ARTICLES
https://doi.org/10.1038/s41591-021-01619-9

nature medicine

# Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning

Caroline Weis[1,2] ✉, Aline Cuénod[3,4], Bastian Rieck[1,2], Olivier Dubuis[5], Susanne Graf[6], Claudia Lang[5], Michael Oberle[7], Maximilian Brackmann[8], Kirstine K. Søgaard[3,4], Michael Osthoff[9,10], Karsten Borgwardt[1,2,11] ✉ and Adrian Egli[3,4,11] ✉

**Early use of effective antimicrobial treatments is critical for the outcome of infections and the prevention of treatment resistance. Antimicrobial resistance testing enables the selection of optimal antibiotic treatments, but current culture-based techniques can take up to 72 hours to generate results. We have developed a novel machine learning approach to predict antimicrobial resistance directly from matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) mass spectra profiles of clinical isolates. We trained calibrated classifiers on a newly created publicly available database of mass spectra profiles from the clinically most relevant isolates with linked antimicrobial susceptibility phenotypes. This dataset combines more than 300,000 mass spectra with more than 750,000 antimicrobial resistance phenotypes from four medical institutions. Validation on a panel of clinically important pathogens, including *Staphylococcus aureus*, *Escherichia coli* and *Klebsiella pneumoniae*, resulting in areas under the receiver operating characteristic curve of 0.80, 0.74 and 0.74, respectively, demonstrated the potential of using machine learning to substantially accelerate antimicrobial resistance determination and change of clinical management. Furthermore, a retrospective clinical case study of 63 patients found that implementing this approach would have changed the clinical treatment in nine cases, which would have been beneficial in eight cases (89%). MALDI-TOF mass spectra-based machine learning may thus be an important new tool for treatment optimization and antibiotic stewardship.**

Antimicrobial-resistant bacteria and fungi pose a serious and increasing threat to the achievements of modern medicine[1,2]. Infections with antimicrobial-resistant pathogens are associated with substantial morbidity, mortality and healthcare costs[3]. Rapid treatment with an effective antimicrobial is critical for the outcome of an infection[4,5]. However, antimicrobial therapy and dosage need to account for the resistance profiles of presumed pathogens, and host-specific factors such as patient age, kidney function, previous medical history and concurrent medication also need to be considered. Early identification of the microbial species causing an infection can improve targeting of therapeutic options based on, for example, the knowledge of intrinsic resistance mechanisms and local epidemiology of resistance[6,7]. However, only a detailed resistance profile permits treatments to be fully optimized. With current culture-based methods, the time from sample collection to resistance reporting can take up to 72 hours, meaning that for a substantial period of time a patient may be receiving an antimicrobial drug with either too narrow or too broad a spectrum[8,9]. To limit the infection-related risk to a patient, broad-spectrum antibiotics are often used. The concept of optimal selection of an antibiotic drug is an important pillar of antibiotic stewardship and has gained significant attention owing to the global emergence and spread of antibiotic-resistant pathogens. A reduction in the time required for a resistance profile to become available will not only substantially improve patient outcomes but would also align well with other goals of antibiotic stewardship[10], including reducing the reliance on broad-spectrum antibiotic treatments, reducing unnecessary broad antibiotic use, and thereby combating the development of antibiotic resistance. In addition, rapid information on antimicrobial resistance may help to accelerate infection prevention measures such as the isolation or cohorting of patients infected with presumed multidrug-resistant pathogens. Polymerase chain reaction (PCR)-based molecular diagnostics may be able to detect single resistance genes directly from patient specimens more rapidly than any culture-based diagnostics. However, such molecular assays are generally narrow-spectrum assays of single gene targets and are associated with problems relating to specificity of resistance that is not genetically mediated (for example, upregulation of efflux pumps), and high cost[11–13].

Matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) mass spectrometry enables rapid microbial species identification. In only a few minutes, MALDI-TOF mass spectrometry can be used to characterize the protein composition of single bacterial or fungal colonies[14–16], and the results are available usually within 24 hours after sample collection[7]. MALDI-TOF mass spectrometry enables precise and low-cost microbial identification,

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland. [2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. [3]Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland. [4]Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland. [5]Viollier AG, Allschwil, Switzerland. [6]Department for Microbiology, Canton Hospital Basel-Land, Liestal, Switzerland. [7]Institute for Laboratory Medicine, Medical Microbiology, Cantonal Hospital Aarau, Aarau, Switzerland. [8]Proteomics, Bioinformatics and Toxins, Spiez Laboratory, Federal Office for Civil Protection, Spiez, Switzerland. [9]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel and University of Basel, Basel, Switzerland. [10]Department of Internal Medicine, University Hospital Basel and University of Basel, Basel, Switzerland. [11]These authors contributed equally: Karsten Borgwardt, Adrian Egli. ✉e-mail: caroline.weis@bsse.ethz.ch; karsten.borgwardt@bsse.ethz.ch; adrian.egli@usb.ch

which has led to the technology becoming the most commonly used method for microbial species identification in clinical microbiology laboratories[17]. MALDI-TOF mass spectrometry has the potential to move beyond the simple identification of an infecting pathogen. The extraction of additional information directly from acquired MALDI-TOF mass spectra data may also enable antimicrobial susceptibility testing. Indeed, a recent study used MALDI-TOF mass spectra to detect markers associated with methicillin resistance in clinical samples of *Staphylococcus aureus*[18]. However, the absence of a comprehensive catalog of marker masses for all potential pathogen and drug combinations has made the translation of such efforts to clinical practice impossible. In this study, we harness the full potential of MALDI-TOF MS to predict antimicrobial resistance through machine learning methods. In this context, previous efforts are rare[19,20] and are hampered by the lack of large, publicly available, high-quality benchmark datasets[21,22].

To develop clinically applicable mass spectra-based antimicrobial resistance prediction approaches, we created the Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra (DRIAMS). DRIAMS is a large-scale, publicly available, high-quality collection of bacterial and fungal MALDI-TOF mass spectra derived from routinely acquired clinical isolates, coupled with the respective laboratory-confirmed antibiotic resistance profile. We used DRIAMS to undertake a large-scale study of the utility of such spectra for antimicrobial resistance prediction, with the aim of improving both patient treatment decisions and antibiotic stewardship.

## Results

**DRIAMS: clinical routine database.** From 2016 to 2018 we assembled a dataset of MALDI-TOF mass spectra from more than 300,000 clinical isolates from four different diagnostic laboratories in Switzerland. The raw dataset consists of a total of 303,195 mass spectra and 768,300 antimicrobial resistance labels and represents 803 different species of bacterial and fungal pathogens. The dataset was processed and organized into four subcollections (DRIAMS-A–D; Fig. 1). DRIAMS-A, the largest collection with 145,341 mass spectra, was collected at the University Hospital Basel (Switzerland) and is used for the main analysis presented in this study. DRIAMS-A contains resistance labels associated with 71 different antimicrobial drugs, for which the number of spectra and antimicrobial resistance ratios can be found in Supplementary Tables 1 and 2. Importantly, the MALDI-TOF mass spectra in DRIAMS-A could be obtained from clinical samples within 24 hours after collection, enabling species identification on a rapid scale as compared with standard phenotypic resistance testing (Extended Data Fig. 1). The complete DRIAMS database is publicly available at https://doi.org/10.5061/dryad.bzkh1899q.

**Machine learning for MALDI-TOF mass spectrometry-based resistance prediction.** To move beyond simple species identification, we preprocessed and binned mass spectra measurement points into fixed bins of 3 Da, ranging from 2,000 Da to 20,000 Da, thus obtaining a 6,000-dimensional vector representation for each sample. The selected bin size is sufficiently large to adequately represent each spectrum while still remaining computationally tractable (for details see Methods). Next, we converted the antimicrobial resistance categories, which are either recorded as susceptible, intermediate, or resistant in the laboratory reports associated with each sample, into a binary label (susceptible versus intermediate or resistant) (for details see Methods). Specifically, we assigned intermediate or resistant samples to the positive class, and susceptible samples to the negative class (in most of the scenarios we considered, the positive class was the minority class). We then split the samples into training and testing datasets, ensuring that all data associated with a specific case were either part of the train dataset or the test dataset, but not both, while keeping a similar antimicrobial class ratio in both the train and test datasets. We used three machine learning approaches for classification, that is, logistic regression, gradient-boosted decision trees (LightGBM), and a deep neural network classifier (multilayer perceptron, MLP), to predict resistance to each individual antimicrobial. The three models were selected because they represent different complexity classes of classifiers (a more in-depth description of these approaches is given in Methods). Subsequently, we report the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC) as performance metrics. AUROC can be understood as the probability of correctly classifying a pair of samples, that is, a resistant or intermediate one and a susceptible one; AUPRC quantifies the ability to correctly detect samples from the smaller of the two classes (that is, the ratio of resistant to intermediate, the resistant/intermediate ratio) while minimizing false discoveries. Overall, we observed that LightGBM and MLP were the best-performing classifiers in terms of AUROC. Figure 1 shows the workflow, from data collection and filtering through to spectra processing and the antimicrobial resistance prediction results.

**Species-specific antimicrobial resistance prediction yields high performance.** We first sought to determine whether the use of species-specific mass spectra in DRIAMS-A would result in high predictive performance. To this end, we performed a focused analysis for three clinically important pathogens: *Staphylococcus aureus*, *Escherichia coli* and *Klebsiella pneumoniae*, all of which are on the World Health Organization priority pathogens list[23]. For each of the three species we selected relevant antibiotics to test based on their clinical usage. We then created a DRIAMS-A subset for each antibiotic, which we further divided into stratified training and testing data as described above. For each of the three species we chose one antibiotic resistance as the major scenario of interest, namely, oxacillin as a marker for methicillin-resistant *S. aureus* (MRSA)[24], and ceftriaxone resistance in *E. coli* and *K. pneumoniae* as a marker for resistance against broad-spectrum beta-lactam antibiotics (for example, extended-spectrum beta-lactamases (ESBLs) or carbapenemases). We then trained a classifier using each model for each of the three major species–antibiotic pairs (Supplementary Table 3). We analyzed to what extent the respective best model was capable of predicting resistance to other antibiotics (Fig. 2), observing a high overall performance in both AUROC and AUPRC; the classifier is therefore capable of providing precise antimicrobial resistance predictions. For *S. aureus*, the prediction of oxacillin resistance reached a high performance, with an AUROC of 0.80 and an AUPRC of 0.49 at a positive class ratio (that is, a resistant/intermediate ratio) of 10.0%. The percentage of positive samples in the test dataset (class ratio) reflects the AUPRC performance of a random binary classifier and therefore represents the baseline to which the classification performance needs to be compared (further details are given in Methods). According to clinical laboratory protocols used in DRIAMS-A, for *S. aureus* strains the reported susceptibilities to beta-lactam antibiotics are inferred from the oxacillin susceptibility test results. We also observed a high performance for *E. coli* and *K. pneumoniae*, in that the prediction of ceftriaxone resistance reached an AUROC of 0.74 in both species, and an AUPRC of 0.30 at a positive class ratio of 10.0% for *E. coli* and an AUPRC of 0.33 at a positive class ratio of 8.2% for *K. pneumoniae*. We would expect the generation of such resistance information within 24 hours to have a substantial impact on treatment adaptation and infection prevention management. Overall, this experiment demonstrated that a species-specific classifier can achieve clinically useful prediction performance with significantly faster determination of antibiotic resistance compared with the laboratory standard of phenotypic resistance determination (Extended Data Fig. 1). We also analyzed to what extent the combination of species identity and mass

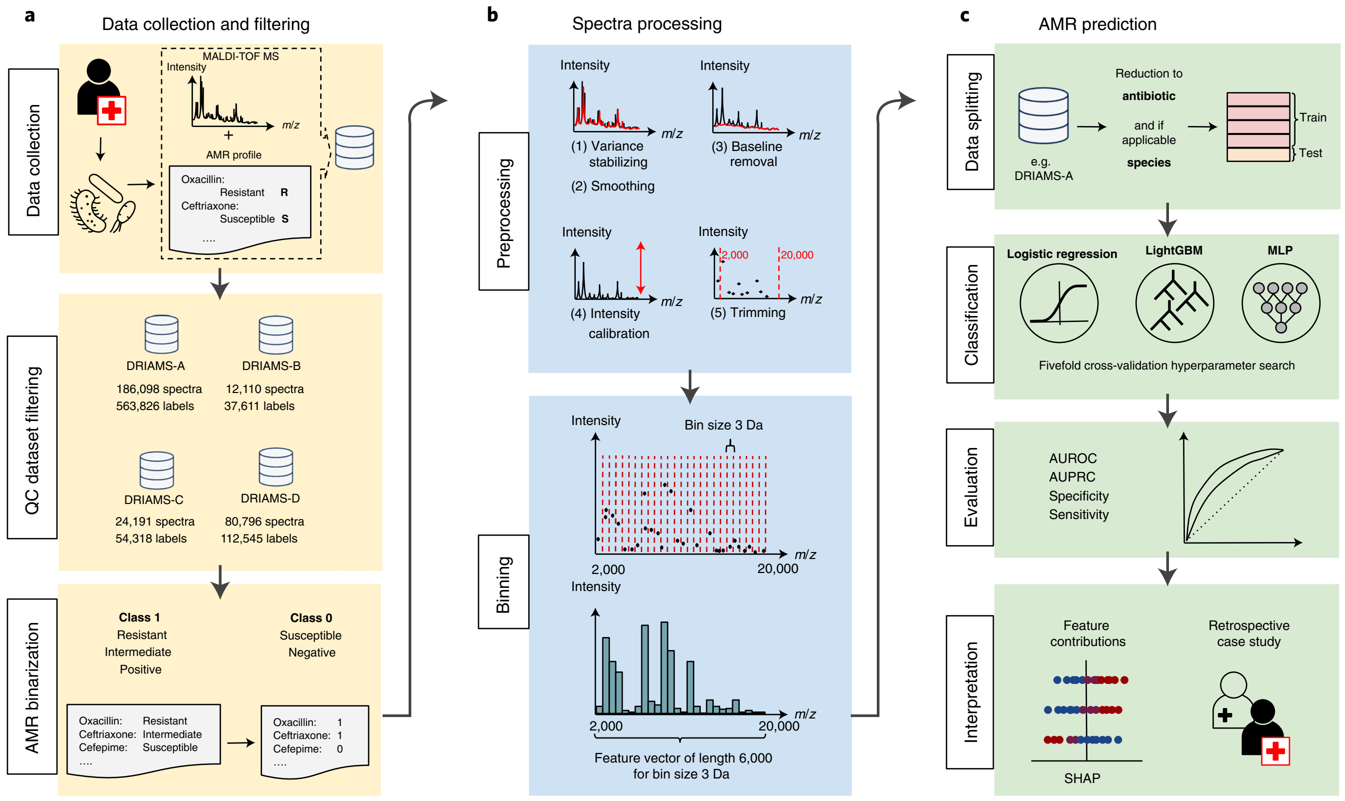


**Fig. 1 | MALDI-TOF mass spectrometry (MS)-based antimicrobial resistance (AMR) prediction workflow. a**, Data collection: samples are taken from infected patients, pathogens are cultured, and their mass spectra and resistance profiles are determined. Spectra and resistance are extracted from the MALDI-TOF MS and laboratory information system; corresponding entries are matched and added to a dataset. Samples are filtered according to workstation. Quality control (QC) dataset filtering: after several QC steps, the datasets are added to DRIAMS. AMR binarization: AMR is defined as a binary classification scenario, with class 1 represented by all labels leading to the antimicrobial not being administered, that is, intermediate or resistant, and positive, while class 0 represents susceptible or negative samples. **b**, Preprocessing: this involves the cleaning of mass spectra. Binning: spectra are binned into equal-sized feature vectors for machine learning. **c**, Data splitting: for the purposes of this study the datasets are reduced to only contain samples belonging to one species. Data are split into 80% training and 20% test, stratified by both antimicrobial class and patient case number. Classification: AMR classifiers are trained with a fivefold cross-validation hyperparameter search, using the classification algorithms logistic regression, LightGBM and a deep neural network classifier (MLP). Evaluation: predictive performance is measured using metrics commonly used in machine learning (AUROC and AUPRC) and by the medical community (specificity and sensitivity). Interpretation: the contribution of individual features to AMR prediction is determined using Shapley values, and clinical impact is assessed using a retrospective case study of samples from the latest 4 months of collected data.

spectrometry information outperforms predictions based on species identity alone. We analyzed AUROC predictive performance for the 42 studied antibiotics (Extended Data Fig. 2). For 31 of them, an AUROC above 0.80 was reached, implying highly accurate predictions. Moreover, for 22 antibiotics, we observed statistically significant improvements in prediction performance using the combined mass spectra in DRIAMS-A as compared with the use of only species information for resistance prediction. The results clearly demonstrate the predictive power of mass spectra-based antimicrobial resistance prediction.

**External datasets improve antimicrobial resistance prediction.** The use of pre-trained machine learning models could expedite uptake of this approach in clinical laboratories already using MALDI-TOF mass spectrometry for species identification. As such, we assessed whether a predictive performance reached using data from one site (for example, one specific hospital) is transferable to other sample collection sites. For the datasets DRIAMS-A–D, each representing one of our four sites, we divided data associated with each case into train and test datasets as described above, and then trained a predictor before testing on each site. We also compared this site-specific training with predictors trained across all sites, and found that site-specific training reaches better predictive performance than across-site validation. With regard to the site-specific training, the large DRIAMS-A dataset is one of the best-performing sites (Fig. 3).

We further investigated whether we could improve prediction for sites where a large dataset is unavailable by leveraging existing large external datasets such as DRIAMS-A. We therefore trained on combinations of training datasets from different sites, including different combinations of the four sites DRIAMS-A–D, and tested on a single external site (that is, DRIAMS-B, -C or -D). Although the transferability of predictive performance from one site to another is an active area of research in the machine learning field of domain adaptation, a recent study[25] has shown that the use of empirical risk minimization by training a single model on pooled data across all training environments often outperforms more complex domain adaptation approaches. The addition of training datasets from other sites to the external site train data was found to be beneficial for validation sites DRIAMS-B and -C (Supplementary Table 4). For the external validation site DRIAMS-D, the best predictive performance was still reached when training exclusively on the site-specific training data. For the external validation site DRIAMS-B, the addition of the large DRIAMS-A dataset proved most beneficial for the scenarios *E. coli* (ceftriaxone) and *K. pneumoniae* (ceftriaxone), while the addition of more data from DRIAMS-C to the training data was beneficial for *S. aureus* (oxacillin).

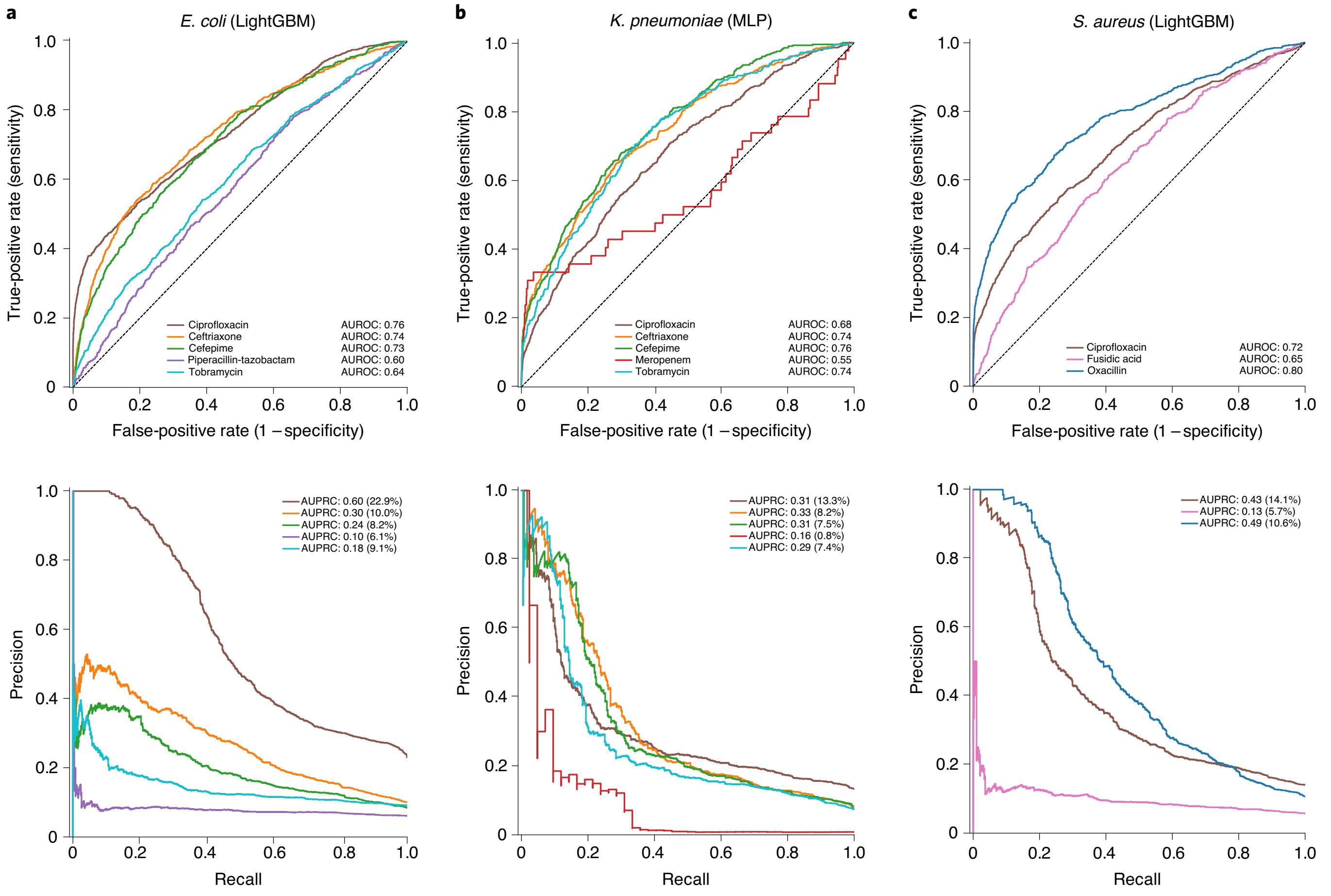


**Fig. 2 | Receiver operating characteristic and precision–recall curves of best-performance AMR prediction models on DRIAMS-A.** The curves were created by appending the scores, while the values displayed in the tables inside the figure stem from reporting the mean for 10 different shuffled stratified train–test splits. **a**, For *E. coli*, the best-performing predictor was that for ciprofloxacin, followed by ceftriaxone, two antibiotics that are critical for indicating an ESBL if resistance is predicted. **b**, For *K. pneumoniae*, cefepime showed the highest performance, with an AUROC of 0.76, also indicative of an ESBL if resistance is predicted. Compared with the other scenarios, its receiver operating characteristic curve has a larger step size, but with more than 500 test samples, the sample size is similar to that of other antibiotics. **c**, Finally, for *S. aureus*, our model performed best for oxacillin, with an AUROC of 0.80. This is particularly relevant give that, for *S. aureus*, the resistance to other beta-lactam antibiotics (including amoxicillin/clavulanic acid and ceftriaxone) is directly derived from oxacillin resistance, indicative of MRSA.

**Species-stratified learning yields superior predictions.** Next, we analyzed whether classifiers can improve the predictive performance by training on a large number of samples from multiple species (as opposed to training on samples from a single species). It is known that different species of bacteria can be resistant to a specific antimicrobial through different mechanisms. For example, resistance against beta-lactam antibiotics in Gram-negative bacteria, such as *E. coli*, may be caused by the production of beta-lactamases, such as CTX-M[26], TEM, and SHV[27,28] or carbapenemases, such as OXA-48[29]. Resistance against beta-lactam antibiotics in Gram-positive bacteria, such as *S. aureus*, can be caused by a penicillinase (blaZ), resulting in a resistance only against penicillin[30], or by an alteration in the penicillin-binding protein (PBP2a), resulting in the MRSA phenotype with resistance against multiple beta-lactam antibiotics[31]. Hence, pooling spectra across species and predicting antimicrobial resistance using the same model regardless of the species poses a more complex learning task than predicting antimicrobial resistance within one specific species. However, stratification by species reduces the number of samples available for training and might therefore lower predictive performance. We assessed the trade-off between the number of available samples and the predictive performance by comparing the performance of a model trained to predict antimicrobial resistance using samples from across all species (ensemble) with that of a collection of models trained separately for single bacterial species (Fig. 4a). In Fig. 4a each point on the curves corresponds to one classifier, trained with the number of samples specified on the *x*-axis. The last, that is, rightmost, point of each curve hence corresponds to the scenario in which all available samples are being used. We observed that training a model for individual species separately led to improved performance for all species, despite the reduction in sample size. Notably, all training samples used to reach the last single-species classification results were also included in the training samples for the last ensemble classifier. The last ensemble classifier therefore had access to at least the same amount of information about the respective species as the last single-species classifier. Nevertheless, it never outperformed the single-species classifiers except for oxacillin resistance in *S. aureus*. Furthermore, a few curves reached a plateau, with the single-species classifier increasing more sharply with the last addition of more training samples. This demonstrates the higher complexity of the ensemble prediction task and the benefit of a larger training dataset, which are critical for capturing different resistance mechanisms.

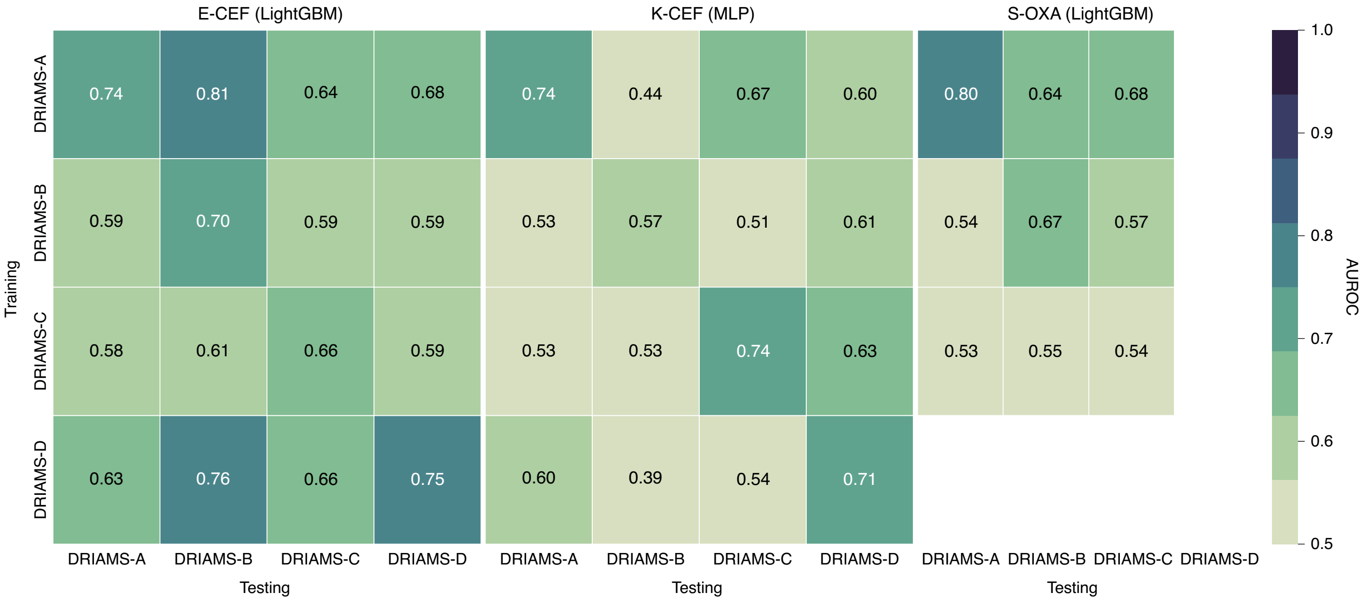


**Fig. 3 | Validation of predictive performance of each scenario trained and tested on DRIAMS-A–D (AUROC).** Shown is the mean AUROC performance of 10 random train–test splits. For comparability, the train–test splits are kept the same for each of the respective four DRIAMS datasets. The values reported on the top right (both training and testing DRIAMS-A) correspond to the values reported in Supplementary Table 3. With the exception of DRIAMS-B *E. coli* (ceftriaxone) (E-CEF), the highest performance is reached when training is performed on the same site as testing. DRIAMS-A and DRIAMS-D exhibit the highest transferability with respect to predictive performance, and overall, transferability seems higher for *E. coli* than for *K. pneumoniae* and *S. aureus*. Due to the different class ratios between test datasets on different sites, AUROC was chosen to enable comparability. K-CEF, *K. pneumoniae* (ceftriaxone); S-OXA, *S. aureus* (oxacillin). For S-OXA no DRIAMS-D data are available.

**Current samples necessary for accurate resistance prediction.** Mass spectra profiles are subject to variations and differences over time, caused by biological differences through the ongoing evolution of the local microbial populations (with new strains being introduced by, for example, traveling), or by technical differences, such as changes after MALDI-TOF mass spectrometer machine maintenance (such as laser replacement and adjustment of internal spectra processing parameters through machine calibration). To guide and encourage further method development, we wanted to illustrate the challenges and limits of mass spectra-based antimicrobial resistance prediction. Hence, we studied whether recent samples are necessary, and whether adding more samples collected at older timepoints would increase predictive performance. We set the latest 4 months of data from DRIAMS-A as a test dataset, and trained classifiers on data collected in 8 month training windows with increasing temporal distance to the test collection window, to simulate the availability of older samples. The training data in the training window were oversampled to match the class ratio in the test data; however, sample sizes could still vary between training windows. We observed a slight decrease in performance with increasing temporal distance between the training and testing data (Fig. 4b) for *E. coli* and *S. aureus*, and a larger decrease for *K. pneumoniae*. We explain this drop by the aforementioned differences that accumulate over time, highlighting the positive effect of having access to recent training samples. Extended Data Fig. 5 also indicates that the reduction in performance training on older datasets could in fact stem from a lower sample size, given that the use of MALDI-TOF mass spectrometry at the DRIAMS-A collection site increased over time.

**Analysis of feature contributions through Shapley values.** Only very few studies have considered full mass spectrum information instead of single peaks for antimicrobial phenotype prediction[19]. We therefore wanted to assess whether predictive performance is primarily driven by only a subset of the peaks, or whether the full spectrum is used. Although this question is partially addressed by the use of feature importance values, their use without additional information can be misleading given that their interpretation is highly contingent on the classifier that was used. Hence, for further analysis, we also calculate the Shapley values, a concept originating from coalitional game theory, which enables the interpretation of model output contributions on both a dataset and per-sample level for each feature[32]. Figure 5 shows the average and per data point Shapley values for the 30 features with the highest average contribution. Given that the tails of the distribution plots for each feature are colored with either the highest or lowest feature value, we see that the predictor is using either the presence of a high intensity value (red) or the absence of any measured intensity (blue) for a positive (resistant/intermediate) class prediction. In the case of *S. aureus* (oxacillin), for the top four mass-to-charge ratio ($m/z$) bins the presence of a peak indicates the positive (resistant/intermediate) class, while for *E. coli* (ceftriaxone) the absence of a peak can also strongly contribute to a positive class prediction. Furthermore, we observe that most of the feature bins with the highest average impact are feature bins with a mass-to-charge ratio ($m/z$) value of less than 10,000 Da (79 out of 90 feature bins in Fig. 5). Most proteins that are reproducibly detected in MALDI-TOF mass spectrometry have a weight less than 10,000 Da[33] and the signal indicates their presence or absence. The distribution of feature importance over all 6,000 features (Extended Data Fig. 4) stemming directly from the classification models indicates that the classifiers utilize the entire range of features.

Most reference studies have focused on oxacillin resistance in *S. aureus* and have identified peaks that were either used to distinguish between methicillin-susceptible *S. aureus* (MSSA) and MRSA or to distinguish between MRSA sub-lineages[34–42]. A subset of these discriminatory peaks was identified to correspond to either constitutively conserved housekeeping genes or to other peptides such as stress proteins or low-molecular-weight toxins[36]. The masses of three identified proteins, 3,007 Da (Delta-toxin), 3,891 Da (uncharacterized protein SA2420.1) and 4,511 Da (uncharacterized protein

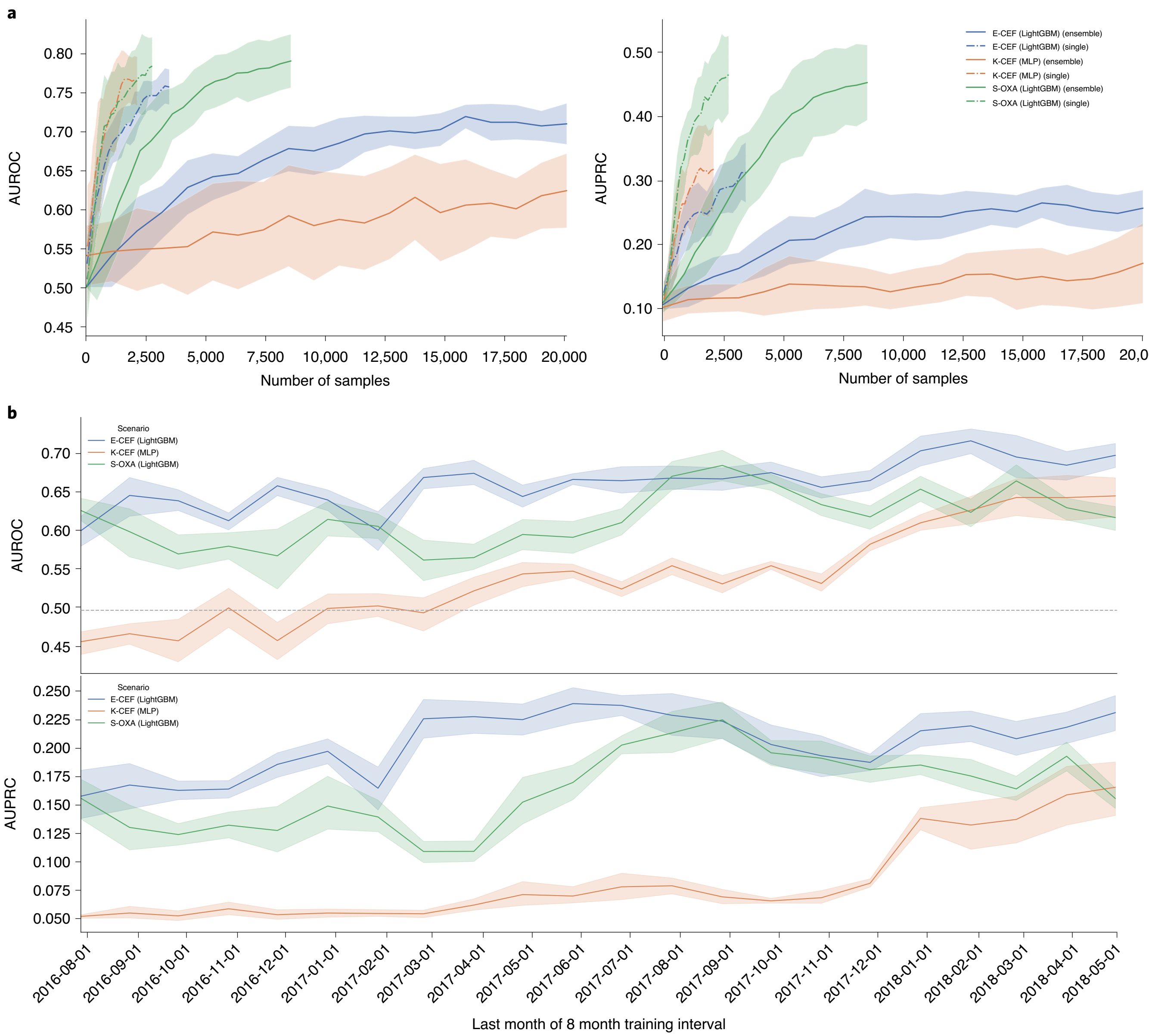


**Fig. 4 | Stability of results with different dataset perturbations. a**, Predictive performance with increasing sample size. AUROC (left) and AUPRC (right) are shown as a function of sample size for complete and species-stratified DRIAMS-A datasets. Experiments were repeated for 10 different shuffled train–test splits. The data are presented as the mean (solid or dashed lines) ± s.d. (shaded region) of these repetitions. Results are shown for the three major scenarios of interest, E-CEF, *E. coli* (ceftriaxone); K-CEF, *K. pneumoniae* (ceftriaxone); and S-OXA, *S. aureus* (oxacillin). At equal sample size, training only on samples from a single species outperforms training in all scenarios. Even for the datasets containing all available samples from the target species (the rightmost points of each curve), the single-species scenario outperforms the ensemble in both E-CEF and K-CEF, while the curves reach a similar predictive performance for S-OXA. **b**, Temporal validation in DRIAMS-A AUROC (upper) and AUPRC (lower) using a sliding 8 month training window with a fixed test set. The test dataset consists of the data collected in the 4 months of May to the end of August 2018. For E-CEF and S-OXA, the predictive performance decreases with increasing temporal distance to the test set, but the fluctuations in the curve are of the same size as the drop over time. The predictive performance for K-CEF decreases more continuously and drastically with increasing temporal distance to the test set.

SAR1012) can be attributed to highly contributing feature bins (Fig. 5 and Supplementary Table 5). A peak at 2,415 *m/z* has previously been identified as MRSA specific[35]. This peak corresponds to the peptide PSM-mec[38], which is encoded on a subset of SCCmec cassettes in close proximity to *mecA*[43,44], which encodes resistance to oxacillin. This peak corresponds to the 83rd highest-ranked feature bin of 2,414–2,417 *m/z* (out of 6,000 feature bins overall) of our respective classifier.

The increased occurrence of multidrug-resistant *E. coli* has been attributed to the spread of a few clonal lineages, in particular to sequence type (ST) 131 (ref. [45]). Previous studies[46,47] have identified ST131 characteristic peaks (8,448 m/z, 8,496 m/z, 11783 m/z), which can be attributed to feature bins being assigned high feature importance and Shapley values by the ceftriaxone *E. coli* classifier.

**Retrospective clinical case study.** To evaluate the clinical benefit of our classifier we evaluated the antibiotic therapy of patients represented in DRIAMS-A, who had invasive serious bacterial infections treated between May and August 2018. We reviewed 416 clinical cases that included positive cultures of *E. coli*, *K. pneumoniae* or

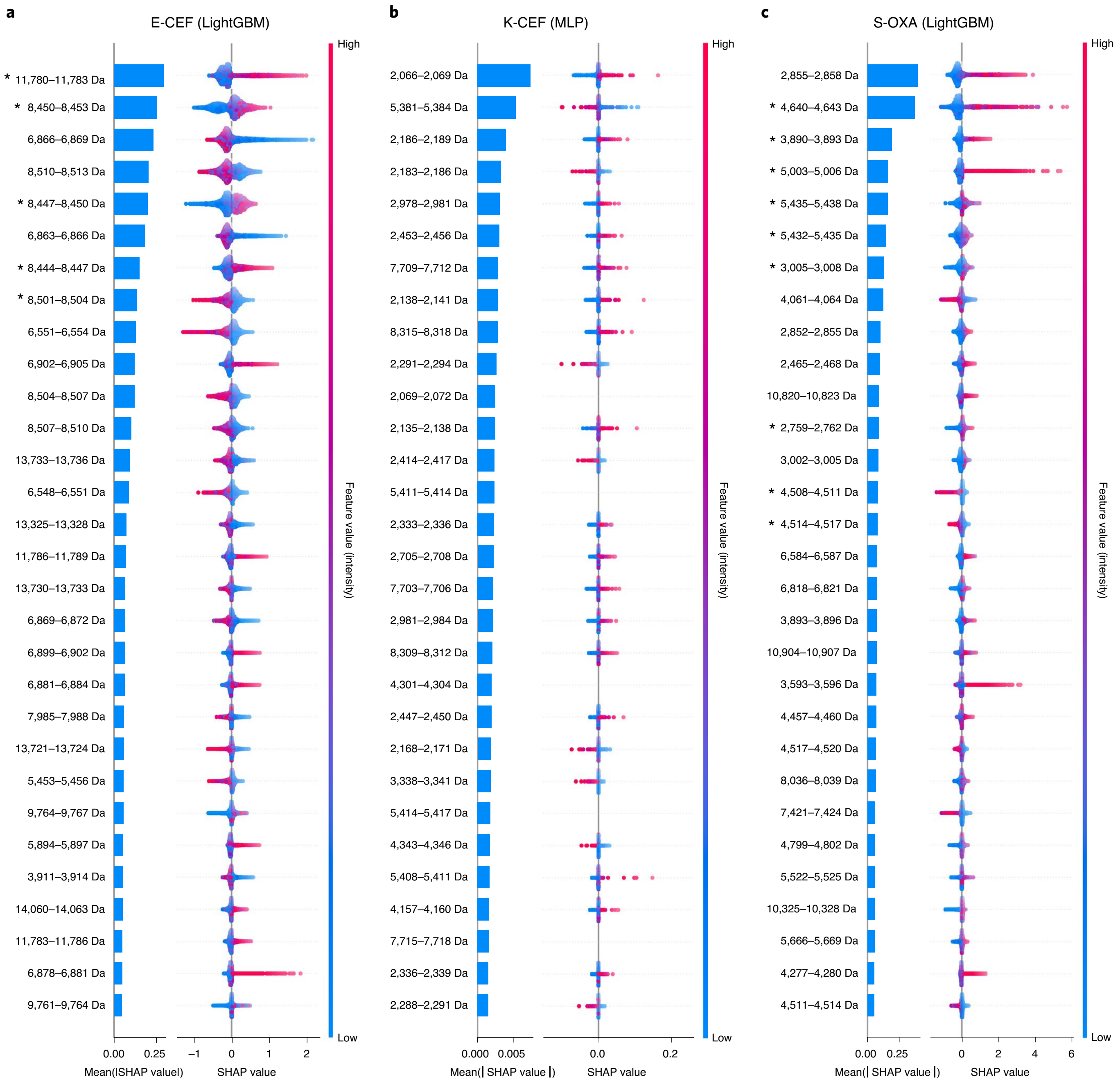


**Fig. 5 | Quantification of feature impact on prediction through analysis of Shapley additive explanations (SHAP) values of the 30 most impactful features. a–c**, For each of the three scenarios E-CEF (LightGBM) (**a**), K-CEF (MLP) (**b**) and S-OXA (LightGBM) (**c**) a barplot on the left indicates the mean Shapley value, that is, the average impact of each feature on the model output magnitude. The scatterplot on the right indicates the distribution of Shapley values, and their impact on the model output, over all test samples. The colors of each test spectrum (according to the color bar: blue for low feature values and red for high feature values) indicate the feature value, that is, the intensity value of the respective feature in the spectrum. The asterisks mark feature bins containing a previously identified protein peak listed in Supplementary Table 5. E-CEF, *E. coli* (ceftriaxone); K-CEF, *K. pneumoniae* (ceftriaxone); S-OXA, *S. aureus* (oxacillin).

*S. aureus* from either blood culture or deep tissue samples. In 63 of these cases an infectious diseases specialist (hereafter referred to as a clinician) was consulted regarding the antibiotic treatment. The consultations occurred after the species had been identified and before the phenotypic antibiotic resistance testing was available (Extended Data Fig. 3). For each case we retrospectively reviewed the recommendations and assessed whether an alternative antibiotic therapy would have been suggested if our classifier had been used at the time at which the MALDI-TOF mass spectrum was acquired.

In 54 of 63 clinical cases the use of the algorithm would not have changed the suggested antibiotic treatment: in 22 cases the clinician suggested de-escalation of the antibiotic regimen to a more narrow-spectrum antibiotic, in 25 cases the suggestion was to continue the current antibiotic regimen, and in seven cases it was to escalate the antibiotic treatment to a broader spectrum antibiotic. The classifier reported an accurate prediction of the antibiotic resistance in 51 of these 54 cases, but given that the decision on antibiotic treatment can be influenced by multiple factors other than the antibiotic resistance of one bacterial species against one antibiotic agent, such as allergy, these did not change the suggested therapy (Fig. 6). In three cases the algorithm predicted susceptibility when phenotypic testing indicated resistance to antibiotics. In none of these three cases, however, would this incorrect prediction have led to a less effective treatment than that suggested without the algorithm. In two of these cases a known MRSA colonization of the patient would have been considered by the clinician, regardless of the prediction of the algorithm. In the third case, *K. pneumoniae* and *E. coli* were both identified in blood culture samples. Given that there were no indications of antibiotic resistance, the clinician would have suggested to keep the current antibiotic treatment against *E. coli* with or without the use of the algorithm, and escalation to a broader spectrum antibiotic was implemented only after phenotypic testing.

In nine cases an alternative antibiotic therapy would have been suggested by the clinician with the use of the classifier at the time of species identification: in seven cases the classifier would have led to a de-escalation of the antibiotic therapy, in one case the use of the algorithm would have changed the suggested treatment to continue the current antibiotic therapy (however, the clinician suggested escalation to a broader spectrum antibiotic agent without the use of the classifier) (Fig. 6), and in one single case the use of the algorithm would have led to an unnecessary escalation of the antibiotic therapy due to a false resistance prediction. In summary, in eight of these nine cases (89%) in which the use of the algorithm would have changed the empiric antibiotic regimen, the classifier correctly predicted susceptibility and this change would have been beneficial and would have promoted antibiotic stewardship, whereas in one case the wrongly predicted resistance would have led to an unneeded escalation of the antibiotic therapy.

## Discussion

We have demonstrated that MALDI-TOF mass spectra-based antimicrobial resistance prediction from routine diagnostic clinical samples is capable of providing accurate predictions within 24 hours after sample collection. This analysis was made possible by collecting the largest real-world clinical dataset of MALDI-TOF mass spectra and corresponding antimicrobial resistance phenotypes. Overall, we observed high predictive performance using calibrated LightGBM and MLP classifiers trained on individual species–antibiotic combinations, such as ceftriaxone resistance in *E. coli* and *K. pneumoniae* and oxacillin resistance in *S. aureus*, obtaining AUROCs larger than 0.70.

We found that the performance of classifiers trained on mass spectra from one site is not generalizable to mass spectra measured at other sites. This may be influenced by many sources, including different phylogenetic strains; a different prevalence of resistance (that is, different class ratios), which can affect predictive performance; or technical variability[48], owing to different machine-specific parameters and settings (that is, batch effects). Similarly, the closer that the time of collection of the training samples is to the time of prediction, the better the predictive power of the trained classifier, probably owing to the same aforementioned reasons. Hence, we would recommend that a clinically applied classifier should be retrained regularly with the most recent data originating from its deployment site. In clinical practice such an algorithm may require regular re-certification. Nonetheless, for individual specific species–antibiotic scenarios, the results suggest that even small sample sizes can lead to high predictive performance.

We demonstrate that to obtain a classifier at a site with a smaller training dataset, combining the available data with an external dataset, such as DRIAMS-A, can increase the training performance. Combining training datasets from different sites increases the sample size, and potentially the coverage of rarer bacterial strains, which improves the predictive performance. However, combining training data originating from different sites also increases the variance in the data, which has the potential to decrease predictive performance. Merging training datasets did not lead to an increased performance on the DRIAMS-D test data, indicating that its outpatient sample pool creates a dataset dissimilar to the hospital datasets. These results emphasize the importance of routine, clinical acquisition of large-scale MALDI-TOF mass spectrometry datasets for antimicrobial resistance prediction: combining large datasets could increase the predictive performance on either prediction site. Furthermore, it is worth noting that all collection sites contributing data to this study are located in a low endemic area for ESBL-producing and MRSA bacteria. Future analysis should assess how data from healthcare centers with a higher burden of antibiotic-resistant bacteria influence the performance of the classifiers.

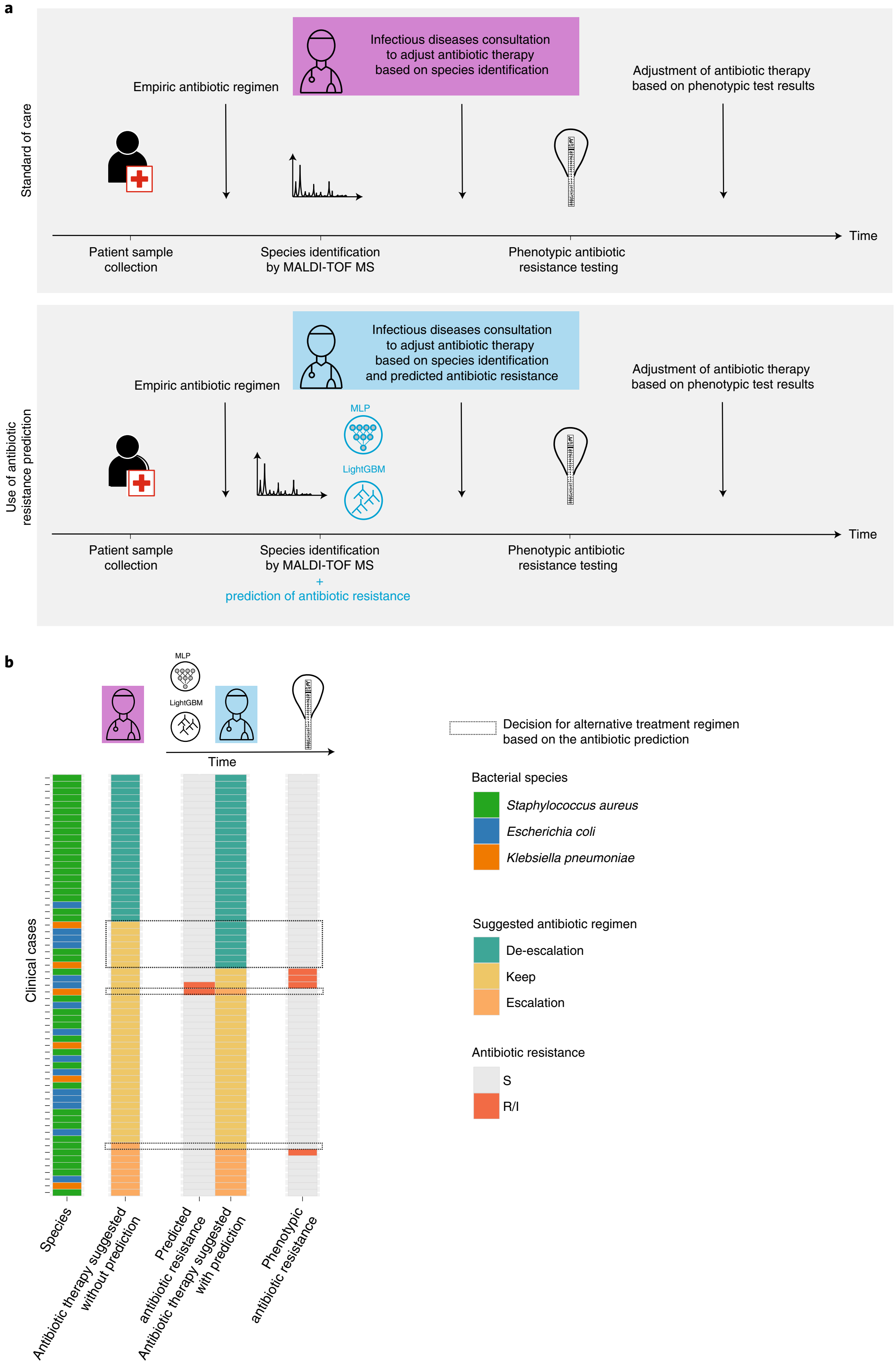
We found the predictive performance of classifiers trained on a single species to be higher than that of classifiers trained on multiple species, indicating the higher complexity of predicting multiple resistance mechanisms. This, combined with the general trend of improved performance if more samples are available, indicates the potential benefits of having access to a large database of MALDI-TOF mass spectra. However, many proteins causing resistance are beyond the effective mass range of MALDI-TOF mass spectra. For example, the penicillin-binding protein in *S. aureus* has a mass of approximately 76,400 Da[49], beta-lactamases in *E. coli* and *K. pneumoniae* weigh approximately 30,000 Da[50–53], and the *E. coli* outer membrane porin OmpC weighs approximately 40,300 Da[54]. Therefore, we propose that our predictor detects resistance-associated changes in the proteome as well as phylogenetic similarity between resistant and susceptible samples. The results also indicate the sample size at which the majority of the information and variance of the samples originating from the DRIAMS-A collection site are covered by the training data, with sample sizes ranging from 2,500 to 5,000 being required to reach the plateau. We therefore suggest collecting a dataset of at least 2,500 samples when working on MALDI-TOF mass spectrometry-based antimicrobial resistance prediction.

Although the antimicrobial resistance classifiers, that is LightGBM and the MLP, are trained and predict resistance labels as a black-box system, analyzing the contribution of each feature bin to the predictive outcome is of utmost importance to explain the antimicrobial resistance predictor decision-making process in a manner that can be interpreted by the user. We therefore determined the feature importance and the Shapley values of each feature bin and compared the results of the highest-weighted bins to known resistance-associated peaks from the literature. The Shapley values indicate that very high or very low feature bin values (corresponding to the presence or absence of a MALDI-TOF mass peak) contribute to the prediction outcome, rather than variations in the feature bin magnitude. This is in line with prior knowledge on MALDI-TOF

**Fig. 6 | Retrospective clinical case study including 63 cases of invasive bacterial infection. a**, Schematic representation of the current standard of care (top row) and the possible use of our classifiers in the clinical workflow (bottom row). **b**, Schematic representation of the review of the 63 cases of invasive bacterial infection. Evaluation of the antibiotic regimen suggested by a clinician without the use of the classifier is shown in column 2; antibiotic resistance predicted by the classifier is shown in column 3; the antibiotic treatment suggested considering the predicted antibiotic resistance, is shown in column 4; and the phenotypically tested antibiotic resistance is shown in column 5. The dashed boxes highlight the cases in which the use of the classifiers would have led to an alternative antibiotic treatment suggestion. De-escalation, change to a more narrow-spectrum antibiotic agent; escalate, change to a broader antibiotic regimen; keep, continue the current antibiotic regimen.

mass spectrometry and confirms that the detection of proteins is responsible for the predictive power, rather than confounding signals or noise.

The literature reference comparison confirms the discriminatory potential of single-feature bins, contributing substantially to our classifiers and also highlighting their generalizability, given that


a
Standard of care
Infectious diseases consultation to adjust antibiotic therapy based on species identification
Empiric antibiotic regimen
Adjustment of antibiotic therapy based on phenotypic test results
Patient sample collection
Species identification by MALDI-TOF MS
Phenotypic antibiotic resistance testing
Time
Use of antibiotic resistance prediction
Infectious diseases consultation to adjust antibiotic therapy based on species identification and predicted antibiotic resistance
Empiric antibiotic regimen
Adjustment of antibiotic therapy based on phenotypic test results
MLP
LightGBM
Patient sample collection
Species identification by MALDI-TOF MS
+
prediction of antibiotic resistance
Phenotypic antibiotic resistance testing
Time
b
MLP
LightGBM
Time
Clinical cases
Species
Antibiotic therapy suggested without prediction
Predicted antibiotic resistance
Antibiotic therapy suggested with prediction
Phenotypic antibiotic resistance
Decision for alternative treatment regimen based on the antibiotic prediction
Bacterial species
*Staphylococcus aureus*
*Escherichia coli*
*Klebsiella pneumoniae*
Suggested antibiotic regimen
De-escalation
Keep
Escalation
Antibiotic resistance
S
R/I

the spectra for these studies were acquired from independent strain collections and on different MALDI-TOF mass spectrometers. Moreover, our classifiers use many more feature bins, for which the discriminatory potential has not previously been identified. An investigation of the protein identity of these yet unknown discriminatory feature bins and their occurrence throughout the respective species would be desirable in the future.

Our retrospective clinical case study shows that our classifier might have a beneficial impact on patient treatment and promote antibiotic stewardship. In 51 of 63 cases, the algorithm supported the treatment regimen suggested by the clinician. In three cases, the inaccurate prediction by our classifier would not have changed the suggested antibiotic regimen, given that the decision is influenced by multiple other factors in addition to the resistance profile towards one antibiotic, such as allergies of the patient, other bacterial species involved in the infection, patient history including the antibiotic profile of previous isolates, and the route of administration of the antibiotic agent. In eight out of 63 cases the accurate prediction by our algorithm would have led to an earlier streamlining of the antibiotic regimen to a more narrow-spectrum antibiotic agent. Similar benefits to antibiotic stewardship have been observed when using genotypic assays such as rapid PCR assays[55]. These findings exemplify the potential of classifiers to optimize antibiotic treatment and assist antibiotic stewardship efforts using real clinical cases. The evaluation of our classifier in prospective clinical studies, on multiple sites with a different prevalence of antimicrobial-resistant bacteria, will be necessary to fully evaluate its clinical impact. Although the prediction of resistance alone would not be used, the prediction may support clinical decision-making that also considers additional patient-related factors.

In summary, our work demonstrates that MALDI-TOF mass spectrometry-based machine learning can provide novel ways to predict antimicrobial resistance in clinically highly relevant scenarios. The results demonstrate the benefit of large sample sizes in predictive performance. Further work could build upon these findings and leverage unlabeled (no antimicrobial resistance profile available) MALDI-TOF mass spectra in DRIAMS for pre-training of a classifier in a semi-supervised fashion before fine-tuning the model on the labeled dataset. In addition to potentially improving the prediction performance, such a training set-up could result in a transfer learning scenario to mitigate batch effects between different collection sites. Although these idiosyncratic challenges need to be overcome, there is also a large potential to improve patient treatment.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-021-01619-9.

Received: 8 August 2020; Accepted: 8 November 2021;
Published online: 10 January 2022

## References

1. World Health Organization. *Global Action Plan on Antimicrobial Resistance* (WHO, 2016).
2. Wise, R. et al. Antimicrobial resistance. Is a major threat to public health. *BMJ* **317**, 609–610 (1998).
3. Cassini, A. et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **19**, 56–66 (2019).
4. Kumar, A. et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**, 1589–1596 (2006).
5. Seymour, C. W. et al. Time to treatment and mortality during mandated emergency care for sepsis. *N. Engl. J. Med.* **376**, 2235–2244 (2017).
6. Huang, A. M. et al. Impact of rapid organism identification via matrix-assisted laser desorption/ionization time-of-flight combined with antimicrobial stewardship team intervention in adult patients with bacteremia and candidemia. *Clin. Infect. Dis.* **57**, 1237–1245 (2013).
7. Osthoff, M. et al. Impact of MALDI-TOF-MS-based identification directly from positive blood cultures on patient management: a controlled clinical trial. *Clin. Microbiol. Infect.* **23**, 78–85 (2017).
8. Banerjee, R. et al. Randomized trial of rapid multiplex polymerase chain reaction-based blood culture identification and susceptibility testing. *Clin. Infect. Dis.* **61**, 1071–1080 (2015).
9. Kommedal, Ø., Aasen, J. L. & Lindemann, P. C. Genetic antimicrobial susceptibility testing in Gram-negative sepsis: impact on time to results in a routine laboratory. *APMIS* **124**, 603–610 (2016).
10. Centers for Disease Control and Prevention. *Core Elements of Antibiotic Stewardship* https://www.cdc.gov/antibiotic-use/core-elements/index.html (2019).
11. Bourdon, N. et al. Rapid detection of vancomycin-resistant enterococci from rectal swabs by the Cepheid Xpert vanA/vanB assay. *Diagn. Microbiol. Infect. Dis.* **67**, 291–293 (2010).
12. Huh, H. J., Kim, E. S. & Chae, S. L. Methicillin-resistant *Staphylococcus aureus* in nasal surveillance swabs at an intensive care unit: an evaluation of the LightCycler MRSA advanced test. *Ann. Lab. Med.* **32**, 407–412 (2012).
13. Cury, A. P. et al. Diagnostic performance of the Xpert Carba-RTM assay directly from rectal swabs for active surveillance of carbapenemase-producing organisms in the largest Brazilian University Hospital. *J. Microbiol. Methods* **171**, 105884 (2020).
14. van Belkum, A., Welker, M., Pincus, D., Charrier, J. P. & Girard, V. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry in clinical microbiology: what are the current issues? *Ann. Lab. Med.* **37**, 475–483 (2017).
15. Croxatto, A., Prod'hom, G. & Greub, G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiol. Rev.* **36**, 380–407 (2012).
16. Dierig, A., Frei, R. & Egli, A. The fast route to microbe identification: matrix assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS). *Pediatr. Infect. Dis. J.* **34**, 97–99 (2015).
17. Hou, T.-Y., Chiang-Ni, C. & Teng, S.-H. Current status of MALDI-TOF mass spectrometry in clinical microbiology. *J. Food Drug Anal.* **27**, 404–414 (2019).
18. Kim, J.-M. et al. Rapid discrimination of methicillin-resistant *Staphylococcus aureus* by MALDI-TOF MS. *Pathogens* **8**, 214 (2019).
19. Weis, C. et al. Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. *Bioinformatics* **36**, i30–i38 (2020).
20. Vervier, K., Mahé, P., Veyrieras, J.-B. & Vert, J.-P. Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. Preprint at https://arxiv.org/abs/1506.07251 (2015).
21. Weis, C. V., Jutzeler, C. R. & Borgwardt, K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clin. Microbiol. Infect.* **26**, 1310–1317 (2020).
22. Wang, H.-Y. et al. A large-scale investigation and identification of methicillin-resistant *Staphylococcus aureus* based on peaks binning of matrix-assisted laser desorption ionization–time of flight MS spectra. *Brief. Bioinform.* **22**, bbaa138 (2021).
23. World Health Organization (WHO). WHO publishes list of bacteria for which new antibiotics are urgently needed. https://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed (2017).
24. Centers for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2019. https://doi.org/10.15620/cdc:82532 (2019).
25. Zhang, H. et al. An empirical framework for domain generalization in clinical settings. In *CHIL '21: Proceedings of the Conference on Health, Inference, and Learning* 279–290 (Association for Computing Machinery, 2021) https://doi.org/10.1145/3450439.3451878
26. Bevan, E. R., Jones, A. M. & Hawkey, P. M. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *J. Antimicrob. Chemother.* **72**, 2145–2155 (2017).
27. Pietsch, M. et al. Molecular characterisation of extended-spectrum β-lactamase (ESBL)-producing *Escherichia coli* isolates from hospital and ambulatory patients in Germany. *Vet. Microbiol.* **200**, 130–137 (2017).
28. Kim, Y.-K. et al. Bloodstream infections by extended-spectrum beta-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* in children: epidemiology and clinical outcome. *Antimicrob. Agents Chemother.* **46**, 1481–1491 (2002).
29. Potron, A., Poirel, L., Rondinaud, E. & Nordmann, P. Intercontinental spread of OXA-48 beta-lactamase-producing Enterobacteriaceae over a 11-year period, 2001 to 2011. *Euro Surveill.* **18**, 20549 (2013).

30. Pereira, L. A., Harnett, G. B., Hodge, M. M., Cattell, J. A. & Speers, D. J. Real-time PCR assay for detection of blaZ genes in *Staphylococcus aureus* clinical isolates. *J. Clin. Microbiol.* **52**, 1259–1261 (2014).
31. Long, S. W. et al. PBP2a mutations causing high-level Ceftaroline resistance in clinical methicillin-resistant *Staphylococcus aureus* isolates. *Antimicrob. Agents Chemother.* **58**, 6668–6674 (2014).
32. Shapley, L. S. 17. A value for n-person games. In *Contributions to the Theory of Games (AM-28)*, Vol. II (eds Kuhn, H. W. & Tucker, A. W.) 307–318 (Princeton University Press, 1953) https://doi.org/10.1515/9781400881970-018
33. Cuénod, A., Foucault, F., Pflüger, V. & Egli, A. Factors associated with MALDI-TOF mass spectral quality of species identification in clinical routine diagnostics. *Front. Cell. Infect. Microbiol.* **11**, 646648 (2021).
34. Camoez, M. et al. Automated categorization of methicillin-resistant *Staphylococcus aureus* clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. *Clin. Microbiol. Infect.* **22**, 161.e1–161.e7 (2016).
35. Josten, M. et al. Identification of agr-positive methicillin-resistant *Staphylococcus aureus* harbouring the class A mec complex by MALDI-TOF mass spectrometry. *Int. J. Med. Microbiol.* **304**, 1018–1023 (2014).
36. Josten, M. et al. Analysis of the matrix-assisted laser desorption ionization–time of flight mass spectrum of *Staphylococcus aureus* identifies mutations that allow differentiation of the main clonal lineages. *J. Clin. Microbiol.* **51**, 1809–1817 (2013).
37. Østergaard, C., Hansen, S. G. K. & Møller, J. K. Rapid first-line discrimination of methicillin resistant *Staphylococcus aureus* strains using MALDI-TOF MS. *Int. J. Med. Microbiol.* **305**, 838–847 (2015).
38. Rhoads, D. D., Wang, H., Karichu, J. & Richter, S. S. The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in staphylococci. *Diagn. Microbiol. Infect. Dis.* **86**, 257–261 (2016).
39. Sauget, M., van der Mee-Marquet, N., Bertrand, X. & Hocquet, D. Matrix-assisted laser desorption ionization–time of flight mass spectrometry can detect *Staphylococcus aureus* clonal complex 398. *J. Microbiol. Methods* **127**, 20–23 (2016).
40. Sogawa, K. et al. Use of the MALDI BioTyper system with MALDI–TOF mass spectrometry for rapid identification of microorganisms. *Anal. Bioanal. Chem.* **400**, 1905–1911 (2011).
41. Wolters, M. et al. MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant *Staphylococcus aureus* lineages. *Int. J. Med. Microbiol.* **301**, 64–68 (2011).
42. Zhang, T. et al. Analysis of methicillin-resistant *Staphylococcus aureus* major clonal lineages by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI–TOF MS). *J. Microbiol. Methods* **117**, 122–127 (2015).
43. Chatterjee, S. S. et al. Distribution and regulation of the mobile genetic element-encoded phenol-soluble modulin PSM-mec in methicillin-resistant *Staphylococcus aureus. PLoS ONE* **6**, e28781 (2011).
44. Hu, Y., Huang, Y., Lizou, Y., Li, J. & Zhang, R. Evaluation of *Staphylococcus aureus* subtyping module for methicillin-resistant *Staphylococcus aureus* detection based on matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Front. Microbiol.* **10**, 2504 (2019).
45. Ludden, C. et al. Genomic surveillance of *Escherichia coli* ST131 identifies local expansion and serial replacement of subclones. *Microb. Genom.* **6**, e000352 (2020).
46. Nakamura, A. et al. Identification of specific protein amino acid substitutions of extended-spectrum β-lactamase (ESBL)-producing *Escherichia coli* ST131: a proteomics approach using mass spectrometry. *Sci. Rep.* **9**, 8555 (2019).
47. Lafolie, J., Sauget, M., Cabrolier, N., Hocquet, D. & Bertrand, X. Detection of *Escherichia coli* sequence type 131 by matrix-assisted laser desorption ionization time-of-flight mass spectrometry: implications for infection control policies? *J. Hosp. Infect.* **90**, 208–212 (2015).
48. Oberle, M. et al. The technical and biological reproducibility of matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) based typing: employment of bioinformatics in a multicenter study. *PLoS ONE* **11**, e0164260 (2016).
49. UniProt. Beta-lactam-inducible penicillin-binding protein. https://www.uniprot.org/uniprot/P07944
50. UniProt. Beta-lactamase OXA-1. https://www.uniprot.org/uniprot/P13661
51. UniProt. Beta-lactamase TEM. https://www.uniprot.org/uniprot/P62593
52. UniProt. Beta-lactamase SHV-24. https://www.uniprot.org/uniprot/Q9S169
53. UniProt. Beta-lactamase CTX-M-1. https://www.uniprot.org/uniprot/P28585
54. UniProt. ompC: outer membrane porin C. https://www.uniprot.org/uniprot/P06996
55. Pickens, C. et al. A multiplex polymerase chain reaction assay for antibiotic stewardship in suspected pneumonia. *Diagn. Microbiol. Infect. Dis.* **98**, 115179 (2020).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Methods

**MALDI-TOF mass spectra acquisition and antimicrobial resistance testing.** We collected data from the daily clinical routine at ISO/IEC 17025 accredited diagnostic laboratories. The study was evaluated by the local ethics committee (IEC 2019-00729). All data used for the machine learning analysis were de-identified prior to analysis. Specifically, all MALDI-TOF mass spectra contained in DRIAMS-A–D were acquired at four microbiological laboratories in Switzerland that provide routine diagnostic services for hospitals and private practices. All laboratories use the Microflex Biotyper System by Bruker Daltonics, which is a widely used MALDI-TOF mass spectrometry system in microbiological routine diagnostics in both North America[16] and Europe[17,18]. The four diagnostic laboratories included in this study are University Hospital Basel (DRIAMS-A), Canton Hospital Basel-Land (DRIAMS-B), Canton Hospital Aarau (DRIAMS-C), and the laboratory service provider Viollier (DRIAMS-D). While the Canton Hospitals Basel-Land and Aarau use the Microflex Biotyper LT/SH System, Viollier uses the Microflex smart LS System. Although these two systems differ in their respective laser gas, they use the same reference spectra database, therefore we included the spectra of both Microflex Biotyper systems. University Hospital Basel uses the two Microflex Biotyper systems in parallel. The species of each mass spectrum was identified using the Microflex Biotyper Database (MBT 7854 MSP Library, BDAL V8.0.0.0_7311-7854, research-use only) included in the flexControl Software (Bruker Daltonics flexControl v3.4). Similar to the mass spectra, antimicrobial resistance profiles were routinely acquired in the same four microbiological laboratories in the same time frames of the dataset. Resistance categories for bacteria were determined using either microdilution assays (VITEK 2, BioMérieux), minimum inhibitory concentration (MIC) stripe tests (Liofilchem) or disc diffusion tests (ThermoFisher Scientific). Resistance categories for yeast were determined using Sensititre Yeast One (Thermofisher). All breakpoint measurements were interpreted to be either susceptible, intermediate or resistant according to EUCAST (European Committee on Antimicrobial Susceptibility Testing)[56] and CLSI (Clinical and Laboratory Standards Institute) (2015 M45; 2017 M60) recommendations. The EUCAST versions used were updated with every EUCAST breakpoints table update and include v6–v8.

**Quality control.** Empty spectra and calibration spectra were excluded from further analysis. This serves to ensure a similar level of data quality for the different sites.

**Matching of MALDI-TOF mass spectra and antimicrobial resistance profiles.** MALDI-TOF mass spectrometry-based antimicrobial resistance prediction requires a dataset containing mass spectra and their corresponding resistance labels, in the form of antimicrobial resistance profiles. To construct such a dataset, MALDI-TOF mass spectrometry and resistance profile measurements belonging to the same microbial isolate must be matched. Given that each site in DRIAMS stores the mass spectra and their corresponding antimicrobial resistance profiles in separate databases, a matching procedure has to be developed for each site.

We use the term 'laboratory report' for the document used to report laboratory measurement results, including antimicrobial resistance profiles, for each patient in clinical care. The species of the specimen is obtained through Bruker Microflex MALDI-TOF mass spectrometry and added to the laboratory report. This decouples laboratory report entry and the mass spectrum; there is no link required between the spectrum file and the laboratory entry after the species is entered. The antimicrobial resistance profiles obtained in their individual experiments are also added to the laboratory report. The laboratory report entries are commonly identified by codes linking them to a patient, or to a unique sample taken from a patient, to which we refer as 'sample ID'. Multiple entries with the same sample ID can exist if several probes were taken from the same patient or several colonies tested from the same probe.

In general, the spectra recorded by the Bruker Microflex systems were labeled with an ambiguous, that is, non-unique, code corresponding to the non-unique sample ID in the laboratory report. MALDI-TOF mass spectra and their corresponding antimicrobial resistance profiles were stored in separate files. In the clinic, MALDI-TOF mass spectra are never intended to be matched up with the laboratory report entries, therefore no established protocols for matching exist. Matching protocols had to be developed uniquely and in an ad-hoc fashion for each labeling system at each institution.

To link mass spectra to their antimicrobial resistance profiles we constructed a unique identifier, using the sample ID and the determined genus of a sample. The rationale behind this strategy is that if multiple sample ID entries exist, this is most probably due to multiple genera being present in the patient samples, leading to several measurements. We omitted samples for which we were unable to construct a unique sample ID–genus pair.

Mass spectra were stored without information on the determined species. Hence, for each spectrum, the species and genus label is determined by re-analyzing the spectra with the University Hospital Basel Bruker library and then matching the spectrum to its corresponding antimicrobial resistance profile using the assigned sample ID and the determined genus. All MALDI-TOF mass spectrometry systems used in this study were maintained according to the manufacturer's standard, and the spectra were routinely acquired using the AutoXecute acquisition mode. The genus is used (instead of species) because it allows for some flexibility between the species assigned to a sample in the laboratory report and the Microflex Biotyper. The species label given in the laboratory report can differ from the species assigned to the corresponding MALDI-TOF mass spectrum by the Microflex Biotyper System given that additional microbiological tests can provide a more accurate label. In what follows, we provide additional details regarding the matching procedure that are specific to each site.

*University Hospital Basel.* Starting in 2015, the spectra were labeled with a 36-position code by the Bruker machine (for example, 022b130c-6c8c-49b5-814d-c1ea8b2e7f93), which we term 'Bruker ID'. This code is guaranteed to be unique for all spectra labeled from one machine. Each antimicrobial resistance profile is labeled with a 6-digit sample ID, which is unique for samples collected in one year. Antimicrobial resistance profiles were collected using the laboratory information system. The laboratory information system includes all entries made for a sample, also entries that have later been corrected and have not been reported or considered for patient treatment. Given that such manual corrections are very rare, the uncertainty in antimicrobial resistance labels is limited. For each year (2015, 2016, 2017 and 2018) there are separate antimicrobial resistance profile tables and folders containing all spectrum samples collected during the corresponding year. We lost 40,569 spectra out of 186,098 by following the aforementioned preprocessing routines (DRIAMS-A).

*Canton Hospital Basel-Land.* The antimicrobial resistance profiles and mass spectra are each labeled with a 6-digit sample ID. The genus shown in each mass spectrum was determined by comparison to the Microflex Biotyper Database (Bruker Daltonics flexControl v3.4); the genus of each antimicrobial resistance profile was stated in the laboratory report. Mass spectra and antimicrobial resistance profiles were merged using the 6-digit sample ID and the genus information.

*Canton Hospital Aarau.* Here, the laboratory report contains the 10-digit sample ID, species label, and antimicrobial resistance profiles of measured samples. This software version did not provide a unique 36-character code for each spectrum, but only a 10-digit sample ID that had to be used to match spectra to the antimicrobial resistance profiles from the laboratory. Given that the sample ID can be shared by different spectra, it cannot be used to uniquely match a species label to an input spectrum. To circumvent this problem, we divided the spectra into 15 batches, each one containing only unique 10-digit sample IDs. Repeated sample IDs were distributed over the batches. These 15 batches were re-analyzed and labeled by the Bruker software, and 15 output files with the given species labeled were created. As a result, the species label for each spectrum in each batch can be determined. The label for each spectrum in the batches can be determined, given that we included only spectra that already had a label in the lab file. Now, each spectrum file has a combined label made up of its 10-digit sample ID and its species label. If this combined label was found to have a unique match in the lab results file, the antimicrobial resistance profile was assigned to the spectrum, otherwise its antimicrobial resistance profile position remained empty and only the spectrum with its species label was added to the dataset. We ignore all spectra that could not be matched to an entry in the laboratory results file (such spectra arise from measurements that do not provide antimicrobial resistance information).

*Viollier.* While all other sites report antimicrobial resistance labels as either 'R' (resistant), 'S' (susceptible), 'I' (intermediate), or as 'positive' or 'negative', the samples provided by Viollier are labeled with precursory measurements, namely the MIC of each antibiotic. We therefore use the breakpoints given in the up-to-date EUCAST guidelines (v9) to convert MIC to RSI.

A total of 80,796 spectra in the `fid` file format are present, identified again through a unique 36-character Bruker ID. The antimicrobial resistance results are identified by a 10-digit sample ID, which are linked to the Bruker IDs in an additional file, the 'linking file'. The main reasons for loss of data in preprocessing are that the antimicrobial resistance results and ID linking files contained significantly fewer entries than the `fid` files present (40,571 and 51,177 respectively), and that following advice by the laboratory personnel, only the 10-digit sample ID could be used for matching to the Bruker ID (which contained a longer version of the laboratory ID). By exclusion of all entries without a unique 10-digit sample ID in both the antimicrobial resistance results and linking files, another significant portion of data was lost. Specifically, there is an overlap of 10,852 filtered entries from the laboratory report file and the linking file. After matching these entries with spectra, 7,771 spectra with 7,720 antimicrobial resistance profiles remained. Spectra without an antimicrobial resistance profile are not used for any supervised learning tasks (such as prediction).

**Hospital hygiene.** The hospital hygiene department specifically screens for multidrug-resistant pathogens to take actions that prevent nosocomial transmission of these. These samples are cultured primarily on selective media containing antibiotics, enabling the growth of resistant strains only.

Growth media have an impact on the bacteria's proteome and thereby on the MALDI-TOF mass spectrum[57]. To avoid that our classifiers recognize media-specific characteristics in the MALDI-TOF mass spectra from the selected media instead of media-independent signatures of non-susceptible bacterial strains, we excluded samples that were collected for the hospital hygiene department from DRIAMS-A for further analysis. The individual sample size per workstation and the predictive performance from MALDI-TOF mass spectra are given in Supplementary Table 6.

**Patient case identification.** For DRIAMS-A, a clinical case was defined as a unique hospital stay, that is, the time frame between the hospital entry and exit of a patient. If a patient was treated at the hospital in 2015 and again in 2018, these were defined as two separate cases. For the retrospective clinical analysis, infections with different bacterial species and different patient isolation materials during the same hospital stay were regarded as different entities, given that different species might require different antibiotic therapy. For DRIAMS-B, DRIAMS-C and DRIAMS-D, no information regarding clinical cases was provided and therefore patient case information is not considered during analysis.

**Dataset characteristics.** All four of the medical institutions are located in Switzerland. Microbial samples in the University Hospital Basel database (that is, DRIAMS-A) mostly originate from patients in the city of Basel and its surroundings. Such patients visit the hospital for either outpatient or inpatient treatment. Samples in the Canton Hospital Basel-Land dataset (that is, DRIAMS-B) primarily originate from the towns surrounding the city of Basel. Patients from the Swiss Canton Aargau seek medical care at the Canton Hospital Aarau (DRIAMS-C). Viollier (DRIAMS-D) is a service provider that performs species identification for microbial samples collected in medical practices and hospitals. Samples originate from private practices and hospitals all over Switzerland.

The DRIAMS-A–D datasets contain data collected as part of the daily clinical routine. All mass spectra measured in a certain time frame are included. The time frame during which each dataset was collected is as follows: DRIAMS-A, 34 months (November 2015–August 2018); DRIAMS-B, 6 months (January 2018–June 2018); DRIAMS-C, 8 months (January 2018–August 2018); and DRIAMS-D, 6 months (January 2018–June 2018).

**Spectral representation.** In the DRIAMS dataset, we include mass spectra in their raw version without any preprocessing, and bin them using several bin sizes. After initial analysis, a bin size of 3 Da was used for all machine learning analyses in this study. This bin size is small enough to allow for separation of mass peaks (for which the exact mass-to-charge position can vary slightly due to measurement noise), while being large enough not to impede computational tractability. The spectra are extracted from the Bruker Flex machine in the Bruker Flex data format. The following preprocessing steps are performed using the R package `MaldiQuant`[58] v1.19: (1) the measured intensity is transformed with a square-root method to stabilize the variance, (2) smoothing using the Savitzky–Golay algorithm with half-window-size 10 is applied, (3) an estimate of the baseline is removed in 20 iterations of the SNIP algorithm, (4) the intensity is calibrated using the total ion current (TIC), and (5) the spectra are trimmed to values in a 2,000–20,000 Da range. For exact parameter values, please refer to the code.

After preprocessing, each spectrum is represented by a set of measurements, each of which is described by its corresponding mass-to-charge ratio and intensity. However, this representation results in each sample having potentially a different dimensionality (that is, cardinality) and different measurements being generally irregularly spaced. Given that the machine learning methods used in this manuscript require their input to be a feature vector of fixed dimensionality, intensity measurements are binned using the bin size of 3 Da. To perform the binning, we partition the $m/z$ axis in the range from 2,000 to 20,000 Da into disjoint, equal-sized bins and sum the intensity of all measurements in the sample (that is, a spectrum) falling into the same bin. Thus, each sample is represented by a vector of fixed dimensionality, that is, a vector containing 6,000 features, which is the number of bins that the $m/z$ axis is partitioned into. We use this feature vector representation for all downstream machine learning tasks.

**Antimicrobial resistance phenotype binarization.** For the machine learning analysis, the values of antimicrobial resistance profiles were binarized during data input to have a binary classification scenario. The categories are based on EUCAST and CLSI recommendations. For tests that report RSI values, resistant (R) and intermediate (I) samples were labeled as class 1, while susceptible (S) samples were labeled as class 0. We grouped samples in the intermediate class together with resistant samples because both types of samples prevent the application of the antibiotic. In EUCAST v6–v8 the intermediate category has higher MICs but, due to safety reasons, in clinical practice this was usually classified as resistance to ensure a suitable safety buffer when dealing with high antibiotic drug concentrations.

**Statistical methods.** If not otherwise indicated, solid lines and performance metrics displayed in figures and tables refer to the mean performance over the test sets of a fivefold cross-validation. Shaded areas and numbers added with ± signs refer to the standard deviation across the respective evaluation metric.

The center line of the boxplot in Extended Data Fig. 1 shows the median, and the lower and upper limits show the first quartile (Q1) and third quartile (Q3), respectively. The lower whisker shows the lowest time requirement until diagnostic result, and the upper whisker shows the largest time recorded (excluding outliers).

The bars in Extended Data Fig. 2 state the mean performance over the test sets of a 10-fold cross-validation. The asterisks marking the bars indicate a statistically significant difference between the reported metrics between all species and species information alone of a two-sided Welch's $t$-test, without assuming equal population variance, at a significance level of 0.05.

**Machine learning methods.** For antimicrobial resistance classification we used a set of three state-of-the-art classification algorithms with different capabilities. These were logistic regression, LightGBM[59] (a modern variant of gradient-boosted decision trees), and a multilayer perceptron deep neural network (MLP). For LightGBM we use the official implementation in the `lightgbm` package, while we use the scikit-learn package for all other models[60]. These models cover a large spectrum of modern machine learning techniques, with logistic regression representing an algorithm from classic statistics, the training process for which can be regularized. LightGBM, by contrast, represents a modern variant of tree-based learning algorithms that focuses specifically on good scalability properties while maintaining high accuracy. Finally, MLPs constitute a simple example of deep learning algorithms. Although they have the highest complexity in terms of computing resources and data requirements than the aforementioned models, deep learning methods can be effective in uncovering complex relationships between input variables.

For each antibiotic, all samples with a missing antimicrobial resistance profile were removed and the machine learning pipeline was applied to the reduced dataset. Samples were randomly split into a training dataset comprising 80% of the samples and a test dataset with the remaining 20%, while stratifying for the class and the species, and ensuring that a sample with a specific patient case is either part of the train dataset or the test dataset, but not both. This step ensures that sample measurements of the same infection (that are probably very similar to each other) are not causing information leakage from training to testing. This is slightly unusual in standard machine learning set-ups, which typically only require stratification by a single class label, but is crucial for our scenarios to guarantee similar prevalence values. To select an appropriate model configuration for a specific task, we use fivefold cross-validation on the training data; in the case that an insufficient number of samples is available, our implementation falls back to a threefold cross-validation on the training dataset to optimize the respective hyperparameters. The hyperparameters are model specific (see below for more details), but always include the choice of an optional standardization step (in which feature vectors are transformed to have zero mean and unit variance). To determine the best-performing hyperparameter set, we optimized the AUROC on the training dataset only. This metric is advantageous in our scenario because it is not influenced by the class ratio and summarizes the performance of correct and incorrect susceptibility predictions over varying classification score thresholds. Having selected the best hyperparameters, we retrain each model on the full training dataset, and use the resulting classifier for all subsequent predictions. Our hyperparameter grid is extensive, consisting of, for example, the choice of different logistic regression penalties (L1, L2, no penalty), the choice of scaling method (standardization or none), and regularization parameters ($C \in \left\{10^{-3},\ 10^{-2},\ \cdots 10^{2},\ 10^{3}\right\}$). For more details please refer to our code (`models.py`).

We implemented all models in Python and published them in a single package (https://github.com/BorgwardtLab/maldi_amr), which we modeled after scikit-learn, a powerful library for machine learning in Python.

**Evaluation metrics.** We report AUROC as the main metric of performance evaluation. The datasets of most antibiotics under consideration have a high class imbalance (20 out of 42 antibiotics have a resistant/intermediate class ratio <20% or >80%). AUROC is invariant to the class ratio of the dataset and therefore permits a certain level of comparability between antibiotics with different class ratios. A pitfall of reporting AUROC in the case of unbalanced datasets, however, is that it does not reflect the performance with respect to precision (or positive predictive value). Therefore, the AUROC can be high while precision is low. To account for this bias, we additionally report the AUPRC; this metric, however, is not used during the training process.

Two other metrics commonly used in clinical research are sensitivity and specificity. Analogous to the receiver operating characteristic curves, we show sensitivity versus specificity curves to illustrate the trade-off between both metrics. Please note commonalities to other metrics: sensitivity, recall and true-positive rate are synonyms and all correspond to the same metric; specificity is a counterpart to the false-positive rate, that is, true-positive rate = 1 − specificity.

*Connection to confusion matrix.* All of the metrics we use here can be derived from the counts in a confusion matrix.

The AUROC shows the true-positive rate (that is, true positive/(true positive + false negative)) against the false-positive rate (that is, false positive/(false positive + true negative)). The AUPRC (as well as the AUROC) is traditionally reported on the minority class. In our scenario, however, although the minority class is the resistant class in most cases, this is not consistent, and for some antibiotics more samples of the resistant will be present. The precision–recall curve shows the recall (true positive/(true positive + false negative)) against the precision (true positive/(true positive + false positive)). The average performance of a random classifier would be 0.5 for AUROC and the percentage of samples of the positive (susceptible) class for AUPRC.

**Shapley values for interpretability analysis.** To improve the interpretability of our classifiers, we calculated Shapley values using the shap package. This package directly supports the explanation of many common machine learning techniques. We used the standard algorithms of the shap package to explain the outputs of our logistic regression and LightGBM models. For the MLP, the gradient-based explanation techniques were unable to be used because of the large memory requirements of the algorithms. We therefore opted to follow common practice and subsample the input dataset, reducing it to 50 barycenters, that is, samples that express most of the variability in the data, via k-means clustering. This enabled us to obtain per-sample Shapley values that contain the relevance of individual features with respect to the overall output of the model.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The full datasets generated during and analyzed during the current study are available in the Dryad repository, https://doi.org/10.5061/dryad.bzkh1899q.

## Code availability

All R and Python scripts can be found in https://github.com/BorgwardtLab/maldi_amr under a BSD 3-Clause License.

## References

56. European Committee on Antimicrobial Susceptibility Testing (EUCAST), European Society of Clinical Microbiology and Diseases. *Clinical Breakpoints and Dosing of Antibiotics* http://www.eucast.org/clinical_breakpoints/ (2021).
57. Schmidt, A. et al. The quantitative and condition-dependent *Escherichia coli* proteome. *Nat. Biotechnol.* **34**, 104–110 (2016).
58. Gibb, S. & Strimmer, K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics* **28**, 2270–2271 (2012).
59. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).
60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

## Acknowledgements

The authors thank O. Grüninger, J. Reist, Y. Mahiddine, D. Lang, C. Straub and M. Schneider for excellent technical assistance in accessing the mass spectra and antimicrobial resistance data at the University Hospital Basel; and B. Rodríguez-Sánchez (Hospital General Universitario Gregorio Marañón, Madrid, Spain), H. Seth-Smith (University Hospital Basel), C. Kaufmann (University Hospital Basel), and V. Pflüger (MabritecAG, Riehen, Switzerland) for valuable advice, technical consultation and feedback throughout the project. The authors also thank X. He (ETH Zurich) for fruitful discussions, and D. Chen (ETH Zurich) and Jacques Schrenzel (Geneva University Hospitals) for assistance with the manuscript and for providing valuable feedback. This study was supported by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung (K.B.), by the two Cantons of Basel through a D-BSSE-Uni-Basel Personalised Medicine grant from the ETH Zurich (PMB-03-17, K.B. and A.E.) and a Doc.Mobility fellowship (A.C.) by the Swiss National Science Foundation (P1BSP3-184342). This study received ethics approval through the local ethics review board (EKNZ number: IEC 2019-00729, 2019-01860 and 2019-00748).

## Author contributions

C.W., B.R. and K.B. designed the machine learning experiments; C.W. and B.R. implemented all experiments of the machine learning analysis; A.E., A.C. and K.K.S. organized data collection; A.C., S.G., O.D., C.L. and M.O. extracted clinical data; A.C. and M.O. performed the retrospective clinical case study; A.C. and C.W. implemented the preprocessing of the datasets DRIAMS-A and DRIAMS-B; C.W. implemented the preprocessing of the datasets DRIAMS-C and DRIAMS-D; M.B. contributed to mass spectrometry data interpretation; M.O. and A.E. provided feedback on the clinical implications of resistance predictions; C.W., B.R. and A.C. designed all display items; C.W., B.R., A.C., K.B. and A.E. wrote the manuscript with the assistance and feedback of all of the co-authors. K.B. and A.E. conceived and supervised the study.

## Competing interests

The authors declare no competing interests.

## Additional information

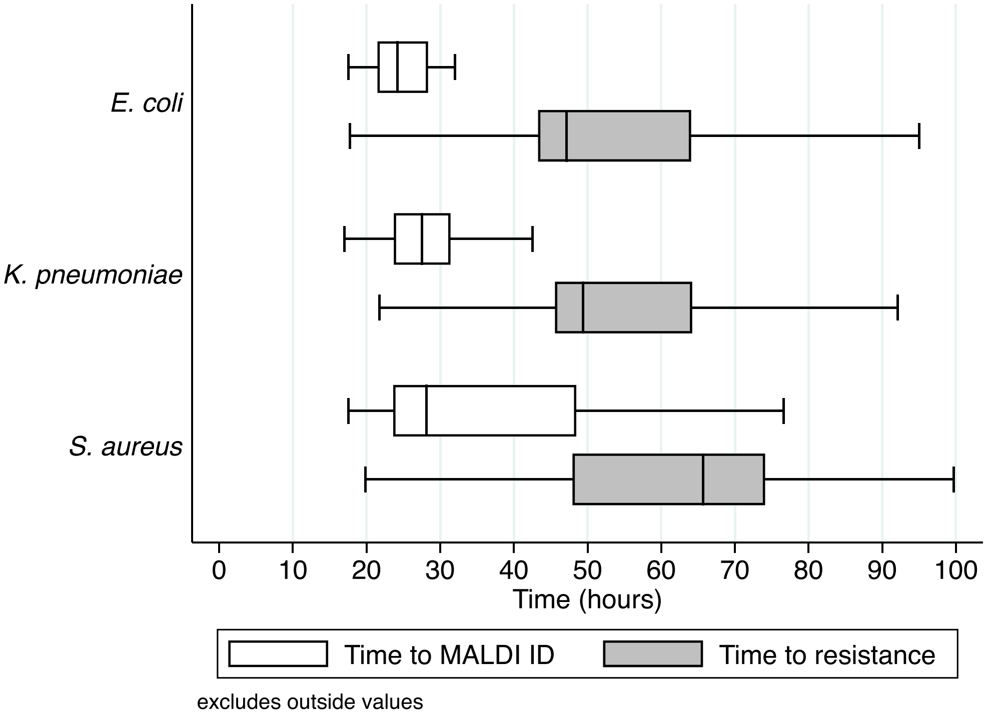
**Extended data** are available for this paper at https://doi.org/10.1038/s41591-021-01619-9.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41591-021-01619-9.

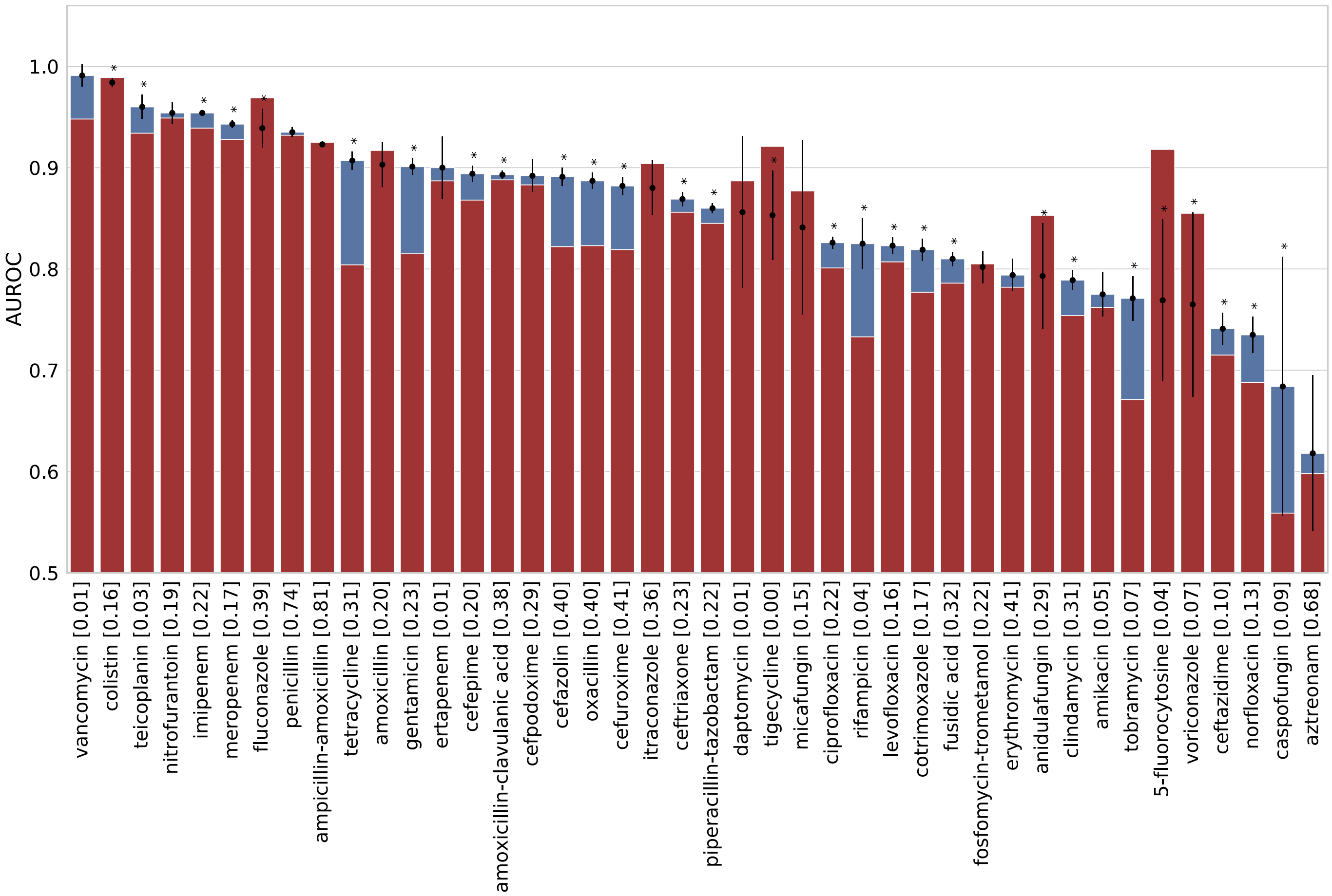
**Correspondence and requests for materials** should be addressed to Caroline Weis, Karsten Borgwardt or Adrian Egli.

**Peer review information** *Nature Medicine* thanks Roman Yelensky and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Alison Farrell was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

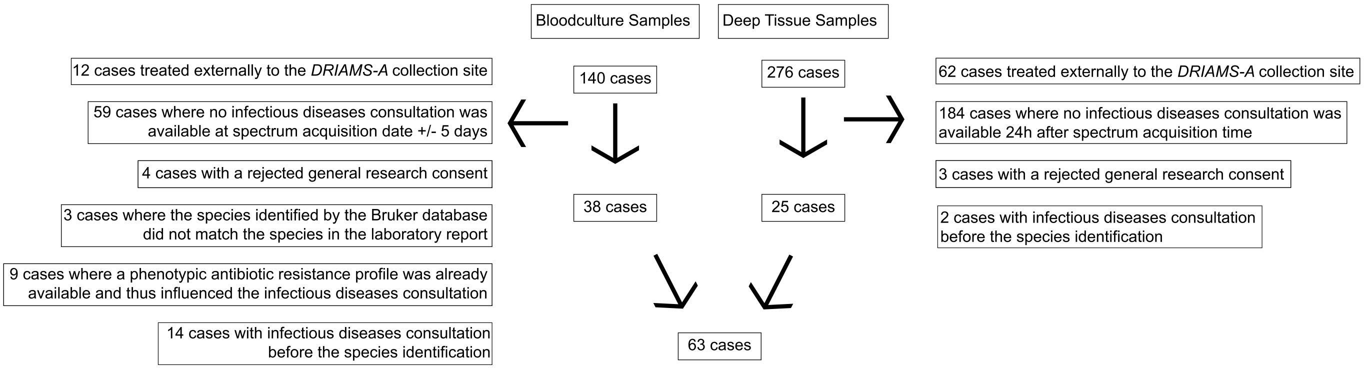
**Reprints and permissions information** is available at www.nature.com/reprints.



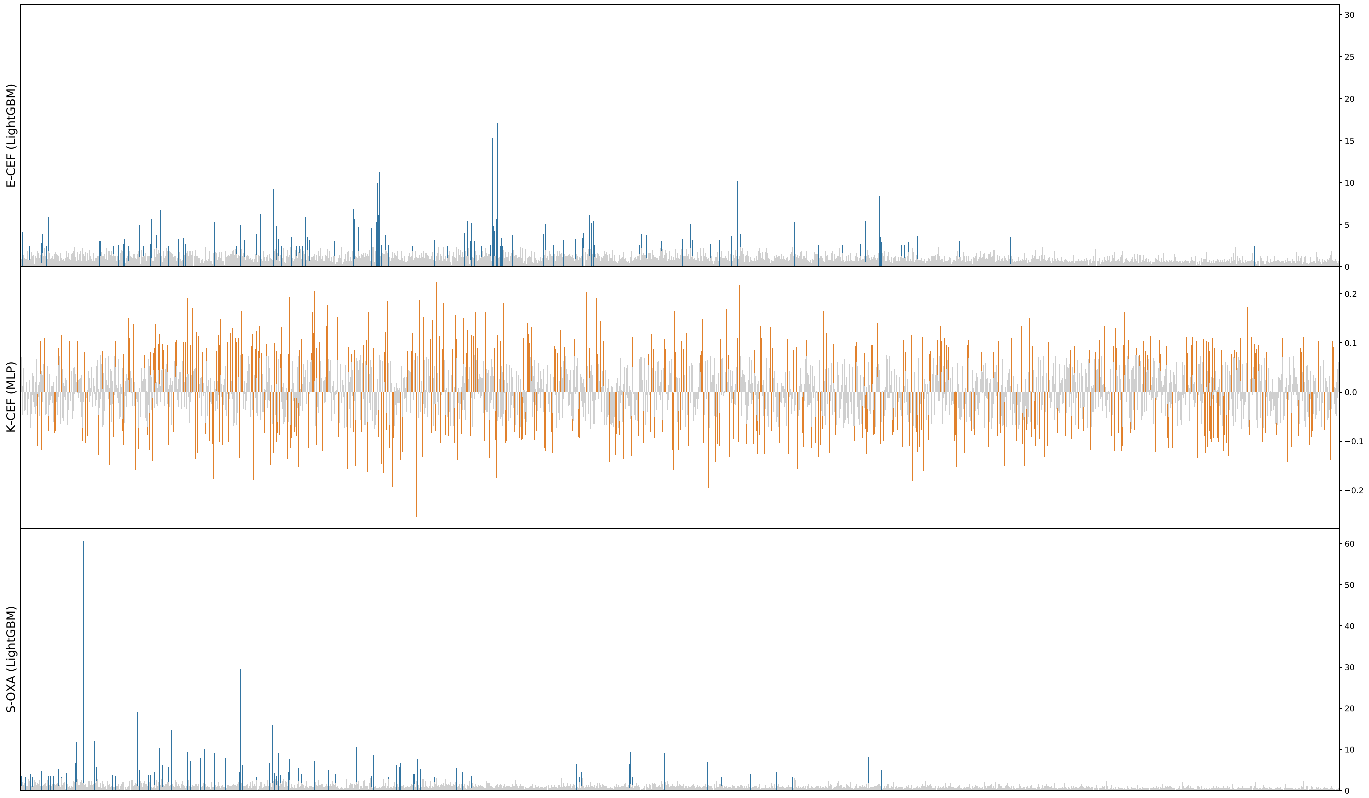
**Extended Data Fig. 1 | Comparison turnaround times between MALDI-TOF MS and resistance.** Time from the entry of a patient sample at the diagnostic laboratory at the *DRIAMS-A* collection site to species identification by MALDI-TOF MS and phenotypic resistance testing for three clinically relevant species: *E. coli* (n = 54), *K. pneumoniae* (n = 66), and *S. aureus* (n = 57). Boxplot shows median and interquartile time ranges in hours, whiskers indicate adjacent values.



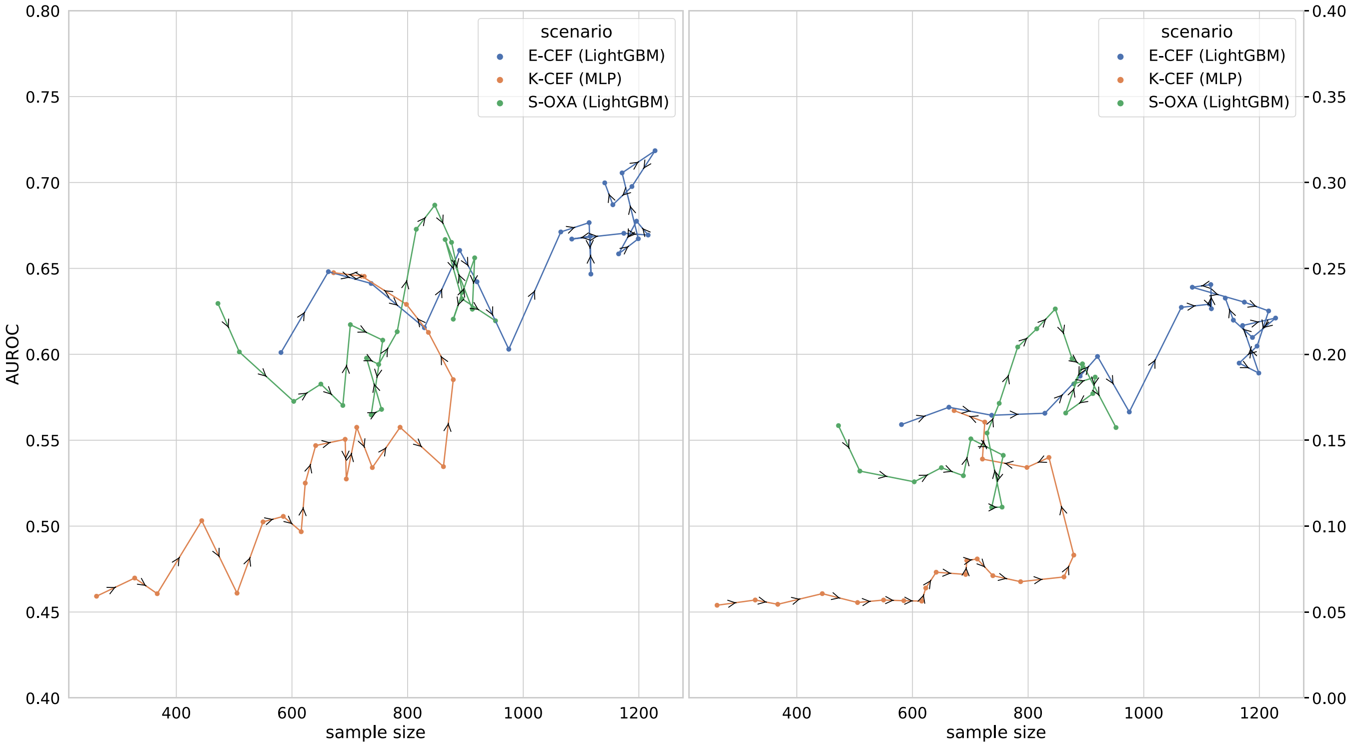
**Extended Data Fig. 2 | Improved antimicrobial resistance prediction based on MALDI-TOF mass spectra combining all species compared to species information alone.** AUROC values of logistic regression classifiers trained on data combining all samples with labels available for each antimicrobial prediction task in *DRIAMS-A*. The blue bars depict predictive performance using spectral data as features. The red bars show the predictive performance when using species label information only. The fractions of resistant/intermediate samples in the training data are indicated in brackets after the antibiotic name. Reported metrics and error bars are the mean and standard deviation of 10 repetitions with different random train–test-splits. The asterisks indicate a statistically significant difference between the reported metrics between all species and species information alone of a two-sided Welch's t-test (not assuming equal population variance) and a significance level of <0.05.


Bloodculture Samples
Deep Tissue Samples
140 cases
276 cases
12 cases treated externally to the *DRIAMS-A* collection site
59 cases where no infectious diseases consultation was available at spectrum acquisition date +/- 5 days
4 cases with a rejected general research consent
3 cases where the species identified by the Bruker database did not match the species in the laboratory report
9 cases where a phenotypic antibiotic resistance profile was already available and thus influenced the infectious diseases consultation
14 cases with infectious diseases consultation before the species identification
62 cases treated externally to the *DRIAMS-A* collection site
184 cases where no infectious diseases consultation was available 24h after spectrum acquisition time
3 cases with a rejected general research consent
2 cases with infectious diseases consultation before the species identification
38 cases
25 cases
63 cases

**Extended Data Fig. 3 | Flowchart inclusion of cases into the retrospective clinical study.** We reviewed 416 clinical cases which had a severe bacterial infection with *K. pneumoniae*, *E. coli* or *S. aureus* between April and August 2018. Cases were excluded if (i) cases were treated external to the *DRIAMS-A* collection site, (ii) no consultation note by a infectious diseases specialist was available within 5 days (for cases with a positive blood culture) or 1 day (for cases with a positive deep tissue sample), (iii) the general research consent was rejected, (iv) the species identified by the Bruker database did not match the species in the laboratory report, (v) the antibiotic resistance profile was already present at the at the time of the infectious diseases consultation and (vi) if the consultation note was written without the knowledge of the species identity. 63 clinical cases were included.



**Extended Data Fig. 4 | Barplot of feature importances of LightGBM and MLP model.** Importance values larger than two times absolute standard deviation are colored in either blue (LightGBM) or orange (MLP). The sign of each feature importance value of the MLP model indicates the association with the positive (positive sign) or negative class. The LightGBM values indicate the contribution to the prediction without direction of association. All three models indicate that a large number of features are relevant for an accurate antimicrobial resistance prediction. The scenario abbreviations follow Fig. 2a.



**Extended Data Fig. 5 | Temporal validation including sample size of training window.** The timepoints correspond to points in Fig. 4 and arrow directions indicate time progression. With time progression both the trends in sample size per 8-month time window and the predictive performance increase. The scenario abbreviations follow Fig. 2a.

nature portfolio

Corresponding author(s): Caroline Weis
Karsten Borgwardt
Adrian Egli

Last updated by author(s): Oct 28, 2021

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

Data collection: MALDI-TOF mass spectra were collected using flexControl Software (Bruker Daltonics flexControl v.3.4) and the species of each mass spectrum was identified using the Microflex Biotyper Database (MBT 7854 MSP Library, BDAL V8.0.0.0_7311-7854 (RUO)).

Data analysis: All R and Python scripts can be found at https://github.com/BorgwardtLab/maldi_amr. No commercial software was used to analyze the data. The software and package versions are: R-3.4.1, Python 3.7.7 (incl. sklearn 0.24.2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets generated during and analyzed during the current study are fully available through the Dryad repository at https://doi.org/10.5061/dryad.bzkh1899q.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | We have not specifically determined a sample size, but accessed all available MALDI-TOF mass spectral data from the included health care institutes. More than 300,000 mass spectra were accessed for this study. We have generated specific analysis to show the effect of increasing sample size (see manuscript for more details). |
| Data exclusions | We excluded measurements for which no respective antimicrobial profile could be obtained. Such sample would not be assigned a label in our analysis. We state the number of excluded spectra in the supplement; the data files list all included spectra. |
| Replication | We used different repetitions (with different random seeds that were chosen after method development had ceased). All experiments are repeated either 5 or 10 times (depending on the data available in each task) to estimate the generalization error. Moreover, fingerprints of the input files are stored with the output of each experiment (which we provide in the associated repository). We made our code public and documented it in order to ensure/facilitate replicability. |
| Randomization | We used a stratified randomized train-test split to split samples (using fixed random seeds to ensure reproducibility). The stratification is a custom implementation ensuring both stratification by resistance class while also ensuring that measurements of one hospital stay from a patient are present either only in train or only in test, to prevent information leakage (see manuscript for more details). We controlled for no other covariates. For the training of all classifiers, we used cross-validation. |
| Blinding | The study data was split into training and validation dataset and additional independent validation dataset accessed from three other healthcare institutions. No information making the patient identifiable is included in the data files; they record the MALDI-TOF MS and resistance testing of microbial isolates together with codes only identifiable from within the protected hospital software system. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study | |
|---|---|---|
| ☒ | □ | Antibodies |
| ☒ | □ | Eukaryotic cell lines |
| ☒ | □ | Palaeontology and archaeology |
| ☒ | □ | Animals and other organisms |
| ☒ | □ | Human research participants |
| ☒ | □ | Clinical data |
| ☒ | □ | Dual use research of concern |

## Methods

| n/a | Involved in the study | |
|---|---|---|
| ☒ | □ | ChIP-seq |
| ☒ | □ | Flow cytometry |
| ☒ | □ | MRI-based neuroimaging |