Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dataset curation and signal heterogeneity shape machine learning classification of innate immune transcriptomic responses

Marina Mancuso

mmancuso@lanl.gov

Los Alamos National Laboratory

Kaitlyn M. Martinez

Los Alamos National Laboratory

Apoorv Shanker

Los Alamos National Laboratory

Sarah Nace

Los Alamos National Laboratory

Kristen Wilding

Los Alamos National Laboratory

Makaela M. Montoya

Los Alamos National Laboratory

Daniel E. Jacobsen

Los Alamos National Laboratory

Harshini Mukundan

Lawrence Berkeley National Laboratory

Jessica Z. Kubicek-Sutherland

Los Alamos National Laboratory

Carrie Manore

Los Alamos National Laboratory

Research Article

Keywords: innate immune response, machine learning, pathogen classification, infection diagnostics, cytokine profiling, differential gene expression, dataset curation

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11197667/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dataset curation and signal heterogeneity shape machine learning classification of innate immune transcriptomic responses

Marina Mancuso[1*], Kaitlyn M. Martinez[1], Apoorv Shanker[2], Sarah Nace[2], Kristen Wilding[3], Makaela M. Montoya[2], Daniel E. Jacobsen[2], Harshini Mukundan[4], Jessica Z. Kubicek-Sutherland[2], Carrie Manore[5]

[1*]Computational Intelligence and Modeling, Los Alamos National Laboratory, Los Alamos, 87545, New Mexico, USA.
[2*]Physical Chemistry and Applied Spectroscopy, Los Alamos National Laboratory, Los Alamos, 87545, New Mexico, USA.
[3*]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, 87545, New Mexico, USA.
[4*]Bioscience Area, Lawrence Berkeley National Laboratory, Berkley, 94720, California, USA.
[5*]Modeling and Observations of Earth Systems, Los Alamos National Laboratory, Los Alamos, 87545, New Mexico, USA.

*Corresponding author(s). E-mail(s): mmancuso@lanl.gov;
Contributing authors: kaitlynm@lanl.gov; apoorv@lanl.gov; snace@lanl.gov; kwilding@lanl.gov; makaela@lanl.gov; djacobsen@lanl.gov; hmukundan@lbl.gov; jzk@lanl.gov; cmanore@lanl.gov;

Abstract

Purpose: Host-response profiling has been proposed as a strategy to support pathogen classification, particularly when direct pathogen detection is delayed. However, host-response classifiers are sensitive to methodological choices, including classifier selection and variation in immune signal strength across sampling conditions.

Methods: This study uses a controlled in vitro host-response system as a testbed to evaluate how dataset curation and signal heterogeneity influence

machine learning classification performance, interpretability, and generalizability. We generated transcriptomic data from cells exposed to two Gram-positive-associated agonists with distinct response magnitudes to represent within-class signal diversity relevant to host-response classification. We compared single-layer neural network (NN), random forest (RF), and generalized linear model (GLM) classifiers using overall and class-based accuracy metrics, and evaluated interpretability using class-based variable importance.

Results: Across datasets, GLM and RF models outperformed NN models, indicating that simpler classifiers may be preferable for small transcriptomic host-response datasets. We further found that feature importance profiles depended strongly on the degree of biological heterogeneity represented in the training data, highlighting a tradeoff between signal-regime-specific interpretability and robustness to within-class variation.

Conclusion: These findings position dataset composition as a central design consideration for host-response classification and have implications for developing more robust classifiers relevant to future infectious disease diagnostics and antimicrobial stewardship.

Keywords: innate immune response, machine learning, pathogen classification, infection diagnostics, cytokine profiling, differential gene expression, dataset curation

1 Background

Antimicrobial resistance is deemed one of the top global public health threats by the World Health Organization [1]. In 2019, antimicrobial resistance resulted in an estimated 1.27 million global deaths, with a particularly high burden among invasive bacterial pathogens in low- and middle-income countries [2]. One contributing factor to antimicrobial resistance is the misuse and overuse of antimicrobials (e.g., antibiotics, antivirals, antifungals) to treat infections [3] [4]. Over time, resistance can develop when incorrect treatments are repeatedly administered. To avoid inappropriate use of antimicrobials, it is essential to accurately classify pathogens as bacterial, viral, parasitic, or fungal [5]. In addition to knowing whether a pathogen is bacterial, it is also important to know whether the bacterium is Gram-positive or Gram-negative [6, 7]. The distinction has important implications for antibiotic selection as Gram-negative bacteria tend to be inherently more resistant to antibiotics due to the reduced permeability of the outer cell membrane [6, 8]. Accurate pathogen classification informs optimal treatment decisions and reduces the search space for pathogen-specific diagnostic tests [9].

One way to improve current pathogen classification methods is to analyze signaling patterns from the host's innate immune system. The innate immune system is the body's first line of defense against invading microbes [10], and its signaling patterns arise from conserved mechanisms of pathogen recognition [11]. The innate immune system detects invading microbes through pathogen associated molecular patterns (PAMPs), which are conserved molecular features shared across broad classes of pathogens. PAMPs are sensed by pattern recognition receptors (PRRs) expressed by

innate immune cells triggering signaling cascades that initiate early immune responses [12]. PAMP-PRR interactions are relatively conserved and occur early during infection, therefore the resulting signaling responses provide a biologically grounded basis for distinguishing different pathogen classes based on their PAMPs (see Supplementary section A for further details on the innate immune system). While the innate immune system's network of over 100 interacting signaling patterns makes this task too complex to decode manually [13], machine learning (ML) can be leveraged to improve the speed and precision of pathogen classification methods by using innate immune signaling data to build classification models.

Others have developed ML models to successfully classify pathogens using deep learning and neural networks [14–19], tree-based methods [14–21], support vector machines [15–21], and regression-based methods [14, 15, 19, 22–24]. These models learned patterns for various types of biological signatures, including transcripts [19, 23, 24], proteins [14, 25], metabolites [14], and whole-genome sequences [20]. These signatures were derived from diverse in vivo systems and clinical human samples, including whole blood [15, 16, 18, 22, 23, 26] or peripheral blood mononuclear cells [22, 24], plasma [14], peritoneal effluent [17], and bronchoalveolar lavage fluid [19]. However, progress in improving host-based pathogen classification using ML remains limited by the complexity of the innate immune system and by modeling and dataset design decisions. We identify three key gaps related to the application of ML to innate immune system data:

In vivo system and clinical human data are often noisy and lack standardization, making it difficult to establish consistent data collection protocols due to heterogeneity across individuals and variation in time since infection, both of which affect the strength of the measured signal [27]. Many prior studies have relied on clinical or in vivo-generated data [14–19, 22, 23, 28] rather than in vitro-generated data [20, 22, 24]. The procedures and conditions under which data are collected (e.g., in vivo vs in vitro) can have downstream effects on ML classifier performance. One challenge of using real human data for training ML models is that patient-specific variability and noise may obscure biologically meaningful signals, limiting accurate classification.

Differential signal strength may not always be captured within a single study, as variation in signal strength among individual samples can affect the training and testing performance of ML models. This limitation may be due to the specialized nature of the data sets on which ML models are trained. Many current studies focus on niche applications such as bloodstream infections [18, 28], pulmonary infections [19], peritonitis [17], or specific protein-protein interactions [21] rather than on broadly applicable host-response patterns. Furthermore, several models did not incorporate healthy or non-infected samples during training [14, 15, 26, 28], and are therefore unable to capture baseline signal variation. Additionally, biological datasets often contain a small number of observations relative to the number of features (i.e., "fat" data). Although small sample sizes are frequently driven by resource-intensive data collection procedures, this limitation underscores the need for thoughtful dataset curation to capture the full range of plausible signals across variables.

Consistency is lacking among ML application, as there is substantial variation in how ML models are applied to innate immune system data. Currently, binary

classification models [14, 15, 18, 22, 23, 26, 28, 29] are used more frequently than multiclass models [17, 19, 24, 25]. Moreover, the choice of ML classifier and hyperparameter selection can significantly impact model accuracy and interpretability. These variations in modeling approaches complicates comparison of conclusions across studies. Consistent, generalizable, and trustworthy methods are needed to interpret and contextualize the black-box nature of ML models.

Here, we use a controlled experimental and computational framework to examine how these gaps affect host-response classification in three ways. First, we establish a protocol for measuring baseline biological signal consistency in a controlled in vitro system. We generate in vitro data from experiments using bacterial-associated agonists to model defined innate immune stimulation. While controlled in vitro systems cannot reproduce the full complexity of infection in vivo, they provide an opportunity to isolate specific sources of variation that are difficult to disentangle in patient-derived datasets. This not only reduces the inherent complexity present in clinical data but also allows for a clearer understanding of the drivers that impact ML accuracy and generalizability. Second, we curate three separate datasets spanning a range of differential signaling strengths to evaluate how within-class signal variability affects classification performance. Finally, we develop a robust framework for tuning, documenting, and validating ML methods that use innate immune signals to classify bacterial-associated host-response states. Using this framework, we evaluate the ability of ML models to distinguish Gram-positive-associated, Gram-negative-associated, and uninfected (non-bacterial-associated) responses within a controlled sandbox for studying ML classification under biologically-motivated signal variation. Motivated by the need for robust host-response classifiers in infectious disease and antimicrobial stewardship [1, 7], this study uses a controlled transcriptomic testbed to show that within-class signal heterogeneity is a critical determinant of ML generalization and interpretability.

2 Methods

This study uses an integrated, interdisciplinary approach that combines experimental data generation with ML modeling. This process is organized into four distinct steps (Figure 1):

1. **Experimental Setup:** generate reproducible, physiologically relevant innate immune signaling data in vitro using PAMPs to mimic bacterial infection (Section 2.1).
2. **Data Curation:** implement a standardized, systematic data curation process to explore the effects of differential gene expression on ML classification models (Section 2.2).
3. **Hyperparameter Evaluation:** tune and select optimal hyperparameters for each ML model type (Section 2.3.1).
4. **Model Evaluation:** train and test ML algorithms to identify patterns of innate immune recognition (Section 2.3.2).

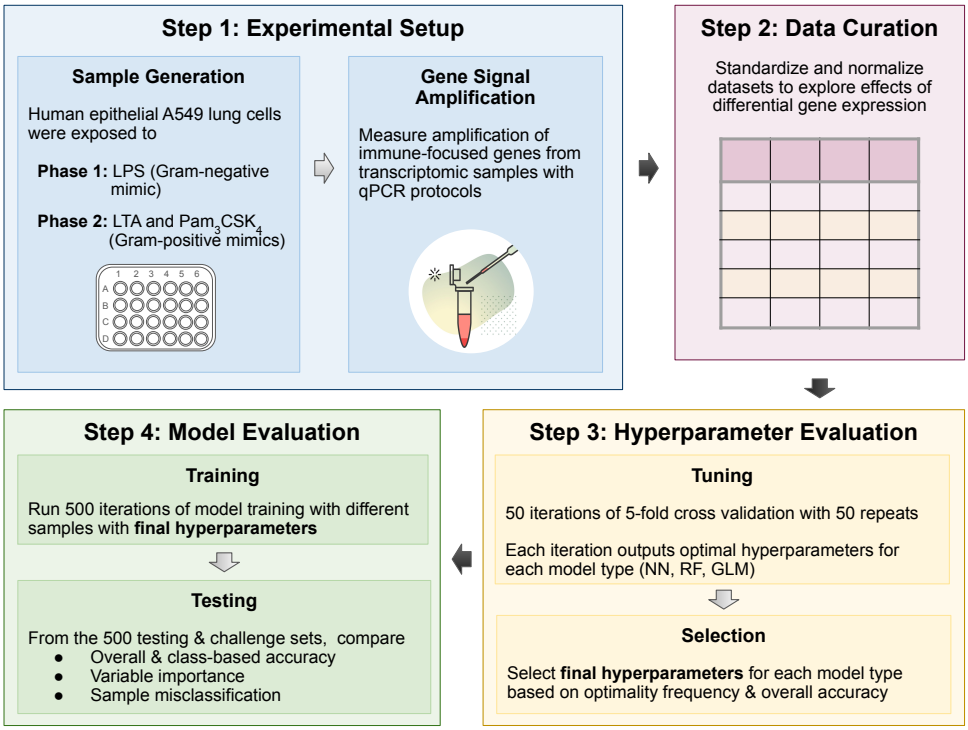


Fig. 1 Schematic overview of the experimental setup, data curation, and machine learning evaluation workflow.

2.1 Experimental data

We used an in vitro microbial mimic exposure assay to generate transcriptomic data reflecting innate immune responses to Gram-positive and Gram-negative bacterial stimuli, and used the resulting transcriptomic gene expression profiles to demonstrate how ML can be applied to classify bacterial subtypes based on innate immune system data. The choice to use transcriptomic data was made to ensure stability of the experimental setup, based on previous data generation work [29, 30]. In total, 195 transcriptomic samples were generated to build ML models for bacterial classification. The experiment was conducted in two phases (Figure 1). The first phase generated 50 Gram-negative bacterial agonists by treating cells with lipopolysaccharide (LPS), a canonical Gram-negative PAMP, to mimic Gram-negative bacterial stimulation. The second phase comprised of generating 64 Gram-positive bacterial mimic exposures, including 32 samples treated with lipoteichoic acid (LTA) and 32 samples treated with Pam_3CSK_4, both of which are representative Gram-positive PAMPs. The remaining 81 samples were untreated and served as non-bacterial controls. Explanations for how LPS, LTA, and Pam_3CSK_4 PAMPs relate to the innate immune system are found in

Section A. Details on the LPS sample generation are described in [30] and Section B.1. Further information on LTA, and Pam_3CSK_4 sample generation can be found in Section B.2.

We used two different types of samples to represent Gram-positive bacteria to examine how variation in signal strength between naturally- and synthetically-treated cells affects ML model performance. Pam_3CSK_4 is associated with heightened a cytokine response compared to LTA [31] (Figure C1). For the remainder of the manuscript, we use "Gram-negative" to refer to LPS-treated samples, "weak Gram-positive" to refer to LTA-treated samples, and "strong Gram-positive" to refer to Pam_3CSK_4-treated samples. In this context, "strong" and "weak" denote differences in transcriptomic signal intensity within the Gram-positive data.

At 24 hours post-treatment, the expression levels of 84 immune-focused genes were measured using quantitative polymerase chain reaction (qPCR) protocols. The same protocol was used for both experimental phases (Figure 1). The two phases differed only in the biological stimulus to which the cells were exposed. As a result, data from both phases could be seamlessly combined using our standardized protocols and procedures.

2.2 Dataset curation

Data curation refers to the way that data are cleaned and processed before being input into a ML model. This process is important to ensure that biologically relevant signals are captured by ML models rather than spurious signals arising from noise and outliers [32]. This process may also include imputation of missing data. The choices made during the data curation process can significantly impact the accuracy, performance, and interpretatability of ML models.

Each sample in the dataset includes a class label (Gram-negative, strong Gram-positive, weak Gram-positive, or untreated) along with the Cycle threshold (C_t) values for 84 immune-focused genes, five housekeeping genes, and three quality control wells. The C_t represents the number of PCR cycles required for gene expression to become detectable above background noise [33]. Further details on C_t normalization are included in Section B.3.

We curated three different data subsets from the full 195 transcriptomic dataset to investigate how differences in Gram-positive signal strength affect ML model performance (Table 1). Recall that Pam_3CSK_4 tends to induce more potent gene expression (lower C_t values) compared to its natural counterpart LTA (Section 2.1, Figure C1, [31]). The C_t values are inversely related to gene expression, so lower C_t values observed with Pam_3CSK_4 and higher C_t values observed with LTA indicate stronger and weaker gene expression, respectively. All three datasets contain observations from the same three class labels: Gram-negative bacterial, Gram-positive bacterial, and untreated. The difference among the datasets is which level of Gram-positive signal strength is included for model training:

1. **Weak-signal dataset (W-dataset):** trained on weak Gram-positive samples (LTA)

2. **Mixed-signal dataset (M-dataset):** trained on both strong and weak Gram-positive samples under a single label (LTA & Pam_3CSK_4)
3. **Strong-signal dataset (S-dataset):** trained on strong Gram-positive samples (Pam_3CSK_4)

Curating these three data subsets for training ML models allows us to examine the downstream effects that data curation choices have on ML models for biological classification. We can evaluate the robustness of our ML models by "challenging" them with data exhibiting the opposite signal strength. Specifically, we test weak Gram-positive samples on the ML models trained on the S-dataset, and test strong Gram-positive data on the ML models trained on W-dataset (Table 1). Models trained on the M-dataset do not include a challenge set because both weak- and strong-signal data are included in their training. The challenge set can be interpreted as a distribution shift across signal magnitude, and is used to test model generalizability. For our host-response classification context, this can simulate real world variability across infection timing, symptom heterogeneity (e.g., symptomatic vs asymptomatic infections), and individual host-response. From an operational perspective, it is important that ML models avoid overfitting to their specific training data and instead have flexibility in handling a variety of signal magnitudes [34, 35].

Table 1 Comparison of the three curated datasets used for ML model training. All datasets are trained on Gram-negative and untreated data. Dataset differences relate to gene signal strength of the Gram-positive samples used for training and testing (e.g., strong and/or weak). The challenge class denotes the Gram-positive signal strength excluded from model training and testing, which is used to further assess model performance.

Dataset	Training/Testing Classes	Challenge Class
W	Gram(-), weak Gram(+), untreated	strong Gram(+)
M	Gram(-), strong & weak Gram(+), untreated	N/A
S	Gram(-), strong Gram(+), untreated	weak Gram(+)

2.3 Machine learning models

Our ML framework compares classification results across three types of ML architectures: single-layer neural networks (NN), random forest (RF), and generalized linear model (GLM) approaches [36–38]. We refer to these as the three "model types". These simple, off-the-shelf ML classifiers were selected based on their high performance in binary classification between Gram-negative and untreated data in our previous work [29]. This previous work showed that adding hidden layers in the NN did not meaningfully improve overall accuracy, and instead complicated model interpretability. This work aims to explore the potential for ML to distinguish bacterial pathogen classes from broadly applicable host-response patterns rather than developing the best model for this purpose. Additional details for each model type are found in Section D.

The ML analysis includes steps 3 and 4 of Figure 1. This process highlights different approaches for training ML classification models to identify patterns of innate immune

recognition. Hyperparameter evaluation involves tuning and selection (Section 2.3.1), and model evaluation involves training and testing of ML models (Section 2.3.2). We used a 70%/30% train/test split for all process steps, as previous work has shown this to be sufficient for transcriptomic data, promoting robust learning while preventing overfitting [29]. This procedure was applied to the three data subsets described in Section 2.2 (Table 1). Specifically, each dataset (W, M, S) was used to train the three model types (NN, RF, GLM) resulting in nine model type/dataset combinations that produced overall and class-based performance metrics. The class-based performance metrics correspond with the three class labels: Gram-negative, Gram-positive, and untreated.

2.3.1 Hyperparameter tuning

The hyperparameters of interest for the NN models include the number of nodes in the hidden layer (size) and the learning rate that controls the speed at which weights are tuned (decay). The hyperparameter of interest for the RF models is the number of variables that are randomly sampled as candidates at each tree split (mtry). Each RF model consists of 500 trees. The GLM models use a multinomial family and explore ranges for the elastic net mixing parameter (alpha) and a general penalty parameter (lambda).

Hyperparameter tuning and selection involved running 50 iterations of 5-fold cross validation (CV) with 50 repeats for each model type (NN, RF, GLM). Specifically, 5-fold cross validation is performed on the training data (70% of the dataset), and this process is repeated 50 times using different samples of the training data for each fold. The model is then tested on the remaining 30% of the dataset. The entire procedure is iterated 50 times to allow different samples to be included in the training data. While our dataset is too small to incorporate a nested CV approach, which applies cross validation on inner and outer folds [39], this process enhances the reliability of our estimates by considering multiple CV iterations on different training sets. This provides a more robust estimate of hyperparameter optimality compared to relying on one tuned estimate from a single training set. Further details on hyperparameter tuning implementation are found in Section E.

Each iteration returns an optimal value for the NN size and decay, RF mtry, and GLM alpha and lambda based on its corresponding testing accuracy. The final hyperparameters are selected from the set of optimal hyperparameters based on the frequency with which each value is identified as optimal and its corresponding overall accuracy on the testing sets. We use "optimal" hyperparameters to refer to the best hyperparameter values identified in each individual iteration, based on that iteration's training data. We use "final" hyperparameters to refer to the selected hyperparameter values used for further model training and evaluation (step 4 of Figure 1). This process is performed separately for each of the three datasets described in Section 2.2. A single size, decay, mtry, alpha, and lambda value is selected as the final set of hyperparameters for each dataset (Table 1).

2.3.2 Model training and evaluation

Model training and evaluation uses the final hyperparameters selected from the hyperparameter tuning and selection steps (Section 2.3.1). For each model type/dataset combination, 500 model iterations were built using 70%/30% train/test splits.

Overall accuracy is calculated as the proportion of correctly classified samples in the testing set across all classes. Class-based (balanced) accuracy is determined for each class label. Class-based accuracy is calculated as the average of sensitivity (true positive rate) and specificity (true negative rate). Individual sample misclassification is calculated as the proportion of times a sample is misclassified in the testing or challenge sets. Variable importance is calculated for each model overall and by class. Further details about the implementation of model training and evaluation are found in Section F.

3 Results

3.1 Hyperparameter tuning

Hyperparameter tuning and selection involved running 50 iterations of 5-fold cross validation with 50 repeats. Each iteration returned an optimal value for the NN size and decay, RF mtry, and GLM alpha and lambda. The final hyperparameters were selected from the set of optimal hyperparameters based on the frequency with which each value was identified as optimal and its corresponding overall accuracy on the testing set (Table 2). Certain model types exhibited more differences in optimal hyperparameters across data set curation strategies than others (Table E1). The RF models show the clearest differences in hyperparameter tuning results across datasets. NN models had the least difference in optimal hyperparameter values across datasets - all datasets preferred a larger number of nodes in the hidden layer and a larger decay value. Hyperparameter tuning results for the GLMs indicate that the combination of alpha and lambda may have more influence on model selection than each parameter alone. We further highlight the tuning results for the RF mtry parameter to demonstrate how data curation impacts optimal hyperparameter selection.

The mtry parameter in RF models denotes the number of variables randomly sampled at each tree split. We explored mtry values between 1 and 30. The W-dataset prefers a moderate number of variables at each tree split, whereas the S-dataset prefers a lower number of variables, and the M-dataset prefers a higher number of variables (Figure 2). We suspect that the S-dataset prefers a lower mtry value because signal variation between class labels is more distinct, and requires fewer variables to be sampled at each split to predict bacterial subtype. The M-dataset may prefer the highest mtry value because greater variation in gene signal levels may require increased model complexity to correctly classify subtype. None of the iterations for any of the datasets returned optimal mtry values below 7. We selected final mtry values of 19 for the W-dataset, 29 for the M-dataset, and 11 for the S-dataset. The most complex RF models are associated with the M-dataset, while the least complex ones correspond to the S-dataset. We refer the reader to Section E and Table E1 for further discussion on the tuning results for the remaining hyperparameters.

Table 2 Test accuracy metrics obtained using the final set of hyperparameter values selected for the three datasets.

Dataset	Model Type	Parameter	Final Value	Median Accuracy	Accuracy Range
W	NN	size	10	0.983	0.917-0.979
W	NN	decay	0.5	0.938	0.833-1.00
W	RF	mtry	19	1.00	0.979-1.00
W	GLM	alpha	1	1.00	0.979-1.00
W	GLM	lambda	0.01	1.00	0.971-1.00
M	NN	size	11	0.931	0.828-1.00
M	NN	decay	0.5	0.931	0.828-1.00
M	RF	mtry	29	0.983	0.966-1.00
M	GLM	alpha	1	1.00	0.948-1.00
M	GLM	lambda	0.005	0.983	0.966-1.00
S	NN	size	11	0.958	0.875-1.00
S	NN	decay	0.5	0.958	0.854-1.00
S	RF	mtry	11	0.979	0.979-1.00
S	GLM	alpha	0.25	1.00	0.958-1.00
S	GLM	lambda	0.01	0.979	0.958-1.00

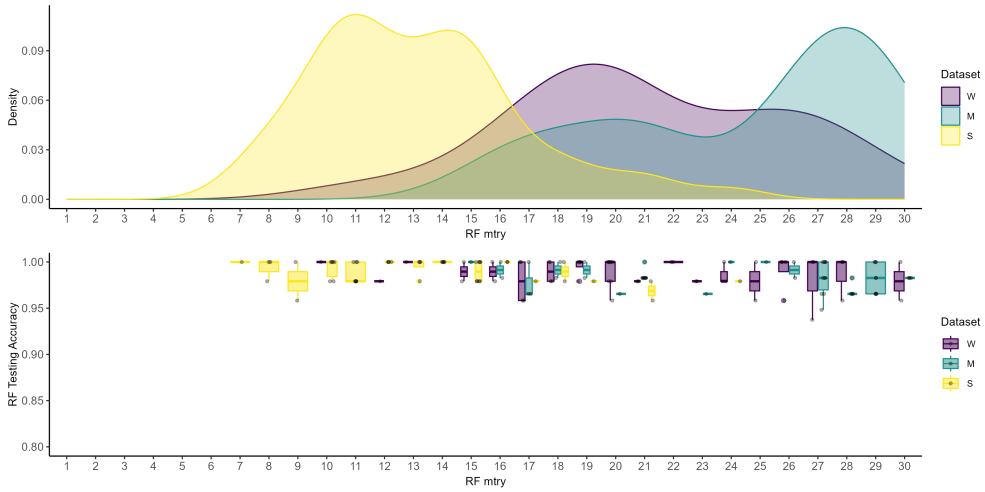


Fig. 2 Random forest's mtry hyperparameter tuning results from 150 models across the three datasets highlight the impact of dataset curation on hyperparameter optimality. (Top) Density plots of the frequency of optimal mtry values and (bottom) boxplots of testing accuracy show differing optimal mtry values across curated dataset. With respect to the number of variables sampled at each tree split, the W-dataset prefers a moderate number of variables, the M-dataset prefers a high number of variables, and the S-dataset prefers a lower number of variables.

Model evaluation is based on the test sets from 500 model iterations. Each iteration implements a 70%/30% train/test split and uses the final hyperparameters selected from the hyperparameter evaluation step (Table 2). We assess model performance using overall and class-based accuracies, as well as individual sample misclassifications.

In addition, variable importance is compared among model types, class labels, and datasets. Variable importance results help inform plausible explanations for sample misclassification across different classifier types and dataset curation approaches.

3.1.1 Overall and class-based testing accuracy

The GLM and RF models significantly outperform the NN models on the testing sets across all three dataset curations (Figure 3). The NN models exhibit the widest testing accuracy distributions for each model type/dataset combination. For overall testing accuracy, the median accuracies for the GLM and RF models are equivalent within each of the three datasets (Table 3). This pattern is also observed for the Gram-negative and untreated class-specific accuracies. For the weak-signal class, the GLMs show the highest median accuracy when trained on the M- and S-datasets and are tied with the RF models when trained on the W-dataset (Table 3). For the strong-signal class, the GLMs show the highest median accuracy when trained on the M-dataset, the RF models show the highest median accuracy when trained on the W-dataset, and the GLM and RF models are tied for the highest median accuracy when trained on the S-dataset.

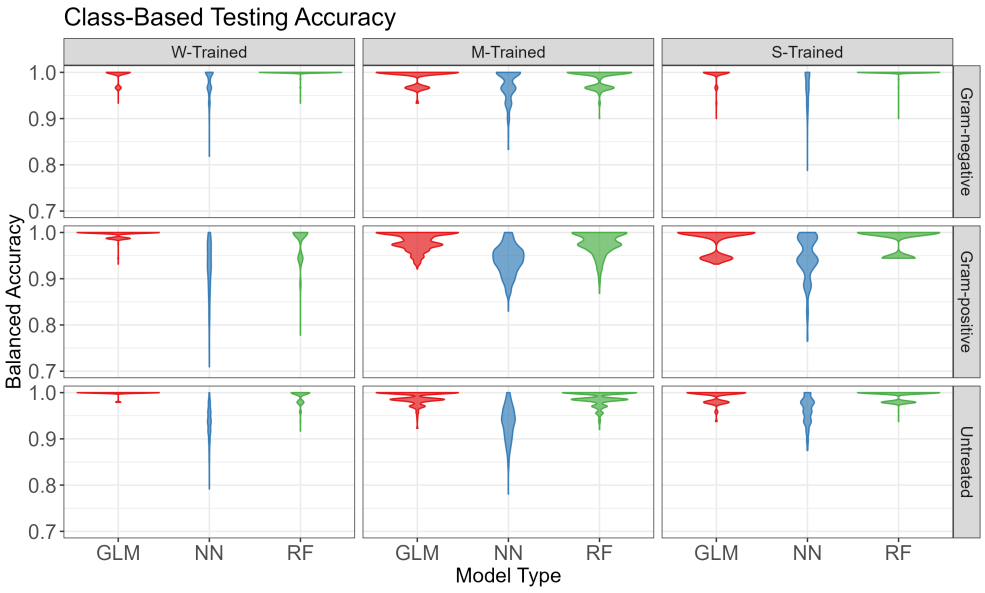


Fig. 3 Distributions of testing accuracy (balanced accuracy) for the three model types organized by class label (rows) and training dataset (columns). Distributions were created from the test sets of 500 model build iterations. The GLM and RF models exhibit consistent performance across datasets and class labels, whereas the NN models show the lowest testing performance across class labels and dataset.

Table 3 Median overall and class-based testing and challenge accuracies for the GLM, NN, and RF models trained on the three datasets reported for each class label. Models trained on the W-dataset are challenged on the strong-signal data, whereas the models trained on the S-dataset are challenged on the weak-signal data.

Dataset	Model	Overall	Gram(-)	Weak Gram(+)	Strong Gram(+)	Untreated
W	GLM	1.00	1.00	1.00	0.938	1.00
W	NN	0.938	0.985	0.932	0.594	0.938
W	RF	1.00	1.00	1.00	0.969	1.00
M	GLM	0.983	1.00	0.987	0.987	0.985
M	NN	0.931	0.967	0.947	0.947	0.938
M	RF	0.983	1.00	0.974	0.974	0.985
S	GLM	1.00	1.00	0.625	1.00	1.00
S	NN	0.958	0.967	0.531	0.944	0.958
S	RF	1.00	1.00	0.25	1.00	1.00

Overall, models trained on the S-dataset achieve the highest median overall testing accuracy among the three datasets. When averaging performance across the three model types within each dataset, the W-dataset yields the highest median accuracy for the Gram-negative class, whereas the S-dataset yields the highest median accuracy for the untreated class. As expected, models trained on the W-dataset achieve the highest median accuracy for the weak-signal class, while models trained on the S-dataset achieve the highest median accuracy for the strong-signal class.

3.1.2 Challenge accuracy for Gram-positive data

Recall that the W-dataset models are challenged on the strong-signal data, while the S-dataset models are challenged on the weak-signal data (Table 1). The W-dataset models perform better on the challenge data than the S-dataset models (Figure 4). In particular, the RF models trained on the S-dataset exhibit the highest performance on the testing set but the lowest performance on the challenge set. This challenge scenario is the only dataset/class label combination in which RF performs worse than NN models (Table 3). Despite having the lowest median overall accuracy and untreated class accuracy, the M-dataset models exhibit a more balanced accuracy between both signal strength magnitudes of the Gram-positive class (Table 3). These results demonstrate the importance of training ML models on data spanning a wide range of signal strengths to enhance model generalization.

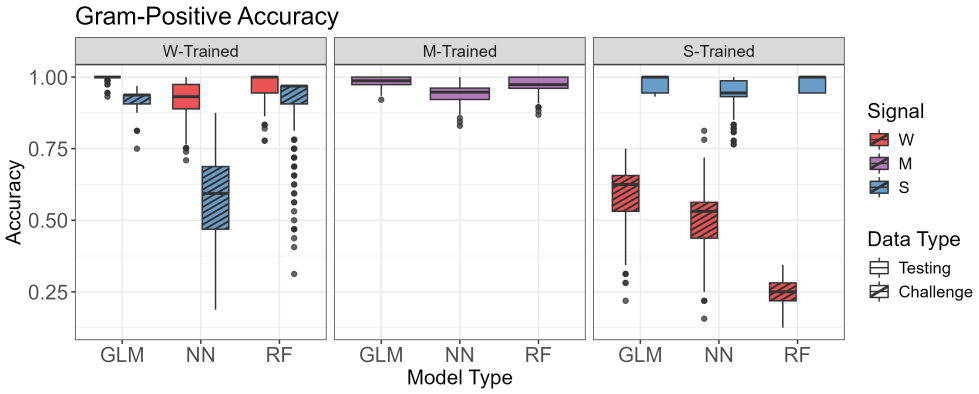


Fig. 4 Distributions of testing (no pattern) and challenge accuracies (striped pattern) for the Gram-positive signal strength magnitudes, organized by training dataset. The testing accuracies for each dataset show moderate-to-high performance across all model types. The challenge accuracies for the W-dataset models (i.e., strong-signal data) show moderate performance for the GLM and RF models and poor performance for the NN models. The challenge accuracies for the S-dataset models (i.e., weak-signal data) show poor performance across all three model types. The M-dataset models exhibit a better balance in accuracy between the two Gram-positive signal strength magnitudes. Note that the M-dataset includes only one boxplot per model type because weak- and strong-signal data were trained under a single label.

3.1.3 Variable importance

Variable importance results differ among model types, dataset curation strategies, and class labels (Figures G6-G9). We also observe variation in gene importance between the two highest performing model types (GLM and RF) for both strong- and weak-signal Gram-positive class labels (Figure 5). We consider "highly important" genes as those that appear in the top 20 variables for importance in at least 90% of the 500 model evaluation iterations for a given model type/dataset combination (i.e., included in the top 20 for at least 450 models). Highly important genes are features that are consistently used to obtain classification prediction. We observe clear differences in the number of highly important genes by model type and dataset curation. RF models tend to identify a greater number of highly important variables compared to GLMs. This may be because the GLMs employ feature selection via L_2 regularization, which is not the case for the RF models. In particular, the GLMs trained on the W- and M-datasets prefer pure L_2 regularization rather than a balance of L_1 and L_2 regularization, as is observed for the S-dataset (Table E1). Consequently, the GLMs trained on the S-dataset tend to yield more highly important variables than those trained on the W- and M-datasets. Descriptions of gene functions and lists of highly important genes are found in Section G and Table G2.

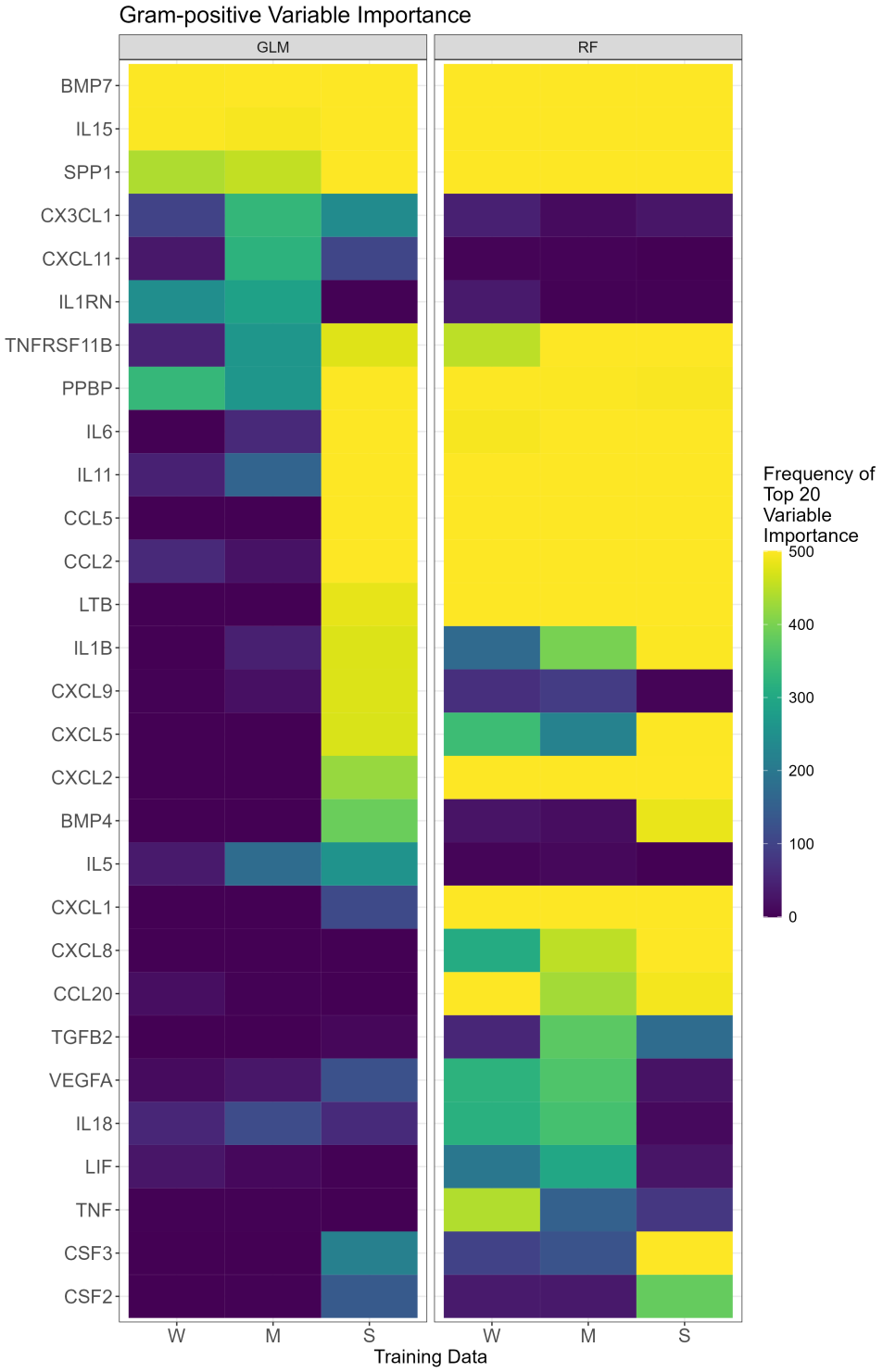


Fig. 5 Variable importance for the Gram-positive class labels in the GLM and RF models, organized by dataset curation. Heatmaps display the frequency with which genes appear among the top 20 variables across 500 models. Genes are ordered according to their top 20 frequency in the GLMs trained on the M-dataset. Models trained on the S-dataset and RF models tend to exhibit more highly important genes.

Fifteen genes are identified as highly important for at least one class label in at least one of the nine model type/dataset combinations: BMP4, BMP7, CCL2, CCL20, CCL5, CSF3, CXCL1, CXCL2, CXCL5, CXCL8, IL11, IL15, IL1B, IL6, and LTB (Figure G10). That is, at least one model type (NN, RF, GLM) trained on at least one dataset (W, M, S) ranks these genes as highly important for each class label (Gram-negative, Gram-positive, untreated). These genes were also identified as important in previously-developed binary Gram-negative vs untreated models [29].

Several genes are consistently highly important across all nine model type/dataset combinations for specific class labels. For example, CCL20 and IL11 are highly important for all model types in the Gram-negative class, BMP7 is highly important for all model types in both strong- and weak-signal Gram-positive classes, and CCL5 and CXCL1 are highly important for all model types in the untreated class. The CCL20, CCL5, and CXCL1 genes were also among the most highly important variables in the previous binary classifiers [29]. Other genes are highly important for only a single class label. For example, BMP2 is uniquely highly important for the Gram-negative class, and CCL24 is uniquely highly important for the untreated class. Other genes are highly important for specific model types. Additional genes show model-specific importance patterns. For example, CCL17, CCL24, and CSF2 are highly important only in NN models, IL18 shows high importance only in GLMs, and IL1B shows high importance only in RF models.

3.1.4 Overall and individual sample misclassification

The sample-specific misclassification rate is defined as the proportion of times that a sample was incorrectly classified in the test or challenge set. Samples are included in the test set for approximately 30% of the 500 model iterations (i.e., around 150 models). Gram-positive samples are included in the challenge set for the opposite signal strength category for each of the 500 model iterations (Table 1). That is, all 32 weak-signal samples are part of the challenge set for the 500 S-dataset models, and all 32 strong-signal samples are part of the challenge set for the 500 W-dataset models. We first present results of overall misclassifications for each dataset, followed by results of individual sample misclassification.

Overall misclassification results are aggregated from sample-specific misclassification rates for each sample of a given class label. For example, the overall Gram-negative misclassification rate for the W-dataset takes the average of the 50 individual Gram-negative sample misclassification rates from the W-dataset. Dataset curation affects both the predicted label and the frequency of sample misclassifications. The Gram-negative and untreated classes show high rates of correct classification across all model type/dataset combinations (Figure 6). The Gram-positive samples (both strong- and weak-signal) are the most frequently misclassified of all classes. When misclassified, the Gram-positive samples are more frequently predicted to be untreated instead of Gram-negative. This may be due to more similarity between Gram-positive and untreated gene expression distributions compared to the Gram-negative ones. In particular, the S-trained models more often classify weak-signal data as untreated because those models were not trained to learn the region that overlaps weak-signal with the untreated baseline.

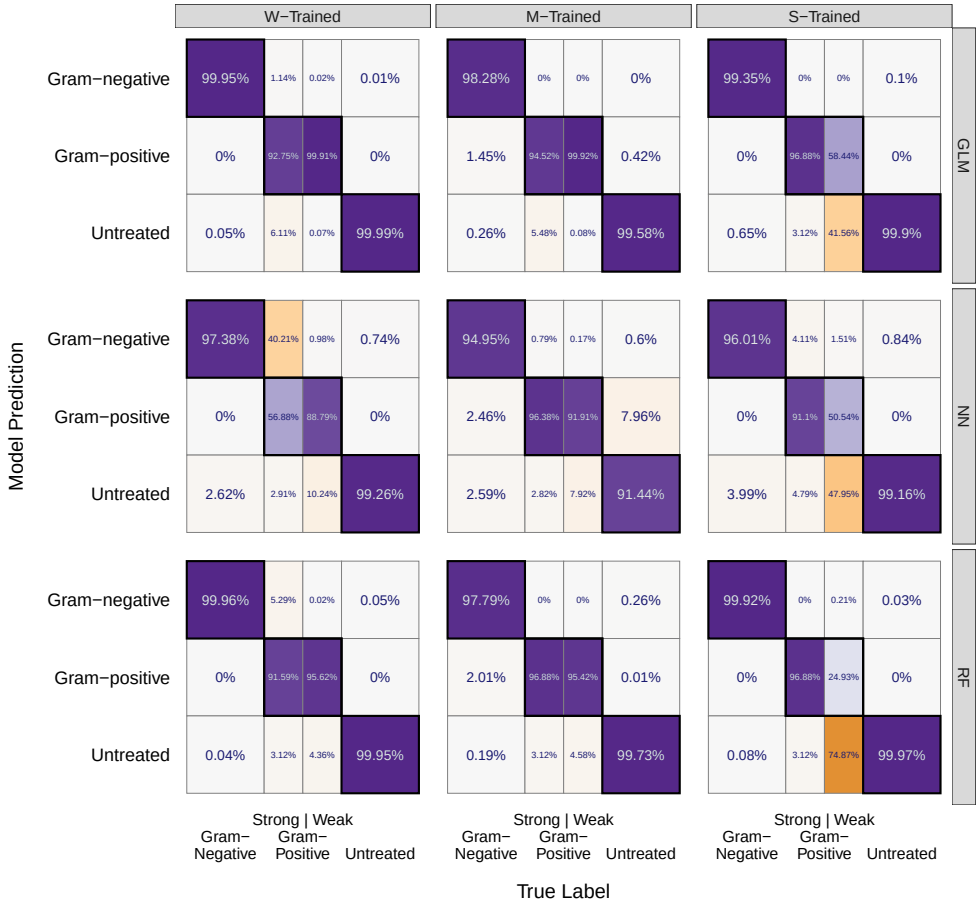


Fig. 6 Confusion matrices for the three dataset curation strategies (columns) and model types (rows) with respect to the true (x-axis) and predicted class labels (y-axis). The purple diagonals for each matrix denote the average percentage of correct label classification across the 500 test sets for each model type/dataset combination. The orange shaded cells denote the percentage of misclassified samples. The Gram-negative and untreated classes show high rates of correct classification across all model type/dataset combinations. Models trained on the S-dataset frequently misclassify the weak-signal data (challenge set) as untreated for all three model types. The NN models trained on the W-dataset frequently misclassify strong-signal data (challenge set) as Gram-negative.

Clear differences in interpretation emerge when we disaggregate misclassification rates by individual samples. For example, the overall misclassification rate for Gram-negative samples of the W-dataset is negligible for the GLM and RF models (Figure 6), but one Gram-negative sample in particular (#135) is misclassified as Gram-positive over 61% of the time for models trained on the W-dataset (Table 4). Some samples exhibit high misclassification rates for certain model type/dataset combinations, or even across all model type/dataset combinations (Figures H11). We highlight a few

examples of highly misclassified samples and use class-based variable importance to assess why these samples may observe higher rates of misclassification than others (Table 5). We primarily focus on results between the GLM and RF models since they have comparable performance relative to the NNs (Section 3.1.1).

Table 4 Gram-positive misclassification rates for each model type/dataset combination for Gram-negative sample 135. Sample 135 has the lowest CXCL5 signal out of the three bacterial class labels (Figure H16). CXCL5 is only highly important for the Gram-negative class in the RF models trained on the S-dataset – the only model type/dataset combination that never misclassified sample 135. Sample 135 was never misclassified as untreated.

Model Type	Training Data	Misclassification Rate
GLM	W	0.611
GLM	M	0.726
GLM	S	0.191
RF	W	0.045
RF	M	1.00
RF	S	0.00

Table 5 Frequently misclassified samples with corresponding predicted class, model type/dataset combination associated with misclassification, and attributes correlating with misclassification.

Sample	True Class	Predicted Class	Model Type-Dataset	Attribute
#135	Gram(-)	Gram(+)	GLM-W, GLM-M, GLM-S, NN-W, NN-M, NN-S, RF-W, RF-M	CXCL5 importance
#152	Strong Gram(+)	Untreated	GLM-W, GLM-M	CCL5, IL6, LTB importance
#158	Strong Gram(+)	Untreated	All	CCL5 importance; Overlapping signal
#54	Weak Gram(+)	Untreated	GLM-S, NN-S, RF-W, RF-M, RF-S	Number of highly important genes

The sample-specific misclassification analysis motivated by class-based variable importance provides insight into how ML models interpret complex biological datasets (Table 5). Although the black-box nature and nonlinear interactions of ML models prevent definitive attribution of misclassification causes, the importance of certain variables may explain the improved classification accuracy for specific samples in particular model type/dataset combinations. For example, Gram-negative sample 135 had the lowest CXCL5 signal out of all class labels (Figure H16), and was only correctly classified in the model type/dataset combination that showed high CXCL5 importance

(RF models trained on S-dataset). Although highly important genes can be informative for class discrimination in most samples, models may have difficulty accurately predicting class labels when a sample's feature values overlap with the distribution of another class (CCL5 for sample 158). The number of highly important variables in a model may also explain misclassification patterns. Models with a larger number of highly important variables may, in some cases, obscure meaningful signals that are more clearly captured by simpler models (sample 54). For certain samples, simpler models trained on fewer variables may achieve improved classification accuracy. Further in-depth discussion of misclassification results of these samples is provided in Section H.

Examining the correlation between variable importance and misclassification can both enhance ML model design as well as motivate directions for future experimental studies. ML model design can be enhanced by up-weighting the importance of genes that are shown to affect correct classification. This can be achieved through a two-step variable selection and model fitting process, where the first step identifies the most important variables for each class, and the second step retrains the models using only the important features identified in the first step (in a way that prevents data leakage). Designing further experiments that specifically target these highly important genes can enhance mechanistic understanding of their roles in innate immunity. This can reveal why these genes are considered important in certain ML classification models. Finally, this process allows us to investigate strategies to best balance model complexity with generalizability. For example, the RF models trained on the S-dataset accurately classified Gram-negative sample 135 and strong Gram-positive sample 152 when the M-dataset models failed to do so, but these same models struggled to accurately predict weak Gram-positive sample 54 (Table 5). Additionally, the M-dataset models exhibited less misclassification of the strong Gram-positive samples compared to the W-dataset models. While the M-dataset models did not outperform either W- or S-dataset models on any specific sample, its training on varied Gram-positive signal strength promoted overall model generalizability.

4 Discussion

4.1 The tradeoff between and generalizability and interpretability

Host response heterogeneity is one of the most confounding factors affecting medical intervention [40], and points to the breadth of host responses across dimensions of age, sex, genetics, and environmental factors [41]. As a result, it may be ideal to train ML models on data with heterogeneous signal to enhance generalizability. In this study, generalizability refers to the model's ability to correctly predict classification when tested on data across experimentally-induced signal-strength shifts. We introduced the challenge set for the W- and S-datasets as a way to test model generalizability for distribution shifts across signal magnitude. This allows us to understand what happens when the training distribution is narrower or broader than the biological variability at testing.

Our weak-signal Gram-positive data could represent samples from asymptomatic infection [42] or from those collected shortly after infection [27]. We saw that the S-trained models performed the poorest when challenged on weak-signal data, and often misclassified these signals as untreated due to the similarity in signal distribution. While the M-trained models showed better generalization across the experimentally-induced signal-strength shifts, this became a tradeoff to model interpretability.

The biological interpretation that can be gleaned from ML models remains dependent on the nature of the curated training dataset. This is most clearly shown by analyzing the GLMs for the Gram-positive class. Some genes show high importance for only the S-dataset and negligible importance for the W- and M-datasets (e.g., IL6, CCL2, CXCL5, etc.). Sometimes, genes that were highly important for the S-dataset and negligibly important for the W-dataset show moderate importance for the M-dataset (e.g., IL11, TNFRSF11B). Other times, genes that are moderately-to-highly important for both S- and W-datasets show less importance in the M-dataset (PPBP). There are even genes that show relatively high importance for the M-dataset that have negligible or low importance in both S- and W-datasets (eg., CX3CL1, CXCL11). However, interpretation was not necessarily reduced in the M-dataset models, but rather shifted from subgroup-specific features to class-level ones. Variable importance from the M-dataset models provides class-level signals, while the W- and S-datasets provide clean, but conditional signals. The homogeneous signal regimes may reveal features specific to LTA-like or Pam_3CSK_4-like responses, but not necessarily features robust to Gram-positive response variation.

Given the nuances between balancing generalizability with interpretability, we identified the following patterns about our small transcriptomic host-response dataset: GLM offers a strong default when interpretability is prioritized, RF provides a useful nonlinear benchmark when the testing data aligns with the signal distribution of the training data, and the single-layer NN did not provide sufficient benefit under these sample size constraints. Mixed-signal training should be considered when robustness to within-class heterogeneity is more important than identifying agonist-specific biomarkers.

4.2 Within-class heterogeneity on generalizability

The hyperparameter tuning process adapts model architecture based on the characteristics of the training data. We showed how the hyperparameter selection can be optimized to improve model generalizability. The M-dataset had the highest signal variability for the Gram-positive class and preferred the highest mtry value in the RF models. The signal level heterogeneity within the Gram-positive class required greater model complexity to correctly classify exposure classes.

The varied within-class heterogeneity across curated datasets further impacted classification accuracy – not just for the Gram-positive class, but also for the downstream classification of Gram-negative and untreated classes. The GLM and RF models show high accuracy across model type/dataset combinations. But when meaningful misclassification occurs for these models, both Gram-negative and weak Gram-positive samples are more often misclassified as untreated. In these false-negative cases, the models are unable to detect an immune response. In contrast, strong Gram-positive

samples are more often misclassified as Gram-negative. In these cases, the models detect an immune response in the strong-signal data, but attribute the signal to the incorrect bacterial subtype. The GLM and RF models maintain a high level of specificity in detecting an immune response, as demonstrated by the untreated class exhibiting the fewest frequent misclassifications across these model types.

4.3 Within-class heterogeneity on interpretability

The amount of within-class heterogeneity also affected variable importance results that point to model interpretability. Models trained on the S-dataset identified a greater number of highly important variables on average compared to models trained on the W- and M-datasets. Some of these variable importance results aligned with our previous findings from binary Gram-negative vs untreated models [29]. However, our previous study did not identify IL11 as a highly ranked gene, whereas the present study found IL11 to be highly important for Gram-negative data across all model type/dataset combinations. Incorporating a third class of Gram-positive bacterial data may have revealed the importance of IL11 in distinguishing Gram-negative bacterial responses. These findings demonstrate that ML model variable importance results can vary substantially depending on how training data are curated.

Assessing individual sample misclassification in the context of ML variable importance can provide insight into both biological interpretation and the black-box nature of ML models. Variable importance results suggested a plausible explanation for why Gram-negative Sample #135 exhibited higher prediction accuracy in classifiers that ranked CXCL5 as a highly important variable. Gram-negative Sample #135 was frequently misclassified as Gram-positive in the models that did not identify CXCL5 as highly important. Sample #135 exhibited the weakest CXCL5 signal among all bacterial samples, and Gram-negative samples generally showed weaker CXCL5 signal levels compared to Gram-positive classes (Figure H16). Because ML models differ fundamentally from mechanistic models, we cannot conclude with certainty which of the selected classifiers most closely reflect known biological response pathways. Nevertheless, our ML models provide insight into pathways that may be important for distinguishing among different types of infection.

4.4 Limitations

The main limitation of this work is that our models classify host-response states induced by bacterial-associated agonists rather than clinical infection or live bacterial exposure. Although this study does not model the full complexity of clinical infection, it addresses a methodological question that is central to how pathogen classifiers operate: how training set composition affects ML model behavior given biological signal within a class.

Another limitation of this work is the relatively small sample size of our dataset. This is a common feature of many biological datasets that require time- and cost-intensive experiments. ML models are generally better suited to large training datasets as small datasets are prone to overfitting. To mitigate this challenge, we resampled the training data used to build our models 500 times during both hyperparameter

tuning and model evaluation steps. As a result, each sample is included in the training data for roughly 70% of the iterations, which helps prevent outlier samples from having too much influence in model outcomes. Within each training iteration, we also incorporated 5-fold cross-validation. Together, this resampling strategy enhances the reliability of our models by providing classification probability distributions with quantified uncertainty.

4.5 Future work

There are various opportunities to optimize the operational capabilities of our ML models from both experimental and computational sides. Our integrated, interdisciplinary approach allows the developments from each side to enhance the efficiency of the other. For example, our ML variable importance results can assist in reducing the panel of immune-focused genes that need to be measured for effective model performance, thereby reducing the costs and time required for subsequent sample generation. Other improvements to experimental data generation include varying exposure times, PAMP doses, cell types, and PAMP stimulation. The current experimental setup exposes human epithelial A549 cells to bacterial PAMPs and measures gene expression 24 hours post-exposure. Future in vitro data generation could expand the range of exposure times to represent differing times since infection, thereby increasing signal strength heterogeneity. Signal strength variation can also be enhanced by exposing cells to different PAMP doses. In addition, new host-response patterns may be revealed from using different cell types for PAMP exposure (e.g., immune cells). We plan to further extend data generation to include viral, fungal, and parasitic agonists. Once baseline gene expression signals for each pathogen class are established with our in vitro system, we can iteratively integrate ML model development with clinical data. These developments will enable construction of a cleaner and more comprehensive dataset than what is achieved with clinical or in vivo data alone.

Opportunities to improve the ML model development side include exploring the use synthetic data, adjusting thresholds for classification probability, and incorporating hierarchical modeling approaches. Exploring the use of synthetic data to augment our models will further mitigate limitations associated with small sample sizes. We can also adjust probability thresholds within the ML models to minimize false-negative rate. In the context of pathogen detection, predicting the presence of infectious microbes when none are present (false positives) may be preferred to failing to detect them when they are present (false negatives). Additional opportunities for model development include incorporating ensemble or hierarchical modeling approaches. An ensemble framework would combine predictions from multiple model types (e.g., GLM and RF) to generate classification outputs. A hierarchical model would predict pathogen class through a multi-step process. For example, a binary model could first detect whether an innate immune response is present (bacterial vs. non-bacterial), followed by a second binary model to distinguish Gram-positive from Gram-negative bacteria. Such a hierarchical modeling framework may be better for mapping model predictions to underlying biological mechanisms. Investigating these extensions will further enhance the clinical utility of ML-based pathogen detection.

5 Conclusion

In this methodological study, we use a controlled in vitro host-response dataset as a testbed to evaluate how dataset curation and signal heterogeneity influence ML classification performance, interpretability, and generalizability. Rather than optimizing a classifier for a single narrowly defined experimental condition, we focused on how within-class host-response variation affects model behavior. By capturing signal variation within Gram-positive data and baseline signals from unexposed control samples, we identified gene response patterns that may have otherwise not been revealed. Specifically, our investigation compared three off-the-shelf ML classifiers (NN, RF, and GLM) for three-way host-response classification models (representing Gram-negative-associated stimulation, Gram-positive-associated stimulation, and untreated controls) and explored how the signal strength of the Gram-positive bacterial samples in the training data affects ML performance. In addition, we examined how models performed when being challenged with data from a held-out Gram-positive-associated signal regime. Our best-performing models achieved high accuracy in distinguishing host-response classes under controlled stimulation conditions, and our robust tuning, training, and testing procedures provided distributions with quantified uncertainty. We evaluated model performance using accuracy, variable importance, and sample misclassification rates, and demonstrated how interpretability and generalizability depend on dataset composition. Finally, we used class-based variable importance results to assess plausible explanations for frequent sample misclassifications. Our results highlight that within-class signal heterogeneity is not merely a source of noise, but a central design consideration for developing generalizable and interpretable ML classifiers from host-response data.

In the context of antimicrobial resistance, host-response classifiers are attractive because they could eventually help distinguish infection classes earlier in the diagnostic process. However, premature deployment of models trained on narrow or homogeneous datasets risks poor generalization when clinical samples contain variable response magnitudes. Our results suggest that dataset design, particularly representation of within-class heterogeneity, should be treated as a central component of model development rather than a secondary pre-processing choice.

Declarations

- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: The datasets generated and analysed during the current study are available in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository, accession number GSE350749 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE350749). The code used to analyze the data is available in github, https://github.com/lanl/bacterial-classification (DOI: 10.5281/zenodo.23327935).
- Competing interests: The authors declare that they have no competing interests.
- Funding: This work was supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated

by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001) and at the Lawrence Berkeley National Laboratory, operated by the University of California for the United States Department of Energy. Funding for this work was provided by the Defense Threat Reduction Agency (CB11015). The views expressed in this article are those of the authors and do not reflect the official policy or position of the U.S. Department of Defense or the U.S. Government.

- Author's contributions: MM contributed to the computational study design, software development, data interpretation and analysis, visualization, and the writing of the original draft. KMM contributed to computational study design, data interpretation and analysis, and assisted with software development, visualization, and writing the original draft. AS led the experimental study design and generated experimental results. SN and KW assisted with data interpretation and analysis. MMM and DEJ assisted with generating experimental results. HM acquired funding and assisted with project management. JZK assisted with supervision of the experimental study design. CM led project management and supervision. All authors assisted with editing and revising the manuscript and have approved this submitted version.
- Acknowledgments: Not applicable

Appendix A The innate immune system

The innate immune system is the body's first line of defense against foreign pathogens [10]. This evolutionarily-conserved system is activated within minutes to hours upon infection and triggers signaling cascades by cytokines and chemokines – proteins that elicit antimicrobial responses [11]. The innate immune system is pathogen-agnostic, meaning it reacts regardless of the presented pathogen. But despite its nonspecific nature, the differential signals from cytokines and chemokines may help distinguish different classes of pathogens. The ability to accurately distinguish pathogen classes assists in narrowing down the search space for specific pathogen testing [43]. This not only enables clinicians to determine an accurate diagnosis with fewer pathogen-specific tests, but also reduces administration of inappropriate treatment that contributes to increasing antimicrobial resistance [1].

Toll-like receptors (TLRs) are key components of the immune system's ability to detect pathogens [13]. They recognize evolutionarily-conserved pathogen-associated molecular patterns (PAMPs) from invading microbes [11]. Stimulation with TLR-ligands induces cytokines and chemokines that regulate immune response [31, 44]. Here, we use "gene signals" or "gene expression" to refer to the cytokine and chemokine amplification levels. Lipopolysacharide (LPS) is a TLR-4 ligand exclusively found in the cell walls of Gram-negative bacteria, such as *E. coli* and *Salmonella*. The functional equivalent of LPS for Gram-positive bacteria is lipoteichoic acid (LTA), which is a TLR-2 ligand. Examples of Gram-positive bacteria include *Staphylococcus aureus* and *Streptococcus pneumoniae*. Knowing whether a bacterial infection is Gram-positive or Gram-negative can have implications for selecting optimal treatment and control methods [7]. Gram-negative bacteria tend to be inherently more resistant to antibiotics

due to the reduced permeability offered by the LPS outer membrane, which prevents antibiotic drugs from reaching the cells [8].

We generate three types of PAMPs from an *in vitro* system to demonstrate how machine learning (ML) can be used to classify subtypes of bacterial pathogens. We use LPS to mimic Gram-negative pathogens, and use LTA and Pam_3CSK_4 to mimic Gram-positive pathogens. While LTA is the natural TLR-ligand that is present in Gram-positive bacteria, Pam_3CSK_4 is its synthetic counterpart that often induces stronger gene signals than LTA [31, 45]. The same pathway is expressed in both LTA and Pam_3CSK_4, but the two components differ in the magnitude of their expressed gene signals. By including a range of variable signal in our data, we can analyze the effect of differential gene expression on ML performance.

We note that our data uses PAMPs to mimic bacterial pathogens rather than the full bacterial cells themselves. However, to facilitate language consistency and help communicate the broader impact of our work, we use the term "Gram-negative" to refer to the LPS-treated samples, "strong Gram-positive" to refer to the Pam_3CSK_4-treated samples, and "weak Gram-positive" to refer to the LTA-treated samples.

Appendix B Experimental data generation and standardization

This experiment involved two phases to generate 195 samples of bacterial mimics. The first phase focused on generating Gram-negative mimics with lipopolysaccharide (LPS) (Section B.1), and the second phase focused on generating Gram-positive mimics with lipoteichoic acid (LTA) and Pam_3CSK_4 (Section B.2). Both phases used human epithelial lung A549 cells that were sourced from American Type Culture Connection (ATCC, CCL-185). The cells were cultured in Kaighn's Modified Ham's Formulation F-12 Media (ThermoFisher Scientific; 21127022) with 10% fetal bovine serum (Sigma Aldrich; F2442-500ML) and 1% v/v penicillin/streptomycin (ThermoFisher Scientific; 15140122).

After first strand synthesis, 96-well Human Cytokines and ChemokinesRT2 Profiler PCR Arrays (Qiagen; PAHS-150ZC-24) using the RT2 qPCR Master Mix (Qia-gen; 330529) were used to process qPCR samples. This array measured RNA levels from 84 different cytokines and chemokines, 5 housekeeping genes, and included 7 control wells for each sample. For each type of sample, mRNA expression was measured for 100 replicates, 50 samples of LPS, 32 samples of LTA, 32 samples of Pam_3CSK_4, and 81 untreated controls from a combination of both sets of experiments. The plates were run on an Applied Biosystems StepOnePlus Real-Time PCR Thermocycler (ThermoFisher Scientific; 4376600) using the following protocol: 95°C for 10 min; 40 cycles: 95°C for 15 sec; 60°C for 1 min [30]. In this process, a fluorescent probe is mixed with each sample allowing the PCR machine to quantify the light intensity of each sample [46]. The C_t value, or cycle threshold, represents the cycle number at which a sample fluoresces more than the background fluorescence [46]. If the C_t value is low, the cytokine or chemokine being detected has a high expression; whereas if the C_t value is high, the cytokine or chemokine being detected has a low expression.

B.1 Gram-negative (LPS) data generation

For the cells used to generate LPS samples, two stocks (frozen at passage #4) were grown separately for one passage. After this first passage, the two cultures were each split into six separate cultures, known as lineages (12 lineages total). These lineages were grown separately to increase the biological variation between them. Five passages of each cell line were cultured in 24-well plates, with each of the 12 lineages cultured in two wells. The day prior to adding to cells, LPS from *P. aeruginosa* (Sigma Aldrich; L9143-25MG) was suspended in water and diluted to 10 μg/mL in fresh media and stored overnight at 4°C. After the cells reached ¿80% confluence, the media of one well from each lineage was replaced with media warmed to 37°C containing 10 μg/mL LPS (treated), while the other well for each lineage was replaced with fresh media warmed to 37°C without LPS (untreated). The samples included in phase 1 have been previously published in [30].

B.2 Gram-positive (LTA and $\mathbf{Pam_3CSK_4}$) data generation

For cells used to generate LTA and Pam_3CSK_4 samples, eight lineages of the cell line were used. Four different non-consecutive passages of each of the lineages were plated into 24-well plates with each of the eight lineages cultured into three wells. The day prior to adding to cells, LTA from *S. aureus* (InvivoGen, tlrl-pslta) and Pam_3CSK_42 – a synthetic triacylated lipopeptide (LP) (InvivoGen, tlrl-pms) – were both suspended in water and diluted to 10 μg/mL in fresh media and stored overnight at 4°C.

On the day prior to adding to cells, the media of one well from each lineage was replaced with media warmed to 37°C containing 10 μg/mL LTA (treated), the second well from each lineage was replaced with media warmed to 37°C containing 10 μg/mL Pam_3CSK_4 (treated). The third well for each lineage was replaced with fresh media warmed to 37°C without LTA or Pam_3CSK_4 (untreated).

B.3 Housekeeping genes and data normalization

In addition to our panel of 84 immune-focused genes, we measured the amplification of five housekeeping genes: ACTB, B2M, GAPDH, HPRT1, and RPLP0. The housekeeping genes are not reactive to LPS, LTA, or Pam_3CSK_4, and should maintain relatively consistent C_t values independent of treatment. We included the housekeeping genes to help normalize C_t values among the different classes for building our ML models.

The B2M housekeeping gene had the lowest treatment affect [29], and was therefore selected to normalize the data. The mean C_t value of the B2M gene was subtracted from the C_t value of the genes. To handle missing observations, undetermined C_t values were replaced with the maximum value of the normalized C_t value of that gene, plus one.

Finally, amplification of three quality controls (HGDC, RTC, and PPC) were tested for PCR efficiancy and contamination. Human genomic DNA contamination (HGDC) was measured to test if any DNA contamination was present, reverse transcription control (RTC) was measured to check for impurities in the RNA sample that would impede reverse transcription, and positive PCR control (PPC) was measured to check for impurities that would impact the PCR process.

Appendix C Strong and weak Gram-positive signal

The C_t values are inversely related to the amount of detected gene expression – we observe lower C_t values with Pam_3CSK_4 and higher C_t values with LTA, indicating stronger and weaker gene expression, respectively. However, to enhance visual intuition, we show the distributions of negative C_t values of Pam_3CSK_4 and LTA. That is, the higher (less negative) -C_t values observed from Pam_3CSK_4 indicate stronger levels of gene expression, whereas the lower (more negative) -C_t values observed from LTA indicate weaker levels of gene expression.

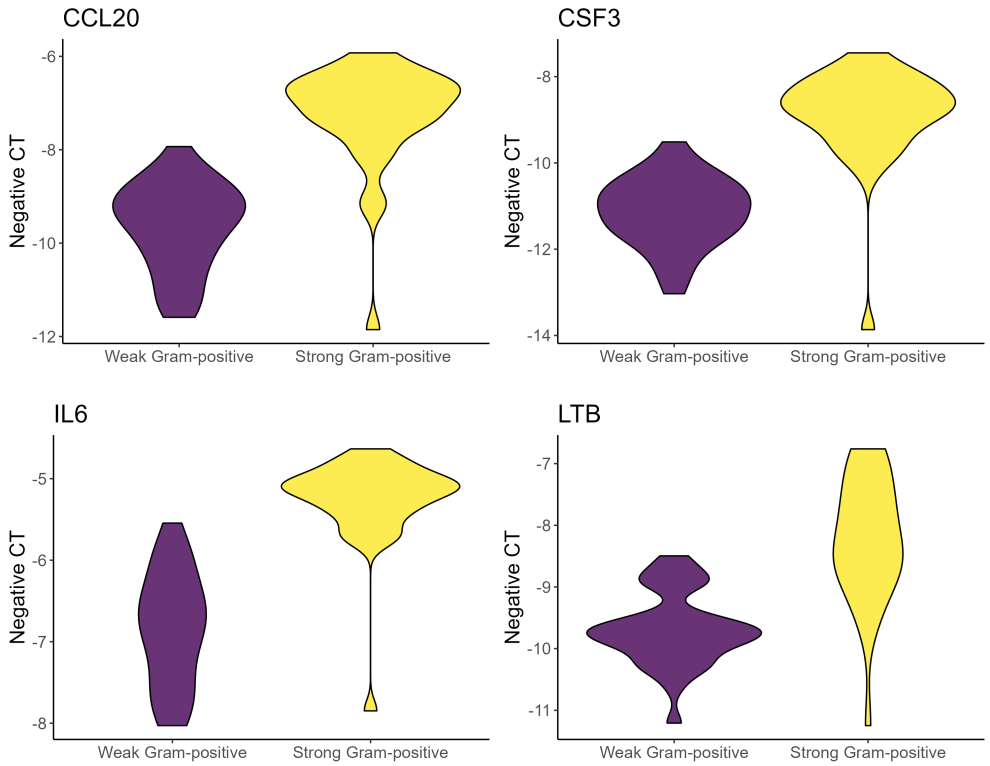


Fig. C1 Gene signal distribution comparison between the weak (LTA) and strong (Pam_3CSK_4) Gram-positive data for CCL20, CSF3, IL6, and LTB. Gene signal is consistently more up-regulated for the strong Gram-positive data. Note that the negative of the C_t values are shown here, where lower values of -C_t corresponds with weaker amounts of gene expression, and higher values of -C_t correspond with stronger gene expression levels.

Appendix D Machine learning models

Neural network (NN): inspired by the brain's neural connectivity. For this study, the input layer consists of the 84 gene expression values (C_t values). This input is manipulated within a single hidden layer, and the output predicts the class label (Gram-negative, Gram-positive, or untreated). Previous work showed that additional hidden layers did not meaningfully improve overall accuracy, and instead complicated model interpretability [29]. The hidden layer manipulates the input via weights that are tuned by backpropagation. In backpropagation, the error is minimized through partial derivative calculations with respect to the weights.

Random forest (RF): uses prediction averages across an ensemble of decision trees. For this study, tree construction is based on the magnitude of C_t values and the resulting prediction for each sample is based on the majority classification label vote across the trees.

Generalized linear models (GLM): transformations of the C_t values are assumed to have a linear relationship with the expectation of the dependent variable (class label). The discrete class labels are linked through the logarithm of the odds of the probability that the class label occurs. The GLMs fit the coefficients to scale the transformed C_t values.

Appendix E Hyperparameter tuning process and results

Implementation: Hyperparameter tuning was performed in R programming software. We use the `train` function from R's `caret` package to build models using each of hyperparameter value. We select the methods "nnet" for the NN models, "rf" for our RF models, and "glmnet" for our GLMs. We use `caret`'s `createDataPartition` function to split our data into training and testing sets while ensuring that we retain the correct class distribution from the full dataset. Testing accuracy was computed with the `predict` function from the `stats` package, along with the `caret`'s `confusionMatrix` function.

Hyperparameter tuning and selection involved running 50 iterations of 5-fold cross validation with 50 repeats for each model type (NN, RF, GLM). The training data (70% of the dataset) was divided into five, mutually exclusive subsets of equal size (folds). The model is trained on the four of the folds and validated on the fifth fold. This process is repeated until each fold is used in validation. Moreover, the cross validation process is repeated 50 times (repeats) using different splits of the training data. That is, the folds of each each repeat comprise of different samples of the training data. Once the model is trained, it is tested on the remaining 30% of the dataset. We repeat this process 50 times (iterations) to allow different samples to be part of the training data. We obtain a more robust and reliable estimate of hyperparameter optimality by averaging performance across multiple folds, repeats, and iterations.

Each iteration returns an optimal value for the NN size and decay, RF mtry, and GLM alpha and lambda. The final hyperparameters are selected from the set of optimal hyperparameters based on the frequency of each value's optimality and the respective overall accuracy on the testing sets. This process is performed separately for each of the data subsets (W-, M-, and S-dataset).

NN size: The size parameter of the neural network indicates the number of nodes in the hidden layer. We explored node sizes of 1-11. Eleven is the maximum node size based on the number of input features. All three datasets prefer a larger number of nodes for the single-layer neural net. We selected a size value of 10 for the W-dataset and size value of 11 for the M- and S-datasets as the final size values, based on the frequency density and testing accuracy distributions (Figure E2). None of the model iterations return an optimal size value less than 3.

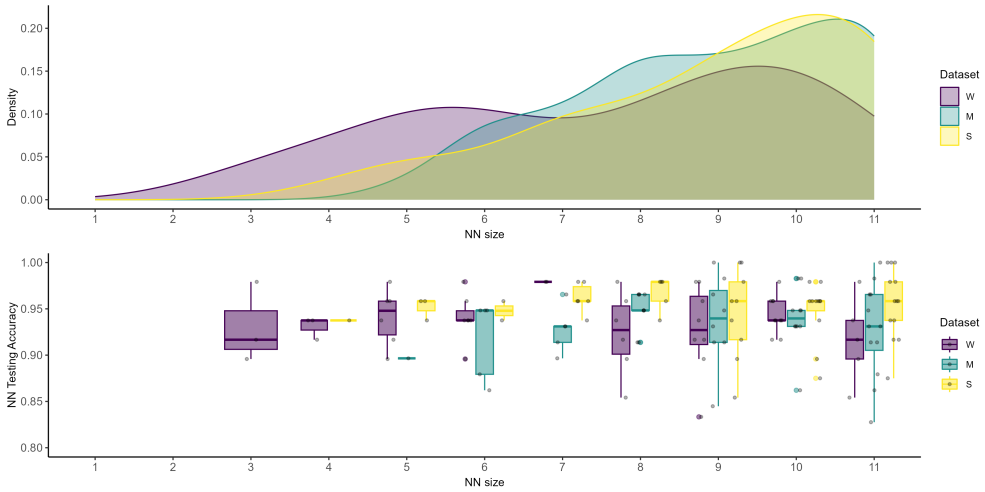


Fig. E2 Size hyperparameter tuning results from 150 neural network models. (Top) Density plot of optimal size frequency. (Bottom) Boxplots of testing accuracy with respect to size value.

NN decay: The decay parameter of the neural network denotes the learning rate for the speed at which the weights are tuned via backpropagation. We explored decay values of 1×10^{-5}, 1×10^{-4}, 0.001, 0.01, 0.05, 0.1, and 0.5. Only decay values of 0.1 and 0.5 are within the set of optimal values (Figure E3). Larger decay values increase the amount of regularization to avoid overfitting and lead to lower weighted coefficients. The decay value of 0.5 is most frequently selected as the optimal value, therefore we selected this as the final decay value for each dataset.

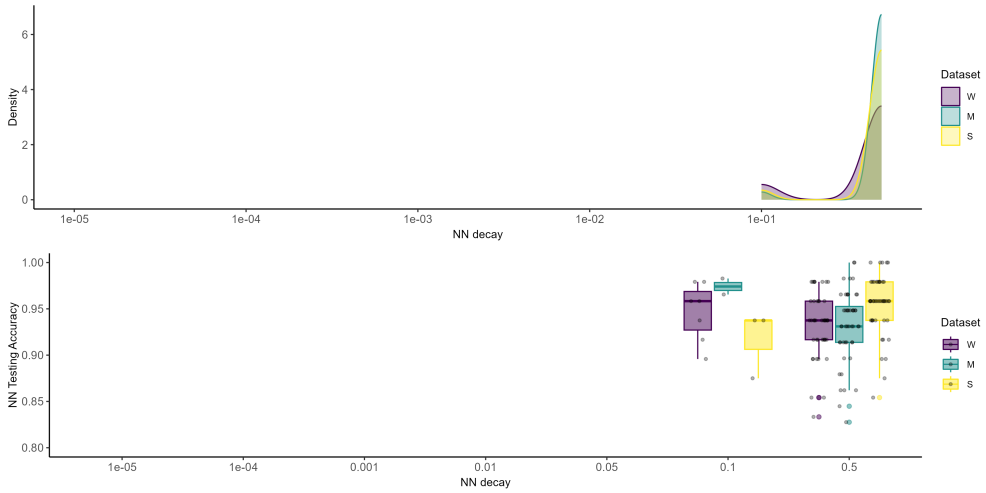


Fig. E3 Decay hyperparameter tuning results from 150 neural network models. (Top) Density plot of optimal decay frequency. (Bottom) Boxplots of testing accuracy with respect to decay value.

GLM alpha: Alpha is the elastic net mixing parameter in the GLMs. We explored alpha values of 0, 0.25, 0.5, 0.75, and 1.00. Alpha ranges from 0 to 1 and controls the balance between L1 (LASSO) and L2 (ridge) regression. An alpha value of 1 indicates pure LASSO regression, which drives the model towards down-selection of highly correlated predictors. By contrast, an alpha value of 0 indicates pure ridge regression, which instead distributes weights among correlated predictors. The W-dataset strongly prefers pure LASSO regression (alpha = 1), and the S-dataset prefers lower-to-moderate values of alpha compared to the M-dataset (Figure E4). We selected final alpha values of 1 for the W- and M-datasets, and 0.25 for the S-dataset. None of the model iterations prefer pure ridge regression (alpha = 0). For the W- and M-datasets, there are less differences in signal strength between the Gram-positive and untreated classes. A lower ratio of signal to noise may prefer pure LASSO regression since there may only be a few variables contributing to different signals.

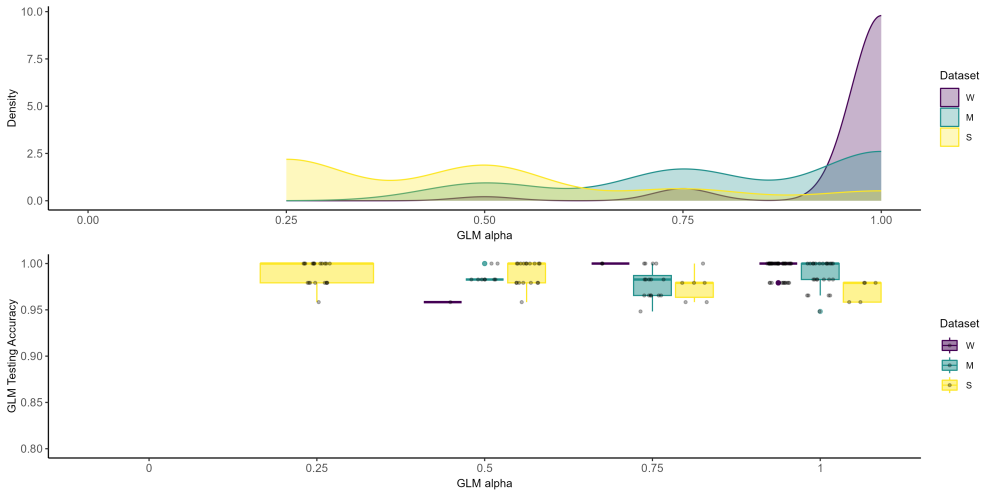


Fig. E4 Alpha hyperparameter tuning results from 150 generalized linear models. (Top) Density plot of optimal alpha frequency. (Bottom) Boxplots of testing accuracy with respect to alpha value.

GLM lambda: Lambda is a general penalty parameter that affects the overall amount of regularization in the GLMs. We explored lambda values of 1×10^{-5}, 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5. The S-dataset prefers larger values of lambda compared to the W- and M-datasets (Figure E5). We selected final lambda values of 0.01 for the W-dataset, 0.005 for the M-dataset, and 0.1 for the S-dataset. Smaller values of lambda lead to less regularization penalty and retain more variables in the model. More variables may be necessary to discern class label in these datasets because the overlap in gene expression among the class labels muddies the signal. This rationale counters the reasoning for preferring pure LASSO regulation, indicating that the combination of alpha and lambda may matter more than each parameter alone. None of the model iterations return optimal lambda values below 1×10^{-4} or above 0.1. This suggests that a moderate amount of regularization is required for successful classification of all three datasets.

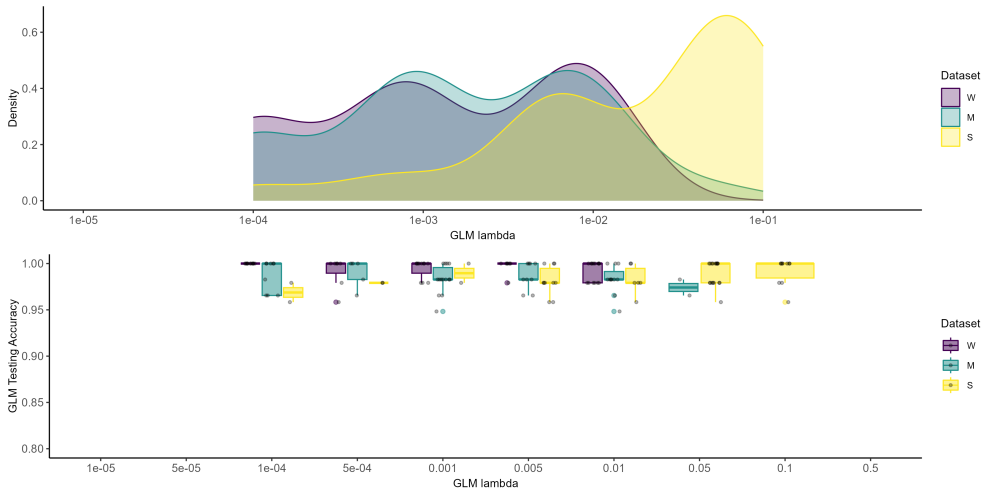


Fig. E5 Lambda hyperparameter tuning results from 150 generalized linear models. (Top) Density plot of optimal lambda frequency. (Bottom) Boxplots of testing accuracy with respect to lambda value.

Appendix F Model training and evaluation implementation

Model training and evaluation uses the final hyperparameters selected from the hyperparameter tuning and selection steps (Section 2.3.1). For each model type (NN, GLM, RF) and dataset (W, M, S) combination, 500 iterations of the models were built using

Table E1 Final hyperparameter selection for the W-, M-, and S-datasets.

Model	Parameter	Range	Final (W)	Final (M)	Final (S)
NN	size	1-11	10	11	11
NN	decay	1e-5, 1e-4, 1e-3, 5e-3, 0.01, 0.05, 0.1, 0.5	0.5	0.5	0.5
RF	mtry	1-30	19	29	11
GLM	alpha	0, 0.25, 0.5, 0.75, 1	1	1	0.25
GLM	lambda	1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 0.01, 0.05, 0.1, 0.5	0.01	5e-3	0.1

70%/30% train/test splits. The `createDataPartition` function was again implemented to split the dataset to maintain the correct label proportions in the training and testing sets. The NNs were built using the `nnet` function of the `nnet` package. The RFs were built using the `randomForest` function from the `randomForest` package and each model uses an ensemble of 500 trees. The GLMs were built using the `glmnet` function from the `glmnet` package and used a multinomial family.

Overall accuracy was calculated based on the proportion of correctly classified samples in the testing set. Class-based accuracy was determined for each classification label (Gram-positive, Gram-negative, untreated) and used the Balanced Accuracy metric obtained from the `confusionMatrix` function. The class-based, or balanced accuracy takes the average of sensitivity (true positive rate) and specificity (true negative rate). Individual sample misclassification was calculated based on the proportion a sample was misclassified in the testing or challenge sets. Model variable importance was calculated for overall and by class label. The `varImp` function from the `caret` package was used to determine variable importance for the GLMs (overall and by class) and overall for the NNs. Variable importance for GLMs is based on the absolute value of the C_t values. Overall variable importance for the NNs uses a combination of the weight values for the nodes in the hidden layer. Class-based variable importance of the NNs uses the the `olden` function from the `NeuralNetTools` package. This measures the relative importance of input variables with respect to hidden layer connection weights. Variable importance for the RFs (overall and by class) use the `importance` function from the `randomForest` package. This uses the mean decrease in accuracy method, which measures how much accuracy decreases when a feature's values are permuted.

Appendix G Variable importance

The selection of 84 immune-focused genes can be divided into several groups according to function (Table G2). We use the following cytokine groupings to relate ML variable importance to biological mechanisms:

- **Chemoattractants:** attract immune cells to sites of infection or injury [47]
- **Anti-inflammatory:** help reduce immune responses and control inflammation [48]
- **Pro-inflammatory:** trigger and amplify immune responses and inflammation [48, 49]
- **Involved in apoptosis:** programmed cell death

- **Growth and differentiation factors:** support development and specialization of different cell types [50, 51]
- **Hematopoietic growth factors:** stimulate the production of blood and immune cells
- **Th1-associated:** promote cell-mediated immune responses; helpful against intracellular pathogens (i.e., viruses and bacteria) [52, 53]
- **Th2-associated:** support humoral immune defense against extracellular pathogens (i.e., parasites, allergens, toxins, extracellular bacteria) [54]
- **Th17-associated:** defend against certain infections and may contribute to inflammation [53, 55, 56]

Functional groupings were verified from the UniProt database [57] with additional references for each group included above.

Table G2 Functional groupings of immune-focused selection of genes. **Bolded** genes are highly important for at least one model type/dataset combination for all class labels. Green-colored genes are highly important for all model type/dataset combinations for the Gram-negative class. Violet-colored genes are highly important for all model type/dataset combinations for the Gram-positive classes (strong- and weak-signal). Blue-colored genes are highly important for the untreated class. Functional groupings were verified from the UniProt database [57]. Functional group descriptions and additional references are included within the text. Note that some genes appear for multiple groups.

Functional Group	Genes
Chemoattractants	CCL1, **CCL2**, CCL3, **CCL5**, CCL7, CCL8, CCL11, CCL13, CCL17, CCL18, CCL19, **CCL20**, CCL21, CCL22, CCL24, CX3CL1, **CXCL1**, **CXCL2**, **CXCL5**, **CXCL8**, CXCL9, CXCL10, CXCL11, CXCL12, CXCL13, CXCL16, IFNA2, IL16, PPBP, XCL1
Anti-inflammatory	IL1RN, IL10, IL13, TGFB2
Pro-inflammatory	**CCL2**, CCL3, **CCL5**, CX3CL1, CXCL8, CXCL9, CXCL10, IFNG, IL1A, **IL1B**, **IL6**, **IL15**, IL17A, IL17F, IL18, **LTB**, MIF, TNF
Involved in apoptosis	FASLG, LTA, TNFRSF11B, TNFSF10
Growth & differentiation factors	ADIPOQ, BMP2, **BMP4**, BMP6, **BMP7**, CNTF, CXCL9, CXCL10, GPI, LIF, MSTN, NODAL, OSM, SPP1, TGFB2, TNFRSF11B, TNFSF11, VEGFA
Hematopoietic growth factors	CD40LG, CSF1, CSF2, **CSF3**, IL3, IL7, **IL11**, **IL15**, THPO, TNFSF11, TNFSF13B
Th1-associated	CXCL10, IFNG, IL2, IL10, IL12A, IL12B, IL18, TNF
Th2-associated	CCL17, IL2, IL3, IL4, IL5, IL10, IL13, IL27
Th17-associated	**CCL20**, **CXCL8**, IL2, **IL6**, IL9, IL17A, IL17F, IL21, IL22, IL23A IL24, IL27

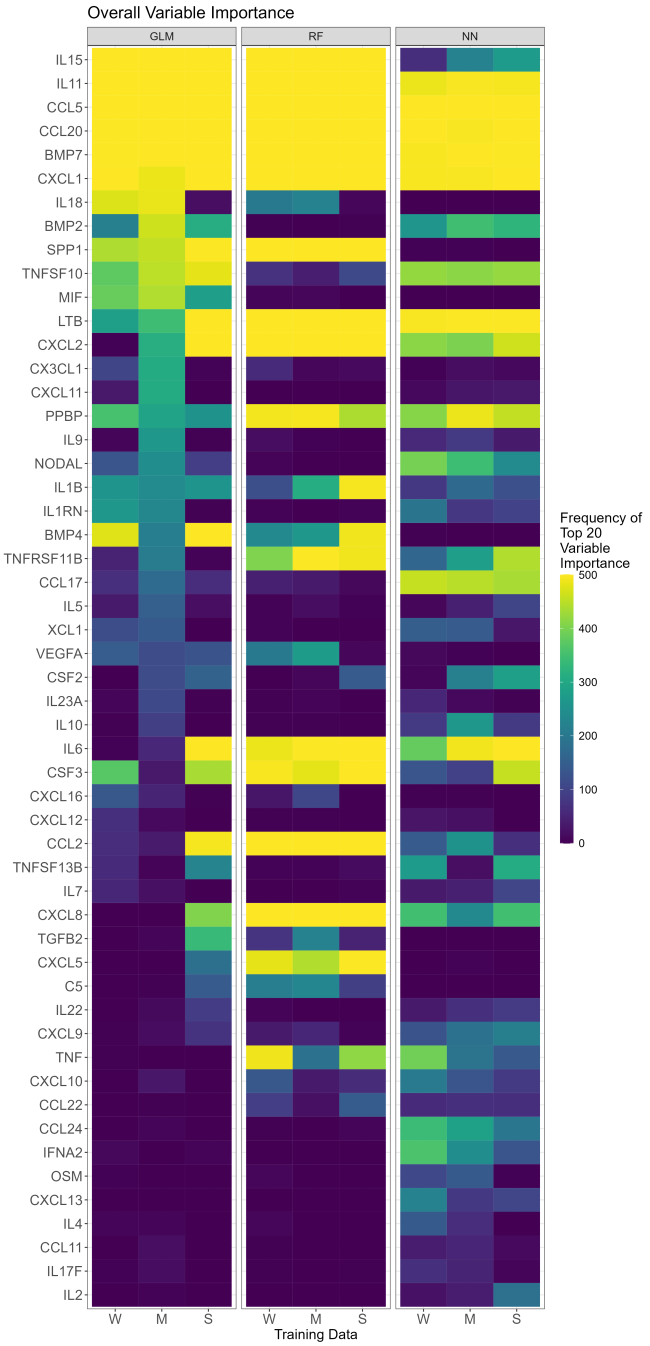


Fig. G6 Overall variable importance organized by dataset (x-axis) and model type (column panels). Heatmaps show the frequency that genes were among the top 20 variables out of 500 models. Genes are ordered based on the top 20 frequency of the GLMs trained on the M-dataset.

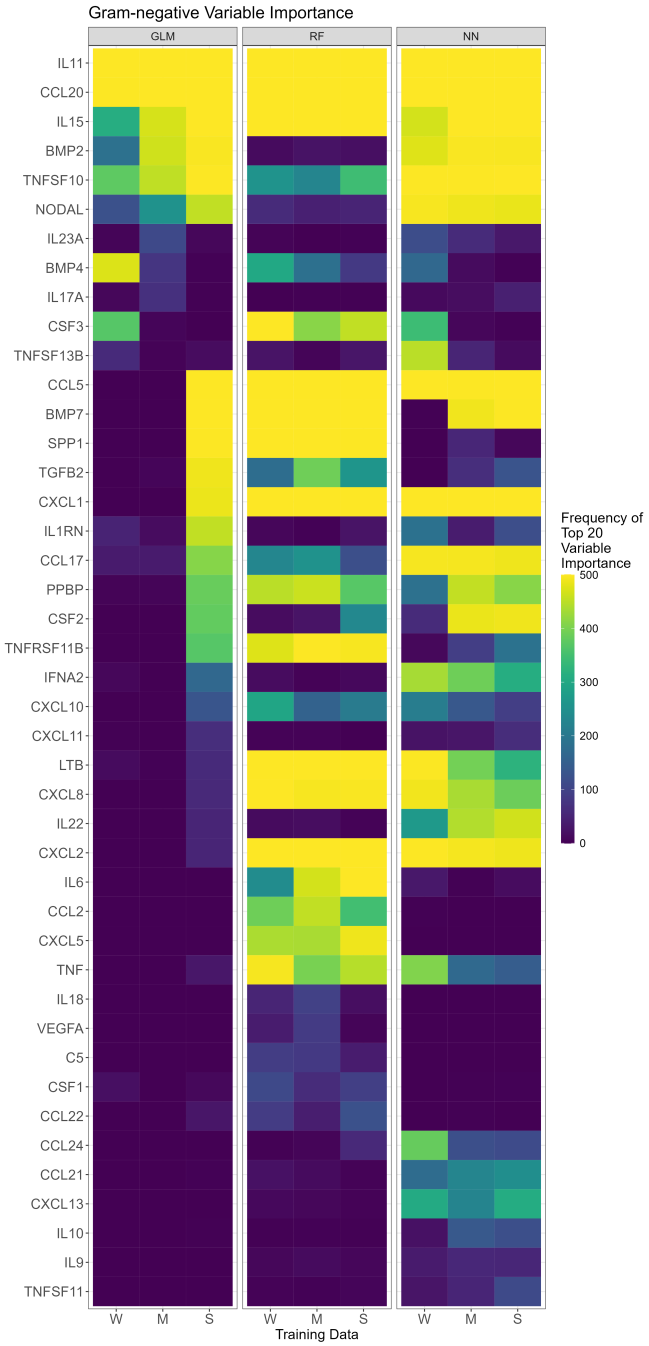


Fig. G7 Variable importance of the Gram-negative class, organized by dataset (x-axis) and model type (column panels). Heatmaps show the frequency that genes were among the top 20 variables out of 500 models. Genes are ordered based on the top 20 frequency of the GLMs trained on the M-dataset.

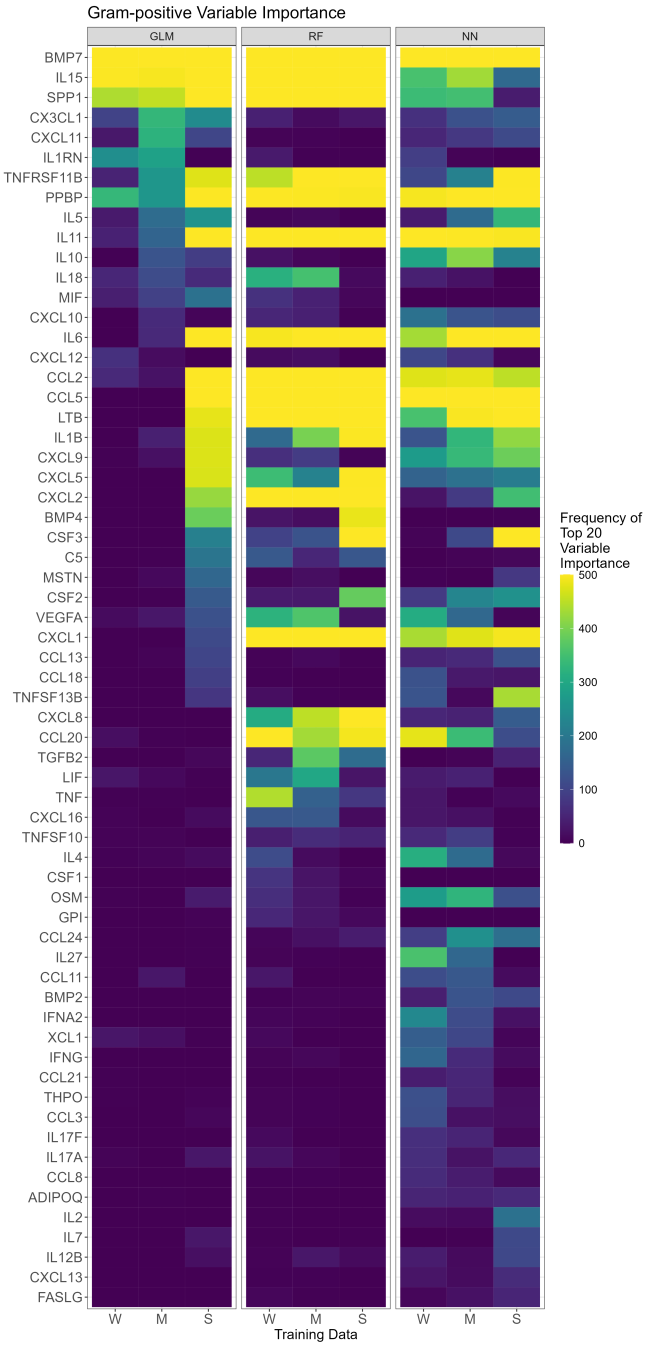


Fig. G8 Variable importance of the Gram-positive classes, organized by dataset (x-axis) and model type (column panels). Heatmaps show the frequency that genes were among the top 20 variables out of 500 models. Genes are ordered based on the top 20 frequency of the GLMs trained on the M-dataset.

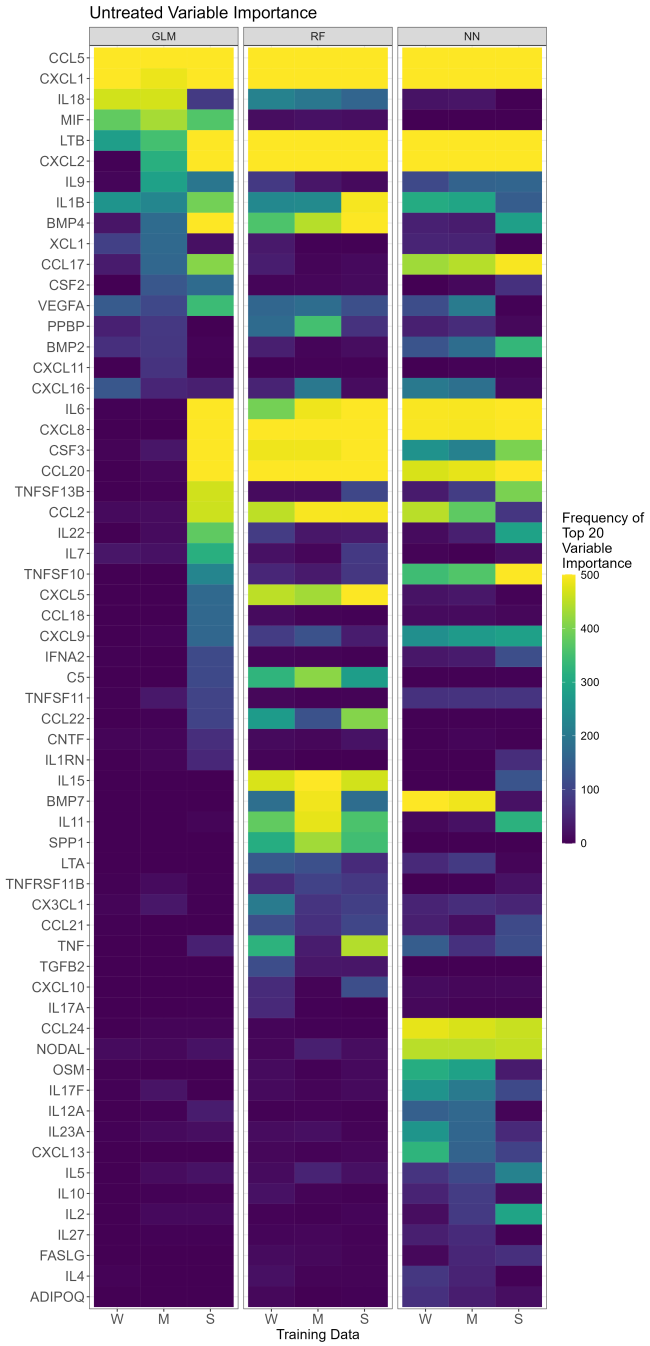


Fig. G9 Variable importance of the untreated class, organized by dataset (x-axis) and model type (column panels). Heatmaps show the frequency that genes were among the top 20 variables out of 500 models. Genes are ordered based on the top 20 frequency of the GLMs trained on the M-dataset.

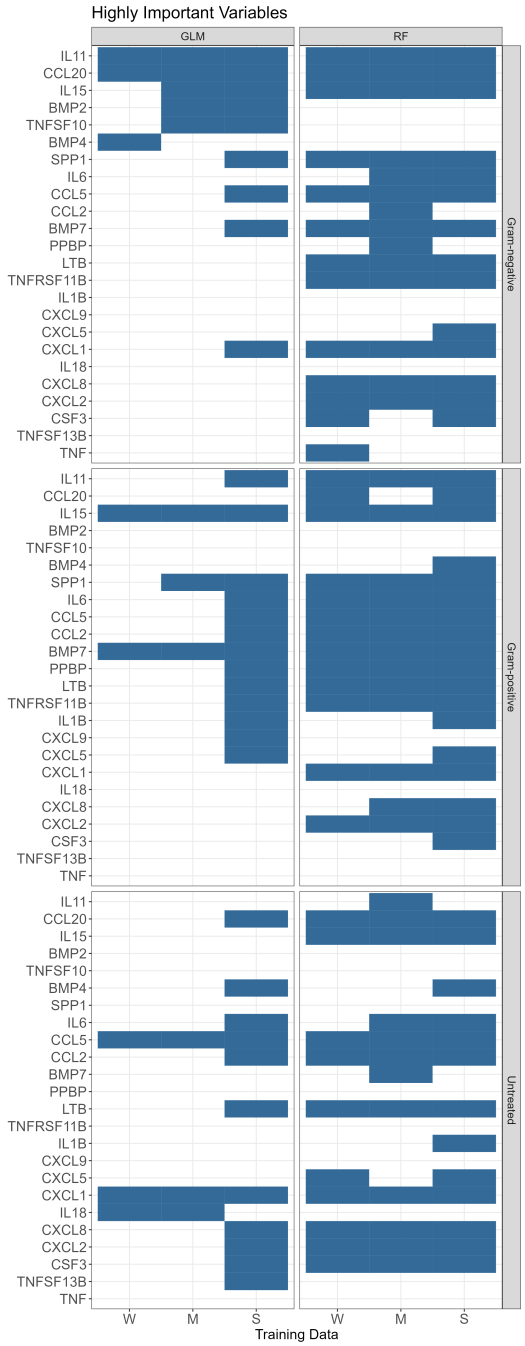


Fig. G10 Highly important variables for the GLM and RF models, organized by dataset (x-axis) and class label (row panels). Genes are considered highly important when ranked in the top 20 variables for at least 450 of the 500 model build iterations (90%).

Appendix H Sample-specific misclassification

We consider samples to be "meaningfully" misclassified if they are misclassified at least 5% of the time for the model type/dataset combination.

Gram-negative class: 28 of 50 samples (56%) are never misclassified for any model type/dataset combination (Figure H12). When only considering the GLM and RF model types, which have the highest test accuracies and comparable performance, only five samples (10%) are meaningfully misclassified in at least one model type/dataset combination (Figure H11). One of these samples is misclassified as Gram-positive, while the other four are misclassified as untreated.

Weak Gram-positive class: 23 of 32 samples (71.9%) are never misclassified when part of the test set (W- and M-datasets, Figure H13). When considering only the GLM and RF model types, only four samples (12.5%) are meaningfully misclassified in the test set (Figure H11). However, these 23 samples are meaningfully misclassified when included in the challenge set (S-dataset). All weak-signal samples (for both test and challenge sets) that are meaningfully misclassified by the GLM and RF models are misclassified as untreated.

Strong Gram-positive class: 16 of 32 samples (50%) are never misclassified when included in the test set (H- or M-datasets, Figure H14). When only considering the GLM and RF model types, only two samples (6.25%) are meaningfully misclassified in the test set (Figure H11). Both of these samples are misclassified as untreated. Additionally, 10 samples (31.25%) are meaningfully misclassified in the challenge set (W-dataset). All of these 10 samples are misclassified as Gram-negative.

Untreated class: all samples are misclassified at least once in at least one model type/dataset combinations (Figure H15). When only considering the GLM and RF model types, only 4 of 86 of these samples ($< 5\%$) are meaningfully misclassified (Figure H11). One of these samples is misclassified as Gram-negative while the other three are misclassified as Gram-positive.

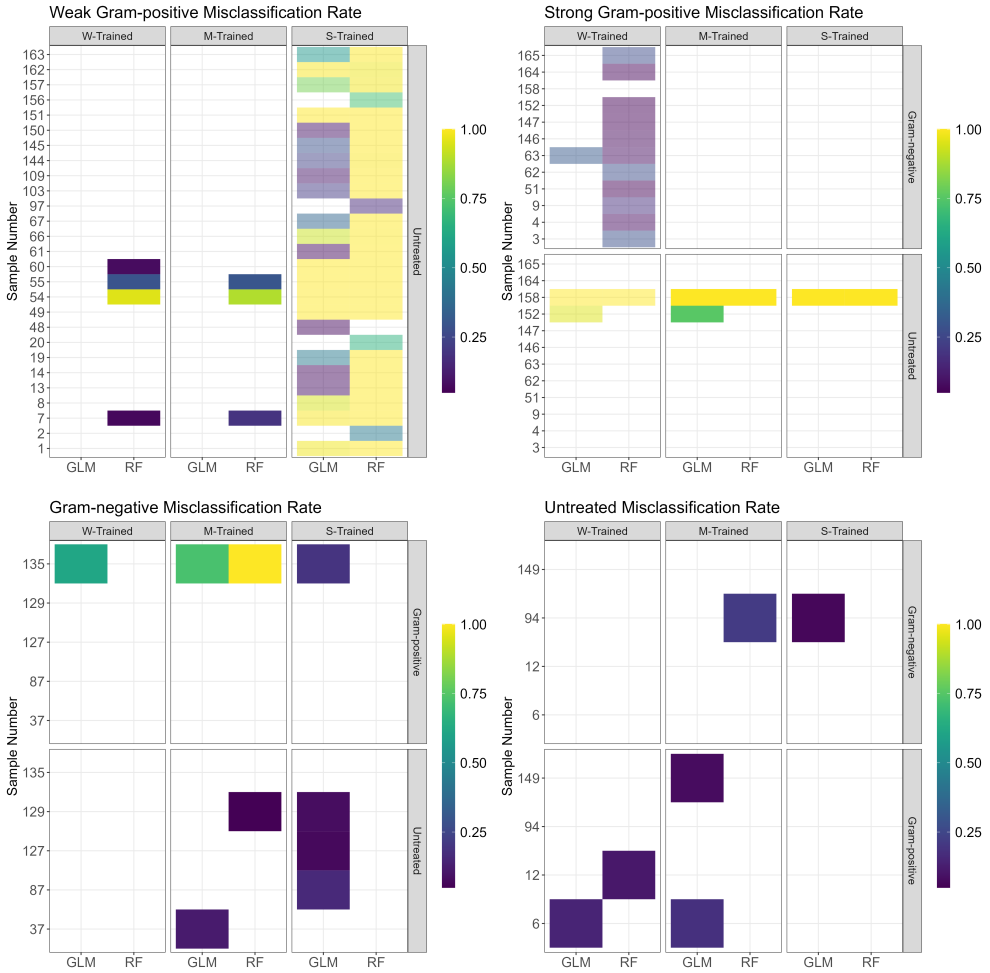


Fig. H11 Sample-specific misclassification rates from GLM and RF models for each class label, organized by predicted class (rows) and dataset curation (columns). Lighter shaded cells indicate the misclassification rates from the challenge set. Only samples that were meaningfully misclassified (i.e., misclassified at least 5% of the time) are shown here.

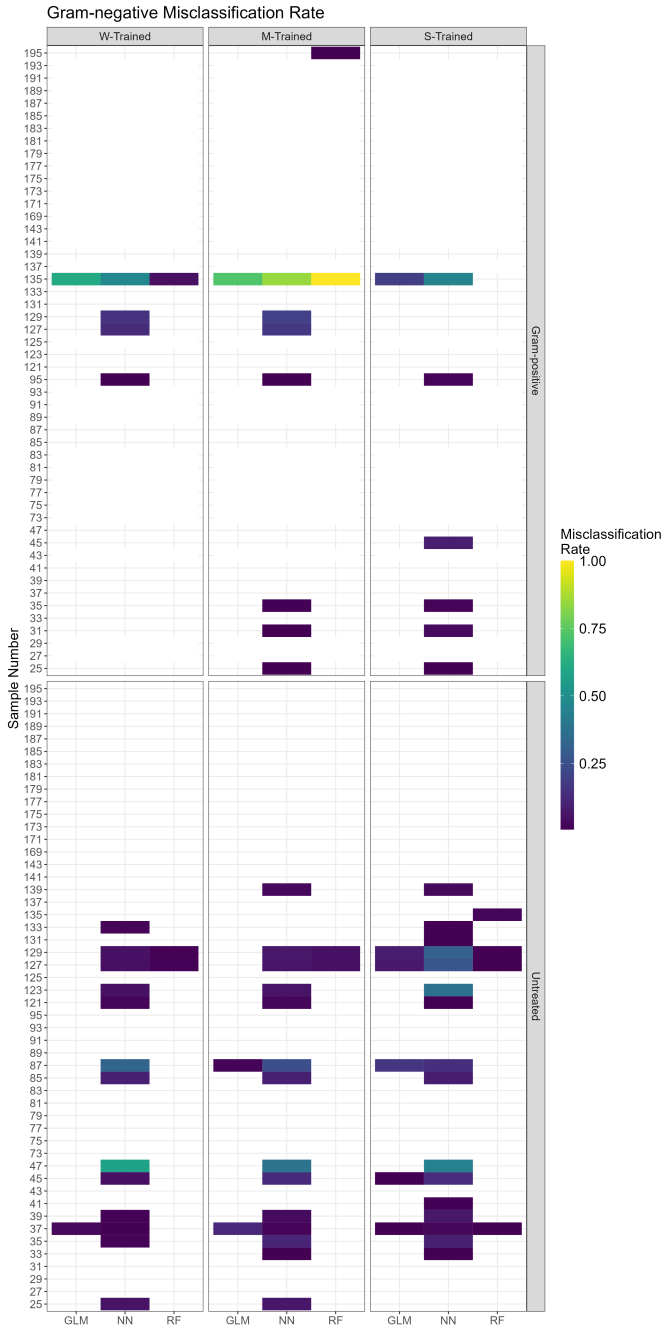


Fig. H12 Sample misclassification rates of the Gram-negative samples for the three model types, organized by prediction class (row panels) and dataset (column panels).

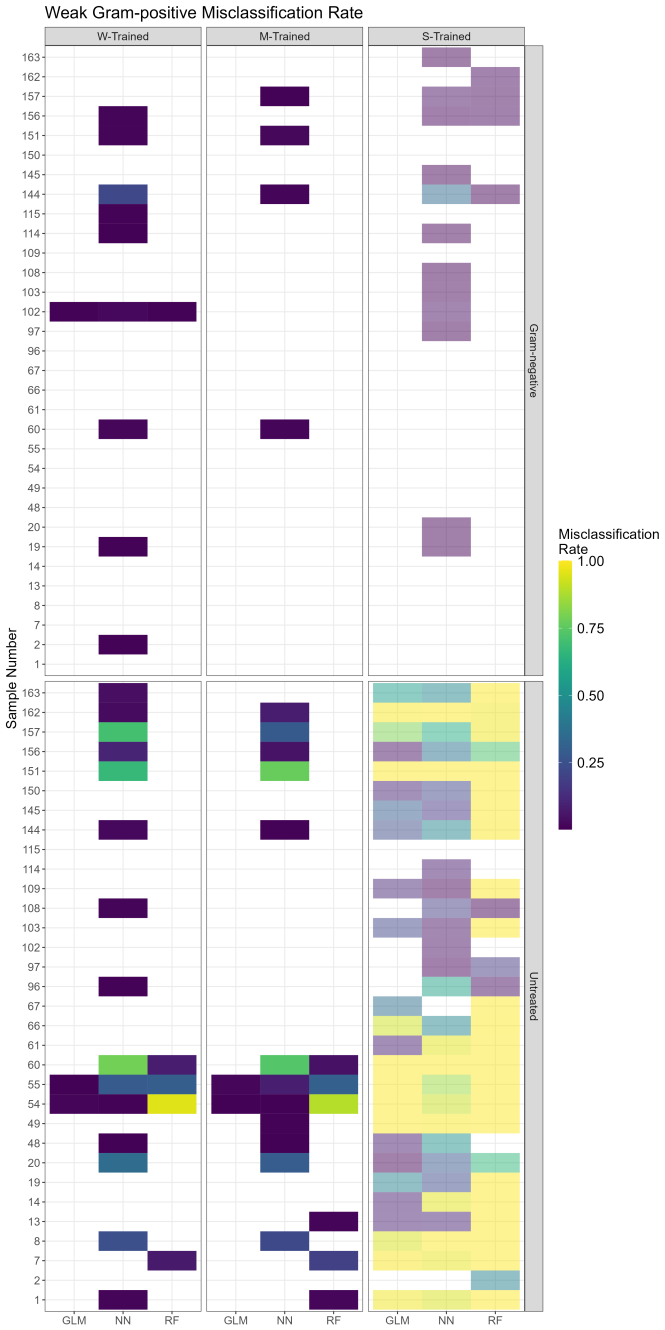


Fig. H13 Sample misclassification rates of the weak Gram-positive samples for the three model types, organized by prediction class (row panels) and dataset (column panels). Lighter shaded cells indicate the misclassification rates from the challenge set (strong-signal data).

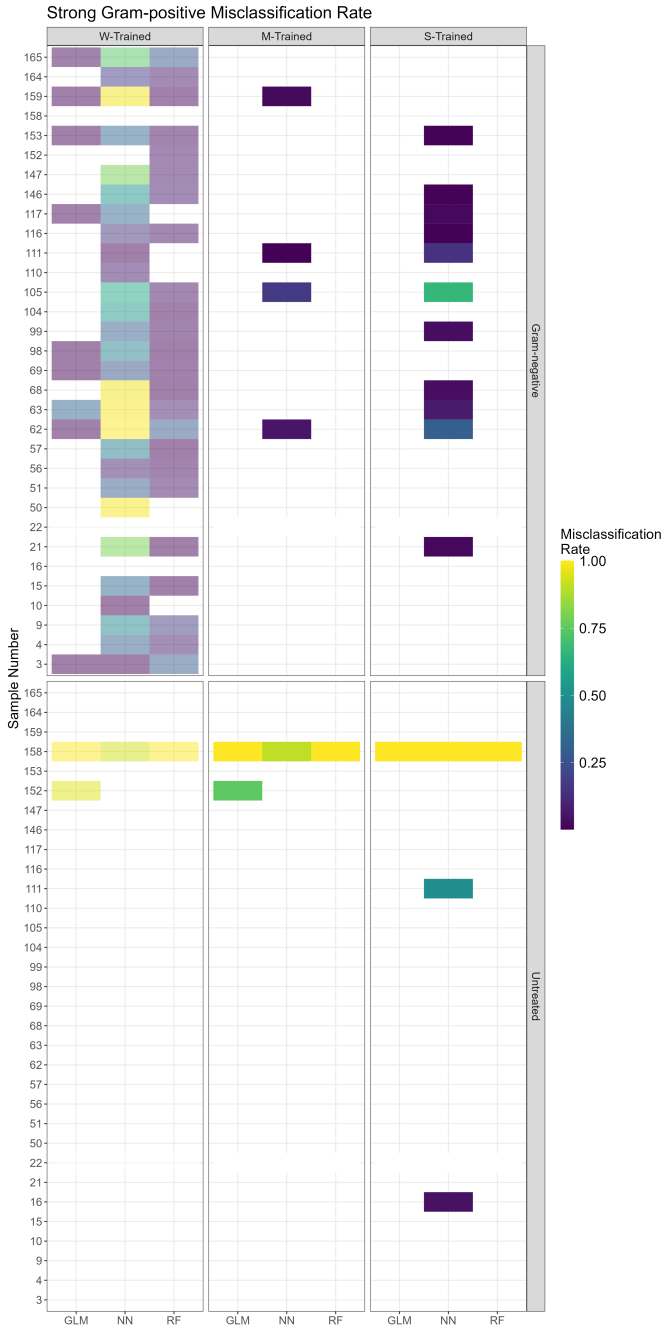


Fig. H14 Sample misclassification rates of the strong Gram-positive samples for the three model types, organized by prediction class (row panels) and dataset (column panels). Lighter shaded cells indicate the misclassification rates from the challenge set (weak-signal data).

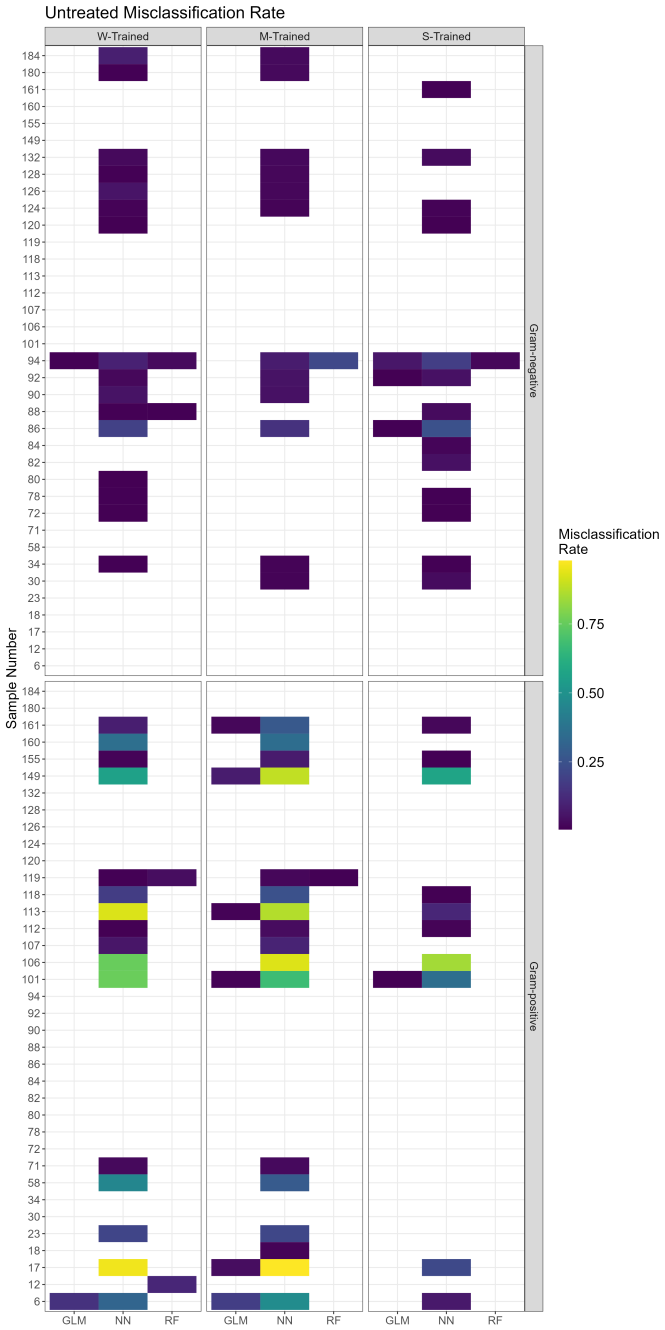


Fig. H15 Sample misclassification rates of the untreated samples for the three model types, organized by prediction class (row panels) and dataset (column panels).

H.1 Gram negative sample #135

Sample 135 is frequently-to-highly misclassified as Gram-positive by all model types trained on the W- and M-datasets and by the GLM and NN models trained on the S-dataset (Figure H12). In contrast, it is never misclassified as Gram-positive by the RF models trained on the S-dataset. This pattern reveals that the models recognize an immune response but fail to correctly distinguish the type of bacterial infection. Sample 135 is never misclassified as untreated. This type of classification uncertainty is clinically relevant and important to report.

CXCL5 is uniquely identified as highly important for the Gram-negative class in the RF models trained on the S-dataset (Figure G10). Recall that C_t values are inversely related to gene expression levels. Sample 135 exhibits the highest C_t value (i.e., the lowest $-C_t$ value) among all bacterial samples (Figure H16). Although the black-box nature and nonlinear interactions of ML models prevent definitive attribution of the cause of misclassification of sample 135, the importance of CXCL5 in the RF models trained on the S-dataset may contribute to improved classification accuracy for this sample.

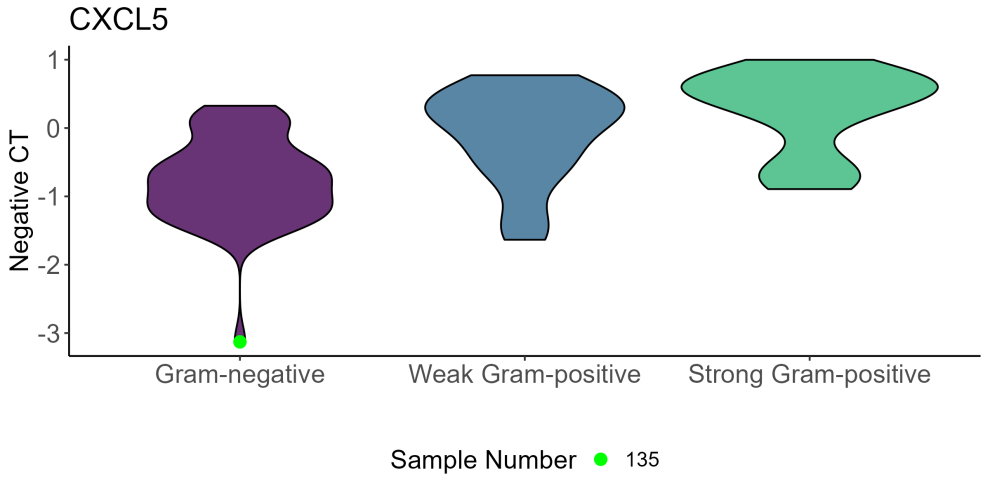


Fig. H16 CXCL5 signal distribution for the Gram-negative, weak Gram-positive, and strong Gram-positive class labels. Note that distributions show the negative of the C_t value, and therefore lower values of the scale correspond to weaker levels of gene expression. Sample 135 is frequently-to-highly misclassified as Gram-positive for all model types trained on the W- and M-datasets and for the GLM and NN models trained on the S-dataset. It is never misclassified as Gram-positive for the RF models trained on the S-dataset. CXCL5 is only highly important for the Gram-negative class in the RF models trained on the S-dataset. Sample 135 also has the lowest CXCL5 signal out of the three bacterial class labels.

H.2 Strong Gram-positive sample #152

Variation in highly important variables across model type/dataset combinations may explain why certain samples are more often correctly classified in some combinations than in others. Sample 152 is frequently misclassified as untreated by the GLMs trained on the W- and M-datasets (Figure H14). CCL5, IL6, and LTB are highly important for the Gram-positive class in all model types trained on the S-dataset, but are not highly important for the Gram-positive class in the GLMs trained on the W- and M-datasets (Figure G10). The upregulation of CCL5, IL6, and LTB in sample 152 exceeds the corresponding distributions observed in untreated samples (Figure H17). The importance of CCL5, IL6, and LTB in models trained on the strong-signal data may explain why sample 152 is more often correctly classified in those models.

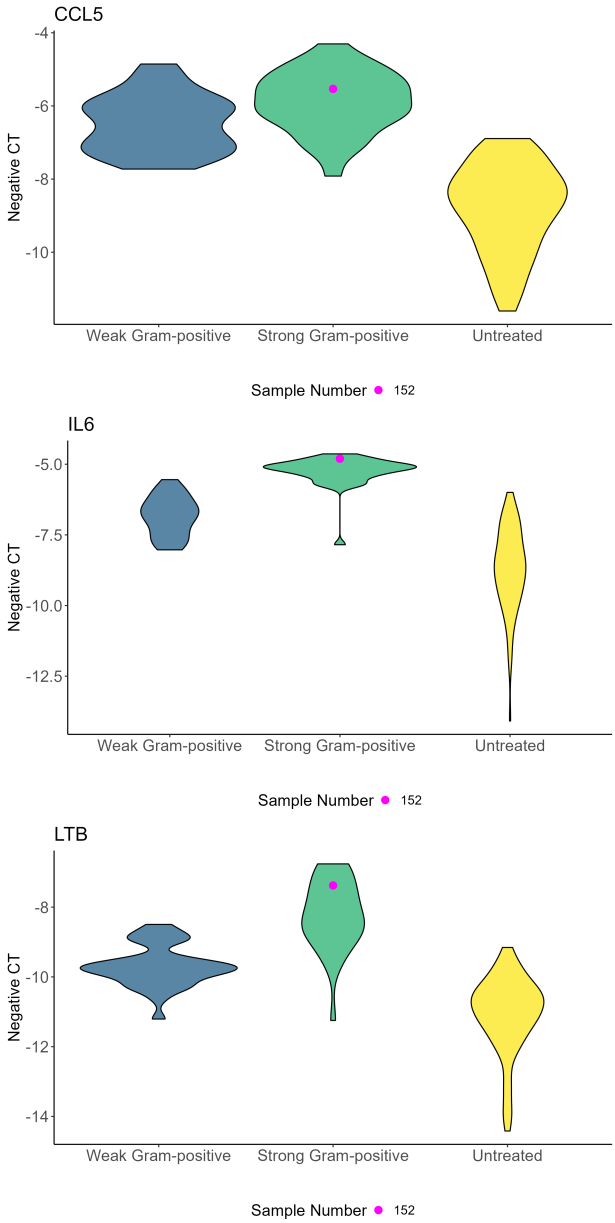


Fig. H17 Signal distributions for the strong Gram-positive and untreated samples for CCL5, IL6, and LTB. Note that distributions show the negative of the C_t value, and therefore higher values of the scale correspond to stronger levels of gene expression. Sample 152 is highly misclassified as untreated for the GLMs trained on the W- and M-datasets. CCL5, IL6, and LTB are highly important for the Gram-positive class for all model types trained on the S-dataset, but are not highly important for the Gram-positive class trained on the GLMs of the W- and M-datasets. The up-regulation of CCL5, IL6, and LTB for sample 152 is higher than the respective distributions of the untreated samples.

H.3 Strong Gram-positive sample #158

Models may have difficulty accurately predicting class labels when a sample's feature values overlap with the distribution of another class, particularly when those features are deemed highly important for classification. Sample 158 is frequently misclassified as untreated across all model type/dataset combinations (Figure H14). CCL5 is highly important for the Gram-positive class in seven out of nine model type/dataset combinations and is also highly important for the untreated class across all model type/dataset combinations (Figure G10). The CCL5 signal for sample 158 is indistinguishable from the range observed in untreated samples (Figure H18). These results suggest that although CCL5 is informative for class discrimination in most samples, overlapping signals can limit its utility for accurately classifying specific samples such as Sample 158.

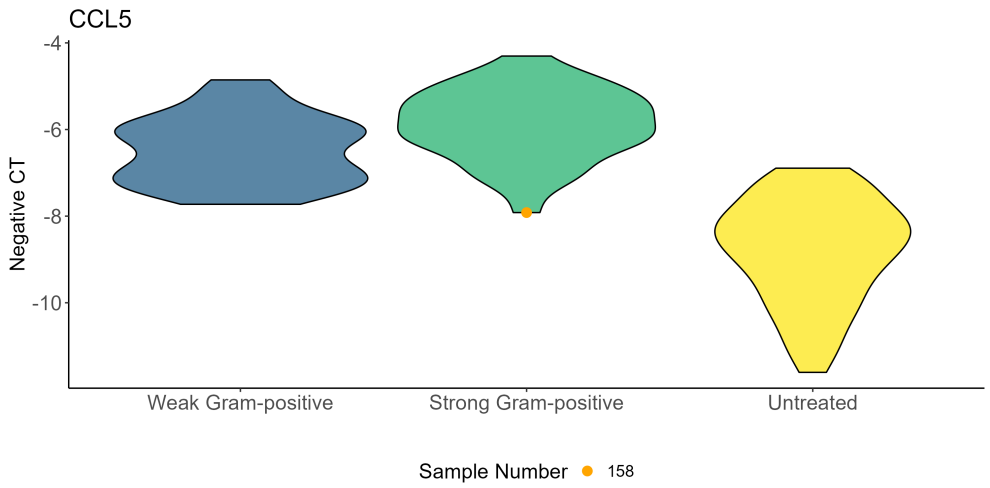


Fig. H18 Signal distribution for the strong Gram-positive and untreated samples for CCL5. Note that distributions show the negative of the C_t value, and therefore lower values of the scale correspond to weaker levels of gene expression. Sample 158 is highly misclassified as untreated for all model type/dataset combinations. CCL5 is highly important for the Gram-positive class in 7 out of 9 model type/dataset combinations, and is highly important for the untreated class for all model type/dataset combinations. The CCL5 signal is indistinguishable from the range of untreated samples.

H.4 Weak Gram-positive sample #54

The number of highly important variables in a model may also explain misclassification patterns. Models with a larger number of highly important variables may, in some cases, obscure meaningful signals that are more clearly captured by simpler models. Sample 54 is frequently misclassified as untreated by the RF models trained on the W- and M-datasets and by all model types trained on the S-dataset (Figure H13). In contrast, Sample 54 is rarely misclassified by the GLM and NN models trained

on the W- and M-datasets. For the W- and M-datasets, no genes are highly important for GLM or NN models that are not also highly important for the RF models (Figure G10). BMP7 is highly important for the Gram-positive class across all nine model type/dataset combinations. The BMP7 signal for Sample 54 is slightly weaker than that observed in untreated samples (Figure H19). These results suggest that, for certain samples, models trained on fewer variables may achieve improved classification accuracy.

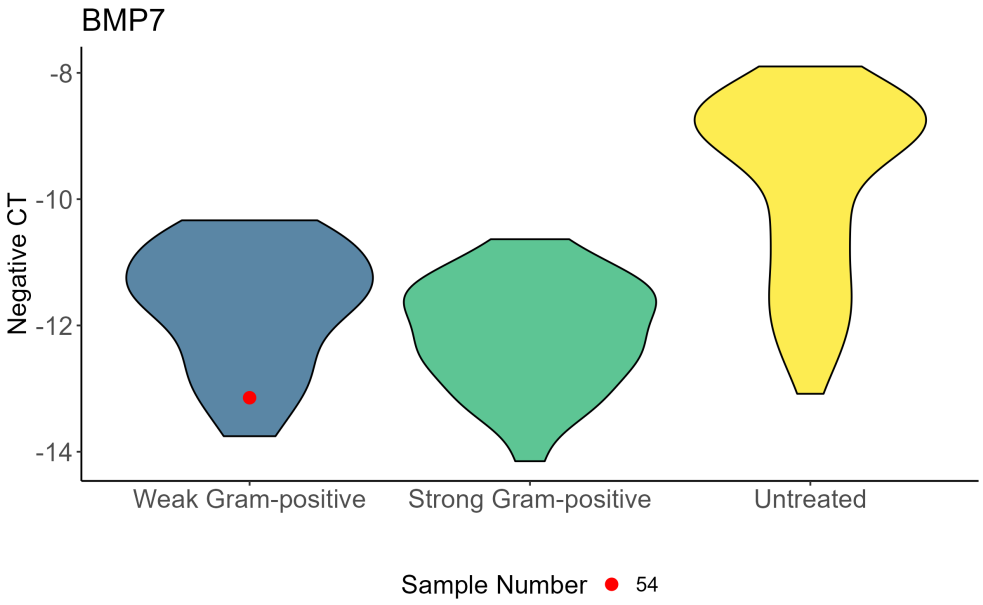


Fig. H19 Signal distribution for the weak Gram-positive and untreated samples for BMP7. Note that distributions show the negative of the C_t value, and therefore lower values of the scale correspond to weaker levels of gene expression. Sample 54 is highly misclassified as untreated for the RF models trained on the W- and M-datasets, and for all model types trained on the S-dataset. However, sample 54 is rarely misclassified for the GLM and NN models of the W- and M-datasets. For the W- and M-datasets, no genes are highly important for GLM or NN models that are not highly important for the RF models. BMP7 is highly important for all model type/dataset combinations for the Gram-positive class. The BMP7 signal for sample 54 is slightly weaker than that from any of the untreated samples. Sample 54 provides evidence to suggest that some samples may have better classification accuracy when trained on fewer variables.

References

[1] Organization, W.H.: Antimicrobial resistance (2023). https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance Accessed 2025-07-29

[2] Kubicek-Sutherland, J.Z., Xie, G., Shakya, M., Dighe, P.K., Jacobs, L.L., Daligault, H., Davenport, K., Stromberg, L.R., Stromberg, Z.R., Cheng, Q., Kempaiah, P., Ong'echa, J.M., Otieno, V., Raballah, E., Anyona, S., Ouma, C., Chain, P.S.G., Perkins, D.J., Mukundan, H., McMahon, B.H., Doggett, N.A.: Comparative genomic and phenotypic characterization of invasive non-typhoidal Salmonella isolates from Siaya, Kenya. PLoS Neglected Tropical Diseases **15**(2), 0008991 (2021) https://doi.org/10.1371/journal.pntd.0008991 . Accessed 2026-01-12

[3] Murray, C.J.L., Ikuta, K.S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S.C., Browne, A.J., Chipeta, M.G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B.H., Kumaran, E.A.P., McManigal, B., Achalapong, S., Agarwal, R., Akech, S., Albertson, S., Amuasi, J., Andrews, J., Aravkin, A., Ashley, E., Babin, F.-X., Bailey, F., Baker, S., Basnyat, B., Bekker, A., Bender, R., Berkley, J.A., Bethou, A., Bielicki, J., Boonkasidecha, S., Bukosia, J., Carvalheiro, C., Castañeda-Orjuela, C., Chansamouth, V., Chaurasia, S., Chiurchiù, S., Chowdhury, F., Clotaire Donatien, R., Cook, A.J., Cooper, B., Cressey, T.R., Criollo-Mora, E., Cunningham, M., Darboe, S., Day, N.P.J., De Luca, M., Dokova, K., Dramowski, A., Dunachie, S.J., Duong Bich, T., Eckmanns, T., Eibach, D., Emami, A., Feasey, N., Fisher-Pearson, N., Forrest, K., Garcia, C., Garrett, D., Gastmeier, P., Giref, A.Z., Greer, R.C., Gupta, V., Haller, S., Haselbeck, A., Hay, S.I., Holm, M., Hopkins, S., Hsia, Y., Iregbu, K.C., Jacobs, J., Jarovsky, D., Javanmardi, F., Jenney, A.W.J., Khorana, M., Khusuwan, S., Kissoon, N., Kobeissi, E., Kostyanev, T., Krapp, F., Krumkamp, R., Kumar, A., Kyu, H.H., Lim, C., Lim, K., Limmathurotsakul, D., Loftus, M.J., Lunn, M., Ma, J., Manoharan, A., Marks, F., May, J., Mayxay, M., Mturi, N., Munera-Huertas, T., Musicha, P., Musila, L.A., Mussi-Pinhata, M.M., Naidu, R.N., Nakamura, T., Nanavati, R., Nangia, S., Newton, P., Ngoun, C., Novotney, A., Nwakanma, D., Obiero, C.W., Ochoa, T.J., Olivas-Martinez, A., Olliaro, P., Ooko, E., Ortiz-Brizuela, E., Ounchanum, P., Pak, G.D., Paredes, J.L., Peleg, A.Y., Perrone, C., Phe, T., Phommasone, K., Plakkal, N., Ponce-de-Leon, A., Raad, M., Ramdin, T., Rattanavong, S., Riddell, A., Roberts, T., Robotham, J.V., Roca, A., Rosenthal, V.D., Rudd, K.E., Russell, N., Sader, H.S., Saengchan, W., Schnall, J., Scott, J.A.G., Seekaew, S., Sharland, M., Shivamallappa, M., Sifuentes-Osornio, J., Simpson, A.J., Steenkeste, N., Stewardson, A.J., Stoeva, T., Tasak, N., Thaiprakong, A., Thwaites, G., Tigoi, C., Turner, C., Turner, P., Doorn, H.R., Velaphi, S., Vongpradith, A., Vongsouvath, M., Vu, H., Walsh, T., Walson, J.L., Waner, S., Wangrangsimakul, T., Wannapinij, P., Wozniak, T., Young Sharma, T.E.M.W., Yu, K.C., Zheng, P., Sartorius, B., Lopez, A.D., Stergachis, A., Moore, C., Dolecek, C., Naghavi, M.: Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet **399**(10325), 629–655 (2022) https://doi.org/10.1016/S0140-6736(21)02724-0 . Accessed 2025-11-18

[4] Andersson, D.I., Hughes, D., Kubicek-Sutherland, J.Z.: Mechanisms and consequences of bacterial resistance to antimicrobial peptides. Drug Resistance

Updates **26**, 43–57 (2016) https://doi.org/10.1016/j.drup.2016.04.002 . Accessed 2026-01-12

[5] Caliendo, A.M., Gilbert, D.N., Ginocchio, C.C., Hanson, K.E., May, L., Quinn, T.C., Tenover, F.C., Alland, D., Blaschke, A.J., Bonomo, R.A., Carroll, K.C., Ferraro, M.J., Hirschhorn, L.R., Joseph, W.P., Karchmer, T., MacIntyre, A.T., Reller, L.B., Jackson, A.F.: Better Tests, Better Care: Improved Diagnostics for Infectious Diseases. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America **57**(Suppl 3), 139–170 (2013) https://doi.org/10.1093/cid/cit578 . Accessed 2025-11-13

[6] Breijyeh, Z., Jubeh, B., Karaman, R.: Resistance of Gram-Negative Bacteria to Current Antibacterial Agents and Approaches to Resolve It. Molecules **25**(6), 1340 (2020) https://doi.org/10.3390/molecules25061340 . Accessed 2025-10-08

[7] Uddin, T.M., Chakraborty, A.J., Khusro, A., Zidan, B.R.M., Mitra, S., Emran, T.B., Dhama, K., Ripon, M.K.H., Gajdács, M., Sahibzada, M.U.K., Hossain, M.J., Koirala, N.: Antibiotic resistance in microbes: History, mechanisms, therapeutic strategies and future prospects. Journal of Infection and Public Health **14**(12), 1750–1766 (2021) https://doi.org/10.1016/j.jiph.2021.10.020 . Accessed 2025-10-08

[8] Saxena, D., Maitra, R., Bormon, R., Czekanska, M., Meiers, J., Titz, A., Verma, S., Chopra, S.: Tackling the outer membrane: facilitating compound entry into Gram-negative bacterial pathogens. npj Antimicrobials and Resistance **1**(1), 17 (2023). Accessed 2025-10-09

[9] Arbefeville, S.S., Timbrook, T.T., Garner, C.D.: Evolving strategies in microbe identification—a comprehensive review of biochemical, MALDI-TOF MS and molecular testing methods. Journal of Antimicrobial Chemotherapy **79**(Suppl 1), 2–8 (2024) https://doi.org/10.1093/jac/dkae275 . Accessed 2025-11-11

[10] Justiz Vaillant, A.A., Sabir, S., Jan, A.: Physiology, Immune Response. In: StatPearls. StatPearls Publishing, ??? (2025). http://www.ncbi.nlm.nih.gov/books/NBK539801/ Accessed 2025-10-08

[11] Mogensen, T.H.: Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. Clinical Microbiology Reviews **22**(2), 240–273 (2009) https://doi.org/10.1128/CMR.00046-08 . Accessed 2025-10-08

[12] Li, D., Wu, M.: Pattern recognition receptors in health and diseases. Signal Transduction and Targeted Therapy **6**, 291 (2021) https://doi.org/10.1038/s41392-021-00687-0 . Accessed 2026-01-12

[13] Kagan, J.C.: Defining the subcellular sites of innate immune signal transduction. Trends in immunology **33**(9), 442–448 (2012) https://doi.org/10.1016/j.it.2012.06.005 . Accessed 2025-10-08

[14] Cao, J., Xiao, Y., Zhang, M., Huang, L., Wang, Y., Liu, W., Wang, X., Wu, J., Huang, Y., Wang, R., Zhou, L., Li, L., Zhang, Y., Ren, L., Qian, K., Wang, J.: Deep Learning of Dual Plasma Fingerprints for High-Performance Infection Classification. Small **19**(7) (2023) https://doi.org/10.1002/smll.202206349 . Accessed 2025-10-06

[15] Gunčar, G., Kukar, M., Smole, T., Moškon, S., Vovko, T., Podnar, S., Černelč, P., Brvar, M., Notar, M., Köster, M., Jelenc, M.T., Osterc, Ž., Notar, M.: Differentiating viral and bacterial infections: A machine learning model based on routine blood test values. Heliyon **10**(8), 29372 (2024) https://doi.org/10.1016/j.heliyon.2024.e29372 . Accessed 2025-07-29

[16] Xie, J., Zheng, X., Yan, J., Li, Q., Jin, N., Wang, S., Zhao, P., Li, S., Ding, W., Cheng, L., Geng, Q.: Deep learning model to discriminate diverse infection types based on pairwise analysis of host gene expression. iScience **27**(6) (2024) https://doi.org/10.1016/j.isci.2024.109908 . Accessed 2025-10-06

[17] Zhang, J., Friberg, I.M., Kift-Morgan, A., Parekh, G., Morgan, M.P., Liuzzi, A.R., Lin, C.-Y., Donovan, K.L., Colmont, C.S., Morgan, P.H., Davis, P., Weeks, I., Fraser, D.J., Topley, N., Eberl, M.: Machine-learning algorithms define pathogen-specific local immune fingerprints in peritoneal dialysis patients with bacterial infections. Kidney International **92**(1), 179–191 (2017) https://doi.org/10.1016/j.kint.2017.01.017 . Accessed 2025-10-06

[18] Zhang, F., Wang, H., Liu, L., Su, T., Ji, B.: Machine learning model for the prediction of gram-positive and gram-negative bacterial bloodstream infection based on routine laboratory parameters. BMC Infectious Diseases **23**(1), 675 (2023) https://doi.org/10.1186/s12879-023-08602-4 . Accessed 2025-11-10

[19] Zhang, T., Deng, Y., Wang, W., Zhao, Z., Wu, Y., Wang, H., Xia, S., Liao, W., Liao, W.: Stacking ensemble learning models diagnose pulmonary infections using host transcriptome data from metatranscriptomics. Scientific Reports **15**, 30516 (2025) https://doi.org/10.1038/s41598-025-15914-9 . Accessed 2025-11-10

[20] Im, H., Hwang, S.-H., Kim, B.S., Choi, S.H.: Pathogenic potential assessment of the Shiga toxin–producing Escherichia coli by a source attribution–considered machine learning model. Proceedings of the National Academy of Sciences **118**(20), 2018877118 (2021) https://doi.org/10.1073/pnas.2018877118 . Accessed 2025-11-17

[21] Sen, R., Nayak, L., De, R.K.: A review on host–pathogen interactions: classification and prediction. European Journal of Clinical Microbiology & Infectious Diseases **35**(10), 1581–1599 (2016) https://doi.org/10.1007/s10096-016-2716-7 . Accessed 2025-10-08

[22] Bodkin, N., Ross, M., McClain, M.T., Ko, E.R., Woods, C.W., Ginsburg, G.S.,

Henao, R., Tsalik, E.L.: Systematic comparison of published host gene expression signatures for bacterial/viral discrimination. Genome Medicine **14**, 18 (2022) https://doi.org/10.1186/s13073-022-01025-x . Accessed 2025-10-09

[23] Ko, E.R., Reller, M.E., Tillekeratne, L.G., Bodinayake, C.K., Miller, C., Burke, T.W., Henao, R., McClain, M.T., Suchindran, S., Nicholson, B., Blatt, A., Petzold, E., Tsalik, E.L., Nagahawatte, A., Devasiri, V., Rubach, M.P., Maro, V.P., Lwezaula, B.F., Kodikara-Arachichi, W., Kurukulasooriya, R., De Silva, A.D., Clark, D.V., Schully, K.L., Madut, D., Dumler, J.S., Kato, C., Galloway, R., Crump, J.A., Ginsburg, G.S., Minogue, T.D., Woods, C.W.: Host-response transcriptional biomarkers accurately discriminate bacterial and viral infections of global relevance. Scientific Reports **13**(1) (2023) https://doi.org/10.1038/s41598-023-49734-6 . Accessed 2025-07-29

[24] Steinbrink, J.M., Liu, Y., Henao, R., Tsalik, E.L., Ginsburg, G.S., Ramsburg, E., Woods, C.W., McClain, M.T.: Pathogen class-specific transcriptional responses derived from PBMCs accurately discriminate between fungal, bacterial, and viral infections. PLOS ONE **19**(12), 0311007 (2024) https://doi.org/10.1371/journal.pone.0311007 . Accessed 2025-11-17

[25] Scherr, T.F., Douglas, C.E., Schaecher, K.E., Schoepp, R.J., Ricks, K.M., Shoemaker, C.J.: Application of a Machine Learning-Based Classification Approach for Developing Host Protein Diagnostic Models for Infectious Disease. Diagnostics **14**(12), 1290 (2024) https://doi.org/10.3390/diagnostics14121290 . Accessed 2025-11-10

[26] Blankley, S., Graham, C.M., Howes, A., Bloom, C.I., Berry, M.P.R., Chaussabel, D., Pascual, V., Banchereau, J., Lipman, M., O'Garra, A.: Identification of the Key Differential Transcriptional Responses of Human Whole Blood Following TLR2 or TLR4 Ligation In-Vitro. PLOS ONE **9**(5), 97702 (2014) https://doi.org/10.1371/journal.pone.0097702 . Accessed 2025-11-17

[27] Miller, T.E., Garcia Beltran, W.F., Bard, A.Z., Gogakos, T., Anahtar, M.N., Astudillo, M.G., Yang, D., Thierauf, J., Fisch, A.S., Mahowald, G.K., Fitzpatrick, M.J., Nardi, V., Feldman, J., Hauser, B.M., Caradonna, T.M., Marble, H.D., Ritterhouse, L.L., Turbett, S.E., Batten, J., Georgantas, N.Z., Alter, G., Schmidt, A.G., Harris, J.B., Gelfand, J.A., Poznansky, M.C., Bernstein, B.E., Louis, D.N., Dighe, A., Charles, R.C., Ryan, E.T., Branda, J.A., Pierce, V.M., Murali, M.R., Iafrate, A.J., Rosenberg, E.S., Lennerz, J.K.: Clinical sensitivity and interpretation of PCR and serological COVID-19 diagnostics for patients presenting to the hospital. The FASEB Journal **34**(10), 13877–13884 (2020) https://doi.org/10.1096/fj.202001700RR . Accessed 2025-10-08

[28] Sharma, V., Patro, S., Choudhary, A., Pradhan, S., Pathi, B.: Combined Predictive Value of Serum Procalcitonin and Platelet Indices in Differentiating Gram-Negative and Gram-Positive Bloodstream Infections. Cureus **17**(9), 91390

(2025) https://doi.org/10.7759/cureus.91390 . Accessed 2025-11-10

[29] Martinez, K.M., Wilding, K., Llewellyn, T.R., Jacobsen, D.E., Montoya, M.M., Kubicek-Sutherland, J.Z., Batni, S., Manore, C., Mukundan, H.: Evaluating the factors influencing accuracy, interpretability, and reproducibility in the use of machine learning classifiers in biology to enable standardization. Scientific Reports **15**(1) (2025) https://doi.org/10.1038/s41598-025-00245-6 . Accessed 2025-05-22

[30] Jacobsen, D.E., Montoya, M.M., Llewellyn, T.R., Martinez, K., Wilding, K.M., Lenz, K.D., Manore, C.A., Kubicek-Sutherland, J.Z., Mukundan, H.: Correlating transcription and protein expression profiles of immune biomarkers following lipopolysaccharide exposure in lung epithelial cells. PloS One **19**(4), 0293680 (2024) https://doi.org/10.1371/journal.pone.0293680

[31] Gambhir, V., Yildiz, C., Mulder, R., Siddiqui, S., Guzzo, C., Szewczuk, M., Gee, K., Basta, S.: The TLR2 agonists lipoteichoic acid and Pam3CSK4 induce greater pro-inflammatory responses than inactivated *Mycobacterium butyricum*. Cellular Immunology **280**(1), 101–107 (2012) https://doi.org/10.1016/j.cellimm.2012.12.001 . Accessed 2025-01-28

[32] Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Coffee, M., Zeng, Z., Li, S., Ting-hao, Huang, Wang, Z., Rehg, J.M., Kautz, H., Zhang, A.: The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning. arXiv (2025). https://doi.org/10.48550/arXiv.2402.12715 . http://arxiv.org/abs/2402.12715 Accessed 2026-01-12

[33] Heid, C.A., Stevens, J., Livak, K.J., Williams, P.M.: Real time quantitative PCR. Genome Research **6**(10), 986–994 (1996) https://doi.org/10.1101/gr.6.10.986

[34] Rabaan, A.A., Tirupathi, R., Sule, A.A., Aldali, J., Mutair, A.A., Alhumaid, S., Muzaheed, Gupta, N., Koritala, T., Adhikari, R., Bilal, M., Dhawan, M., Tiwari, R., Mitra, S., Emran, T.B., Dhama, K.: Viral Dynamics and Real-Time RT-PCR Ct Values Correlation with Disease Severity in COVID-19. Diagnostics **11**(6), 1091 (2021) https://doi.org/10.3390/diagnostics11061091 . Accessed 2026-01-12

[35] Charilaou, P., Battat, R.: Machine learning models and over-fitting considerations. World Journal of Gastroenterology **28**(5), 605–607 (2022) https://doi.org/10.3748/wjg.v28.i5.605 . Accessed 2026-01-12

[36] Zivich, P.N., Naimi, A.I.: A primer on neural networks. American Journal of Epidemiology **194**(6), 1473–1475 (2024) https://doi.org/10.1093/aje/kwae380 . Accessed 2026-01-12

[37] Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324 . Accessed 2026-01-12

[38] Arnold, K.F., Davies, V., Kamps, M., Tennant, P.W.G., Mbotwa, J., Gilthorpe, M.S.: Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. International Journal of Epidemiology **49**(6), 2074–2082 (2020) https://doi.org/10.1093/ije/dyaa049 . Accessed 2026-01-12

[39] Wainer, J., Cawley, G.: Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Systems with Applications **182**, 115222 (2021) https://doi.org/10.1016/j.eswa.2021.115222 . Accessed 2026-01-12

[40] Davenport, E.E., Burnham, K.L., Radhakrishnan, J., Humburg, P., Hutton, P., Mills, T.C., Rautanen, A., Gordon, A.C., Garrard, C., Hill, A.V.S., Hinds, C.J., Knight, J.C.: Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. The Lancet Respiratory Medicine **4**(4), 259–271 (2016) https://doi.org/10.1016/S2213-2600(16)00046-1 . Accessed 2026-02-05

[41] Horst, R., Jaeger, M., Smeekens, S.P., Oosting, M., Swertz, M.A., Li, Y., Kumar, V., Diavatopoulos, D.A., Jansen, A.F.M., Lemmers, H., Toenhake-Dijkstra, H., Herwaarden, A.E., Janssen, M., Molen, R.G., Joosten, I., Sweep, F.C.G.J., Smit, J.W., Netea-Maier, R.T., Koenders, M.M.J.F., Xavier, R.J., Meer, J.W.M., Dinarello, C.A., Pavelka, N., Wijmenga, C., Notebaart, R.A., Joosten, L.A.B., Netea, M.G.: Host and environmental factors influencing individual human cytokine responses. Cell **167**(4), 1111–112413 (2016) https://doi.org/10.1016/j.cell.2016.10.018 . Accessed 2026-02-05

[42] Sfikakis, P.P., Verrou, K.-M., Ampatziadis-Michailidis, G., Tsitsilonis, O., Paraskevis, D., Kastritis, E., Lianidou, E., Moutsatsou, P., Terpos, E., Trougakos, I., Chini, V., Manoloukos, M., Moulos, P., Pavlopoulos, G.A., Kollias, G., Hatzis, P., Dimopoulos, M.A.: Blood Transcriptomes of Anti-SARS-CoV-2 Antibody-Positive Healthy Individuals Who Experienced Asymptomatic Versus Clinical Infection. Frontiers in Immunology **12**, 746203 (2021) https://doi.org/10.3389/fimmu.2021.746203 . Accessed 2025-10-08

[43] Institute of Medicine (US) Committee on Emerging Microbial Threats to Health in the 21st Century, Smolinski, M.S., Hamburg, M.A., Lederberg, J.: Pathogen Discovery, Detection, and Diagnostics. In: Microbial Threats to Health: Emergence, Detection, And Response. National Academies Press (US), ??? (2003). https://www.ncbi.nlm.nih.gov/books/NBK221492/ Accessed 2025-10-08

[44] Long, E.M., Millen, B., Kubes, P., Robbins, S.M.: Lipoteichoic Acid Induces Unique Inflammatory Responses when Compared to Other Toll-Like Receptor 2 Ligands. PLoS ONE **4**(5), 5601 (2009) https://doi.org/10.1371/journal.pone.0005601 . Accessed 2025-01-22

[45] Watanabe, S., Zenke, K., Muroi, M.: Lipoteichoic Acid Inhibits

Lipopolysaccharide-Induced TLR4 Signaling by Forming an Inactive TLR4/MD-2 Complex Dimer. The Journal of Immunology **210**(9), 1386–1395 (2023) https://doi.org/10.4049/jimmunol.2200872 . Accessed 2025-10-08

[46] Adams, G.: A beginner's guide to RT-PCR, qPCR and RT-qPCR | The Biochemist | Portland Press (2020). https://doi.org/10.1042/BIO20200034 . https://portlandpress.com/biochemist/article/42/3/48/225280/A-beginner-s-guide-to-RT-PCR-qPCR-and-RT-qPCR Accessed 2025-08-14

[47] Paul, F., Pellegrini, S., Uzé, G.: IFNA2: The prototypic human alpha interferon. Gene **567**(2), 132–137 (2015) https://doi.org/10.1016/j.gene.2015.04.087 . Accessed 2025-08-14

[48] CUSABIO: Get an Overview of Inflammatory Cytokines & Inflammatory Response- CUSABIO. https://www.cusabio.com/cytokines/Inflammatory-Response.html Accessed 2025-08-14

[49] PubChem: LTB - lymphotoxin beta (human). https://pubchem.ncbi.nlm.nih.gov/gene/LTB/human Accessed 2025-08-14

[50] PubChem: CNTF - ciliary neurotrophic factor (human). https://pubchem.ncbi.nlm.nih.gov/gene/CNTF/human Accessed 2025-08-14

[51] GENCODE: Human Gene GPI (ENST00000356487.11) from GENCODE V48. https://genome.ucsc.edu/cgi-bin/hgGene?db=hg38&hgg_chrom=chr19&hgg_gene=ENST00000356487.11&hgg_start=34353329&hgg_end=34402412&hgg_type=knownGene Accessed 2025-08-14

[52] CUSABIO: Get an Overview of Th1 Cytokines- CUSABIO. https://www.cusabio.com/cytokines/Th1-Cytokines.html Accessed 2025-08-14

[53] Mitamura, Y., Nunomura, S., Furue, M., Izuhara, K.: IL-24: A new player in the pathogenesis of pro-inflammatory and allergic skin diseases. Allergology International **69**(3), 405–411 (2020) https://doi.org/10.1016/j.alit.2019.12.003 . Accessed 2025-08-14

[54] CUSABIO: Get an Overview of Th2 Cytokines- CUSABIO. https://www.cusabio.com/cytokines/Th2-Cytokines.html?srsltid=AfmBOopXd17x4qNqAnFF8xkVyjXXXZXOYoztSOIzbH0gmXCF3NX8FTrF Accessed 2025-08-14

[55] CUSABIO: Get an Overview of Th17 Cytokines- CUSABIO. https://www.cusabio.com/cytokines/Th17-Cytokines.html Accessed 2025-08-14

[56] Carl, J.W., Bai, X.-F.: IL27: Its Roles in the Induction and Inhibition of Inflammation. International Journal of Clinical and Experimental Pathology **1**(2), 117–123 (2008). Accessed 2025-08-14

[57] The UniProt Consortium: UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Research **53**(D1), 609–617 (2025) https://doi.org/10.1093/nar/gkae1010 . Accessed 2025-08-14