in the training data as it is likely to co-occur with several other mutations. In the most extreme case, classical k-mer-based classifiers are expected to fail on mutations that were not seen in the training set. Automated mutation scoring has the advantage that it can estimate functional effects on mutations that were not seen in the training set, contingent on the accuracy of the underlying mutation scores. Our findings suggest that this is likely the case for many genes involved in antimicrobial resistance. Any signal to be picked by an ML algorithm from the original DNA k-mer feature space can only be extracted from the labeled training WGS-AST dataset. Contrary to the k-mer counts, the extended feature sets and especially the engineered PROVEAN mutation scoring is based on data from a much larger database of non-redundant sequences used by BLAST across many different organisms. This is a practical example of the utilization of a large corpus of unlabeled data for supervised learning, a strategy successfully used in machine learning [31].

Our results demonstrate that the addition of PROVEAN scoring features or, in general, any biologically relevant metadata can be beneficial for building WGS-AST models. In particular, models may improve for pathogens that either have complex resistance phenotypes that cannot be explained by a small number of specific mutations [32], or have higher baseline mutational rates and can acquire more mutations in general, e.g., hypermutator phenotypes. The usage of the biologically meaningful feature spaces has an additional advantage of high clinical interpretability of model decisions. Further research into the utility of these engineered features is needed to fully understand their role in WGS-AST models. The current study has explored their benefit across a large set of models without any model specific optimizations.

The results presented in this study constitute a small but significant contribution towards the overall goal of improving patients' treatment outcome in the clinic via the application of WGS-AST models. Further research into model stability, confidence, interpretability, and how to best apply predictive models in the clinic is required. A successful implementation of a WGS-AST system will likely also incorporate relevant biological knowledge of rules-based models.

4. Materials and Methods

4.1. Data Retrieval

Genome assemblies and associated resistance/susceptibility profiles for nine clinically relevant pathogens (Acinetobacter baumannii, Citrobacter freundii, Enterobacter cloacae, Escherichia coli, Klebsiella pneumoniae, Proteus mirabilis, Pseudomonas aeruginosa, Serratia marcescens, and Staphylococcus aureus) were obtained from ARESdb [13], which contains WGS and AST data from proprietary and public sources [3,8,16,33–37]. Compound names imported from public data sources were corrected via regular expression base matching as defined in Supplementary Table S3. Minimum inhibitory concentration (MIC) values were translated into susceptible/intermediate/resistant (S/I/R) interpretative categories via clinical breakpoints according to CLSI 29 standards [38]. Intermediate phenotypes were treated as resistant. In cases where the lowest measured MIC value was equal to the intermediate MIC breakpoint, the value was interpreted as susceptible. Sequencing reads were quality trimmed and filtered using Trimmomatic v0.39 [39] and de novo assembled using SPAdes v3.13.1 [20]. Completeness of the assembled genomes was assessed using BUSCO v5.2.2 [40] and QUAST v5.0.2 [41]. Assemblies which did

not meet the QC criteria of N50 $\geq$ 5000, L50 $\leq$ 500, BUSCO completeness $\geq$ 75%, and BUSCO duplication $\leq$ 7% were filtered out. Organism–compound datasets with fewer than 100 susceptible and 100 resistant isolates were excluded. Filtered datasets were partitioned into training and test sets (80%:20%) using a genome-distance-based method [17]. This dataset partitioning method is designed to reduce similarity between the training and the test dataset. Datasets that did not contain at least 10 resistant and at least 10 susceptible isolates in the test set were excluded from the study. A detailed list of publicly available samples used for training and testing of particular models is provided in Supplementary Table S4.

4.2. Antimicrobial Compounds

The WGS-AST data are listed in Section 4.1 for the following 30 antimicrobial compounds that passed the data extraction criteria: amoxicillin and clavulanic acid (AMC), amikacin (AMI), ampicillin (AMP), ampicillin and sulbactam (AMP-SLB), aztreonam (AZT), ceftazidime (CAZ), cefazolin (CFZ), ciprofloxacin (CIP), clindamycin (CLI), ceftriaxone (CRO), cefotaxime (CTX), cefuroxime (CXM), ertapenem (ERT), erythromycin (ERY), cefepime (FEP), cefoxitin (FOX), gentamicin (GEN), imipenem (IMI), levofloxacin (LEV), meropenem (MER), meticillin (MET), minocycline (MIN), moxifloxacin (MOX), nitrofurantoin (NIT), oxacillin (OXA), benzylpenicillin (PEN), piperacillin and tazobactam (PTZ), tetracycline (TET), tobramycin (TOB), and sulfamethoxazole and trimethoprim (TSU).

4.3. Machine Learning Feature Generation

The following feature types were used for XGB models: nucleotide k-mers, amino acid k-mers, protein occurrence counts, multilocus sequence types, AMR marker presence, and assembly quality control metrics. DNA k-mers of a length of 21 nucleotides were obtained by KMC 3.1.0 [42]. To obtain AA k-mers and protein counts, PROKKA v1.14.1 [43] was run on each genomic assembly file and the resulting FASTA files were converted to AA-kmers of length 16 and protein presence count matrices. MLSTs were obtained from ARESdb and encoded using one-hot encoding. The following features were used as assembly QC metrics: the number of contigs, the length of the largest contig, the assembly length, GC content, N50, N75, L50, and L75. AMR markers in ARESdb [13] were clustered with CD-HIT [44] (parametrized as -c 0.9 -n 5 -aS 0.8 -aL 0.8). Each assembly was searched for against the database of AMR markers in ARESdb and marker clusters associated with found markers were noted. The presence of clusters was encoded in a one-hot feature encoding. Presence indicators of known AMR markers were considered as additional features; however, unlike rules-based models, no prior information about any marker-phenotype relations was provided for model training. Any existing marker-phenotype correlations had to be picked automatically by the ML algorithm purely from the training data, in the same fashion as for any other features. To exclude any possibility of biases introduced by common feature selection on the full dataset, all features for prediction on the test sets were generated separately only at the prediction time.

4.4. Proteome-Wide Scoring of Functional Alterations

On top of the features listed in the previous section, 4 specific mutation scores based on PROVEAN [23] predictions were determined for each protein cluster in the training dataset. Training set genomes were first annotated by PROKKA and clustered by the name of the identified gene. All identified proteins of a given name were then clustered with CD-HIT [44] (with parameters -c 0.9 -aL 0.9 -aS 0.9 -n 5) to obtain a set of protein clusters. In each protein cluster the reference sequence was defined as the most frequent sequence of the given cluster. Let p be a protein sequence belonging to a cluster Cj with a reference sequence r_j. For each such a protein p, a set of its amino acid variants against the reference sequence r_j in HGVS nomenclature [45] was determined using a tool called Mutanalyzer [46]. Let us name that set of variants $HGVS(p, r_j)$. The union of all variants of

the reference sequence r_j, $\cup_p HGVS(p, r_j)$, was then scored using PROVEAN. We defined a score *M(p)* of the protein p that belonged to the cluster with the reference r_j as

$$M(p) = \min_x \{ \text{PROVEAN}(x, r_j) \mid x \in \text{HGSV}(p, r_j) \}. \quad (1)$$

For any assembly A_i and any sequence cluster C_j the following four scores are then defined:

$$S_{ij}^{\min} = \min_p \{ M(p) \mid p \in A_i \wedge p \in C_j \} \quad (2)$$

$$S_{ij}^{\max} = \max_p \{ M(p) \mid p \in A_i \wedge p \in C_j \} \quad (3)$$

$$L_{ij} = \max_p \{ |p| \mid p \in A_i \wedge p \in C_j \} \quad (4)$$

$$N_{ij} = |\{ p \mid p \in A_i \wedge p \in C_j \}| \quad (5)$$

If the set $\{p \mid p \in A_i \wedge p \in C_j\}$ was an empty set then we defined each of the above terms as zero. The above four terms (2)–(5) for each cluster can be respectively interpreted as a mutation score of the most altered protein, a mutation score of the most native-like protein, the longest protein in the cluster, and the number of proteins in the cluster. If there were N clusters found by CD-HIT in the training set then the above procedure defined *4N* numeric features for each assembly. The most expensive part of the mutation scores calculations was running PROVEAN itself. The computing costs can be reduced significantly if the union of all mutations of a given cluster are calculated first and then evaluated together in a single PROVEAN call. PROVEAN scores mutations by searching the reference sequence against the non-redundant BLAST database [47]. PROVEAN version 1.1.5 together with BLAST version 2.2.31 and the corresponding BLAST non-redundant database were used in this study. These BLAST searches were the most expensive part of the training data generation, but importantly the BLAST results could be cached at training time and reused at prediction time. This allowed the generation of predictions within seconds, especially since only a small subset of CD-HIT clusters needed to be evaluated at prediction time.

4.5. Feature Filtering

The feature matrix was prepared considering all training assemblies of a given pathogen. Zero-variance training features were removed and features with identical values across all training samples were de-duplicated, keeping one representative. Subsequently, for each organism and antimicrobial compound, a subset of the organism's full count matrix for which S/R class information of the given compound was available was extracted. The feature space was then condensed by univariate feature selection before training. Features were tested for independence from the S/R category using the χ^2 test as implemented in scikit-learn. The resulting p-values were corrected for false discovery by the Benjamini–Hochberg procedure [48] and filtered by a q-value threshold of 0.05. For *P. aeruginosa* datasets, relatively few features were correlated with phenotype at the default q-value threshold, thus the q-value threshold for *P. aeruginosa* datasets was increased to 0.2. Of the features passing the filtering steps, at most half a million features with the highest log-odds ratio were retained.

4.6. Model Training

Extreme gradient boosting was used for training predictive models of antimicrobial resistance from WGS data for a set of nine clinically relevant pathogens. XGB models were trained using XGBoost 1.2.0 via the provided Python 3 bindings. Only the training datasets were used for training. All models were trained with the following XGB hyperparameters: min_child_weight: 0, max_depth: 14, gamma: 0.0001, subsample: 0.721, colsample_bytree: 0.4947, colsample_bylevel: 0.5366, max_delta_step: 2, lambda: 2.394, learning_rate: 0.0485, n_estimators: 360, objective: binary:logistic, eval_metric: logloss. To counter label imbalance in the training dataset, the training isolates were weighted with weights calculated

such that the total weight of susceptible and resistant isolates was equal, capped to the range 0.1 and 10 for strongly imbalanced datasets. To ensure model stability, the number of estimators in each model was controlled via the early stopping strategy by monitoring performance in an internal tenfold CV of the training set, with the early_stopping_rounds parameter set to 15. The final model was then trained on the complete training set with the number of estimators determined from the early stopped CV run. The above hyperparameters were optimized by several rounds of automated hyperparameter search using a genetic search algorithm combined with manual optimization on a selected subset of several *K. pneumoniae* and *P. aeruginosa* models. During the hyperparameter search, the XGB model performance was found to be reasonably robust to hyperparameter changes and models trained with hyperparameters optimized for individual models of *P. aeruginosa* and *K. pneumoniae* were found to have a similar performance to the models trained with the default set of hyperparameters (data not shown). The performance of trained models was evaluated on the test datasets; the internal tenfold CV was not used for model evaluation in any way. Feature Importance Scoring: The importance of individual model features was evaluated by SHAP values [49]. For each test set assembly, SHAP values for the 20 most important features were recorded. Relative importance of different feature types in individual pathogens was determined by aggregating the SHAP values over all test assemblies of all the pathogens' models (Figure 3).

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms222313049/s1.

Author Contributions: Conceptualization, P.M., L.L. and S.B.; methodology, P.M.; software, P.M. and L.L.; validation, P.M.; formal analysis, P.M.; investigation, P.M.; resources, L.L.; data curation, P.M. and L.L.; writing—original draft preparation, P.M.; writing—review and editing, S.B.; visualization, P.M.; supervision, A.M., S.B. and T.R.; project administration, S.B.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Vienna Business Agency, grant number 24478239.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: L.L., P.M., S.B. and A.M. are employed by Ares Genetics GmbH. The authors declare no conflict of interest between this study and the company.

References

1. O'Neill, J. The Review on Antimicrobial Resistance (Chaired by Jim O'Neill). Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. 2016. Available online: https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf (accessed on 1 December 2021).
2. Břinda, K.; Callendrello, A.; Cowley, L.; Charalampous, T.; Lee, R.S.; MacFadden, D.R.; Kucherov, G.; O'Grady, J.; Baym, M.; Hanage, W.P. Lineage calling can identify antibiotic resistant clones within minutes. *bioRxiv* **2018**, *40*, 3204. [CrossRef]
3. Bradley, P.; Gordon, N.C.; Walker, T.M.; Dunn, L.; Heys, S.; Huang, B.; Earle, S.; Pankhurst, L.J.; Anson, L.; De Cesare, M.; et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **2015**, *6*, 10063. [CrossRef]
4. Avdic, E.; Wang, R.; Li, D.X.; Tamma, P.D.; Shulder, S.E.; Carroll, K.C.; Cosgrove, S.E. Sustained impact of a rapid microarray-based assay with antimicrobial stewardship interventions on optimizing therapy in patients with Gram-positive bacteraemia. *J. Antimicrob. Chemother.* **2017**, *72*, 3191–3198. [CrossRef] [PubMed]
5. Banerjee, R.; Teng, C.B.; Cunningham, S.A.; Ihde, S.M.; Steckelberg, J.M.; Moriarty, J.P.; Shah, N.D.; Mandrekar, J.N.; Patel, R. Randomized Trial of Rapid Multiplex Polymerase Chain Reaction–Based Blood Culture Identification and Susceptibility Testing. *Clin. Infect. Dis.* **2015**, *61*, 1071–1080. [CrossRef]
6. Li, X.; Zhang, Z.; Liang, B.; Ye, F.; Gong, W. A review: Antimicrobial resistance data mining models and prediction methods study for pathogenic bacteria. *J. Antibiot.* **2021**, *74*, 838–849. [CrossRef] [PubMed]
7. Pesesky, M.W.; Hussain, T.; Wallace, M.; Patel, S.; Andleeb, S.; Burnham, C.A.D.; Dantas, G. Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data. *Front. Microbiol.* **2016**, *7*, 1887. [CrossRef] [PubMed]

8. Mahfouz, N.; Ferreira, I.; Beisken, S.; von Haeseler, A.; Posch, A.E. Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: A systematic review. *J. Antimicrob. Chemother.* **2020**, *75*, 3099–3108. [CrossRef] [PubMed]
9. Bortolaia, V.; Kaas, R.S.; Ruppe, E.; Roberts, M.C.; Schwarz, S.; Cattoir, V.; Philippon, A.; Allesoe, R.L.; Rebelo, A.R.; Florensa, A.F.; et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **2020**, *75*, 3491–3500. [CrossRef] [PubMed]
10. Zankari, E.; Allesøe, R.; Joensen, K.G.; Cavaco, L.M.; Lund, O.; Aarestrup, F.M. PointFinder: A novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.* **2017**, *72*, 2764–2768. [CrossRef] [PubMed]
11. Feldgarden, M.; Brover, V.; Gonzalez-Escalona, N.; Frye, J.G.; Haendiges, J.; Haft, D.H.; Hoffmann, M.; Pettengill, J.B.; Prasad, A.B.; Tillman, G.E.; et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **2021**, *11*, 12728. [CrossRef] [PubMed]
12. Alcock, B.P.; Raphenya, A.R.; Lau, T.T.Y.; Tsang, K.K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A.-L.V.; Cheng, A.A.; Liu, S.; et al. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **2020**, *48*, D517–D525. [CrossRef] [PubMed]
13. Ferreira, I.; Beisken, S.; Lueftinger, L.; Weinmaier, T.; Klein, M.; Bacher, J.; Patel, R.; von Haeseler, A.; Posch, A.E. Species identification and antibiotic resistance prediction by analysis of whole-genome sequence data by use of ARESdb: An analysis of isolates from the unyvero lower respiratory tract infection trial. *J. Clin. Microbiol.* **2020**, *58*, e00273-20. [CrossRef]
14. Drouin, A.; Giguère, S.; Déraspe, M.; Marchand, M.; Tyers, M.; Loo, V.G.; Bourgault, A.-M.; Laviolette, F.; Corbeil, J. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genom.* **2016**, *17*, 754. [CrossRef] [PubMed]
15. Aun, E.; Brauer, A.; Kisand, V.; Tenson, T.; Remm, M. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.* **2018**, *14*, e1006434. [CrossRef]
16. Nguyen, M.; Brettin, T.; Long, S.W.; Musser, J.M.; Olsen, R.J.; Olson, R.D.; Shukla, M.P.; Stevens, R.L.; Xia, F.F.-F.; Yoo, H.; et al. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. *Sci. Rep.* **2018**, *8*, 421. [CrossRef] [PubMed]
17. Lüftinger, L.; Májek, P.; Beisken, S.; Rattei, T.; Posch, A.E. Learning from Limited Data: Towards Best Practice Techniques for Antimicrobial Resistance Prediction From Whole Genome Sequencing Data. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 1–9. [CrossRef] [PubMed]
18. Drouin, A.; Letarte, G.; Raymond, F.; Marchand, M.; Corbeil, J.; Laviolette, F. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* **2019**, *9*, 4071. [CrossRef] [PubMed]
19. Wood, D.E.; Lu, J.; Langmead, B. Improved metagenomic analysis with Kraken 2. *Genom. Biol.* **2019**, *20*, 257. [CrossRef]
20. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
21. ValizadehAslani, T.; Zhao, Z.; Sokhansanj, B.A.; Rosen, G.L. Amino Acid k-mer Feature Extraction for Quantitative Antimicrobial Resistance (AMR) Prediction by Machine Learning and Model Interpretation for Biological Insights. *Biology* **2020**, *9*, 365. [CrossRef] [PubMed]
22. Tunstall, T.; Portelli, S.; Phelan, J.; Clark, T.G.; Ascher, D.B.; Furnham, N. Combining structure and genomics to understand antimicrobial resistance. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 3377–3394. [CrossRef] [PubMed]
23. Choi, Y.; Sims, G.E.; Murphy, S.; Miller, J.R.; Chan, A.P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **2012**, *7*, e46688. [CrossRef] [PubMed]
24. Choi, Y. A Fast Computation of Pairwise Sequence Alignment Scores between a Protein and a Set of Single-Locus Variants of Another Protein. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, Orlando, FL, USA, 7–10 October 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 414–417.
25. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
26. Langendonk, R.F.; Neill, D.R.; Fothergill, J.L. The Building Blocks of Antimicrobial Resistance in Pseudomonas aeruginosa: Implications for Current Resistance-Breaking Therapies. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 307. [CrossRef] [PubMed]
27. Moya, B.; Juan, C.; Alberti, S.; Pérez, J.L.; Oliver, A. Benefit of Having Multiple ampD Genes for Acquiring β-Lactam Resistance without Losing Fitness and Virulence in Pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* **2008**, *52*, 3694–3700. [CrossRef] [PubMed]
28. Choi, U.; Lee, C.-R. Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in Escherichia coli. *Front. Microbiol.* **2019**, *10*, 953. [CrossRef] [PubMed]
29. Liu, Y.-F.; Yan, J.-J.; Lei, H.-Y.; Teng, C.-H.; Wang, M.-C.; Tseng, C.-C.; Wu, J.-J. Loss of outer membrane protein C in Escherichia coli contributes to both antibiotic resistance and escaping antibody-dependent bactericidal activity. *Infect. Immun.* **2012**, *80*, 1815–1822. [CrossRef]
30. Tenover, F.C.; Filpula, D.; Phillips, K.L.; Plorde, J.J. Cloning and sequencing of a gene encoding an aminoglycoside 6′-N-acetyltransferase from an R factor of Citrobacter diversus. *J. Bacteriol.* **1988**, *170*, 471–473. [CrossRef]

31. Mitchell, T.M. *The Role of Unlabeled Data in Supervised Learning BT-Language, Knowledge, and Representation*; Larrazabal, J.M., Miranda, L.A.P., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 103–111.
32. Simner, P.J.; Beisken, S.; Bergman, Y.; Posch, A.E.; Cosgrove, S.E.; Tamma, P.D. Cefiderocol Activity Against Clinical Pseudomonas aeruginosa Isolates Exhibiting Ceftolozane-Tazobactam Resistance. *Open Forum Infect. Dis.* **2021**, *8*, ofab311. [CrossRef]
33. Wattam, A.R.; Davis, J.J.; Assaf, R.; Boisvert, S.; Brettin, T.; Bun, C.; Conrad, N.; Dietrich, E.M.; Disz, T.; Gabbard, J.L.; et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* **2017**, *45*, D535–D542. [CrossRef] [PubMed]
34. Bethesda (MD): National Database of Antibiotic Resistant Organisms (NDARO), National Center for Biotechnology Information. Available online: https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/ (accessed on 1 December 2021).
35. Karp, B.E.; Tate, H.; Plumblee, J.R.; Dessai, U.; Whichard, J.M.; Thacker, E.L.; Robertson Hale, K.; Wilson, W.; Friedman, C.R.; Griffin, P.M.; et al. National Antimicrobial Resistance Monitoring System: Two Decades of Advancing Public Health Through Integrated Surveillance of Antimicrobial Resistance. *Foodborne Pathog. Dis.* **2017**, *14*, 545–557. [CrossRef] [PubMed]
36. Kos, V.N.; Déraspe, M.; Mclaughlin, R.E.; Whiteaker, J.D.; Roy, P.H.; Alm, R.A.; Corbeil, J.; Gardner, H. The Resistome of Pseudomonas aeruginosa in Relationship to Phenotypic Susceptibility. *Antimicrob. Agents Chemother.* **2015**, *59*, 427–436. [CrossRef] [PubMed]
37. Harris, P.N.A.; Peleg, A.Y.; Iredell, J.; Ingram, P.R.; Miyakis, S.; Stewardson, A.J.; Rogers, B.A.; McBryde, E.S.; Roberts, J.A.; Lipman, J.; et al. Meropenem versus piperacillin-tazobactam for definitive treatment of bloodstream infections due to ceftriaxone non-susceptible Escherichia coli and Klebsiella spp (the MERINO trial): Study protocol for a randomised controlled trial. *Trials* **2015**, *16*, 24. [CrossRef] [PubMed]
38. Wayne, P. *Performance Standards for Antimicrobial Susceptibility Testing*, 29th ed.; CLSI supplement, M100; Wayne, P., Ed.; Clinical and Laboratory Standards Institute: Annapolis Junction, MD, USA, 2019.
39. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
40. Manni, M.; Berkeley, M.R.; Seppey, M.; Simão, F.A.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **2021**, *38*, 4647–4654. [CrossRef]
41. Mikheenko, A.; Prjibelski, A.; Saveliev, V.; Antipov, D.; Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **2018**, *34*, i142–i150. [CrossRef] [PubMed]
42. Kokot, M.; Dlugosz, M.; Deorowicz, S. KMC 3: Counting and manipulating k-mer statistics. *Bioinformatics* **2017**, *33*, 2759–2761. [CrossRef]
43. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef]
44. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
45. Dunnen, J.T.D.; Antonarakis, S.E. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum. Mutat.* **2000**, *15*, 7–12. [CrossRef]
46. Vis, J.K.; Vermaat, M.; Taschner, P.E.M.; Kok, J.N.; Laros, J.F.J. An efficient algorithm for the extraction of HGVS variant descriptions from sequences. *Bioinformatics* **2015**, *31*, 3751–3757. [CrossRef] [PubMed]
47. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
48. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
49. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.