nature portfolio

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

Supplementary information

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

In the format provided by the authors and unedited

antimicrobial class	number of samples in class	average positive sample ratio	antimicrobials
OTHER BETA-LACTAM ANTIBACTERIALS	174826	0.23	cefuroxime, cefepime, ceftriaxone, cefazolin, meropenem_without_meningitis, cefpodoxime, ertapenem, meropenem_with_pneumonia, meropenem, ceftazidime, cefixime, imipenem, meropenem_with_meningitis, cefoxitin_screen, aztreonam
BETA-LACTAM ANTIBACTERIALS, PENICILLINS	107128	0.5	penicillin_with_meningitis, penicillin_with_endokarditis, penicillin, ampicillin-amoxicillin, amoxicillin-clavulanic acid_uncomplicated_hwi, penicillin_with_other_infections, penicillin_without_endokarditis, amoxicillin, oxacillin, penicillin_with_pneumonia, amoxicillin-clavulanic acid, piperacillin-tazobactam
OTHER ANTIBACTERIALS	83998	0.102	fusidic acid, nitrofurantoin, colistin, metronidazole, fosfomycin-trometamol, vancomycin_grd*, daptomycin, linezolid, vancomycin, teicoplanin, teicoplanin_grd*, fosfomycin
QUINOLONE ANTIBACTERIALS	58054	0.214	moxifloxacin, levofloxacin, ciprofloxacin, norfloxacin, quinolones
AMINOGLYCOSIDE ANTIBACTERIALS	46799	0.109	tobramycin, gentamicin, amikacin, gentamicin_high_level, aminoglycosides
SULFONAMIDES AND TRIMETHOPRIM	26640	0.183	cotrimoxazole
MACROLIDES, LINCOSAMIDES AND STREPTOGRAMINS	23099	0.357	clarithromycin, azithromycin, erythromycin, clindamycin
TETRACYCLINES	22377	0.142	minocycline, tetracycline, doxycycline, tigecycline
DRUGS FOR TREATMENT OF TUBERCULOSIS	10976	0.049	rifampicin, rifampicin_1mg-l**
ANTIMYCOTICS FOR SYSTEMIC USE	5594	0.17	amphotericin b, micafungin, 5-fluorocytosine, caspofungin, voriconazole, anidulafungin, fluconazole, itraconazole, posaconazole
ANTIBIOTICS FOR TOPICAL USE	3815	0.007	mupirocin
AMPHENICOLS	203	0.138	chloramphenicol
total	**563509**	**0.244**	

Supplementary Table 1: Categorisation into antimicrobial class, average number of spectra and average resistance positive class ratio of 71 antimicrobials contained in *DRIAMS-A*. The values for each individual antimicrobial can be found in **Suppl. Tab. 2**. * vancomycin_grd and teicoplanin_grd is a MIC Strip Test used at University Hospital Basel in very rare cases to detect glycopeptide intermediate *S. aureus*; ** '1mg_l' indicates the concentration of rifampicin, when MIC are measured in liquid culture as it is routinely done for *Mycobacterium tuberculosis* (MTB). These entries of non-MTB species were entered incorrectly into the laboratory information system.

species	antibiotic	scenario abbreviation	model	AUROC	AUPRC
E. coli	ceftriaxone	E-CEF	LightGBM	**0.74 ± 0.02**	**0.30 ± 0.03**
			LR	0.70 ± 0.03	0.24 ± 0.02
			MLP	0.68 ± 0.03	0.22 ± 0.03
K. pneumoniae	ceftriaxone	K-CEF	LightGBM	0.67 ± 0.03	0.24 ± 0.05
			LR	0.68 ± 0.04	0.26 ± 0.07
			MLP	**0.74 ± 0.04**	**0.33 ± 0.07**
S. aureus	oxacillin	S-OXA	LightGBM	**0.80 ± 0.03**	**0.49 ± 0.06**
			LR	0.75 ± 0.04	0.37 ± 0.07
			MLP	0.79 ± 0.03	0.46 ± 0.10

Supplementary Table 3: A. Comparison of performance of three machine learning models. Metrics are reporting the mean ± standard deviation for 10 different shuffled stratified train–test splits. For *S. aureus* (oxacillin) and *E. coli* (ceftriaxone), the best predictive performance is reached with LightGBM; for *K. pneumoniae* (ceftriaxone) with the multi-layer perceptron (MLP). Additionally, the abbreviations for species-antibiotic scenarios are introduced.

A.

Test site: DRIAMS-B (class 1 ratio 20.9%)			Train site: DRIAMS-A	Train site: DRIAMS-B	Train sites: DRIAMS-A DRIAMS-B	Train sites: DRIAMS-A DRIAMS-C DRIAMS-D*	Train sites: DRIAMS-A DRIAMS-B DRIAMS-C DRIAMS-D*
species	**antibiotic**	**model**					
E. coli	ceftriaxone	LightGBM	0.60 ± 0.13	0.55 ± 0.20	**0.66 ± 0.12**	0.63 ± 0.14	0.62 ± 0.11
K. pneumoniae	ceftriaxone	MLP	0.23 ± 0.12	0.29 ± 0.14	**0.37 ± 0.17**	0.20 ± 0.11	0.33 ± 0.18
S.aureus	oxacillin	LightGBM	0.24 ± 0.12	0.30 ± 0.20	0.45 ± 0.24	0.25 ± 0.11	**0.48 ± 0.18**

B.

Test site: DRIAMS-C (class 1 ratio 16.3%)			Train site: DRIAMS-A	Train site: DRIAMS-C	Train sites: DRIAMS-A DRIAMS-C	Train sites: DRIAMS-A DRIAMS-B DRIAMS-D*	Train sites: DRIAMS-A DRIAMS-B DRIAMS-C DRIAMS-D*
species	**antibiotic**	**model**					
E. coli	ceftriaxone	LightGBM	0.31 ± 0.06	0.34 ± 0.06	0.39 ± 0.05	0.35 ± 0.06	**0.40 ± 0.06**
K. pneumoniae	ceftriaxone	MLP	0.34 ± 0.09	0.42 ± 0.11	0.42 ± 0.11	0.25 ± 0.07	**0.44 ± 0.09**

C.

Test site: DRIAMS-D (class 1 ratio 10.0%)			Train site: DRIAMS-A	Train site: DRIAMS-D	Train sites: DRIAMS-A DRIAMS-D	Train sites: DRIAMS-A DRIAMS-B DRIAMS-C	Train sites: DRIAMS-A DRIAMS-B DRIAMS-C DRIAMS-D
species	**antibiotic**	**model**					
E. coli	ceftriaxone	LightGBM	0.29 ± 0.09	**0.48 ± 0.05**	0.43 ± 0.06	0.34 ± 0.08	0.43 ± 0.07
K. pneumoniae	ceftriaxone	MLP	0.09 ± 0.03	**0.23 ± 0.06**	0.21 ± 0.05	0.11 ± 0.04	0.20 ± 0.04

Supplementary Table 4: Union experiment results for *DRIAMS-B, DRIAMS-C* and *DRIAMS-D* in AUPRC For each site the dataset was split into training and test; only the training sets are combined in the union training sets. **A.** Results on *DRIAMS-B* (AUPRC). Neither the large external dataset *DRIAMS-A* nor the internal datasets at the target site alone reach the best performance on the target site test data. A union of the target training data and large external datasets is able to reach significant improvements over the target site performance. (*) For S-OXA no *DRIAMS-D* data is available. The results for test sites *DRIAMS-C* and *DRIAMS-D* are listed in **Suppl. Tab. 3**.
B. Results for test site *DRIAMS-C*. The predictive performance benefits from combining the target site train data from *DRIAMS-C* with the large *DRIAMS-A* datasets, compared to training one either train site alone. For two scenarios adding *DRIAMS-B* further improves the performance slightly. **C.** Results for test site *DRIAMS-D*. The behaviour differs from the other test sites in that training exclusively on the test site DRIAMS-D leads to the best predictive performance. No samples for *S. aureus* (oxacillin) available for *DRIAMS-D*.

species	feature bin [m/z]	rank (feature importance)	rank (Shapley values)	Ref: exact m/z value given	Ref: method discriminatory peak identification	Ref: number of strains	Ref: target
S. aureus	2759-2762	14	12	2762[50]	*Genetic algorithm via ClinProTools*	290	Dominant lineages within MRSA (ST5)
				2760[48]	*Support Vector Machine*	160	Discrimination between MSSA and MRSA
S. aureus	3005-3008	13	7	3007[44]	NA	401	Main Clonal Lineages (CC1)
S. aureus	3890-3893	4	3	3891[45]	Visual Examination	600	CC5, CC97
				3891[44]	NA	401	Main Clonal Lineages (CC5, CC25)
S. aureus	4508-4511	19	14	4511[49]	Visual examination	85	Major lineages within MRSA (CC45, CC30)
				4511[44]	NA	401	CC30, CC45, CC398, ST88
S. aureus	4514-4517	12	15	4514[42]	*Supervised neural network* via ClinProTools	82	MRSA Clonal Complexes (CC398)
S. aureus	4640-4643	2	2	4641[49]	Visual examination	85	Major lineages within MRSA (CC8, CC22)
				4641[48]	*Support Vector Machine*	160	Discrimination between MSSA and MRSA
S. aureus	5003-5006	3	4	5002[49]	Visual examination	85	Major lineages within MRSA (CC22)
				5004[42]	*Supervised neural network* via ClinProTools	82	MRSA Clonal Complexes (CC22)
				5002[45]	Visual examination	600	CC22
				5002[44]	NA	401	Main Clonal Lineages (CC22)
S. aureus	5432-5435 5435-5438	6, 7	6, 5	5437[49]	Visual examination	85	Major lineages within MRSA (CC5, CC45, CC22, CC8. ST1, ST15. ST80)
				5440[47]	Support Vector Machine	626	MSSA CC98
E. coli	8501-8504	10	8	8496[55]	Support Vector Machine	109	ST131
E. coli	8444-8447 8447-8450 8450-8453	9, 5, 3	7, 5, 2	8448[54]	*Peak Statistic Calculation* in ClinProTools	197	ST131
E. coli	11780-11783	1	1	11783[54]	*Peak Statistic Calculation* in ClinProTools	197	ST131

Supplementary Table 5: Highly ranked feature bins which have been identified in previous studies. We performed a thorough literature research including studies which reported peaks falling into feature bins receiving high weight of our classifier and report relevant peaks. 'NA': not available as this study did not determine new discriminatory peaks, but focussed on identifying previously known peaks[44]. 'Target' which subgroups within the species were aimed to be discriminated against.

species	antibiotic		hospital hygiene	blood	deep tissue	genital	respiratory	stool	urine	varia	total
E. coli	ceftriaxone	n	659	1190	1073	24	364	5	1473	173	4961
		%	13.3	24.0	21.6	0.5	7.3	0.1	29.7	3.5	100.0
		recall	88.8 ± 0.3	77.8 ± 0.7	57.6 ± 1.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	83.5 ± 0.5	0.0 ± 0.0	
K. pneumoniae	ceftriaxone	n	229	273	204	5	268	15	1790	76	2860
		%	8.0	9.6	7.1	0.2	9.4	0.5	62.6	2.7	100.0
		recall	79.1 ± 0.5	62.4 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	46.5 ± 0.8	0.0 ± 0.0	99.6 ± 0.1	3.8 ± 2.6	
S.aureus	oxacillin	n	379	708	1356	34	517	0	187	609	3790
		%	10.0	18.7	35.8	0.9	13.6	0.0	4.9	16.1	100.0
		recall	89.9 ± 0.5	52.3 ± 2.2	90.8 ± 0.7	0.0 ± 0.0	33.4 ± 1.0	NA	1.8 ± 1.5	2.2 ± 2.0	

Supplementary Table 6: Overview of sample distribution and distinguishability over the workstations in *DRIAMS-A.* For each scenario, the number of samples and the percentage of the total number of samples is stated for each workstation. Additionally, a multi-class logistic regression classifier was trained to predict the workstation based on the MALDI-TOF spectrum, and the predictive performance is reported by the average recall. Generally, we observe a high recall for all workstations if a substantial sample size is present; particularly for hospital hygiene and blood, substantiating prior knowledge that the growth medium can be detected from the MALDI-TOF mass spectra (see **Methods**).