frontiers in Cellular and Infection Microbiology

BRIEF RESEARCH REPORT
published: 15 February 2021
doi: 10.3389/fcimb.2021.610348

Learning From Limited Data: Towards Best Practice Techniques for Antimicrobial Resistance Prediction From Whole Genome Sequencing Data

Lukas Lüftinger[1,2], *Peter Májek*[1], *Stephan Beisken*[1], *Thomas Rattei*[2] *and Andreas E. Posch*[1*]

[1] Ares Genetics GmbH, Vienna, Austria, [2] Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

OPEN ACCESS

Edited by:
Adrian Egli,
University Hospital of Basel, Switzerland

Reviewed by:
Helena M. B. Seth-Smith,
University Hospital of Basel, Switzerland
Xiaowei Zhan,
University of Texas Southwestern Medical Center, United States
Samuel A. Shelburne,
University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Andreas E. Posch
andreas.posch@ares-genetics.com

Specialty section:
This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: 25 September 2020
Accepted: 11 January 2021
Published: 15 February 2021

Citation:
Lüftinger L, Májek P, Beisken S, Rattei T and Posch AE (2021) Learning From Limited Data: Towards Best Practice Techniques for Antimicrobial Resistance Prediction From Whole Genome Sequencing Data. Front. Cell. Infect. Microbiol. 11:610348. doi: 10.3389/fcimb.2021.610348

Antimicrobial resistance prediction from whole genome sequencing data (WGS) is an emerging application of machine learning, promising to improve antimicrobial resistance surveillance and outbreak monitoring. Despite significant reductions in sequencing cost, the availability and sampling diversity of WGS data with matched antimicrobial susceptibility testing (AST) profiles required for training of WGS-AST prediction models remains limited. Best practice machine learning techniques are required to ensure trained models generalize to independent data for optimal predictive performance. Limited data restricts the choice of machine learning training and evaluation methods and can result in overestimation of model performance. We demonstrate that the widely used random k-fold cross-validation method is ill-suited for application to small bacterial genomics datasets and offer an alternative cross-validation method based on genomic distance. We benchmarked three machine learning architectures previously applied to the WGS-AST problem on a set of 8,704 genome assemblies from five clinically relevant pathogens across 77 species-compound combinations collated from public databases. We show that individual models can be effectively ensembled to improve model performance. By combining models *via* stacked generalization with cross-validation, a model ensembling technique suitable for small datasets, we improved average sensitivity and specificity of individual models by 1.77% and 3.20%, respectively. Furthermore, stacked models exhibited improved robustness and were thus less prone to outlier performance drops than individual component models. In this study, we highlight best practice techniques for antimicrobial resistance prediction from WGS data and introduce the combination of genome distance aware cross-validation and stacked generalization for robust and accurate WGS-AST.

Keywords: machine learning, genomics, antimicrobial resistance, antibiotics, whole genome sequencing (WGS)

INTRODUCTION

Antimicrobial resistance (AMR) is a rising global threat to human health. To ensure the continued efficacy of antimicrobial compounds, prudent use of this resource is crucial (O'Neill, 2016). Accurate determination of antimicrobial resistance *via* antimicrobial susceptibility testing (AST) is crucial to ensure optimal patient treatment as well as to inform antibiotic stewardship and outbreak monitoring.

In this context, resistance predictions from WGS data may effectively complement phenotypic AST: The time-to-result (TTR) of WGS-based workflows is effectively governed by the continuously decreasing cost and runtime of genome sequencing, while phenotypic testing is ultimately limited by the pathogen's growth rate (Bradley et al., 2015; Břinda et al., 2018). Machine learning (ML) algorithms are increasingly applied for prediction of AMR from WGS data (WGS-AST). Recently described WGS-AST techniques use nucleotide k-mer representations of genome assemblies or raw sequencing data, attempting to learn differences in k-mer counts or presence/absence patterns that correlate with shifts in susceptibility to a target antibiotic (Drouin et al., 2016; Aun et al., 2018; Nguyen et al., 2018a; Drouin et al., 2019). This data-driven approach does not require expert knowledge of AMR mechanisms or prior information on AMR genes, and can thus also be applied towards learning of models for novel antibiotics and unknown resistance mechanisms. Other representations of genomic data, such as amino acid k-mers or protein variants have been used for WGS-AST model training as well (Kim et al., 2020; Valizadehaslani et al., 2020).

Challenges arise, however, when learning is not based on features derived from validated, curated AMR markers for the resistance phenotype in question. For example, the significant impact of population structure when applying ML algorithms to WGS-AST data has been noted before (Hicks et al., 2019). Performance of ML models evaluated on isolates from the same experiment as the training data tends to be significantly higher than performance on isolates sampled from independent data sources. Due to limited availability of WGS data coupled with AST information, the performance of WGS-AST models is usually evaluated by cross-validation (CV). Most commonly this is performed using a random splitting criterion, i.e., by dividing samples randomly (Davis et al., 2016; Nguyen et al., 2018a; Drouin et al., 2019). Performance measures obtained by random CV can however only be assumed valid for the larger population if the sample-generating process yields approximately independent and identically distributed (i.i.d.) samples (Ruppert, 2004). This assumption is violated in data points generated by evolutionary processes, which are correlated as a function of the recency of their last common ancestor. This includes, for example, data pertaining to gene function (Tabe-Bordbar et al., 2018) or protein structure (AlQuraishi, 2019), but also whole genomes. By random splitting, similar samples in an existing dependence structure, e.g., evolutionary distance, may be split into the training and test set of CV. This causes the model to overfit by learning features that are spuriously correlated with the phenotype, features which are also present in the test set due to the violated assumption of independence. (Roberts et al., 2017) For example, k-mers mapping to the replication machinery of a resistance cassette-carrying plasmid vector may be highly correlated with resistance due to the prevalence of the plasmid in resistant isolates, despite not contributing to resistance itself. A model overfit to this population by inclusion of such spurious correlations may fail unexpectedly on a population of isolates where the resistance cassette has integrated into the genome. Biological datasets with low sample count but a high number of features further increase the potential of dependence structures and the risk of overfitting (Clarke et al., 2008), and are known to be susceptible to overestimation of model performance by random CV (Roberts et al., 2017).

Ultimately, applying a trained model to multiple large and independently sampled datasets is the gold standard for gauging model generalizability, though this is currently impractical for WGS-AST. To estimate generalization performance in the absence of additional data, blocking CV techniques can be used. Blocking CV seeks to split data into pre-defined similar groups of samples, thus reducing the splitting of dependence structures into the training and test sets of CV (Valavi et al., 2019).

Another significant challenge in achieving robust WGS-AST models with high predictive accuracy is selection of an appropriate learning algorithm. High dimensionality and a low number of training samples constrain the selection of suitable choices. In this study we selected three established learning algorithms which have previously been applied to the WGS-AST problem, and exhaustively benchmarked them across a set of five clinically relevant pathogens (*A. baumannii, E. coli, K. pneumoniae, P. aeruginosa and S. aureus*) and a total of 77 species-compound combinations. We also investigated the possibility of improving model accuracy and robustness by ensembling different learning algorithms such as majority vote and stacked generalization (Wolpert, 1992). This commonly used set of techniques has, to the best of our knowledge, not been explored in the context of antimicrobial resistance prediction from WGS data.

RESULTS

Random CV May Overestimate WGS-AST Model Generalizability

To assess the impact of data splitting techniques on performance estimates of WGS-AST models, we trained extreme gradient boosting (Chen and Guestrin, 2016) models under random and genome distance-aware CV. Genome distance-aware CV attempts to improve independence of test sets by segregating samples based on a known dependence structure in the data, namely genome similarity (see Methods). This mirrors the application of the trained model towards independently sampled datasets, in the absence of actual new data.

Genome assemblies coupled with AST information were obtained from public databases (see Methods) for five human

pathogens (*A. baumannii*, *E. coli*, *P. aeruginosa*, *K. pneumoniae* and *S. aureus*) and a total set of 77 organism/compound combinations. Data was split into 5 CV folds by either a random or genome distance-aware splitting criterion. Random CV splitting was repeated 10 times while varying the random seed to enable significance estimation (see **Supplementary Methods Section 3**). Extreme gradient boosting (XGB) machine learning models were trained on nucleotide k-mer representations of each of the resulting training sets (see Methods) and evaluated on the corresponding test sets.

Of the 77 investigated organism/compound pairs, 60 exhibited significantly higher balanced accuracy (bACC) estimates for random CV than for genome distance-aware CV (**Figure 1**). The average bACC estimated by random CV was 4.45% greater than that of distance-aware CV, indicating that performance estimates by random CV are likely to overestimate the true performance of WGS-AST models on unseen, independent data sampled from a population that is not comprehensively represented in the training data. The observed effect is congruent with published findings of the generalization properties of WGS-AST models applied to independently sampled data (Hicks et al., 2019). To empirically demonstrate that performance estimates by random CV are prone to be overoptimistic we trained XGB models on the full set of *P. aeruginosa* samples and evaluated them on an independent dataset of 140 samples (Ferreira et al., 2020) (see **Supplementary Figure S1**). On average, bACC of the trained XGB models on this test set was 10.12% lower than estimated by random CV. Distance-aware CV provided more conservative estimates while not completely rescuing the overestimation bias, likely due to novel AMR mechanisms associated with the independent dataset (see *Discussion*).

Benchmarking of Machine Learning Algorithms for WGS-AST

We selected three machine learning algorithms for prediction of antimicrobial resistance from WGS data represented as nucleotide k-mer profiles: extreme gradient boosting (XGB) (Chen and Guestrin, 2016), elastic net regularized logistic regression (ENLR) (Friedman et al., 2010), and set covering machine (SCM) (Marchand and Shawe-taylor, 2000). All selected algorithms were recently reported to perform well on the WGS-AST task (Aun et al., 2018; Nguyen et al., 2018a; Drouin et al., 2019; Ferreira et al., 2020; Lees et al., 2020).

Selected algorithms were benchmarked across a set of five clinically relevant bacterial pathogens and a total of 77 organism/compound combinations (**Figure 2A**). Predictive performance across evaluated algorithms was similar, with a median difference between the strongest and weakest model for an organism/compound combination of 4.22% bACC (**Figure 2B**). ENLR, XGB, and SCM algorithms yielded the model with the highest bACC for 34, 28, and 15 datasets, respectively. Despite their characteristically low complexity and high interpretability, SCM models outperformed the more complex ENLR and XGB models on several datasets, particularly when few resistant isolates were available (**Figure 2C**).

Model Stacking Improves Predictive Performance and Robustness of Individual ML Models

To improve predictive performance, we then employed stacking, a model ensembling technique. The ENLR algorithm was used to train a metamodel which learned to optimally combine predictions of individual component XGB, ENLR and SCM

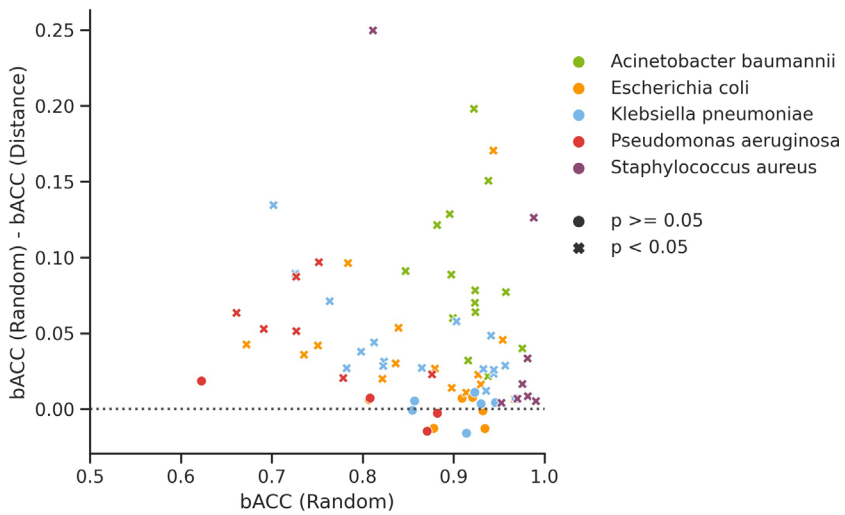

FIGURE 1 | Difference in balanced accuracy (bACC) of XGB models trained and evaluated under random CV and genome distance-aware CV for all considered organism/compound pairs. Significance thresholds are the probability of obtaining bACC estimates as low or lower than the ones from genome distance-aware CV when sampling from a normal distribution fitted to 10 random CV replicates obtained with different random seeds.

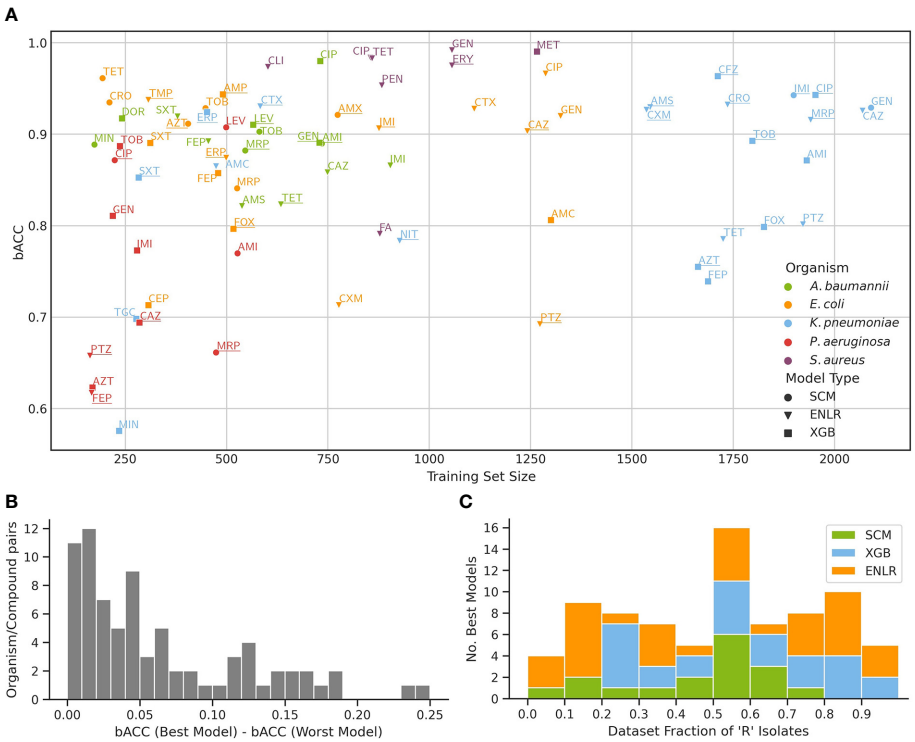

FIGURE 2 | Benchmark of three ML algorithms on the prediction of antimicrobial resistance from WGS data. **(A)** Predictive performance of models for each organism/compound pair as a function of training set size. For each pair, performance of a model with the highest bACC is shown, and underlined if the stacking model outperformed it. The mapping of compound names to compound abbreviations is given in **Supplementary Table S4**. **(B)** Distribution of bACC differences between the models with highest and lowest bACC for all organism/compound pairs. **(C)** Number of top performing models from each algorithm as a function of the fraction of resistant isolates in the training set.

models (**Figure 3** and Methods). We compared the stacked model with a simpler ensembling approach based on the majority vote of all component models. On average, stacked models improved over the sensitivity and specificity of their component models by 1.77% and 3.20%, respectively. The stacking model was found to be the best model by bACC of outer CV in 30 out of 77 organism/compound combinations, outperforming individual component models and the majority vote ensemble. To gauge robustness, we considered a model to have encountered a failure mode if it exhibited a drop in bACC of more than 5.00% compared to the best model for that organism and compound. The stacked models encountered failure modes in 3 out of 77 cases, thus exhibiting superior robustness compared to component models and the simple majority vote ensemble (**Table 1**).

Failure Modes of Component Models and Biological Interpretation

We selected two organism/compound pairs with large differential performance among component models and investigated the biological underpinnings of observed failure modes by annotating k-mers mapping to known AMR biomarkers (Ferreira et al., 2020). For practical reasons, we investigated the models trained in the CV fold exhibiting the largest differential performance and considered only the top 10 most impactful features of each model (see **Supplementary Tables 8** and **9**).

For the combination agent piperacillin and tazobactam (PTZ) in *Klebsiella pneumoniae*, the SCM model exhibited a drop of on average 10% bACC in comparison to XGB and ENLR models. This drop was due to decreased specificity of the SCM model, caused by the model making a comparably larger number of false resistance calls (see **Supplementary Table 6**). Of the four features learned by the model, two mapped to known AMR markers *gyrA* and *catB3*, involved in fluoroquinolone and phenicol resistance, respectively, with no known function in PTZ resistance (Bunny et al., 1995; Drlica and Zhao, 1997). This indicates a strong reliance of the model on features which are spuriously correlated with the phenotype. Conversely, the corresponding XGB model learned multiple k-mers mapping to *blaKPC* beta-lactamase genes, known to confer resistance to piperacillin (Bush and Jacoby, 2010). The stacking model incorporating this SCM model learned to fully disregard the predictions of the SCM model in favor of ENLR and XGB predictions (see **Supplementary Table 7**).

Conversely, for tobramycin (TOB) in *Acinetobacter baumannii*, XGB and ENLR exhibited reduced bACC, mostly due to failure to identify resistant samples in one CV fold. The

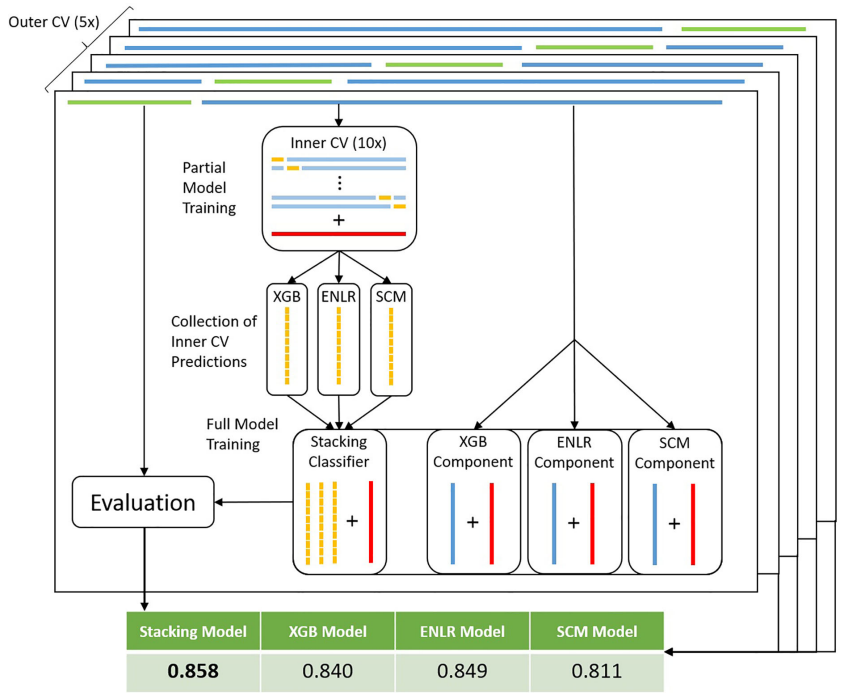

FIGURE 3 | Workflow for model stacking with nested CV. For each training data set in the outer CV loop (dark blue bars on top) complete with true resistance status of samples (red bars), an inner CV loop is run (light blue bars). The full set of predictions (yellow bars) obtained from the test sets of the inner CV are used to train a stacking model to ideally combine predictions from each of the components. At the same time, full component models are trained on the training data set (blue bars within component models). Subsequently, predictions are made by all full component models on the test dataset (green bars on top). Predictions are made by the stacking model using the component model predictions as input features. Finally, performance metrics are obtained by scoring predictions of each model type against the true resistance status of test set samples.

TABLE 1 | Summary statistics of model performance (averaged over organisms and compounds) and number of top-1 placements and failure modes (a more than 5% drop in bACC compared to the best performing model) per organism and compound combination.

Algorithm	bACC	Sensitivity	Specificity	# Top-1 Rankings (bACC)	# Failure Modes Encountered
ENLR	0.849	0.807	**0.890**	15	9
XGB	0.840	0.813	0.866	13	13
SCM	0.811	0.805	0.818	8	31
Majority vote	0.846	0.818	0.873	17	10
Stacking	**0.858**	**0.826**	**0.890**	**30**	**3**

Best metrics in boldface.

SCM model performed consistently well. Feature analysis showed that two of the only three features considered by the SCM model could be mapped to N-Acetyltransferase genes *aadB* and *aacC5*, known to confer resistance to aminoglycosides (Shaw et al., 1993; Cox et al., 2015). The XGB and ENLR models learned a high number of features (512 and 6351, respectively), indicating potential overfitting. In the top 10 features of each, only XGB exhibited interpretable features, namely *aacA16*, an aminoglycoside acetyltransferase, and *msrE*, conferring resistance to erythromycin (Sharkey and O'Neill, 2018). The stacking model learned to assign the highest weight to the SCM component, thereby achieving second place performance after the individual SCM itself (see **Supplementary Table 7**).

DISCUSSION

Random CV May Overestimate WGS-AST Model Generalizability

We demonstrate on a large collection of public datasets that special care must be taken when applying machine learning techniques to the WGS-AST problem. Two common properties of genomics datasets, namely high dimensionality (Clarke et al., 2008) and sparse and biased sampling of the underlying data distribution, invalidate default design choices such as random dataset partitioning for evaluation of generalizability.

Awareness of the issue of splitting data for WGS-AST ML is developing; a recent study (Aytan-Aktug et al., 2020) used

genome clustering based on a similarity threshold, splitting only full clusters into different CV folds together. This approach to data partitioning is also widely used in gene- and protein-based deep learning, where generally only a single training, validation, and test dataset are used (AlQuraishi, 2019; Strodthoff et al., 2019). While grouping by a similarity threshold increases biological meaningfulness and independence of data splits (potentially further reducing performance overestimation), it may cause strongly disbalanced CV fold sizes, especially in a small data regime. The genomic distance-aware method proposed in this work by design generates equally sized folds and aims at maximizing the sample independence across the folds. **Supplementary Figure S3** shows how the proposed method partitions public *P. aeruginosa* samples used in this work.

Similarly, hierarchical clustering has been used for removal of highly clonal genomes from the dataset (Nguyen et al., 2018b), though mainly due to computational considerations. While deduplication is likely to reduce the impact of dependence structures in the training data, the large dimensionality and sparsity of AMR information in a genome represented as k-mer counts makes finding a useful deduplication criterion tricky, especially if the goal is for the model to learn unknown AMR mechanisms.

Of note, data splitting methods controlling for population structure are expected to provide performance estimates differing from random splitting under two conditions: significant population structure must exist in the training dataset, and causal AMR mechanisms must be correlated with population structure. Datasets of closely related samples (not reflecting the true diversity of the underlying population), and datasets containing homogeneously distributed AMR mechanisms, allow only limited insight into possible performance drops due to novel AMR mechanisms associated with distinct populations. Thus, such techniques may still overestimate performance on independently sampled datasets to varying degrees.

Ultimately, a comprehensive assessment of the impact of different clustering and deduplication strategies on model generalizability estimates may be valuable. However, to not only overcome overestimation of performance but to raise predictive accuracy beyond FDA requirements for AST devices (FDA, 2009) and hasten application of WGS-AST models in a diagnostic setting, a greater depth and width of training and test data will be required.

Benchmarking of Machine Learning Algorithms for WGS-AST

Comparing three different ML algorithms, we find that no single algorithm is clearly superior using the respectively chosen feature space, model parametrization and evaluation criteria. While training set size was positively correlated with performance of all investigated algorithms (see **Supplementary Figure S2**), both species identity and antibiotic compound class clearly influenced classifier performance. Previously established findings regarding the significant challenge in providing accurate AMR predictions for *P. aeruginosa* have been affirmed by this work (Aun et al., 2018). Likewise, we obtain high accuracy predictions for *S. aureus* and most antibiotic compounds in *E. coli*, reflecting earlier results obtained with approaches operating on curated sets of AMR markers instead of nucleotide k-mers (Bradley et al., 2015; Moradigaravand et al., 2018). A notable example of the influence of the compound class on prediction accuracy is the consistently high performance of models for resistance to the fluoroquinolones ciprofloxacin (CIP) and levofloxacin (LEV), which is strongly determined by single nucleotide polymorphisms to the DNA gyrase gene *gyrA* and topoisomerase IV gene *parC* (Jacoby, 2005).

Model Stacking Improves Predictive Performance and Robustness of Individual ML Algorithms

Several WGS-AST machine learning techniques have been described in the scientific literature. We demonstrate that individual ML algorithms, while performing similarly on average, are susceptible to different failure modes when applied to the WGS-AST problem, such that no single algorithm is clearly preferable for all organism and compound combinations. We illustrate that a stacking ensemble improves predictive performance and robustness, largely beyond that of any of its component models.

It has been suggested that the use of a diverse set of learning algorithms improves predictive accuracy of ensembling models (Kuncheva and Whitaker, 2003). While we systematically benchmarked three algorithms previously reported to perform well on the problem at hand, adding additional ML architectures to the stack is straightforward and may be a promising next step to further improve predictive accuracy and robustness, even in the absence of additional data. Conversely, we note that in settings where model interpretability is of overriding importance, for example in biomarker discovery, individual highly interpretable models such as the SCM may be preferred over complex model ensembles.

Conclusion

We describe the choice of ML model evaluation strategy and architecture as key aspects affecting model performance and generalizability based on publicly available WGS-AST data sets. To facilitate WGS-AST across organism-compound combinations and translation into clinical practice, applying best practice machine learning techniques and further complementing of publicly available WGS-AST data is important.

MATERIALS AND METHODS

Data Retrieval

Genome assemblies and associated resistance/susceptibility profiles for five clinically relevant pathogens (*A. baumannii, E. coli, K. pneumoniae, P. aeruginosa*, and *S. aureus*) were obtained from public data sources (See **Supplementary Tables 1** and **2**) (Karp et al., ; NCBI NCBI, ; Kos et al., 2015; Wattam et al., 2016; Nguyen et al., 2018a; Mahfouz et al., 2020). Minimum inhibitory

concentration (MIC) values, if present, were interpreted (S/I/R) *via* clinical breakpoints according to CLSI 29 standards (Wayne, 2019). Intermediate phenotypes were treated as resistant for model training and evaluation. Isolates with MIC values less than or equal to a dilution step in the intermediate range (meaning that the MIC interpretive category was ambiguous according to CLSI 29 standards) were treated as susceptible. Data was filtered to pass assembly QC metrics (Ferreira et al., 2020). Finally, only organism-compound pairs were included for which at least 50 susceptible and resistant isolates as well as 200 isolates in total could be retrieved (see **Supplementary Tables 1–3**). Using these cut-offs, a total number of 8704 genome assemblies were retrieved.

Genome assemblies used for evaluation of CV estimates on an independent dataset (Ferreira et al., 2020) were obtained from NCBI (PRJNA553678). AST data were obtained from the authors.

Data Partitioning for Training and Evaluation

Models were trained and evaluated in a nested 10x/5x cross-validation scheme, whereby the inner 10x cross-validation was used to obtain the training features for the stacking model (**Figure 2**).

Genome-distance-based cross-validation folds were created for each species individually such that genome distance was maximized between the test sets of folds (see **Supplementary Methods Section 1**). In short, for all assemblies of each organism, a distance matrix was computed with Mash v2.2 (Ondov et al., 2016). From the distance matrix, two seed samples with the largest genomic distance among them were identified. Subsequently, for each remaining sample, the minimal distance to either of the seeds was computed. Additional seed samples up to the number of desired CV folds were added by selecting samples with the highest minimal distance to existing seeds. Finally, all remaining samples were assigned to seed samples iteratively by assigning to each seed the sample with the lowest genomic distance. The generated five sample groups of even size were used as input to CV. Randomly split CV folds for comparison were created using scikit-learn (Pedregosa et al., 2011).

Feature Creation and Feature Selection

For XGB and ENLR models, feature extraction and selection were performed according to the following procedure. For all training assemblies of each organism, a count matrix of overlapping k-mers of length 15 was built using KMC 3.1.0 (Kokot et al., 2017). Zero-variance k-mers were removed. Out of all k-mers having identical count profiles across training isolates, only a single representative k-mer was retained. Subsequently, for each organism and relevant 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 k-mer feature space was then condensed by univariate feature selection before application of machine learning. K-mers were tested for independence from the S/R category using the χ^2 test as implemented in scikit-learn and filtered by a p-value of $p < 0.05$. Of the k-mers passing this filtering step, at most 1.5 million k-mers with the highest log-odds ratio were retained. For SCM models, k-mer features of length 31 were created from assemblies with Kover2 according to the supplied manual. To exclude the possibility of biases introduced by common feature selection on the full dataset, features for prediction on the test sets of the outer cross-validation were created only at prediction time.

Model Training

We trained extreme gradient boosting (XGB), elastic net regularized logistic regression (ENLR) and set covering machine (SCM) models for prediction of antimicrobial susceptibility from WGS data for a set of five clinically relevant pathogens. A fixed set of hyperparameters was used across all organisms and compound pairs, except for the number of trees in the model which was tuned *via* internal CV. We explored the choice of CV method for hyperparameter optimization and found that the performance estimated by the outer CV method is relatively insensitive to the choice of the inner CV method (see **Supplementary Figure S4**) and thus used a distance-based splitting criterion for internal CV of both XGB and ENLR methods. ENLR models were trained using the glmnet_python package, version 0.2.0 (Friedman et al., 2010), and the hyperparameters lambda and alpha were tuned *via* an internal CV. Set covering machine models were trained with the Kover2 package, version 2.0.3 (Drouin et al., 2019) according to the supplied manual and using risk-bound hyperparameter selection (see **Supplementary Methods Sections 4** and **5**).

Individual models were combined into a stacked model (Wolpert, 1992), with ENLR serving as the learning algorithm. Classically, stacking is achieved using a disjunct mixing set, whereby the predictions of component models on the mixing set serve as the input features on which the stacking classifier is trained. Due to the limited amount of available data, this was achieved here by training partial component models in an inner 10x (distance-based) CV loop (**Figure 3**). Predictions of component models on all test sets were then concatenated into the training features of the stacking model. Predictions with the stacked model were made on the prediction output of the individual, full component models (XGB, ENLR, and SCM) (see **Supplementary Methods Section 2**).

Model Evaluation

Component ML models as well as the stacking model were evaluated in the outer CV loop by predicting the MIC interpretive category (susceptible or resistant) on samples in the test set. Confusion matrices were summed up from outer CV folds. Performance of trained models was evaluated on the balanced accuracy (bACC) metric (Brodersen et al., 2010), as this metric allows evaluation of a model on imbalanced datasets. The bACC is furthermore related to the arithmetic mean of very major error (VME) and major error (ME), two performance criteria commonly applied to AST testing methods. Models created by the individual algorithms (XGB, ENLR, SCM), the

majority vote ensemble model and the stacking model were ranked by counting the number of other models achieving higher bACC on each organism/compound pair.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

LL, PM, SB, TR, and AP devised the study design. LL and PM wrote the code, performed experiments, and analyzed the resulting data. LL wrote the first draft of the manuscript. LL, PM, and SB wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Austrian Research Promotion Agency (FFG) (grants 866389, 874595, and 879570).

ACKNOWLEDGMENTS

We thank Thomas Weinmaier for help with data retrieval, Michael Ante for fruitful discussion of the statistical analysis of results, and Anna Yuwen for critical reading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2021.610348/full#supplementary-material

REFERENCES

AlQuraishi, M. (2019). ProteinNet: A standardized data set for machine learning of protein structure. *BMC Bioinf.* 20, 1–10. doi: 10.1186/s12859-019-2932-0

Aun, E., Brauer, A., Kisand, V., Tenson, T., and Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput. Biol.* 14, 1–17. doi: 10.1371/journal.pcbi.1006434

Aytan-Aktug, D., Clausen, P. T. L. C., Bortolaia, V., Aarestrup, F. M., and Lund, O. (2020). Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. *mSystems* 5, 1–15. doi: 10.1128/msystems.00774-19

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* 6, 1–14. doi: 10.1038/ncomms10063

Břinda, K., Callendrello, A., Cowley, L., Charalampous, T., Lee, R. S., MacFadden, D. R., et al. (2018). Lineage calling can identify antibiotic resistant clones within minutes. *bioRxiv* 403204, 455–464. doi: 10.1101/403204

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *Proc. Int. Conf. Pattern Recognit.* 3121–3124. doi: 10.1109/ICPR.2010.764

Bunny, K. L., Hall, R. M., and Stokes, H. W. (1995). New mobile gene cassettes containing an aminoglycoside resistance gene, aacA7, and a chloramphenicol resistance gene, catB3, in an integron in pBWH301. *Antimicrob. Agents Chemother.* 39, 686–693. doi: 10.1128/AAC.39.3.686

Bush, K., and Jacoby, G. A. (2010). Updated functional classification of β-lactamases. *Antimicrob. Agents Chemother.* 54, 969–976. doi: 10.1128/AAC.01009-09

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. doi: 10.1145/2939672.2939785

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., et al. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nat. Rev. Cancer* 8, 37–49. doi: 10.1038/nrc2294

Cox, G., Stogios, P. J., Savchenko, A., and Wright, G. D. (2015). Structural and molecular basis for resistance to aminoglycoside antibiotics by the adenylyltransferase ANT(2″)-Ia. *MBio* 6, 1–9. doi: 10.1128/mBio.02180-14

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., et al. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci. Rep.* 6, 1–12. doi: 10.1038/srep27930

Drlica, K., and Zhao, X. (1997). DNA gyrase, topoisomerase IV, and the 4-quinolones. *Microbiol. Mol. Biol. Rev.* 61, 377–392. doi: 10.1128/.61.3.377-392.1997

Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., et al. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics* 17, 1–15. doi: 10.1186/s12864-016-2889-6

Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., and Laviolette, F. (2019). Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-40561-2

FDA (2009). *Guidance for Industry and FDA Class II Special Controls Guidance Document : Antimicrobial Susceptibility Test (AST) Systems Preface Public Comment : Additional Copies*. Available at: https://www.fda.gov/medical-devices/guidance-documents-medical-devices-and-radiation-emitting-products/antimicrobial-susceptibility-test-ast-systems-class-ii-special-controls-guidance-industry-and-fda (Accessed December 7, 2020).

Ferreira, I., Beisken, S., Lueftinger, L., Weinmaier, T., Klein, M., Bacher, J., et al. (2020). 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.* 58, 1–11. doi: 10.1128/JCM.00273-20

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Software* 33, 1–22. doi: 10.1016/j.expneurol.2008.01.011

Hicks, A. L., Wheeler, N., Sánchez-Busó, L., Rakeman, J. L., Harris, S. R., and Grad, Y. H. (2019). Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLoS Comput. Biol* 15, 1–21. doi: 10.1101/607127

Jacoby, G. A. (2005). Mechanisms of resistance to quinolones. *Clin. Infect. Dis.* 41, S120–S126. doi: 10.1086/428052

Karp, B. E., Tate, H., Plumblee, J. R., Dessai, U., Whichard, J. M., Thacker, E. L., et al. (2017). National Antimicrobial Resistance Monitoring System: Two Decades of Advancing Public Health Through Integrated Surveillance of Antimicrobial Resistance. *Foodborne Path. Dis.* 14, 545–557. doi: 10.1089/fpd.2017.2283

Kim, J., Greenberg, D. E., Pifer, R., Jiang, S., Xiao, G., Shelburne, S. A., et al. (2020). VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *PLoS Comput. Biol.* 16, e1007511. doi: 10.1371/journal.pcbi.1007511

Kokot, M., Dlugosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761. doi: 10.1093/bioinformatics/btx304

Kos, V. N., Déraspe, M., Mclaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., et al. (2015). The Resistome of Pseudomonas aeruginosa in Relationship to Phenotypic Susceptibility. *Antimicrob. Agents Chemother.* 59, 427–436. doi: 10.1128/AAC.03954-14

Kuncheva, L.II, and Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* 51, 181–207. doi: 10.1049/ic:20010105

Lees, J. A., Galardini, M., Wheeler, N. E., Horsfield, S. T., and Parkhill, J. (2020). Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *MBio* 11, 1–22. doi: 10.1128/mBio.01344-20

Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A., and Posch, A. E. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. *J. Antimicrob. Chemother.* 75, 3099–3108. doi: 10.1093/jac/dkaa257

Marchand, M., and Shawe-taylor, J. (2000). The Set Covering Machine. *J. Mach. Learn. Res.* 1, 723–746. doi: 10.1162/jmlr.2003.3.4-5.723

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Precise prediction of antibiotic resistance in Escherichia coli from full genome sequences. *PLoS Comput. Biol.* 14, 2–17. doi: 10.1101/338194

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R. D., et al. (2018a). Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-017-18972-w

Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., et al. (2018b). Using machine learning to predict antimicrobial minimum inhibitory concentrations and associated genomic features for nontyphoidal Salmonella. *J. Clin. Microbiol.* 57, 380782. doi: 10.1128/JCM.01260-18

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0997-x

O'Neill, J. (2016). Tackling Drug-Resistant Infections Globally. *J. Pharm. Anal.* 6, 71–79. doi: 10.1016/j.jpha.2015.11.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography (Cop)* 40, 913–929. doi: 10.1111/ecog.02881

Ruppert, D. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *J. Am. Stat. Assoc.* 99, 567–567. doi: 10.1198/jasa.2004.s339

Sharkey, L. K. R., and O'Neill, A. J. (2018). Antibiotic Resistance ABC-F Proteins: Bringing Target Protection into the Limelight. *ACS Infect. Dis.* 4, 239–246. doi: 10.1021/acsinfecdis.7b00251

Shaw, K. J., Rather, P. N., Hare, R. S., and Miller, G. H. (1993). Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiol. Rev.* 57, 138–163. doi: 10.1128/mmbr.57.1.138-163.1993

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., et al (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 48, D9–D16. doi: 10.1093/nar/gkz899

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2019). Universal Deep Sequence Models for Protein Classification. *bioRxiv* 704874, 1–11. doi: 10.1101/704874

Tabe-Bordbar, S., Emad, A., Zhao, S. D., and Sinha, S. (2018). A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci. Rep.* 8, 1–11. doi: 10.1038/s41598-018-24937-4

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol. Evol.* 10, 225–232. doi: 10.1111/2041-210X.13107

Valizadehaslani, T., Zhao, Z., Sokhansanj, B. A., and Rosen, G. L. (2020). Amino acid K-mer feature extraction for quantitative antimicrobial resistance (AMR) prediction by machine learning and model interpretation for biological insights. *Biol. (Basel)* 9, 1–92. doi: 10.3390/biology9110365

Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., et al. (2016). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* 45, 535–542. doi: 10.1093/nar/gkw1017

Wayne, P. (2019). *Performance standards for antimicrobial susceptibility testing. 29th ed. CLSI supplement M100* (Wayne, PA: Clinical and Laboratory Standards Institute).

Wolpert, D. (1992). Stacked Generalization. *Neural Networks* 5, 241–259. doi: 10.1016/S0893-6080(05)80023-1

Conflict of Interest: LL, PM, SB, and AP are employed by Ares Genetics GmbH.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.