New Phytologist

Research

Machine learning-based identification of general transcriptional predictors for plant disease

Jayson Sia[1]*, Wei Zhang[2,3]*, Mingxi Cheng[1], Paul Bogdan[1,4] and David E. Cook[2]

[1]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA; [2]Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA; [3]Institute for Integrative Genome Biology, University of California, Riverside, CA 92521, USA; [4]Center for Complex Particle Systems (COMPASS), University of Southern California, Los Angeles, USA

Authors for correspondence:
Paul Bogdan
Email: pbogdan@usc.edu

David E. Cook
Email: decook@ksu.edu

Received: *15 May 2024*
Accepted: *10 October 2024*

New Phytologist (2025) **245:** 785–806
doi: 10.1111/nph.20264

Key words: *Arabidopsis thaliana*, feature selection, general stress response, machine learning, network science, plant–pathogen interaction, predictive biology.

Summary

- This study investigated the generalizability of *Arabidopsis thaliana* immune responses across diverse pathogens, including *Botrytis cinerea*, *Sclerotinia sclerotiorum*, and *Pseudomonas syringae*, using a data-driven, machine learning approach.
- Machine learning models were trained to predict disease development from early transcriptional responses. Feature selection techniques based on network science and topology were used to train models employing only a fraction of the transcriptome. Machine learning models trained on one pathosystem where then validated by predicting disease development in new pathosystems.
- The identified feature selection gene sets were enriched for pathways related to biotic, abiotic, and stress responses, though the specific genes involved differed between feature sets. This suggests common immune responses to diverse pathogens that operate via different gene sets.
- The study demonstrates that machine learning can uncover both established and novel components of the plant's immune response, offering insights into disease resistance mechanisms. These predictive models highlight the potential to advance our understanding of multigenic outcomes in plant immunity and can be further refined for applications in disease prediction.

Introduction

Plants rely on an innate immune system to counter infection. A hallmark of the plant innate immune system is the detection of invasion through diverse receptor catalogues that survey both the apoplastic and cytoplasmic compartments (Cook *et al.*, 2015; Bentham *et al.*, 2020). At the cell surface, membrane anchored pattern recognition receptors (PRRs), in the form of receptor-like kinases and receptor-like proteins, have been well described for their ability to detect a wide variety of ligands to initiate immune responses (Zipfel *et al.*, 2004; Liebrand *et al.*, 2013; Liu *et al.*, 2022). Perception of invasion at the cell surface can trigger a number of cellular responses, including ion flux, reactive oxygen species production, posttranslational modifications, and transcriptional reprogramming collectively contributing to plant defense (Couto & Zipfel, 2016; Wan *et al.*, 2019). Intracellular receptors, referred to as nucleotide-binding and leucine-rich repeat receptor (NLR) domain-containing proteins, detect non-self and modified-self ligands to initiate a plant defense response (Jones & Dangl, 2006; Jubic *et al.*, 2019). Structural data shows that different types of activated NLRs form multimer protein complexes that likely aid immunity through calcium signaling and small molecule generation (Wang *et al.*, 2019; Ma *et al.*, 2020; Martin *et al.*, 2020). Following defense activation, phytohormones play a substantial role in mediating plant immunity (Pieterse *et al.*, 2012; Aerts *et al.*, 2021). While many phytohormones participate in plant immunity, the salicylic acid (SA) and jasmonic acid (JA) pathways are central to diverse biotic interactions (Kazan & Lyons, 2014). The SA and JA response pathways are generally seen to be antagonistic and differentially effective against microbes that display different host interaction strategies, from those that interact with living cells, termed biotrophs, to those that promote host cell-death, termed necrotrophs (Glazebrook, 2005). Cross talk between phytohormone pathways and their network structure likely serve an important role in plants ability to respond to diverse and changing biotic and abiotic challenges. Collectively, the extracellular and intracellular immune receptors, downstream transcriptional responses, and the function of phytohormone pathways form a coordinated immune response (Tsuda *et al.*, 2009; Tintor *et al.*, 2013; Ngou *et al.*, 2021; Yuan *et al.*, 2021).

Plant immune receptors detect specific immunogenic ligands of diverse origins, but it is less clear how plants integrate diverse signals to achieve immune responses (Tsuda & Somssich, 2015). An open question concerns the extent to which the immune responses are fine-tuned to specific ligands that enact a tailored

*These authors contributed equally to this work.
[Correction added on 19 December, after first online publication: the affiliations for the author Paul Bogdan have been updated].

defense response to the invader (Tsuda *et al.*, 2009; Bjornson *et al.*, 2021). Analysis of plant transcriptional responses to diverse pathogens and ligands suggests there is both overlap and divergence to differing immunogenic signals. For instance, *Arabidopsis thaliana* seedlings responding to either a plant cell wall-derived oligogalacturonides or bacterial-derived flagellin peptide, elicit similar early transcriptional responses that diverged with time (Denoux *et al.*, 2008). Also in *A. thaliana*, early transcriptional responses were largely overlapping in response to seven diverse immunogenic ligands, but also, the FLS2–flg22 interaction had a substantial number of unique transcriptional responses (Bjornson *et al.*, 2021). Early transcriptional response of *A. thaliana* to genetically diverse *Botrytis cinerea* isolates showed that wild-type (WT) Col-0 ecotype displayed significant transcriptional variation for a number of transcripts involved in hormone related signaling and PRR responses (Zhang *et al.*, 2017). Despite the diverse transcriptional responses to the genetically diverse pathogen isolates, final disease development outcomes are rather limited (Zhang *et al.*, 2017), which can be conceptualized as many roads leading to the same place. Such a buffered immune response to diverse inputs may be mediated by interconnected immune subnetworks, providing robustness to the diversity of microbial interactors (Tsuda *et al.*, 2009; Kim *et al.*, 2014; Hillmer *et al.*, 2017). It is clear that activation of different PRRs do not provide identical transcriptional outputs (Li *et al.*, 2016), but immune signaling can channel responses from diverse inputs into largely overlapping outputs (Tsuda *et al.*, 2009; Zhang *et al.*, 2017; Bjornson *et al.*, 2021). The systems view of plant immunity suggests a highly interconnected network involving receptor-mediated detection, intracellular signaling, hormone cross talk, and cellular output leading to an immune response (Tsuda & Somssich, 2015; Katagiri, 2018; Mishra *et al.*, 2019; Aerts *et al.*, 2021; Delplace *et al.*, 2022).

Machine learning (ML) represents a broad class of algorithms designed to optimize a function that relates input to output data. Highly predictive ML algorithms, such as random forests (Chen & Ishwaran, 2012) and support vector machines (Ben-Hur *et al.*, 2008) are well developed, and require less data abstraction through hidden states, making their output and learning human discernable. This contrasts deep learning techniques of the last decade that rely on highly abstracted feature selection techniques, able to predict the most complex interactions, but obfuscate human interpretation (Choo & Liu, 2018). Depending on the application, ML for biology may tend toward traditional ML approaches that have fewer variables and greater interpretability to aid mechanistic understanding or hypothesis generation (Xu & Jackson, 2019). There are also significant challenges in the application of ML to biological datasets because of their relatively limited size compared to other data domains (Greener *et al.*, 2021). In general, if there are not orders of magnitude more observations than predictors, complex deep learning models can underperform compared to more traditional ML algorithms (Wang *et al.*, 2021).

We sought to further explore the connection between transcriptional immune responses and disease output in order to identify general patterns that predict final disease outcome using publicly available data. The initial research focused on the *A. thaliana–B. cinerea* pathosystem because the system is characterized by multigenic, small effect interactions, not dominated by effector–NLR interactions (Finkers *et al.*, 2007; Soltis *et al.*, 2019). This is an important consideration in looking for convergent immune responses across diverse pathosystems. Additionally, *B. cinerea* is a broad host-range necrotrophic pathogen that employs host-agnostic cell wall degrading enzymes, toxins, cell-death-inducing proteins and small RNA to manipulate host defense, while plants rely on PRRs, phytochemicals, and JA defense pathways to limit disease (AbuQamar *et al.*, 2017; Bi *et al.*, 2023). Additionally, previous research has resulted in datasets large enough to employ ML (Zhang *et al.*, 2017). Using a data-driven ML analysis pipeline, we tested if an ensemble of post-infection transcriptional responses could inform final disease outcome across a set of diverse pathosystems. For this case, ML is a powerful approach that does not require preselecting candidate genes or pathways, and it can capture complex, nonlinear patterns across the whole transcriptome that collectively contribute to disease development, a complex multigenic trait. This approach can capture far more diverse, and likely biologically relevant, transcriptional patterns compared to traditional co-expression or differential expression analysis. Additionally, we employed a range of feature selection techniques, including those built from network theory and network geometry, to identify specific sets of genes providing accurate disease prediction across pathosystems. Thus, our approach is a novel application of ML and network science to plant-immunology, resulting in the discovery of genes not previously associated with plant immunity or pathogen response that may capture a general plant immune response.

Materials and Methods

Machine learning tasks

To understand the relationships between transcriptional response and disease outcome, we used the state-of-the-art supervised ML algorithms to train models that can accurately predict plant disease severity from gene expression data. The classification task is to predict the disease severity (classes) based on the gene expression profile of dual or sole species. We used plant disease phenotypic data, fungal colonized lesion area in plant-fungal pathosystems or bacteria growth in plant-bacteria pathosystems, as labeled multiclass data. The transcriptomic data derived from dual species or solely from plant host or pathogen serve as input data. To identify the crucial gene set that control disease development, we performed different feature selection methods on the training plant gene expression datasets and applied the selected genes on different trained models to evaluate performance improvement. The multi-step workflow starts from plant disease phenotypic data and RNA-Seq read counts data, followed by ML and feature selection methods, and outputs disease predictions (Fig. 1a).

Disease phenotypic and transcriptomic data acquisition

We used the previously published plant disease phenome and transcriptome data involving 1164 *A. thaliana* (L.) and *B. cinerea*

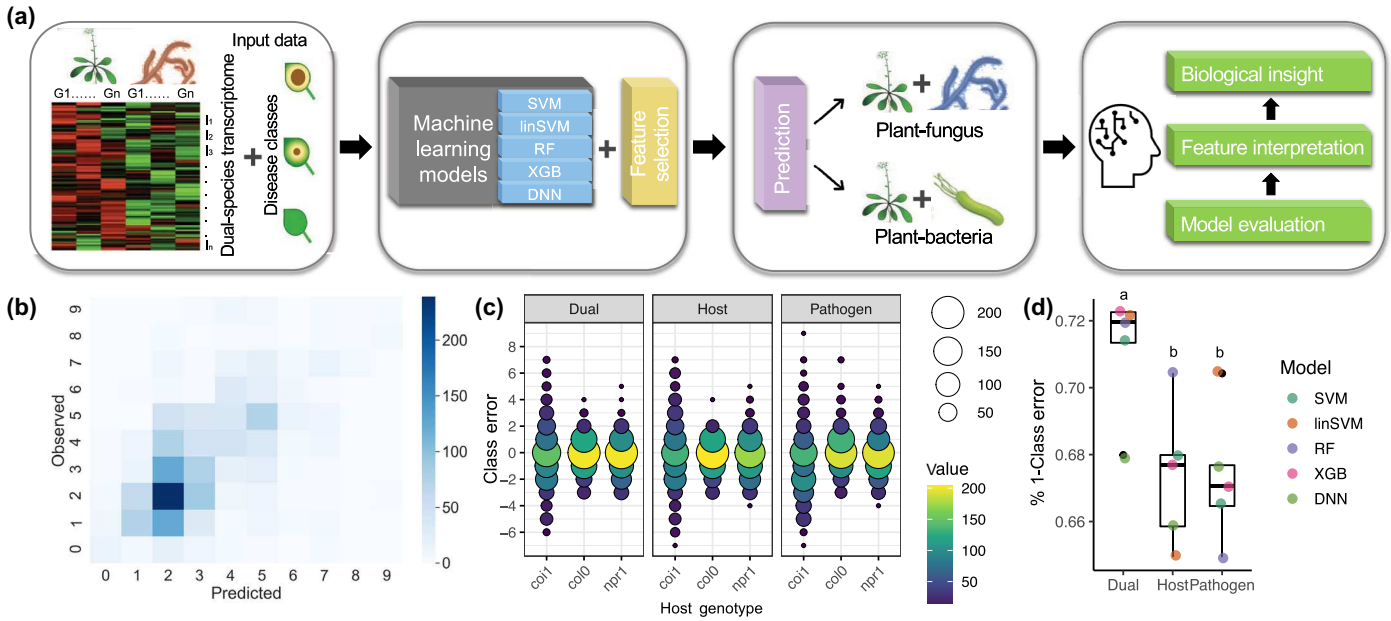

Fig. 1 Machine learning (ML) can accurately predict later plant disease outcomes from earlier dual-species transcriptome. (a) Schematic overview of interpretable ML strategies on plant immune elements in plant–pathogen pathosystems. (b) Heatmap of the observed and predicted disease classes based on the dual-species whole transcriptome data. (c) Class errors calculated by predicted and measured plant disease classes on Arabidopsis-*Botrytis cinerea* transcriptome data (Ngou *et al*., 2021). (d) The host genotypes coi1 and npr1 refer to the gene deletion mutant lines and col0 refers to the wild-type line [20]. (d) The 1-class error accuracy of predicted plant disease prediction is shown as box plots. The box plots depict the interquartile range (IQR) of data, representing the 25th to 75th percentile of data, and the median is depicted as a dark line inside of each box. The box whiskers depict the range of data defined by 1.5 times the IQR, and any outlier data points outside this range are shown as black circles. Individual results for all five ML models are shown as colored circles indicated in the key. The five ML models are support vector machine (SVM, teal dot), linear SVM (linSVM, orange dot), Random Forest (RF, purple dot), XG Boost (XGB, pink dot), and deep neural network (DNN, green dot). The data are grouped by results using both the host and pathogen transcriptome (dual), or solely the plant transcriptome (host), or the fungal transcriptome (pathogen). The letters above box blots indicate similarity groupings based on statistical significance (one-way ANOVA, $P < 0.05$).

(Pers.) interaction pathosystems (Zhang *et al*., 2017, 2019). In brief, the *Arabidopsis–Botrytis* dataset include four replicates of 97 *Botrytis* isolates infecting three *Arabidopsis* (Col-0) genotypes WT with complete immunity, and two immune compromised lines, *npr1* and *coi1*, compromised in SA and JA mediated defense response, respectively. The plant disease phenotypic data, fungal colonized leaf area, were measured at 72 h postinoculation. The transcriptomic data were collected at 16 h postinoculation and raw RNA-Seq data are available from the National Center for Biotechnology Information (NCBI) under BioProject no. PRJNA473829 and accession no. SRP149815.

The validation set of plant–fungal pathogen interaction was derived from *Arabidopsis* and *Sclerotinia sclerotiorum* (Lib.) interactome (Badet *et al*., 2017). The disease phenotypic data were collected at 24 h postinoculation, and tissues at the edge of developed necrotic lesions were collected for RNA extraction. Data were accessed through NCBI Gene Expression Omnibus (GEO) accession GSE106811. The validation set of plant–bacteria pathogen interaction was derived from *Arabidopsis* and *Pseudomonas syringae* (Van Hall) interactome (Nobori *et al*., 2018). The disease phenotypic data were measured as the bacteria growth at 48 h postinoculation. The transcriptomic data were collected at 6 h postinoculation and are available through NCBI GEO GSE103442. We note that there was a problem with miss-labeled metadata samples for the pretreatment samples. Only samples pretreated with flg22 were retained for initial analysis, the rest of the samples receiving pretreatment with chitin or SA were dropped due to ambiguous labeling in the metadata file on SRA.

Data processing

To perform the multi classification-learning task, we divided the *Botrytis* infecting *Arabidopsis* training disease phenotypic data into 10 classes based on the disease severity to generate multiclass labels. We consider the largest lesion from the training dataset corresponding to the highest disease class (Class 9). A lesion size of 0 is mapped to the lowest disease class (Class 0). Then, we divided the lesion data into 10 equal lesion bin sizes and labeled each training sample accordingly. To obtain comparable disease class label for the plant-fungal pathogen test data, the labels for healthy samples and infected samples were transformed as disease class 1 and 6 based on the distribution patterns of lesion area measured from *Arabidopsis–Sclerotinia* pathosystems and *Arabidopsis–Botrytis* pathosystems (Supporting Information Fig. S1). Similarly, disease class labels were assigned based on the comparison of distributions of lesion area from *Arabidopsis–Botrytis* pathosystems and bacteria growth measured from *Arabidopsis–Pseudomonas* pathosystems (Figs S1, S2).

To avoid the bias caused by the multiple mapped reads to both plant host and pathogen reference genomes, we used dual-species

reference genome to map the RNA-Seq reads derived from the pathosystems (Aprianto *et al.*, 2016; O'Keeffe & Jones, 2019). Briefly, we first generated a dual-species reference genome by concatenating the *Arabidopsis* TAIR10 reference genome (Berardini *et al.*, 2015) with either of the three pathogen reference genomes, *B. cinerea* (strain B05.10, build ASM83294v.1) (Van Kan *et al.*, 2017), *S. sclerotiorum* (strain 1980 UF-70, build ASM185786v.1), or *P. syringae* (strain *P. syringae* pv *tomato* DC3000, build GCF_000007805.1) (Winsor *et al.*, 2016), respectively. Raw RNA-Seq reads were mapped to the corresponding combined dual-species reference genome by STAR (Dobin *et al.*, 2013). Raw gene counts were normalized as transcripts per million (TPM) (Wagner *et al.*, 2012). For the *Arabidopsis–Botrytis* dataset, individual sequencing libraries were removed if they had < 30% uniquely mapped reads. Annotated genes in from *Arabidopsis* were removed from the count table if the gene had read counts < 200 from across all sample libraries. After quality control, there were 1102 samples and 20 340 expressed *Arabidopsis* genes, and 8761 expressed *Botrytis* genes included in the *Arabidopsis–Botrytis* training dataset. The same criteria were used to assess the two validation datasets (*A. thaliana–S. sclerotiorum* and *A. thaliana–P. syringae*), and both sets passed quality control. To be consistent, only count data for the 20 340 *Arabidopsis* genes used in training were retrieved from the validation data.

The individual ML models were trained on the gene expression values of the *Arabidopsis–Botrytis* dataset. Data were split into 70% training and 30% test sets. Since the range of gene expression values vary significantly per gene, feature scaling is needed to ensure that the contribution of each feature is not biased toward larger numerical values. In addition, scaling the features ensure faster gradient descent convergence for some of the ML models. Here, we scaled the features by removing the mean and scaling to a unit variance (Han *et al.*, 2012). Additionally, the *Arabidopsis–Botrytis* dataset is highly imbalanced with respect to the lesion size/disease class (Fig. S1). Lower disease classes (class 0–3) are more represented in the data than higher disease classes (class 6–9). To mitigate this data imbalance, we utilized SMOTE (synthetic minority oversampling technique) to oversample the minority data classes to have an even representation of the classes during training (Chawla *et al.*, 2002). This technique increases the minority class examples by synthesizing new data stochastically from the existing training class sample space. Note that it is important that the training/validation/test splits should be done before any preprocessing steps (e.g. data standardization) to avoid 'data peeking', which can over-inflate test prediction performance. For example, data standardization of the validation data should be done based on the aggregate statistics of the training data only. If the average or data statistics is based on the entire dataset (including validation and test sets), in effect we have considered information from the test data.

Supervised model training

According to the multiclassification task and the data characteristics in this study, we selected supervised ML strategy to build the learning models that can learn from the transcriptome input derived from dual species or sole species and can accurately predict the correct phenotypic disease class outcome in a plant host and pathogen interaction pathosystems. A total of five popular ML algorithms for supervised learning were tested, including two support vector machine (SVM)-based algorithms, two decision tree-based algorithms, and a deep neural network algorithm. All the models are trained on the *Arabidopsis–Botrytis* pathosystems dataset. Hyper-parameters tuning was performed for each model using scikit-learn's GridSearchCV. The following parameters are tuned for the following models: learning rate for DNN, learning rate and max_depth for XGBoost, number of trees and max_depth for RF, and gamma for the SVM. Support vector machine is a supervised learning algorithm that can efficiently perform both linear and nonlinear classifications on data for robust prediction. It can construct a hyperplane or set of hyperplanes in a high-dimensional feature space to achieve the largest distance to the nearest training data point of any class. Although SVMs are naturally used for binary class tasks, the algorithms can be developed to apply on multiclassification tasks by reducing the multiclass task to several binary problems (Noble, 2006). Linear SVM is a linear classifier that is used for linearly separable data into classes by using several straight lines. We used support vector machine (Radial Basis Function (RBF) Kernel), SVC(), and linear SVM, SVC(kernel = 'linear'), from SCIKIT-LEARN package (Noble, 2006; Pedregosa *et al.*, 2011). The parameters for SVM and linear SVM are based on the default setting.

Decision tree-based algorithms generally have high performance on small-to-medium structured data, and are a popular and robust approach for various ML tasks because they are invariant to feature scaling and transformation, and independence of irrelevant features ensure constructing inspectable models (Breiman *et al.*, 2017). Decision trees make decisions using a graph to represent all possible solutions queried by certain conditions. Random forests, also called random decision forests, are decision tree-based ensemble algorithms that perform classification tasks using a bagging strategy to build a multitude of decision trees (forest) where only a subset of features is randomly selected to build a forest or decisions are collected from some trees (Ho, 1995). Therefore, the accuracy of such ensembled decisions by random forests is generally higher than models built on single random decision tree. However, random forests show biases in data including categorical variables with high-variation levels and/or correlated variables, which are common cases in transcriptomic data (Strobl *et al.*, 2007). We used random forests (RandomForestClassifier) for learning modeling training from SKLEARN PYTHON package with parameters n_estimators = 100, max_depth = 30, random_state = 0. The choice of 100 trees for the RF was based on our results running tests using the *Arabidopsis* only dataset. The 1-class error accuracy for $n = 100$ was 0.70; $n = 1000$ was 0.72; and $n = 10\,000$ was 0.71. Therefore, increasing the number of trees two orders of magnitude did not substantially improve model performance, but it did significantly increase computational run time. The $n = 100$ estimators was used to balance performance and time. The random forest was also used for ranking the feature importance for feature extraction.

Extreme gradient boosting (XGBoost) is another decision tree-based ensemble-learning algorithm that uses an optimized gradient boosting algorithm for classification, regression, and ranking tasks. Gradient boosting builds sequential models by using gradient descent algorithm to minimize the errors from previous models while increase the influences of high-performance models (Chen & Guestrin, 2016). Such strategy can solve the drawbacks of random forest-based models learning from data with high-varied level of categorical variables and/or correlated variables. Based on gradient boosting framework, XGBoost is further optimized to avoid overfitting and/or bias through several ways, including parallel processing, tree-pruning, handing missing values and regularization (Ma *et al.*, 2020). Furthermore, it dramatically improves the computing power for boosted tree algorithms. We used XGBoost (*XGBClassifier*) for modeling training from XGBoost python package with default parameters. The ranked features generated from the XGBoost model are used for feature selection.

We also included an artificial neural network-based model among our supervised learning algorithms for comparison. We used KERAS package to build the modular neural network available from TENSORFLOW package (Abadi *et al.*, 2016) to build a deep neural network (DNN) model with a substantial credit assignment path of depth 2 (with batch normalization, 512 dense hidden layer with ReLU activation function, dropout = 0.5, dense 10-class output layer with softmax activation function). The cross-entropy loss and the Adam optimizer were used with a learning rate of 0.001. The trained neural network model could directly transform the raw whole transcriptome input into a more abstract and composite intermediate feature representation and extract such higher-level intermediate features to predict disease severity.

Feature selection

To better understand genes contributing to plant disease development, we performed feature selection on expressed genes from the whole *A. thaliana* transcriptome. Feature selection methods were assessed by using the feature selection gene sets in each trained model to observe performance improvement or decay. The full set of feature-selected genes by the different methods can be found in Table S1. The plant whole transcriptome data served as a baseline control using the random forest model with a prediction 1-class error accuracy of 70.4%. We randomly selected 100 plant transcripts as negative control with the highest 1-class error accuracy of 65.7% by SVM model. We explored domain knowledge-based methods to obtain the plant genes associated with pathogen defense. The list of 130 defense genes is based on previously published literature characterized as having defense function in the *Arabidopsis–Botrytis* pathosystem. The term 'response to biotic stimulus' from Gene Ontology was used to identified 1182 *Arabidopsis* genes (The Arabidopsis Information Resource (TAIR), 2023), of which, 912 had detectable expression in the *Arabidopsis–Botrytis* transcriptome. Previously published *Arabidopsis* genes that were either positively or negatively correlated with plant disease lesion areas were selected as feature selection sets PosCorr_786 and NegCorr_762 (Zhang *et al.*, 2017, 2019). The RF_100 and XGB_500 feature selection sets were generated from the plant gene set based on the ranked importance during model training.

To obtain the gene set based on the gene co-expression network topology, we first constructed the plant gene co-expression network using the *Arabidopsis* transcriptome data. We then calculated the network structure parameters, including node degree (Degree_500) measuring the number of connections between a node (gene) and its neighbors, betweenness centrality (Btwns_500) measuring the number of shortest paths between all pairs of nodes in the network which pass through the focal node. Apart from the classical network metrics, we used for node ranking and feature selection the fractal dimension centrality (FDC_3000) and node fractal dimension (NFD) centrality (NFD_3000), which captures the topological features and the degree of complexity and heterogeneity in generating rules of complex networks (Xiao *et al.*, 2021). Additionally, we also used the bipartite graph-based feature selection method to extract the *Arabidopsis* genes based on their co-expression relationships during *B. cinerea* infection. First, the bipartite network is created from the plant–pathogen co-expression dataset. Then, a unipartite network projection on the *Arabidopsis* genes is extracted from the bipartite network based on the frequency of shared connection to the *Botrytis* gene set (Pavlopoulos *et al.*, 2018). We assessed feature selection membership using an UpSet plot from package UPSET (Conway *et al.*, 2017). Chi-square test of independence were used to evaluate dependence between the feature selection sets and two gene sets previously implicated in GWAS host response, or general response to immunogenic peptides (Bjornson *et al.*, 2021). We further characterized the potential biological significance of seven feature selection sets, namely NFD, Bipartite, Betweenness, Degree, XGB, RF, and FDC. Each feature selection set was analyzed for GO enrichment of Biological Process pathways, using the SHINYGO web application (v.0.80) (Ge *et al.*, 2020), with *A. thaliana* as species, FDR cutoff 0.05, 20 pathways shown, and min and max pathway size of 2 and 5000, respectively. Pathways were selected based on FDR and sorted by Fold Enrichment for Chart and Tree figures. Pathways in Trees were colored based on higher GO groupings as determined by QUICKGO web application (v.2024-04-29) (Binns *et al.*, 2009). Two feature selection sets, XGB and Bipartite, did not have significant GO Biological Process enrichment. To analyze across all seven sets, we analyzed the proportional gene membership for high-level GP terms as identified by SHINYGO. For each gene set, the proportional rank of the high-level GO categories was determined as the proportion of the number of genes identified in the category over the number of total genes in the category as determined from AMIGO2 web application (v.2.5.17) filtering for *A. thaliana* (Binns *et al.*, 2009; Carbon *et al.*, 2009; The Gene Ontology Consortium *et al.*, 2023). The rank of each of the 20 high-level GO categories was determined for each feature selection, and the Rank distribution within each high-level GO categories across feature selection sets was determined and plotted using pandas and matplotlib of PYTHON3 (Hunter, 2007; Team T Pandas Development, 2024).

Evaluation

We measure the model performance based on the classification errors. Since $a+1/-1$ classification error is within the tolerable error range, we used the 1-class error True Positive (TP_1) to define the accuracy of the predicted disease classes. We used a confusion matrix approach to evaluate the performance of the multiclassification task of ML models and feature selection methods. Each prediction falls into one of the four cases: true positive (TP_1), false positive (FP), true negative (TN), and false negative (FN). To describe the distribution of predictions across each of these categories, we calculated a variety of performance matrices, including Accuracy, Precision, Recall, F1_Score, and mean squared error (MSE).

$$\text{Classification Error(CE)} = (\text{Predicted disease class} - \text{Observed disease class})$$

$$\text{Adjusted Classification Error(ACE)} = \begin{cases} 0 & \text{if CE} \in \{-1, 01\} \\ \text{CE} & \text{otherwise} \end{cases}$$

$$TP_1 \text{ (1-class error TP)}$$

$$TP_1 = \begin{cases} 1 & \text{if CE} \in \{-1, 0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Accuracy} = \frac{TP_1 + TN}{TP_1 + FP + TN + FN}$$

$$\text{Precision} = \frac{TP_1}{TP_1 + FP}$$

$$\text{Recall} = \frac{TP_1}{TP_1 + TN}$$

$$\text{F1 Score} = \frac{(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} = \frac{TP_1}{TP_1 + 0.5 \times (FN + FP)}$$

$$\text{Mean squared error} = \frac{1}{n} \times \sum\nolimits_n \text{ACE}^2$$

Results

Machine learning can predict disease outcomes from transcriptomics of host–pathogen interactions

We sought to test the hypothesis that transcriptional patterns during the early stages of plant infection are predictive for final disease outcomes. To gain understanding from the modeling, and to account for data size, we mainly focused on ML algorithms with fewer variables and interpretable components. To identify gene sets that are robust predictors of plant disease, we utilized feature selection and cross validation on multiple pathosystems (Fig. 1a). A total of five ML algorithms were trained on *B. cinerea* infecting *A. thaliana* (Zhang *et al.*, 2017, 2019; Fig. 1a). The data included 96 diverse *B. cinerea* isolates infecting *A. thaliana* ecotype Col-0, along with a salicylic acid (SA) receptor mutant, *npr1* that is defective in SA-induced defense (Mou *et al.*, 2003; Spoel *et al.*, 2003; Wu *et al.*, 2012), and a bioactive jasmonic acid (JA) receptor mutant, *coi1* that normally participates in E3 ubiquitin ligase mediated activation of JA defense responses (Xu *et al.*, 2002; Thines *et al.*, 2007). Transcriptional responses during host infection were measured by RNA-Seq at 16 h postinoculation and lesion size for each interaction was measured at 72 h postinoculation (Fig. 1a; Zhang *et al.*, 2017). Transcriptional responses (i.e. processed RNA-Seq per gene) were used as predictors and final lesion size as the response to SVM with linear and RBF kernels (Noble, 2006), two decision tree-based algorithms, random forest (RF) (Ho, 1995) and extreme gradient boost (XGB) (Chen & Guestrin, 2016), and a DNN (LeCun *et al.*, 2015) (see the Materials and Methods section for details). Our goal was to develop models that can predict disease outcomes across pathosystems, which for our training data was lesion size, but in many other systems disease results in wilting, curling, stunting, discoloration or other phenotypes unrelated to a lesion development. To address this, we reasoned that converting lesion size values to 10 disease classes would provide a framework to convert any disease outcome dataset into disease classes to apply the model across pathosystems. To assess how converting disease lesion size to disease classes impacted the results, we performed a comparison. Transcriptional data from both host and pathogen were used to train two regression models on disease lesion size as a function of transcription using the XGB algorithm and a DNN architecture. Plotting the residual error for the DNN and XGB showed acceptable prediction performance, with a mean square error (MSE) of 111.60 and root mean square error (RMSE) of 10.77 for the DNN, and MSE of 88.83 and RMSE of 9.42 for XBG (Fig. S3A,B). Given that we wanted to compare the regression results to classification, we considered the regression predictions that were (+/−) the discretization bin size to be correct, resulting in an accuracy of 63% for the DNN and 77% for XGB (Fig. S3A,B). To assess the fit of the linear regression, the measured vs predicted lesion sizes were plotted and R^2 calculated, which indicated that the XBG model provided a better fit, explaining 48% of the variance in the data compared to 37% explained by the DNN model (Fig. S3C).

To convert the continuous lesion size into classes, we used a data-driven approach to create balanced disease (i.e. lesion) classes and to normalize the distribution of observations (Fig. S1; see the Materials and Methods section for details). Disease class predictions using both the host and pathogen transcriptomes showed a clear association between the observed and predicted values for the holdout test data (Figs 1b, S4). To quantify the results, we calculated the difference between the observed disease class and the predicted disease class, which we termed class error. For example, a class error of zero means the observed and predicted classes were the same, while a class error of 3 means the observed disease class was three classes higher than predicted (i.e. the prediction underestimated disease outcome). Host genotype

significantly influenced disease class prediction, and infected *npr1* plants produced more similar predictions compared to infected WT *A. thaliana* than did infected *coi1* plants (Fig. 1c; Table S2). This result likely reflects the substantial shift in disease outcomes seen for *coi1* infection (Fig. S5), and is consistent with the previously reported importance of jasmonic acid defense pathway against *B. cinerea* infection (Zhang *et al.*, 2017). The difference in prediction accuracy between host genotypes was consistent when both the plant and fungal transcriptomes were modeled together (i.e. dual), or if only one organism's transcriptome data were used (Fig. 1c; Table S3). This suggests ML algorithms are sensitive to genetic perturbation, and the impact of the *coi1* mutation on transcriptional response and disease development were more difficult to predict. To quantify model performance, we considered predictions within one disease class of the observed to be correct and refer to this throughout the research as 1-class error accuracy. Interestingly, across all host genotypes and ML approaches, the 1-class error accuracy was significantly higher when models were trained using transcriptional data from both the host and the pathogen compared to either alone (72% compared to $< 68\%$, respectively) (Fig. 1d). Comparing these results to those obtained from regression (Fig. S3), the proxy estimates for accuracy were similarly *c.* 70% averaged across models for either regression or classification.

We further tested if ML models are useful for the task of predicting disease from transcriptional data. Given that the dataset had a high occurrence of disease class two observations (Figs S1, S4), we tested if this class imbalance could account for model performance, in which the models labeled samples with the most common class. To test this, we calculated the 1-class error prediction metrics for a naïve model in which all test data were assigned a predicted disease class of two (Table S4). This naïve model had an accuracy of 58%, while the five ML models had an average accuracy of 68%, ranging from a low of 66% to a high of 70% (Table S4). To further demonstrate that the developed models were performing well and identifying useful patterns in the data, disease classes for the test data were randomly drawn from the distribution of disease classes, assigned to a test sample, and then performance metrics were computed. This random assignment and assessment was repeated for 100 iterations and allowed us to build a distribution of possible outcomes given the data. The performance metrics of this iterative shuffling experiment showed an average accuracy of 43% ($\pm 2\%$) (Table S4; Fig. S6), which is again substantially lower than that observed for the five developed ML models. In one final test on the utility of ML for predicting disease outcomes, we tested perturbed data in which the disease class was intentionally changed to an incorrect label. For this, we altered an increasing percentage of the training data disease classes, from 25% up to 75% of samples, to contain an incorrect label by more than one disease class (Fig. S7). The test data for prediction remained unchanged, only the training data for model building where altered. The resulting confusion matrices showed a decrease in performance as the percentage of mislabeled training data increased, especially above 50% mislabeled data (Fig. S7). This trend was also apparent for one-class error assessment, where having 50% or 75% mislabeled training data substantially degraded model performance. It was interesting that while model performance decreased at 25% mislabeled training data, the impact was modest and overall prediction performance was similar to the model trained on correctly labeled data (Fig. S7). Overall, these results show that ML is a useful class of algorithms for modeling disease outcomes from transcriptional data. The models are performing better than chance and are not reliant on class imbalance, and the models are also sensitive to data manipulation, all of which supports the view that the models are able to identify salient data patterns for predicting disease outcomes.

Looking more deeply into the classification results using the real unaltered data, all measures of model performance were either statistically similar or better using the combined host–pathogen transcriptome data compared to using only one organism's transcriptome (Figs 1d, S8). For analysis using host–pathogen transcriptomes, the XGBoost model performed the highest with a prediction 1-class error accuracy of 72.3% (Fig. 1d). The RF model provided the highest 1-class error accuracy (70.4%) when only considering the host transcriptome, while the results from the linear SVM provided the highest 1-class error accuracy (70.5%) for pathogen alone analysis (Fig. 1d). Collectively, these results show that modeling the response of both species together provides the most accurate prediction of plant disease outcome. This reflects the importance of both actors in determining disease development and highlights the dynamic and complex nature of dual-species interactions. For subsequent model testing across different pathosystems, we proceeded with only using the host plant genes as predictors as these transcripts are common between diverse datasets, while pathogen gene sets change. Also, we continued with converting disease outputs data into disease classes, as the results suggest that we obtained similar prediction results, and this framework allows us to use the model in systems with diverse disease outcomes.

Feature selection to identify general predictors of plant disease

We seek to identify a subset of transcripts that were both predictive of disease outcome, and that might also reflect meaningful biological mechanisms contributing to plant disease. Feature selection is a common practice in ML to help reduce the large predictors (p), small observation (n) problem associated with high-dimensional data (Clarke *et al.*, 2009; Altman & Krzywinski, 2018) common in genotype-to-phenotype studies. A total of 12 feature selection approaches were employed using a range of techniques and feature set sizes to specify sets of transcripts used to train the ML algorithms and evaluate disease prediction. The feature selection approaches fall under the following six techniques – expert domain knowledge, statistical correlation, ML feature importance, co-expression network measures, co-expression network geometry, and a technique termed bipartite graph analysis (see the Materials and Methods section for details). For most feature selection techniques, the size of the feature set (e.g. how many genes to select) is not known. To experimentally determine this, we evaluated the impact of feature set

size on model performance across seven feature set sizes ranging from 10 elements to the whole *A. thaliana* transcriptome (Fig. S9). The results indicate that both feature set size and model type impacted prediction accuracy (ANOVA, $P < 2e\text{-}16$; Table S5), and increasing the number of genes from 10 to 500 dramatically improved model performance (Fig. S9). We

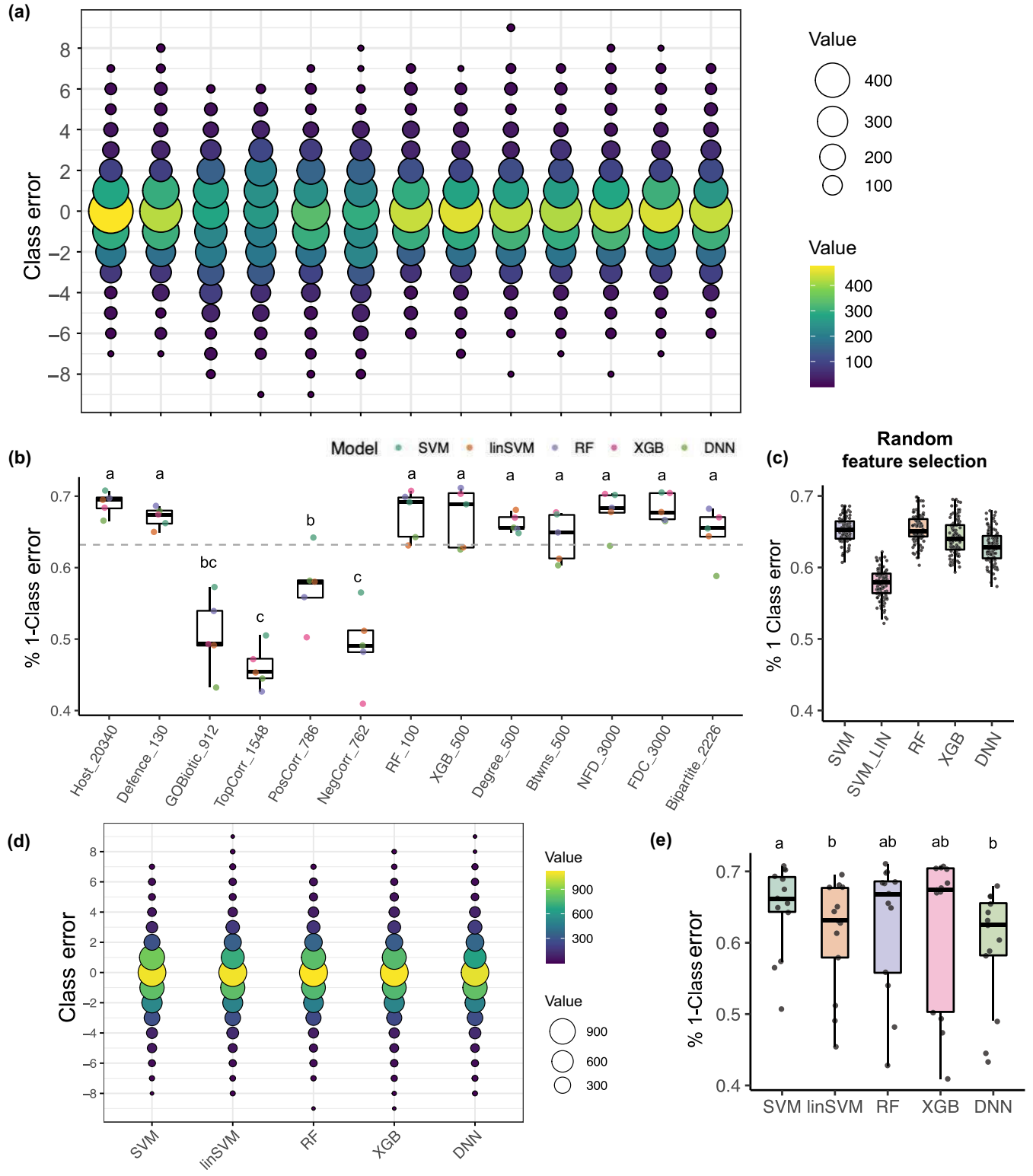

Fig. 2 Evaluation of feature selection methods and machine learning (ML) models on *Arabidopsis* infected by *Botrytis*. (a) Class error of predicted and observed plant disease class. The resulting class error for each sample was counted and shown as a bubble plot, where the bubble size and color indicate the number of samples in that class. The *x*-axis labels correspond to those shown in (b). (b) The 1-class error accuracy of plant disease predictions by feature selection method is shown as box plots. The box plots depict the interquartile range (IQR) of data, representing the 25th to 75th percentile of data, and the median is depicted as a dark line inside of each box. The box whiskers depict the range of data defined by 1.5 times the IQR. Individual results for the five ML models are shown as colored circles indicated by the key above the plot. The five ML models are support vector machine (SVM, teal dot), linear SVM (linSVM, orange dot), Random Forest (RF, purple dot), XG Boost (XGB, pink dot), and deep neural network (DNN, green dot). The *x*-axis samples are the *Arabidopsis thaliana* transcriptome (Host_20340), Gene Ontology (GO) defense-related genes (Defence_130), Gene Ontology biotic interaction-related genes (GOBiotic_912), the top correlated genes (TopCorr_1548), positively correlated genes (PosCorr_786), negatively correlated genes (NegCorr_762), genes from Random Forest feature importance (RF_100), genes from XG Boost feature importance (XGB_500), genes from the calculated network node degree (Degree_500), genes from the calculated network node betweenness (Btwns_500), genes from the network calculated node fractal dimension centrality (NFD_3000), genes from the network calculated fractal dimension centrality (FDC_3000), and genes from the bipartite graph selection (Bipartite_2226). (c) The 1-class error accuracy of the five ML models based on 100 randomly selected features over 100 iterations represented as box plots as described in (b). Each dot indicates the 1-class error accuracy from one set of 100 randomly selected genes for training and testing. The dashed line at 0.632 in (b) represents the average accuracy from random feature selection across all five ML models, provided for comparison. (d) Class error of predicted and observed plant disease class as shown in (a). Results are shown for all feature selection sets grouped by the five ML models. (e) The 1-class error accuracy of predicted plant disease class of data shown in (d). The letters above box blots indicate similarity groupings based on statistical significance (one-way ANOVA, $P < 0.05$). The box plots represent data as described in (b).

considered a number of set sizes for each feature selection method (Fig. S10), and the best feature set size for each feature selection approach as the inflection point for model performance across the tested feature set sizes (Fig. S11). For further evaluation, only a single set size for each feature selection technique was used, indicated as the number associated with the feature selection name (Table S1). Evaluating disease prediction based on class error showed substantial variation in the agreement between observed and predicted disease classes, with both feature selection and ML model choice contributing to prediction performance (Figs 2a, S10; Table S6). Predictions using the entire *A. thaliana* transcriptome resulted in the most zero class error predictions (i.e. no difference between predicted and observed), followed by the co-expression network degree set and the XGBoost feature importance set (Fig. 2a). Assessing the models for 1-class error accuracy, eight of the feature selection sets performed statistically similar compared to predictions made using the entire *A. thaliana* transcriptome (Fig. 2b). This included an expert knowledge set, Defense, based on previous characterization of *A. thaliana–B. cinerea* interaction (Corwin *et al.*, 2016; Zhang *et al.*, 2017), the two feature importance sets from RF and XGBoost, the two sets based on co-expression network measures, the two sets based on co-expression geometry, and the bipartite graph analysis (Fig. 2b). The goal to identify causal associations underlying feature sets must be assessed against random statistical associations in the data. To address this, we created random feature sets of 100 genes sampled from the *A. thaliana* genome, trained and evaluated model performance for the 5 ML algorithms, and repeated this process 100 times selecting new sets of random genes. Interestingly, we see that for four of the ML models, random gene feature sets can predict disease outcomes with *c.* 60–68% 1-class error accuracy (Fig. 2c). The SVM and RF had the highest average accuracies for the random gene feature sets at 65.7% and 65.4%, respectively (Fig. 2c). We interpret this result to show the power of ML models at identifying patterns in data, and in this case, biological meaning is not a prerequisite for predictive power, as has been noted previously (Koo & Ploenzke, 2021). Comparing the 1-class error accuracy of the feature selection sets vs the results from many random sets of genes, many of the feature selection sets had a higher average 1-class error accuracy than the random gene set, including Defense (67.0%), random forest (RF) (67.4%), XGBoost (67.1%), Degree (66.2%), Betweenness (64.3%), node fractal dimension (NFD) (67.9%), fractal dimension (FDC) (68.4%), and Bipartite (64.8%) (Fig. 2b). Assessing the ML models across all feature selections sets showed that the linear SVM and DNN consistently had lower prediction performance (Fig. 2d,e).

Pretrained models can predict disease outcomes for a new plant–fungal interaction

To further assess if specific feature selected genes can generally predict disease outcomes, we tested the models on new data not used for training. An independent dataset of RNA-Seq collected from *Arabidopsis* infected with *S. sclerotiorum* and noninoculated control plants (Badet *et al.*, 2017) was used to predict disease outcomes. The original disease classification was measured on a 1–6 scale, which we transformed to a 0–9 scale to be compatible with our pretrained models (see the Materials and Methods section for details). The full analysis used the 65 trained models from the *A. thaliana–B. cinerea* interaction dataset, consisting of each of our five ML algorithms trained for each of the 12 feature selection sets plus the model trained on the entire *A. thaliana* transcriptome. As the transcriptomes for *B. cinerea* and *S. sclerotiorum* are different, we only used models trained on host transcripts. Measuring class error across the models for each feature selection list showed that the bipartite feature selection list had the most zero class error predictions, followed by the GO Biotic response feature selection list (Fig. 3a). While many of the feature selection sets had a wide range of class errors, 10 feature selection sets had the majority of predictions within one disease outcome class (i.e. between 1 and −1 class error; GOBiotic (63.3%), TopCorr (70.0%), PossCorr (63.3%), NegCorr (90.0%), XGB (73.3%), Degree (70.0%), Betweenness (56.7%), NDF (53.3%), FDC (56.6%), Bipartite (70.0%)) (Fig. 3a,b). Model assessment showed a wide variation of performance, driven by both the ML model and feature selection (Table S7). For example, the average 1-class error accuracy across feature selection sets ranging from

c. 30% to 100% (Figs 3b, S12). These results are also reflected in the 100 rounds of random feature selection assessment, showing a wide range of possible results across models, with the RF and SVM having the highest average 1-class error accuracy (Fig. 3c).

The two treatments, mock vs inoculated, had a significant impact on model performance, where uninoculated plants tended to have a positive class error while inoculated plants tended to have a negative class error, indicating an over and under prediction of

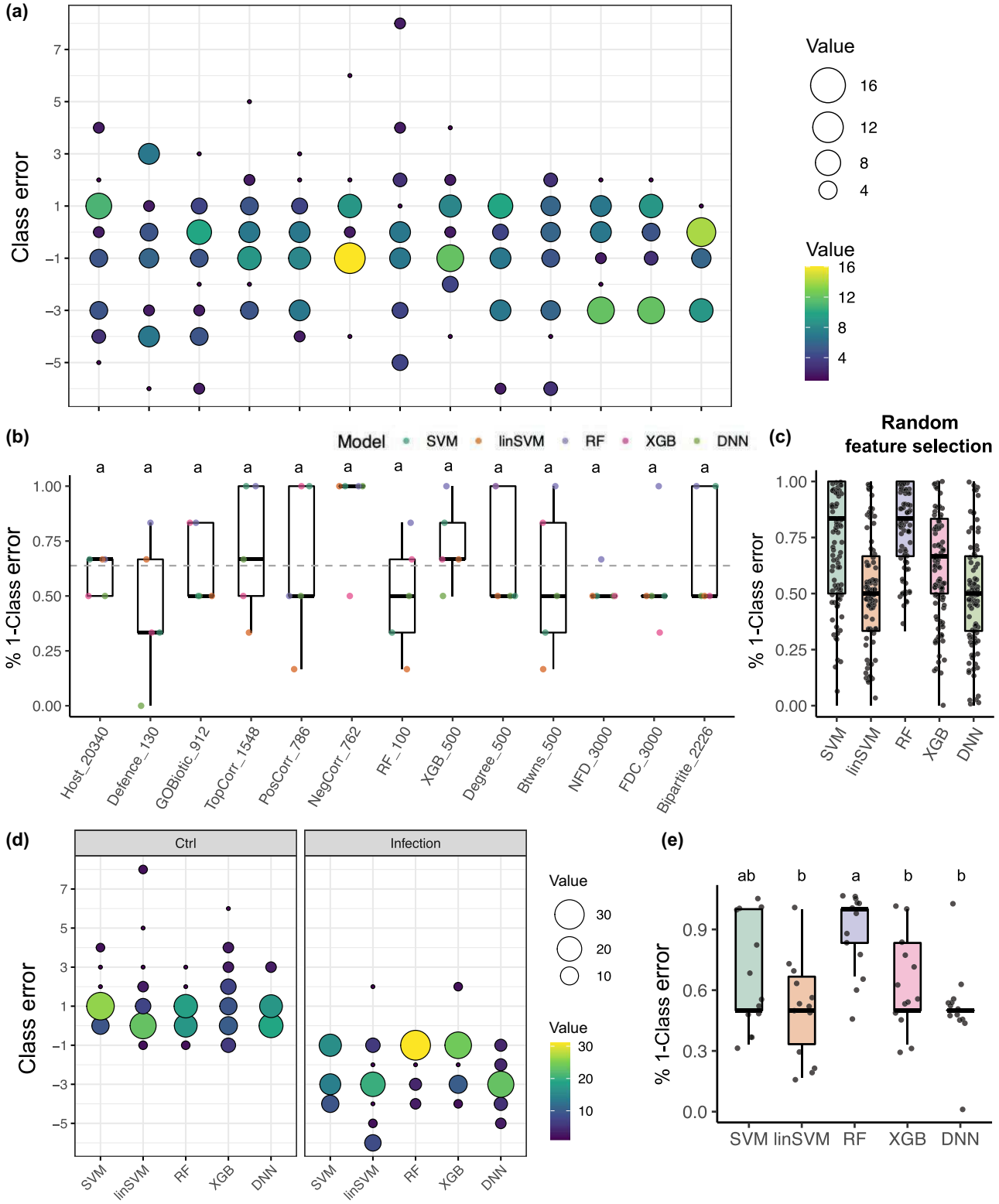

Fig. 3 Trained machine learning (ML) models coupled with feature selection methods can accurately predict plant disease class on *Arabidopsis* infected by *Sclerotinia sclerotiorum*. (a) Class error of predicted and observed plant disease class. The resulting class error for each sample was counted and shown as a bubble plot, where the bubble size and color indicate the number of samples in that class. The *x*-axis labels correspond to those shown in (b). (b) The 1-class error accuracy of plant disease predictions by feature selection method is shown as box plots. The box plots depict the interquartile range (IQR) of data, representing the 25th to 75th percentile of data, and the median is depicted as a dark line inside of each box. The box whiskers depict the range of data defined by 1.5 times the IQR. Individual results for the five ML models are shown as colored circles indicated by the key above the plot. The five ML models are support vector machine (SVM, teal dot), linear SVM (linSVM, orange dot), Random Forest (RF, purple dot), XG Boost (XGB, pink dot), and deep neural network (DNN, green dot). (c) The 1-class error accuracy of the five ML models based on 100 randomly selected features over 100 iterations represented as box plots as described in (b). Each dot indicates the 1-class error accuracy from one set of 100 randomly selected genes for training and testing. The dashed line at 0.638 in (b) represents the average accuracy from random feature selection across all five ML models, provided for comparison. (d) Class errors calculated by predicted and measured plant disease classes on *Arabidopsis* transcriptome of control (Ctrl) uninoculated plants or infected (Infection) by *S. sclerotiorum* by five ML models. Results are shown for all feature selection sets grouped by the five ML models. (e) The letters above box blots indicate similarity groupings based on statistical significance (one-way ANOVA, $P < 0.05$). The box plots represent data as described in (b).

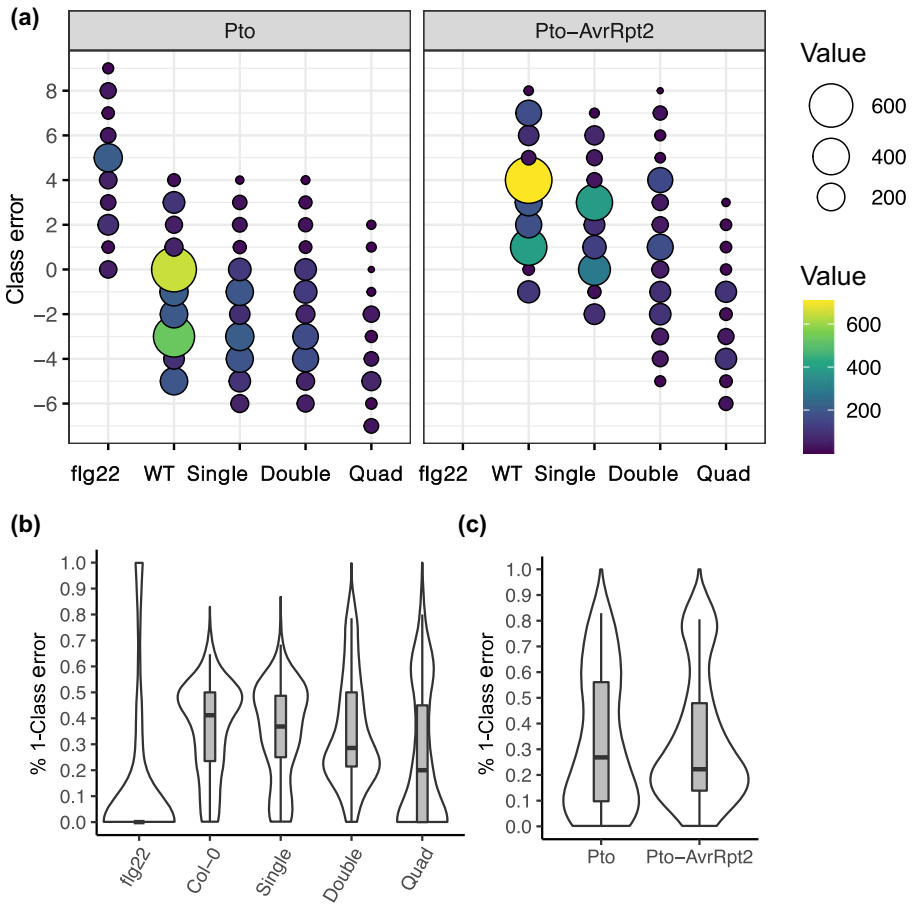

Fig. 4 Genetic perturbations for immunity and virulence hamper generalization for bacteria disease prediction. (a) Class error of predicted and observed plant disease shown as a bubble plot. Individual host treatments were grouped by immune priming (flg22), wild-type Col-0 (WT), or single, double, or quadruple mutation in Col-0. Interactions with Pto *Pseudomonas syringae* shown in left plot, and infection with Pto *P. syringae* expressing the avirulence gene *AvtRpt2* shown on right. Infection by Pto-AvtRpt2 did not use immune priming as a treatment. The results were combined from predictions using all five machine learning (ML) models and feature selection sets previously described. (b) The 1-class error accuracy of predicted plant disease class grouped by host treatment. The data are represented as a violin plot showing the range and distribution of the data, along with a box plot that depicts the interquartile range (IQR) of data, representing the 25th to 75th percentile of data, the median is depicted as a dark line inside of each box, and the whiskers depict the range of data defined by 1.5 times the IQR. (c) The 1-class error accuracy of predicted plant disease class grouped by pathogen treatment shown as a violin and box plot as described in (b).

disease development, respectively (Fig. 3d). The RF model had the highest 1-class error accuracy across all feature selection sets, followed by predictions from the SVM algorithm (Fig. 3e). These results showed that ML models pretrained on reduced feature selection gene sets from the *A. thaliana–B. cinerea* dataset, could correctly predict disease outcomes for an independent dataset of the same host infected with a different fungal pathogen, *S. sclerotiorum*. However, we note that the RF model had 100% prediction 1-class error accuracy for 7 of 12 feature selection sets, and 5 of those 8 feature selection sets provided the highest 1-class error accuracy disease predictions on the original *A. thaliana–B. cinerea* dataset (Figs 2b, 3b). This suggests the ability to deliver highly accurate disease outcome predictions for independent pathosystems using ML, and that feature selection may allow a significant reduction in the number of predictor terms while maintaining model performance.

ML models can predict disease outcome for cross-kingdom pathogen attack

To rigorously investigate the modeling results, we extended our predictions to another independent dataset, *Arabidopsis* infected

with the bacterium *P. syringae* (Nobori *et al.*, 2018). The purpose of this analysis was twofold. First, changing host infection from an eukaryotic to prokaryotic pathogen broadly tests the robustness of the trained models and therefore their learned patterns. This addresses the issue of statistical pattern matching common in big data (i.e. overfitting), in which predictions are good

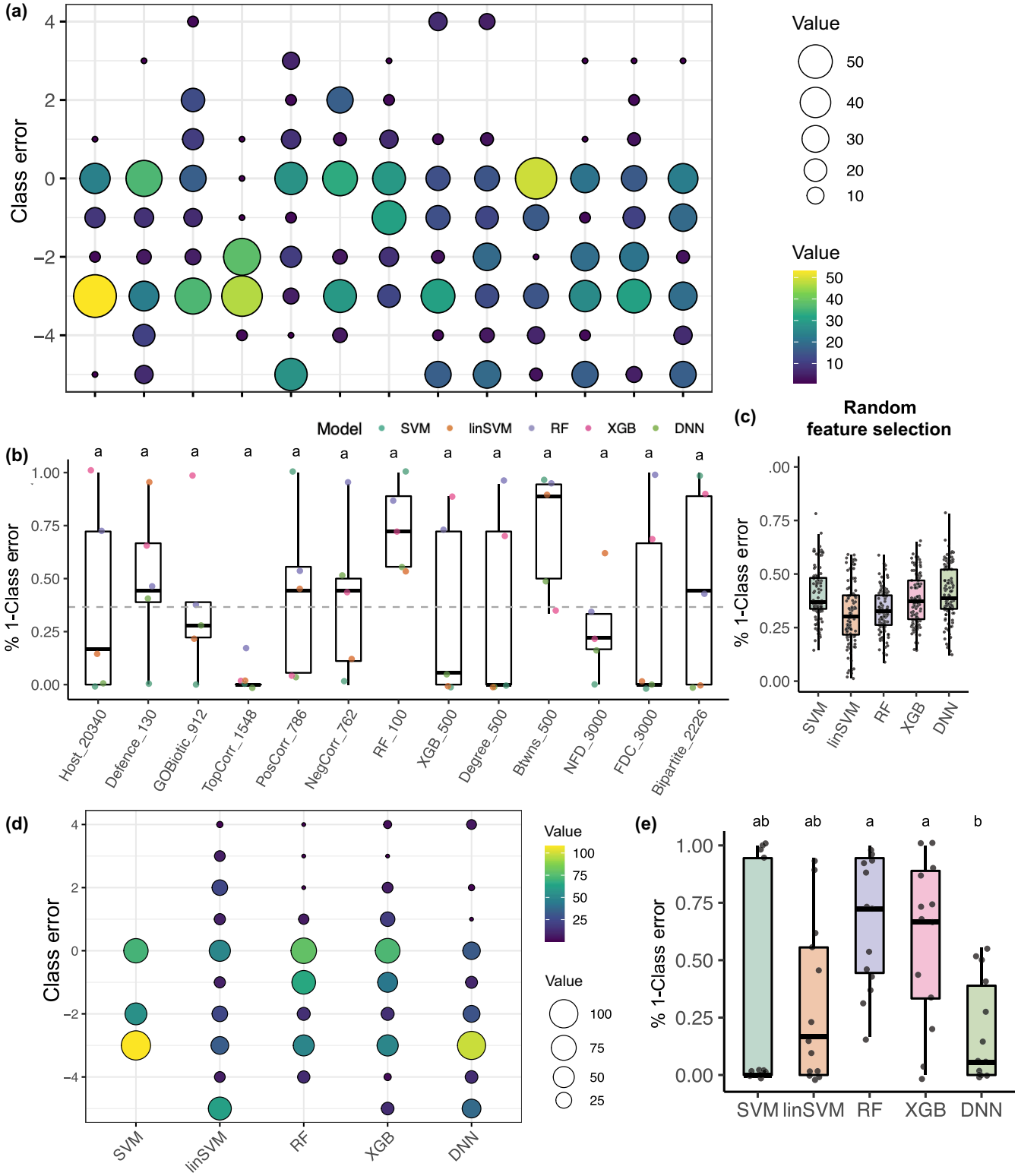

Fig. 5 Trained machine learning (ML) models can predict disease outcomes for new pathosystems. (a) Class error of predicted and observed plant disease class for *A. thaliana* infected by the bacterial pathogen *P. syringae*. The resulting class error for each sample was counted and shown as a bubble plot, where the bubble size and color indicate the number of samples in that class. The *x*-axis labels correspond to those shown in (b). (b) The 1-class error accuracy of plant disease predictions by feature selection method are shown as box plots. The box plots depict the interquartile range (IQR) of data, representing the 25th to 75th percentile of data, and the median is depicted as a dark line inside of each box. The box whiskers depict the range of data defined by 1.5 times the IQR. Individual results for the five ML models are shown as colored circles indicated by the key above the plot. The five ML models are support vector machine (SVM, teal dot), linear SVM (linSVM, orange dot), Random Forest (RF, purple dot), XG Boost (XGB, pink dot), and deep neural network (DNN, green dot). (c) The 1-class error accuracy of the five ML models based on 100 randomly selected features over 100 iterations represented as box plots as described in (b). Each dot indicates the 1-class error accuracy from one set of 100 randomly selected genes for training and testing. The dashed line at 0.366 in (b) represents the average accuracy from random feature selection across all five ML models, provided for comparison. (d) Class errors calculated by predicted vs observed disease classes for all observations by the five ML models. Results are shown for all feature selection sets grouped by the five ML models. (e) The 1-class error accuracy of predicted plant disease class for data shown in (d). The letters above box blots indicate similarity groupings based on statistical significance (one-way ANOVA, $P < 0.05$). The box plots represent data as described in (b).

for their original data, but poor for new datasets. If our models were overfit to the original dataset, then model performance for bacterial infection would be poor. Alternatively, if our trained models learned general host predictors of disease outcome, the models should perform well on the new system, related to the hypothesis that plants contain general transcriptional responses to diverse microbes. Second, this approach addresses an area of ML termed transfer learning, which leverages the knowledge gained from one task to complete a different but related task. In this case, instead of building a new predictive model from scratch for *P. syringae* infection, we transfer our pretrained models to this new dataset. The limits and demonstration of transfer learning in biological systems is not well developed, and this offered a good test case. The *A. thaliana–P. syringae* dataset comprised 27 different treatments made up of combinations of genetic differences in the host and pathogen, as well as immune priming (Fig. S2; Nobori *et al.*, 2018). The specific responses, such as immune priming and combinatorial immune mutants, would not be expected to produce transcriptional responses similar to the original training dataset, and therefore, prediction results would likely not be accurate for such interactions. To test this, disease predictions were compared to the observed predictions and both host genotype and pathogen genotype impacted predictions (Table S8). The dataset was split for *P. syringae* expressing or not expressing AvrRpt2, and into *A. thaliana* WT Col-0, immune primed Col-0, and single, double or quadruple Col-0 immune mutants, and include all 5 ML models trained on the 12 different feature selection sets (Fig. 4a). The interaction between WT host and pathogen produced substantially more zero class error predictions (32.8%), and predictions within one error class (46.11%), consistent with our expectations for better predictive performance (Fig. 4a). Compared to WT Col-0 infection with virulent *P. syringae*, disease severity was overestimated for the immune primed infection (i.e. flg22 pre-exposure), and underestimated on hosts with compromised immunity (Fig. 4a). Clearly, immune priming and genetic perturbation of host immunity negatively impacted predictive performance. The impact of immune priming and combinatorial genetic perturbations was also seen using 1-class error accuracy assessment, showing that non-WT interactions negatively impacted model performance (Fig. 4b). The 1-class error accuracy was also lower when *A. thaliana* was infected by an avirulent *P. syringae* strain expressing *AvrRpt2* (Fig. 4c). Collectively, these results show that disease predictions for WT compatible host–pathogen interactions were more accurate than predicting incompatible or host immune compromised outcomes. This suggests that genetic perturbations and NLR based immunity caused a transcriptional profile that was too dissimilar from the original training data for accurate prediction. Therefore, only data for virulent *P. syringae* infecting WT Col-0 (13 observations) were used for further analysis.

To understand how our *A. thaliana–B. cinerea* trained models with feature selection perform on *A. thaliana–P. syringae* WT interaction data, disease predictions were generated. Assessment through class error showed variable results across feature selection set trained models (Fig. 5a). The feature set based on network analysis node betweenness gave the most zero class errors predictions (Fig. 5a). Predictions using feature selection sets Defence (53.3%), RF (68.1%), and Betweenness (74.8%), provided at least half of their respective predicted disease classes within one-class error of the observed value. Quantifying 1-class error accuracy also indicated variable performance across the ML models (Fig. 5b). These results reflect the complexity of the prediction problem, which uses host transcriptome responses of only a small subset of genes from early infection of a fungal pathogen to predict disease outcomes for bacterial infection. Nonetheless, 6 of the 12 feature selection sets produced an average 1-class error accuracy greater than the average random feature set (36.6%) (Fig. 5b,c): Defence (48.9%), PosCorr (42.2%), NegCorr (40.0%), RF (74.4%), Betweenness (72.2%), and Bipartite (46.7%). Also, 4 of the 8 feature selection sets that performed above the random feature sets for *A. thaliana–P. syringae* were the same that provided the highest 1-class error accuracy predictions from the original *A. thaliana–B. cinerea* training (Figs 2b, 4b). This is an indicator of robust performance for these feature selection sets. Looking at individual model performance, the distribution of predicted class errors and 1-class error accuracy shows differences between ML algorithms (Fig. 5d,e). When analyzed across all feature selection sets, the RF model had the highest prediction performance, average 1-class error accuracy of $65.4 \pm 28.0\%$ (mean $\pm$ SD), followed by XGBoost, $58.5 \pm 34.4\%$ (mean $\pm$ SD) (Fig. 5d,e). Overall, the change from a eukaryotic to prokaryotic pathogen did not preclude accurate disease prediction, suggesting that a common transcriptional response could be modeled that predicted disease outcomes across a range of interactions.

Feature selection identified known and underexplored genes involved in plant defense

We aimed to understand the biological significance of underlying genes of the feature selection sets. To integrate the analysis across all three pathosystems, we re-focused our analysis on results from only the RF and XGBoost algorithms that proved most accurate for *P. syringae* predictions (Fig. 5e). Disease predictions across the three pathosystems were reassessed for the RF and XGB trained models for each feature selection sets (Fig. 6a). The

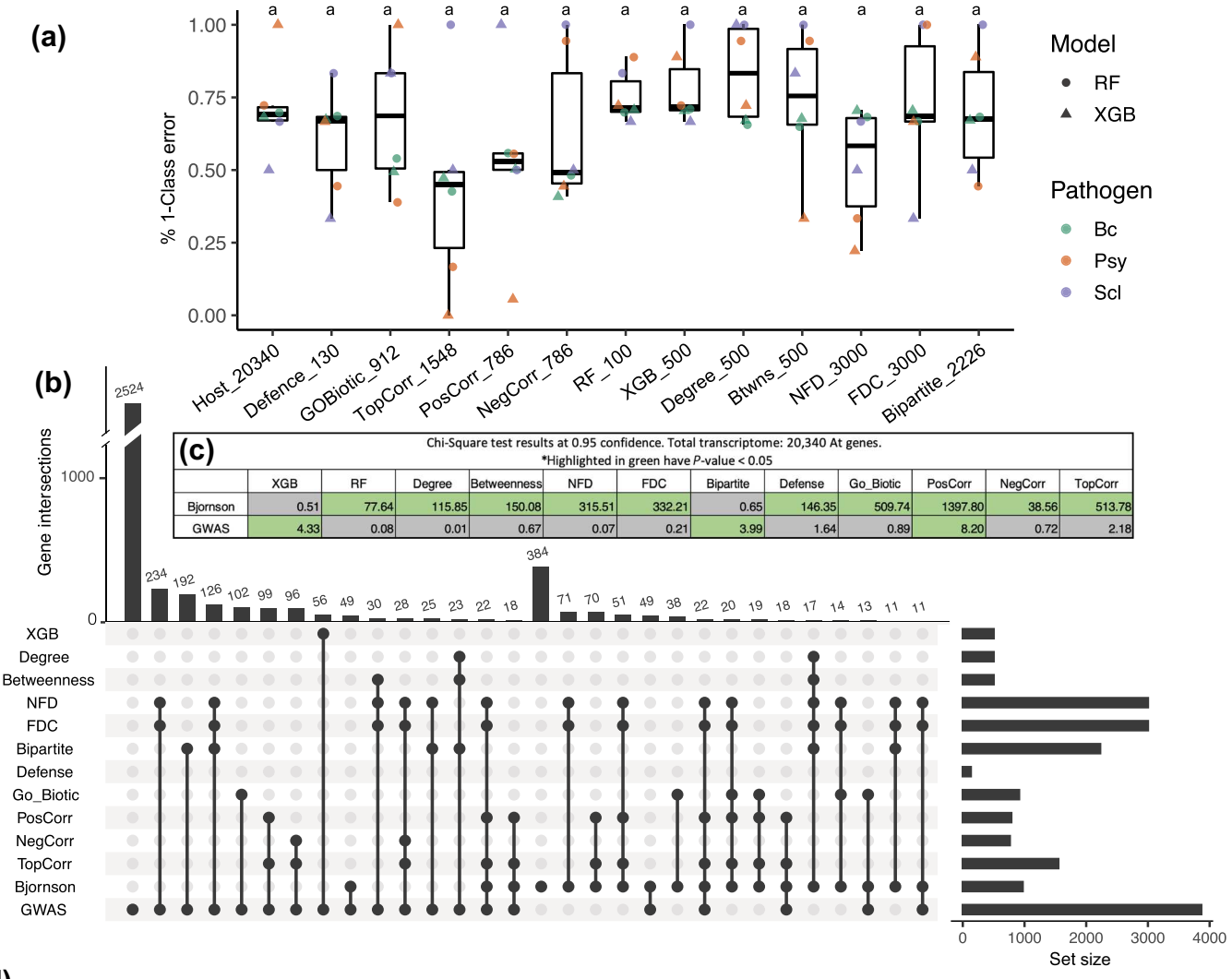

Fig. 6 Uncovering general predictors of the plant immune response. (a) The 1-class error accuracy of plant disease predictions by feature selection method are shown as box plots. The box plots depict the interquartile range (IQR) of data, representing the 25th to 75th percentile of data, and the median is depicted as a dark line inside of each box. The box whiskers depict the range of data defined by 1.5 times the IQR. Individual results from the two machine learning (ML) models are shown, Random Forest (RF, circle) and extreme gradient boost (XGB, triangle), from the three datasets, *Botrytis cinerea* (Bc, green), *Pseudomonas syringae* (Psy, orange), and *Sclerotinia sclerotiorum* (Scl, purple). The letters above the box plots indicate no significant difference (One-way ANOVA, $P < 0.05$). (b) UpSet plot of *Arabidopsis thaliana* gene features based on different feature selection lists. This shows the top 15 set intersections between the different feature selection lists and the Bjornson (Bjornson *et al*., 2021) and GWAS (Corwin *et al*., 2016) gene lists. (c) Results of chi-squared test of independence at 0.95 confidence for the different feature selection lists vs the Bjornson and GWAS gene lists. Chi-squared results that were statistically significant at *P*-value of 0.05 are highlighted in green. Gene sets tested for independence were XG Boost (XGB), RF, network node degree (Degree), network node betweenness (Betweenness), node fractal dimension (NFD), fractal dimension centrality (FDC), bipartite graph (Bipartite), Gene Ontology (GO) defense-related genes (Defense), Gene Ontology biotic-interaction related genes (GOBiotic), positively correlated genes (PosCorr), negatively correlated genes (NegCorr), and the top correlated genes (TopCorr). (d) Left, Line plot showing the proportional rank of the top 20 common high-level GO terms identified across the seven feature selection gene sets. The gene sets are identified on the *x*-axis, while the individual GO terms are shown to the right. Middle, the distribution of proportional ranks for each of the GO term groups. The box plots depict the IQR of data, representing the 25th to 75th percentile of data, and the median is depicted as a gray line inside of each box. The box whiskers depict the range of data defined by 1.5 times the IQR and each individual rank is shown as a black point. Not all feature selection sets had the same top 20 GO terms, resulting in missing ranks from being displayed in the Left line plot. The top-ranked GO term for three feature selection sets, regulation of biological quality, was added to the plot as black points in their corresponding rank position and labeled above the plot.

average 1-class error accuracy for all feature selection sets is higher than the random feature selection set (Fig. 6a) (random feature selection accuracies for RF (32.4%) and XGB (38.4%)). Remarkably, three feature selection sets, XGB (79.1%), Degree (80.7%), and Betweenness (77.4%), have an average 1-class error accuracy of over 75% across all three pathosystems (Fig. 6a). The transcriptional response of each gene set per host infection indicated that these genes are not necessarily the most highly expressed genes during infection (Fig. S13A), nor are they significantly more induced as a group during infection compared to noninfected (Fig. S13B,C). These results are interesting because most transcriptional analysis of the host immune response involves identification of highly expressed genes or genes differentially expressed between healthy and infected plants. Comparing the individual gene composition for each feature selection set showed they were largely nonoverlapping sets, indicating that the different methods identified different genes (Fig. 6b). Two gene sets from previously published literature were used as controls to understand and validate the results. One control gene set contained 2524 genes previously identified through genome-wide association study (GWAS) of *A. thaliana* genotypes infected by *B. cinerea* (Corwin *et al*., 2016). The UpSet plot showed that the GWAS had the largest overlap with the NFD and FDC gene sets, as well as the Bipartite set (Fig. 6b). A chi-square test of independence showed that the GWAS gene set significantly overlapped the XGBoost ($\chi^2 = 4.33$, $P = 0.037$), bipartite graph ($\chi^2 = 8.47$, $P = 0.004$), and positive correlation ($\chi^2 = 3.98$, $P = 0.045$) feature sets (Fig. 6c; Table S9). We also compared the feature selection sets to 970 genes identified as common early transcriptionally responsive genes in *A. thaliana* to diverse immunogenic elicitors (Bjornson *et al*., 2021), referred to, here, as the Bjornson set. The Bjornson set also had the largest overlap with the NFD and FDC gene sets, as well as the Top and Positive correlation sets (Fig. 6b). A chi-square test of independence showed that the Bjornson set had statistically significant overlap with nearly all the identified feature selection sets (10 of 12 sets, $P < 0.001$, see Table S10 for numbers), except for the XGBoost and the bipartite feature selection sets (Fig. 6b).

To understand the functional relevance of the feature selection gene sets, gene ontology (GO) enrichment analysis for Biological Function was performed for the seven feature selection sets derived from ML and geometric graph analysis (Fig. 6d). There was significant GO pathway enrichment for five of the feature selection sets (FDR cutoff 0.05, Tables S11–S15), but not for the XGB or bipartite gene sets. For the two feature selection sets based on network metrics of the gene co-expression graphs, Betweenness and Degree, we found that the roughly top 20 overrepresented pathways had some overlap, but also many differences (Figs S14, S15). For instance, the Betweenness set is the most overrepresented for two pathways related to indole-containing compound biosynthesis/metabolism, followed by two pathways related to photosynthesis (Fig. S14). The top seven enriched GO biological function pathways for the Degree gene set are predominately related to photosynthesis and ATP synthesis, followed by indole-containing compound biosynthesis/metabolism (Fig. S15). The remaining significantly enriched pathways for the Betweenness feature selection set were all related to response to stress/chemical/biotic/external stimulus (Fig. S14). The two feature selection sets based on fractal geometric analysis of the gene co-expression graphs, NFD and FDC, identified a very similar top 20 significantly enriched pathways (Figs S16, S17). Pathways related to photosynthesis were again the most significantly enriched, followed by pathways related to response to oxygen levels. The NFD and FDC sets were also significantly enriched for pathways related to response to stress/-chemical/biotic/external stimulus, as well as organic substance metabolism/biosynthesis (Figs S16, S17). The relatively smaller RF feature selection gene set of 100 genes was significantly enriched for 11 GO biological function pathways (Fig. S18). The GO biological function pathways with the highest fold enrichment were related to response to organic cyclic compounds, defense response to bacterium, and response to biotic/external stimulus (Fig. S18). These results show that while there was rather limited overlap between the individual genes across the feature selection sets (Fig. 6b), there were many overlapping significantly enriched GO-term Biological Process pathways, including related

to process such as photosynthesis, organic metabolism, and response to stimulus. We further used the high-level GO terms in each of the seven feature selection sets to rank the top 20 terms based on the proportion to the entire category (Table S16). This allowed incorporation of the XGB and Bipartite feature selection sets that did not have statistical enrichment but could be used for rank assessment of GO terms across gene sets. The summary rank plots showed that high-level GO categories related to response to external and biotic stimulus were on average the most highly ranked categories based on the feature selection gene sets (Fig. 6d). The rank plots also identified where there were substantial differences between feature selection sets. For example, the fifth ranked GO term for the Bipartite gene set was related to developmental process involved in reproduction, which was ranked much lower in the other sets (Fig. 6d). The XGB feature selection set showed many rank order changes, such as macromolecular organization was rank 2, response to abiotic stimulus was rank 3, and response to biotic stimulus was rank 11, which were all different ranks than the average ranking for the respective term across the other feature selection sets.

There was no concordance for the top proportionally ranked terms across feature selection sets. For instance, the top proportionally ranked term for the XGB, Betweenness, and Degree feature selection sets was regulation of biological quality (Fig. 6d; Table S16), but the term was not in the top 20 for any of the other feature selection sets. Interestingly, while this term was proportionally ranked one for each of the three feature selection sets, the XGB feature selection genes that overlap this GO term are unique compared to the Betweenness or Degree, while the Betweenness and Degree feature selection genes that overlap this GO term had a more substantial overlap (Fig. S19). Looking more deeply into the 22 genes uniquely identified in the XGB feature selection set that overlap the regulation of biological quality GO term, a number of known or potentially underexplored genes and pathways related to plant defense were identified. For instance, one identified gene, *A. thaliana* HopM interactor 7 (*AtMIN7*, At3g43300), is an effector target and required for *A. thaliana* immunity to the bacterial pathogen *P. syringae* (Nomura *et al.*, 2006, 2011). Molecularly, AtMIN7 is an adenosine diphosphate ribosylation factor (ARF) guanine nucleotide exchange factor (GEF), which functions in early endosomal vesicle trafficking (Tanaka *et al.*, 2009). Importantly, while AtMIN7 was identified for its role in plant immunity to bacteria, the homolog in wheat is functionally required for resistance to *Fusarium graminearum*, the causal agent of Fusarium head blight (Machado Wood *et al.*, 2021). Another gene implicated in plant defense that was part of this unique subset of 22 genes from the XGB feature selections related to the regulation of biological quality was *callose synthase 7* (*CalS7*, At1g06490), which is a phloem-specific callose synthase required for normal plant growth and development, but also provides callose deposition during wound response (Xie *et al.*, 2011). Another two of the 22 genes have been implicated in plant response to phosphate starvation. This includes *alfin-like 6* (*AL6*, At2g0270) a plant homeodomain (PHD) containing protein that can function as a histone chemical modification reader and contributes to transcriptional regulation in response to phosphate levels (Chandrika *et al.*, 2013), but also in response to the plant hormone jasmonic acid (Vélez-Bermúdez & Schmidt, 2021). The other gene, *phosphate deficiency response 2* (*PDR2*, At5g23630), encodes a P_5-type ATPase localized to the endoplasmic reticulum and is required for normal root patterning and growth under low-phosphate conditions (Ticconi *et al.*, 2009). Lastly, two other genes in this list are involved with mRNA maturation mediated by the spliceosome complex. This includes the small nuclear ribonucleoprotein *SM-like4* (*LSM4*, At5g27720) involved in pre-mRNA alternative splicing (Zhang *et al.*, 2011), and At2g42330, which has not been extensively characterized but is annotated to be involved in spliceosomal complex disassembly. The regulation of LSM4 through protein methylation has recently been shown to impact alternative splicing and is required for normal Arabidopsis response to abiotic stress and infection by *Pseudomonas* (Agrofoglio *et al.*, 2024). Additionally, the interaction of LSM4 with a metacaspase (AtMC3), was previously reported as required for normal mRNA processing and resistance to *P. syringae* (Wang *et al.*, 2021). This indicates that our analytic approach identified the splicesome and alternative splicing as an underexplored mechanism contributing to plant stress and defense responses, likely as a general, not specific, response pathway. Collectively, these results indicate that our ML approach and feature selection techniques identified highly relevant genes involved in plant defense response using a novel method not previously applied to plant-microbe biology. These results support the hypothesis that the feature selection gene sets represent collections of genes broadly predictive of Arabidopsis disease development, and likely contain genes or represents pathways that can be more fully explored for their role in plant disease development.

Discussion

Plants possess two distinct classes of immune receptors, membrane receptors surveying the extracellular space, and cytosolic receptors sensing microbial activity (Ngou *et al.*, 2022). Evidence indicates that the plant immune system can be conceptually thought of as an integrated functional unit, with cross talk between immune receptor classes and disease outputs providing synergy, redundancy, and specificity (Tsuda *et al.*, 2009; Dong *et al.*, 2015; Yuan *et al.*, 2021). Despite our growing knowledge of the plant immune system, there are few tools or approaches that can predict a plant's general immune performance. Current research and improvement approaches rely heavily on specific single-gene interactions, which have proved effective, but are hampered by their intensive time requirements and lack of generalization. Additionally, such genetic resistance has proven relatively short lived when deployed at scale (Kiyosawa, 1982), although considerable effort is being put into stacking immune receptors to increase efficacy in the field (Pradhan *et al.*, 2015; Ghislain *et al.*, 2019; Luo *et al.*, 2021). As plant sciences and crop improvement move to engineered systems approaches (Shigenaga *et al.*, 2017; Marchal *et al.*, 2022; Vuong *et al.*, 2023), greater understanding and predictive power for processes such as plant immunity, abiotic stress tolerance, growth, and nutrient

utilization are needed. Systems approaches leveraging increasingly common big datasets, analyzed by interdisciplinary teams, can help provide the knowledge and framework to achieve these goals.

Here, we addressed general plant immunity by looking for early transcriptional patterns that are predictive of disease outcomes across a range of pathogen attack. Our hypothesis was that plants have a common early signaling response to a range of immunogenic signals that lead to different degrees of defense output. We tested this hypothesis by training models on RNA-Seq data from one pathosystem, and independently testing the trained models on new input data from different pathosystems. Such a model should only be predictive of disease outcome in the new system if the model learned general patterns of early RNA-Seq response indicative of final disease outcome. We further refined our approach, and addressed the underlying biology, by identifying 10 subsets of host genes using feature selection techniques in order to train the ML models with fewer parameters. A significant finding from this research is that ML models trained using only a fraction of the total host transcriptome, from 0.5% to 15% of total genes, were able to accurately predict disease outcomes across fungal and bacterial pathosystems, including both necrotrophic and biotrophic pathogens. This is important because many ML models do not perform well on new datasets, showing a lack of generalization. We interpret this predictive performance across diverse datasets to reflect a general disease-related transcriptional response that was captured by the models. The results indicate that different feature selection gene sets, that have largely nonoverlapping membership, can independently provide high-predictive power across diverse pathosystems. This result is consistent with the idea of immune response canalization, and that immune signaling is a robust network with many paths to a similar response (Tsuda *et al.*, 2009; Zhang *et al.*, 2017). Another interpretation is that given a high number of predictors (genes) there are many combinatorial sets that can be used for predicting the outcome. This does warrant caution from overinterpreting the significance of a specific gene in a given gene set, as the result may more broadly reflect important pathways, but the individual gene may not be a major determinant of the outcome. The applications of ML to large biological datasets remains an active area of investigation and much work is needed to understand the limits of interpretation.

There are many novel aspects to our ML interrogation of the plant immune response. Our approach overcomes general limitations for the most common analysis pipelines, such as differential gene expression (DGE), cluster or module analysis, and network construction utilizing simple co-expression (i.e. correlation). Such approaches fail to capture nonlinear relationships, have limited predictive power, and have limited generalizability. For instance, commonly used DGE simply reports transcript levels that are higher or lower between conditions, but the approach fails to integrate patterns or relationships between transcripts, which collectively account for a response of interest. Gene co-expression networks integrate relationships between transcripts, but networks constructed using linear correlations will miss nonlinear relationships and the approach assumes that transcripts with similar patterns are functional related. By contrast, ML models can identify more diverse transcriptional patterns that collectively correspond to phenotype development. The ML models employed here capture nonlinear dynamics, are not limited to identifying transcriptional outliers, and inherently integrate multigenic patterns. Our approach utilized disease classes instead of continuous scale disease measurements so that models could be applied across pathosystems with diverse disease phenotypes. The use of classification can also aid interpretability over a continuous scale that may require expert knowledge of the pathosystem. Classification systems are also more robust to outliers in training data as the prediction is not influenced by the magnitude of the data. A drawback of classification is the loss of information through discretization, in which different measurements are combined into a single class. Our approach to use 10 disease classes helped to balance information loss with generalization across systems, and our use of a 1-class error assessment reflects the likely small overall difference between adjacent disease classes. We acknowledge that using a 1-class error estimate over estimates model performance. If a research project intends to distinguish or predict small differences in disease outcomes, our models would not be appropriate. This is also largely a reflection of the dataset used for model development. The dataset was skewed to having a high number of observations in a disease class in the range of one to three. We showed through analysis that the models performed much better than randomly assigning a disease outcome, demonstrating the utility of ML for this application, but the issue of data balance is a challenge for biological data. Our approach was limited to using only host transcriptional data for cross-pathosystem predictions. Future efforts to use more diverse training data organized in such a manner that allows more biological diversity and microbial data to be included in the prediction could further aid in generalization and understanding. Our results clearly show that ML is a valuable modeling approach that can identify salient patterns in large data to make predictions. An interesting finding is how robust the ML models were, as evidenced by high-prediction accuracy even with 25% mislabeled data, using only a fraction of the transcriptome, and across diverse pathosystems. The underlying basis for this robustness is not clear, but could be related to both the biology of plant–microbe interactions and the use of ML on large datasets. The polygenic nature of plant disease manifestation means that a substantial number of transcripts collectively contribute to the outcome, and therefore, a large combinatorial set of transcriptional patterns can each be used for prediction. Feature selection may help identify key genes or pathways, but it may also remove key hubs important for a particular disease outcome. We also note that while our models performed better than null control models, average performance across diverse trials often showed average accuracy measurements in the mid to upper seventies. To increase accuracy, training models on smaller gene sets with known direct impact for a specific system, higher resolution disease phenotyping, and deeper sampling may produce higher accuracy models. However, this may be at the cost of model generalization across pathosystems. The development of a very large pretrained model using diverse data, followed by specific pathosystem fine-tuning may provide a solution that balances these considerations. The transfer of trained models across biological

systems or questions is underdeveloped, but is important in order to leverage large datasets, especially from model systems. Future efforts and more sophisticated methods of transfer learning will aid in ML for small datasets and may provide a novel means to compare and contrast systems. Along with advancements in computer science, further developing experimental approaches will be important, such as using a pan-genome approach or the development of a meta-pathosystem design, in which multiple species are used in data collection for model development. A further consideration for future research is the use of transcriptional data from nonbiotic interactions to create a 'noninfected' class, or other approaches to create null-models to aid in interpretation or gene identification.

Another contribution of this research is demonstrating how to link biological questions to the deployment of ML to enhance our understanding and interpretation. A concern for ML in biology is the identification of biologically irrelevant patterns. This can occur during model training, where patterns particular to a given dataset are learned (i.e. overfit). We reasoned that if our ML models were overfit to *A. thaliana–B. cinerea* interactions, they would have limited predictive power for the new pathosystem data. Since this was not the case, our conclusions is that models that performed well across pathosystems, trained on feature selection gene sets, reflect biologically meaningful patterns for general plant response to infection. These results provide insights into the biology of plant–microbe interactions. For instance, while the best feature selection approaches did not identify the same genes, they did identify similar GO biological processes. This again highlights the robust nature of the immune systems, that many genes making up a pathway response play an important role in disease development. It was also quite striking how many of the most significantly overrepresented GO-term biological pathways were related to response to stress, biotic, abiotic, and external stimulus. This indicates that the feature selection techniques were not identifying statistical noise in the data, but accurately identifying biologically relevant genes. This likely contributes to the models being able to predict disease across the various pathosystems. This point is highlighted by the XGB ML feature selection set identifying *AtMIN7* as an important gene from the *A. thaliana–B. cinerea* training data. This gene was originally identified as a target of a *P. syringae* type three secreted effector (Nomura *et al.*, 2011), but has been shown to more broadly impact plant disease development to fungal pathogens as well (Machado Wood *et al.*, 2021), possibly through its role in leaf cuticle or stomatal development (Zhao *et al.*, 2020). The identification of *AtMIN7* from a fungal infection datasets highlights that the plant immune system evolved to defend against all classes of pathogens, and while there are specific responses, there is also substantial molecular overlap between host responses to diverse pathogens. This represents a conceptual shift that can help identify pathways and networks for improvement to develop more general and robust plant immune responses. Another example is the identification of genes involved in phosphate response. The phosphate starvation pathway is a key network regulating plant response and interaction with mycorrhizal symbionts (Shi *et al.*, 2021), and has more recently been implicated as a key hub coordinating nutrition and defense responses and community membership of the root microbiota (Castrillo *et al.*, 2017). The interaction between phosphate response and defense appears to be molecularly integrated by two key proteins, Phosphate Transporer1 (PHT1) and the receptor-like cytoplasmic kinase Botrytis-Induced Kinase 1 (BIK1) (Dindas *et al.*, 2022). Plant phosphate response has even been identified as contributing to resistance to insect herbivory through induction of the jasmonate pathway (Khan *et al.*, 2016). Further understanding and exploiting key pathways that can broadly impact plant growth and development are critical to future crop improvement. This research highlights that ML and big data are important tools to provide insights and generate hypotheses. Future research to understand how identified, but uncharacterized genes, are related to plant immunity is important to broaden our understanding of general plant immunity and provide additional breeding targets for crop improvement.

There is increasing interest in predictive biology and engineered systems to benefit society. We provide here an example of mining publicly available data to understand general components of plant disease development. This approach is general and scalable, allowing for the interrogation of other datasets and integration of new data as it becomes available. It will be of great interest to go from predicting disease outcomes from RNA-Seq to predicting general disease outcome from DNA sequence. Results presented here further our understanding of complex plant–microbe interactions and offer a framework for future research.

Acknowledgements

The authors wish to thank Dr Kenichi Tsuda for providing original bacteria growth phenotypic data and Sylvain Raffaele for providing comments on *S. sclerotiorum* disease index. This work is supported by the National Science Foundation (NSF) through the Models for Uncovering Rules and Unexpected Phenomena in Biological Systems (MODULUS) (award no. MCB-1936800 to DEC and MCB-1936775 to PB), United State Department of Agriculture-National Institute of Food and Agriculture (USDA-NIFA) (award no. 2018-67013-28492) to DEC, and awards to PB through the NSF Career Award (CPS/CNS-1453860), the NSF awards (CCF-1837131, CNS-1932620, and CMMI-1936624), the NSF award No. 2243104 under the Center for Complex Particle Systems (COMPASS), the DARPA Young Faculty Award and Director's Fellowship Award (N66001-17-1-4044), and the U.S. Army Research Office (ARO) under Grant No. W911NF-23-1-0111. [Correction added on 19 December, after first online publication: the funding details for the author Paul Bogdan have been updated in the preceding sentence.] The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied by the Defense Advanced Research Projects Agency, the National Science Foundation, or the Department of Defense.

Competing interests

None declared.

Author contributions

All the authors contributed to the development of the project. JS and MC were responsible for data preprocessing and model construction. JS and WZ conducted data analysis with contributions by MC, PB and DEC. All authors contributed to data interpretation and writing the manuscript. JS and WZ contributed equally to this work.

ORCID

Paul Bogdan https://orcid.org/0000-0003-2118-0816
Mingxi Cheng https://orcid.org/0000-0002-8070-6665
David E. Cook https://orcid.org/0000-0002-2719-4701
Jayson Sia https://orcid.org/0000-0002-5790-2801
Wei Zhang https://orcid.org/0000-0002-5092-643X

Data availability

All datasets used in this study are publicly available. The *A. thaliana* and *B. cinerea* transcriptome data are available through NCBI Bioproject PRJNA473829 (Zhang *et al.*, 2017, 2019). The *A. thaliana* and *S. sclerotiorum* transcriptome data are available through NCBI GEO accession GSE106811 (Badet *et al.*, 2017). The *A. thaliana* and *P. syringae* transcriptome data are available through NCBI GEO GSE103442 (Nobori *et al.*, 2018). All codes are fully open source and available via GitHub at https://github.com/jcsia/AtBotML.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M *et al.* 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv*: 1603.04467.

AbuQamar S, Moustafa K, Tran LS. 2017. Mechanisms and strategies of plant defense against *Botrytis cinerea*. *Critical Reviews in Biotechnology* **37**: 262–274.

Aerts N, Pereira Mendes M, Van Wees SCM. 2021. Multiple levels of crosstalk in hormone networks regulating plant defense. *The Plant Journal* **105**: 489–504.

Agrofoglio YC, Iglesias MJ, Perez-Santángelo S, de Leone MJ, Koester T, Catalá R, Salinas J, Yanovsky MJ, Staiger D, Mateos JL. 2024. Arginine methylation of SM-LIKE PROTEIN 4 antagonistically affects alternative splicing during Arabidopsis stress responses. *Plant Cell* **36**: 2219–2237.

Altman N, Krzywinski M. 2018. The curse(s) of dimensionality. *Nature Methods* **15**: 399–400.

Aprianto R, Slager J, Holsappel S, Veening J-W. 2016. Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. *Genome Biology* **17**: 198.

Badet T, Voisin D, Mbengue M, Barascud M, Sucher J, Sadon P, Balagué C, Roby D, Raffaele S. 2017. Parallel evolution of the POQR prolyl oligo peptidase gene conferring plant quantitative disease resistance. *PLoS Genetics* **13**: e1007143.

Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. 2008. Support vector machines and kernels for computational biology. *PLoS Computational Biology* **4**: e1000173.

Bentham AR, De la Concepcion JC, Mukhi N, Zdrzałek R, Draeger M, Gorenkin D, Hughes RK, Banfield MJ. 2020. A molecular roadmap to the plant immune system. *Journal of Biological Chemistry* **295**: 14916–14935.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**: 474–485.

Bi K, Liang Y, Mengiste T, Sharon A. 2023. Killing softly: a roadmap of *Botrytis cinerea* pathogenicity. *Trends in Plant Science* **28**: 211–222.

Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**: 3045–3046.

Bjornson M, Pimprikar P, Nürnberger T, Zipfel C. 2021. The transcriptional landscape of *Arabidopsis thaliana* pattern-triggered immunity. *Nature Plants* **7**: 579–586.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 2017. *Classification and regression trees*. London, UK: Routledge.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, The AmiGO Hub, The Web Presence Working Group. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**: 288–289.

Castrillo G, Teixeira PJPL, Paredes SH, Law TF, de Lorenzo L, Feltcher ME, Finkel OM, Breakfield NW, Mieczkowski P, Jones CD *et al.* 2017. Root microbiota drive direct integration of phosphate stress and immunity. *Nature* **543**: 513–518.

Chandrika NNP, Sundaravelpandian K, Yu S, Schmidt W. 2013. ALFIN – LIKE 6 is involved in root hair elongation during phosphate deficiency in Arabidopsis. *New Phytologist* **198**: 709–720.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**: 321–357.

Chen TQ, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, 13–17 August 2016, 785–794. [WWW document] URL http://arxiv.org/abs/1603.02754 [accessed 5 October 2022].

Chen X, Ishwaran H. 2012. Random forests for genomic data analysis. *Genomics* **99**: 323–329.

Choo J, Liu S. 2018. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications* **38**: 84–92.

Clarke B, Fokoué E, Zhang HH. 2009. Variability, information, and prediction. In: Clarke B, Fokoue E, Zhang HH, eds. *Principles and theory for data mining and machine learning*. New York, NY, USA: Springer, 1–52.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**: 2938–2940.

Cook DE, Mesarich CH, Thomma BPHJ. 2015. Understanding plant immunity as a surveillance system to detect invasion. *Annual Review of Phytopathology* **53**: 541–563.

Corwin JA, Copeland D, Feusier J, Subedy A, Eshbaugh R, Palmer C, Maloof J, Kliebenstein DJ. 2016. The quantitative basis of the Arabidopsis innate immune system to endemic pathogens depends on pathogen genetics. *PLoS Genetics* **12**: e1005789.

Couto D, Zipfel C. 2016. Regulation of pattern recognition receptor signalling in plants. *Nature Reviews Immunology* **16**: 537–552.

Delplace F, Huard-Chauveau C, Berthomé R, Roby D. 2022. Network organization of the plant immune system: from pathogen perception to robust defense induction. *The Plant Journal* **109**: 447–470.

Denoux C, Galletti R, Mammarella N, Gopalan S, Werck D, De Lorenzo G, Ferrari S, Ausubel FM, Dewdney J. 2008. Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings. *Molecular Plant* **1**: 423–445.

Dindas J, DeFalco TA, Yu G, Zhang L, David P, Bjornson M, Thibaud MC, Custódio V, Castrillo G, Nussaume L *et al.* 2022. Direct inhibition of phosphate transport by immune signaling in Arabidopsis. *Current Biology* **32**: 488–495.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dong X, Jiang Z, Peng Y-L, Zhang Z. 2015. Revealing shared and distinct gene network organization in Arabidopsis immune responses by integrative analysis. *Plant Physiology* **167**: 1186–1203.

Finkers R, van den Berg P, van Berloo R, ten Have A, van Heusden AW, van Kan JAL, Lindhout P. 2007. Three QTLs for *Botrytis cinerea* resistance in tomato. *Theoretical and Applied Genetics* **114**: 585–593.

Ge SX, Jung D, Yao R. 2020. SHINYGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**: 2628–2629.

Ghislain M, Byarugaba AA, Magembe E, Njoroge A, Rivera C, Román ML, Tovar JC, Gamboa S, Forbes GA, Kreuze JF *et al.* 2019. Stacking three late blight resistance genes from wild species directly into African highland potato varieties confers complete field resistance to local blight races. *Plant Biotechnology Journal* **17**: 1119–1129.

Glazebrook J. 2005. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annual Review of Phytopathology* **43**: 205–227.

Greener JG, Kandathil SM, Moffat L, Jones DT. 2021. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology* **23**: 40–55.

Han JW, Kamber M, Pei J. 2012. Data transformation and data discretization. In: *Data mining concepts and techniques, 3rd edn.* Waltham, MA, USA: Morgan Kaufmann Publishers, 111–118.

Hillmer RA, Tsuda K, Rallapalli G, Asai S, Truman W, Papke MD, Sakakibara H, Jones JDG, Myers CL, Katagiri F. 2017. The highly buffered Arabidopsis immune signaling network conceals the functions of its components. *PLoS Genetics* **13**: e1006639.

Ho TK. 1995. Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition.* Montreal, QC, Canada: IEEE Computer Society Press, 278–282. [WWW document] URL http://ieeexplore.ieee.org/document/598994/ [accessed 5 October 2022].

Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering* **9**: 90–95.

Jones JDG, Dangl JL. 2006. The plant immune system. *Nature* **444**: 323–329.

Jubic LM, Saile S, Furzer OJ, El Kasmi F, Dangl JL. 2019. Help wanted: helper NLRs and plant immune responses. *Current Opinion in Plant Biology* **50**: 82–94.

Katagiri F. 2018. Review: plant immune signaling from a network perspective. *Plant Science* **276**: 14–21.

Kazan K, Lyons R. 2014. Intervention of phytohormone pathways by pathogen effectors. *Plant Cell* **26**: 2285–2309.

Khan GA, Vogiatzaki E, Glauser G, Poirier Y. 2016. Phosphate deficiency induces the jasmonate pathway and enhances resistance to insect herbivory. *Plant Physiology* **171**: 632–644.

Kim Y, Tsuda K, Igarashi D, Hillmer RA, Sakakibara H, Myers CL, Katagiri F. 2014. Mechanisms underlying robustness and tunability in a plant immune signaling network. *Cell Host & Microbe* **15**: 84–94.

Kiyosawa S. 1982. Genetics and epidemiological modeling of breakdown of plant disease resistance. *Annual Review of Phytopathology* **20**: 93–117.

Koo PK, Ploenzke M. 2021. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence* **3**: 258–266.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521**: 436–444.

Li B, Meng X, Shan L, He P. 2016. Transcriptional regulation of pattern-triggered immunity in plants. *Cell Host & Microbe* **19**: 641–650.

Liebrand TWH, van den Berg GCM, Zhang Z, Smit P, Cordewener JHG, America AHP, Sklenar J, Jones AME, Tameling WIL, Robatzek S *et al.* 2013. Receptor-like kinase SOBIR1/EVR interacts with receptor-like proteins in plant immunity against fungal infection. *Proceedings of the National Academy of Sciences, USA* **110**: 10010–10015.

Liu Z, Hou S, Rodrigues O, Wang P, Luo D, Munemasa S, Lei J, Liu J, Ortiz-Morea FA, Wang X *et al.* 2022. Phytocytokine signalling reopens stomata in plant immunity and water loss. *Nature* **605**: 332–339.

Luo M, Xie L, Chakraborty S, Wang A, Matny O, Jugovich M, Kolmer JA, Richardson T, Bhatt D, Hoque M *et al.* 2021. A five-transgene cassette confers broad-spectrum resistance to a fungal rust pathogen in wheat. *Nature Biotechnology* **39**: 561–566.

Ma S, Lapin D, Liu L, Sun Y, Song W, Zhang X, Logemann E, Yu D, Wang J, Jirschitzka J *et al.* 2020. Direct pathogen-induced assembly of an NLR immune receptor complex to form a holoenzyme. *Science* **370**: eabe3069.

Machado Wood AK, Panwar V, Grimwade-Mann M, Ashfield T, Hammond-Kosack KE, Kanyuka K. 2021. The vesicular trafficking system component MIN7 is required for minimizing *Fusarium graminearum* infection. *Journal of Experimental Botany* **72**: 5010–5023.

Marchal C, Pai H, Kamoun S, Kourelis J. 2022. Emerging principles in the design of bioengineered made-to-order plant immune receptors. *Current Opinion in Plant Biology* **70**: 102311.

Martin R, Qi T, Zhang H, Liu F, King M, Toth C, Nogales E, Staskawicz BJ. 2020. Structure of the activated ROQ1 resistosome directly recognizing the pathogen effector XopQ. *Science* **370**: eabd9993.

Mishra B, Kumar N, Mukhtar MS. 2019. Systems biology and machine learning in plant–pathogen interactions. *Molecular Plant–Microbe Interactions* **32**: 45–55.

Mou Z, Fan W, Dong X. 2003. Inducers of plant systemic acquired resistance regulate NPR1 function through redox changes. *Cell* **113**: 935–944.

Ngou BPM, Ahn H-K, Ding P, Jones JDG. 2021. Mutual potentiation of plant immunity by cell-surface and intracellular receptors. *Nature* **592**: 110–115.

Ngou BPM, Ding P, Jones JDG. 2022. Thirty years of resistance: zig-zag through the plant immune system. *Plant Cell* **34**: 1447–1478.

Noble WS. 2006. What is a support vector machine? *Nature Biotechnology* **24**: 1565–1567.

Nobori T, Velásquez AC, Wu J, Kvitko BH, Kremer JM, Wang Y, He SY, Tsuda K. 2018. Transcriptome landscape of a bacterial pathogen under plant immunity. *Proceedings of the National Academy of Sciences, USA* **115**: E3055–E3064.

Nomura K, Debroy S, Lee YH, Pumplin N, Jones J, He SY. 2006. A bacterial virulence protein suppresses host innate immunity to cause plant disease. *Science* **313**: 220–223.

Nomura K, Mecey C, Lee Y-N, Imboden LA, Chang JH, He SY. 2011. Effector-triggered immunity blocks pathogen degradation of an immunity-associated vesicle traffic regulator in Arabidopsis. *Proceedings of the National Academy of Sciences, USA* **108**: 10774–10779.

O'Keeffe KR, Jones CD. 2019. Challenges and solutions for analysing dual RNA-seq data for non-model host–pathogen systems. *Methods in Ecology and Evolution* **10**: 401–414.

Pavlopoulos GA, Kontou PI, Pavlopoulou A, Bouyioukos C, Markou E, Bagos PG. 2018. Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* **7**: giy014.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al.* 2011. SCIKIT-LEARN: machine learning in PYTHON. *Journal of Machine Learning Research* **12**: 2825–2830.

Pieterse CMJ, Van der Does D, Zamioudis C, Leon-Reyes A, Van Wees SCM. 2012. Hormonal modulation of plant immunity. *Annual Review of Cell and Developmental Biology* **28**: 489–521.

Pradhan SK, Nayak DK, Mohanty S, Behera L, Barik SR, Pandit E, Lenka S, Anandan A. 2015. Pyramiding of three bacterial blight resistance genes for broad-spectrum resistance in deepwater rice variety, Jalmagna. *Rice* **8**: 19.

Shi J, Zhao B, Zheng S, Zhang X, Wang X, Dong W, Xie Q, Wang G, Xiao Y, Chen F *et al.* 2021. A phosphate starvation response-centered network regulates mycorrhizal symbiosis. *Cell* **184**: 5527–5540.

Shigenaga AM, Berens ML, Tsuda K, Argueso CT. 2017. Towards engineering of hormonal crosstalk in plant immunity. *Current Opinion in Plant Biology* **38**: 164–172.

Soltis NE, Atwell S, Shi G, Fordyce R, Gwinner R, Gao D, Shafi A, Kliebenstein DJ. 2019. Interactions of tomato and *Botrytis cinerea* genetic diversity: parsing the contributions of host differentiation, domestication, and pathogen variation. *Plant Cell* **31**: 502–519.

Spoel SH, Koornneef A, Claessens SMC, Korzelius JP, Van Pelt JA, Mueller MJ, Buchala AJ, Métraux JP, Brown R, Kazan K *et al.* 2003. NPR1 modulates cross-talk between salicylate- and jasmonate-dependent defense pathways through a novel function in the cytosol. *Plant Cell* **15**: 760–770.

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**: 25.

Tanaka H, Kitakura S, De Rycke R, De Groodt R, Friml J. 2009. Fluorescence imaging-based screen identifies ARF GEF component of early endosomal trafficking. *Current Biology* **19**: 391–397.

Team T Pandas Development. 2024. pandas-dev/pandas: Pandas. *Zenodo*. doi: 10.5281/zenodo.10957263.

The Arabidopsis Information Resource (TAIR). 2023. The Arabidopsis Information Resource. [WWW document] URL https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGO_and_PO_Annotations%2FGene_Ontology_Annotations [accessed 24 July 2022].

The Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL *et al.* 2023. The Gene Ontology knowledgebase in 2023. *Genetics* **224**: iyad031.

Thines B, Katsir L, Melotto M, Niu Y, Mandaokar A, Liu G, Nomura K, He SY, Howe GA, Browse J. 2007. JAZ repressor proteins are targets of the SCFCOI1 complex during jasmonate signalling. *Nature* **448**: 661–665.

Ticconi CA, Lucero RD, Sakhonwasee S, Adamson AW, Creff A, Nussaume L, Desnos T, Abel S. 2009. ER-resident proteins PDR2 and LPR1 mediate the developmental response of root meristems to phosphate availability. *Proceedings of the National Academy of Sciences, USA* **106**: 14174–14179.

Tintor N, Ross A, Kanehara K, Yamada K, Fan L, Kemmerling B, Nürnberger T, Tsuda K, Saijo Y. 2013. Layered pattern receptor signaling via ethylene and endogenous elicitor peptides during Arabidopsis immunity to bacterial infection. *Proceedings of the National Academy of Sciences, USA* **110**: 6211–6216.

Tsuda K, Sato M, Stoddard T, Glazebrook J, Katagiri F. 2009. Network properties of robust immunity in plants. *PLoS Genetics* **5**: e1000772.

Tsuda K, Somssich IE. 2015. Transcriptional networks in plant immunity. *New Phytologist* **206**: 932–947.

Van Kan JAL, Stassen JHM, Mosbach A, Van Der Lee TAJ, Faino L, Farmer AD, Papasotiriou DG, Zhou S, Seidl MF, Cottam E *et al.* 2017. A gapless genome sequence of the fungus *Botrytis cinerea*. *Molecular Plant Pathology* **18**: 75–89.

Vélez-Bermúdez IC, Schmidt W. 2021. Chromatin enrichment for proteomics in plants (ChEP-P) implicates the histone reader ALFIN-LIKE 6 in jasmonate signalling. *BMC Genomics* **22**: 845.

Vuong UT, Iswanto ABB, Nguyen Q-M, Kang H, Lee J, Moon J, Kim SH. 2023. Engineering plant immune circuit: walking to the bright future with a novel toolbox. *Plant Biotechnology Journal* **21**: 17–45.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* **131**: 281–285.

Wan W-L, Fröhlich K, Pruitt RN, Nürnberger T, Zhang L. 2019. Plant cell surface immune receptor complex signaling. *Current Opinion in Plant Biology* **50**: 18–28.

Wang J, Hu M, Wang J, Qi J, Han Z, Wang G, Qi Y, Wang HW, Zhou JM, Chai J. 2019. Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science* **364**: eaav5870.

Wang P, Fan E, Wang P. 2021. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters* **141**: 61–67.

Wang S, Xue M, He C, Shen D, Jiang C, Zhao H, Niu D. 2021. AtMC1 associates with LSM4 to regulate plant immunity through modulating pre-mRNA splicing. *Molecular Plant–Microbe Interactions* **34**: 1423–1432.

Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FSL. 2016. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. *Nucleic Acids Research* **44**: D646–D653.

Wu Y, Zhang D, Chu JY, Boyle P, Wang Y, Brindle ID, de Luca V, Després C. 2012. The Arabidopsis NPR1 protein is a receptor for the plant defense hormone salicylic acid. *Cell Reports* **1**: 639–647.

Xiao X, Chen H, Bogdan P. 2021. Deciphering the generating rules and functionalities of complex networks. *Scientific Reports* **11**: 22964.

Xie B, Wang X, Zhu M, Zhang Z, Hong Z. 2011. *CalS7* encodes a callose synthase responsible for callose deposition in the phloem. *The Plant Journal* **65**: 1–14.

Xu C, Jackson SA. 2019. Machine learning and complex biological data. *Genome Biology* **20**: 76.

Xu L, Liu F, Lechner E, Genschik P, Crosby WL, Ma H, Peng W, Huang D, Xie D. 2002. The SCFCOI1 ubiquitin-ligase complexes are required for jasmonate response in Arabidopsis. *Plant Cell* **14**: 1919–1935.

Yuan M, Jiang Z, Bi G, Nomura K, Liu M, Wang Y, Cai B, Zhou JM, He SY, Xin XF. 2021. Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature* **592**: 105–109.

Zhang W, Corwin JA, Copeland D, Feusier J, Eshbaugh R, Chen F, Atwell S, Kliebenstein DJ. 2017. Plastic transcriptomes stabilize immunity to pathogen diversity: the jasmonic acid and salicylic acid networks within the Arabidopsis/Botrytis pathosystem. *Plant Cell* **29**: 2727–2752.

Zhang W, Corwin JA, Copeland DH, Feusier J, Eshbaugh R, Cook DE, Atwell S, Kliebenstein DJ. 2019. Plant–necrotroph co-transcriptome networks illuminate a metabolic battlefield. *eLife* **8**: e44279.

Zhang Z, Zhang S, Zhang Y, Wang X, Li D, Li Q, Yue M, Li Q, Zhang YE, Xu Y *et al.* 2011. Arabidopsis floral initiator SKB1 confers high salt tolerance by regulating transcription and pre-mRNA splicing through altering histone H4R3 and small nuclear ribonucleoprotein LSM4 methylation. *Plant Cell* **23**: 396–411.

Zhao Z, Yang X, Lü S, Fan J, Opiyo S, Yang P, Mangold J, Mackey D, Xia Y. 2020. Deciphering the novel role of AtMIN7 in cuticle formation and defense against the bacterial pathogen infection. *International Journal of Molecular Sciences* **21**: 5547.

Zipfel C, Robatzek S, Navarro L, Oakeley EJ, Jones JDG, Felix G, Boller T. 2004. Bacterial disease resistance in Arabidopsis through flagellin perception. *Nature* **428**: 764–767.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Distribution of training and test splits for the standardized lesion class data from *Arabidopsis–Botrytis* interaction.

Fig. S2 Distribution of bacteria growth derived from *Arabidopsis–Pseudomonas syringae* interaction.

Fig. S3 Regression analysis of lesion data.

Fig. S4 Heatmap of the observed vs predicted disease classes by ML model for host–pathogen transcriptome data.

Fig. S5 Distribution of lesion sizes during *Arabidopsis–Botrytis* interaction.

Fig. S6 Performance metrics for test label shuffling.

Fig. S7 Performance analysis with varying percentages of mislabeled classes.

Fig. S8 Evaluation of ML models based on host, fungal or dual-species transcriptomes derived from *Arabidopsis–Botrytis* interaction.

Fig. S9 Changing the size of the feature set impacts ML performance.

Fig. S10 Evaluation of set size and feature selection methods impacting prediction performance during *Arabidopsis–Botrytis* interaction.

Fig. S11 Set size and feature selection threshold selection.

Fig. S12 Assessment of ML models and feature selection methods on a new dataset collected from *Arabidopsis–Sclerotinia sclerotiorum* interaction.

Fig. S13 Transcript quantification for feature selection sets.

Fig. S14 Biological process GO enrichment and term relatedness for Betweenness 500.

Fig. S15 Biological process GO enrichment and term relatedness for Degree 500.

Fig. S16 Biological process GO enrichment and term relatedness for NFD 3000.

Fig. S17 Biological process GO enrichment and term relatedness for FDC 3000.

Fig. S18 Biological process GO enrichment and term relatedness for RF 100.

Fig. S19 Gene overlap for top proportionally ranked GO term, regulation of biological quality, for three feature selection sets.

Table S1 Gene lists derived from 12 feature selection methods.

Table S2 Results from ANOVA for predicted disease class errors of *Arabidopsis–Botrytis* infection.

Table S3 One-way ANOVA for evaluation parameters by transcriptome data types of *Arabidopsis–Botrytis* infection.

Table S4 One-class error prediction metrics for naïve, shuffled, and five ML models.

Table S5 Two-way ANOVA on evaluation parameters by effects of ML models and feature size of *Arabidopsis–Botrytis* infection.

Table S6 Two-way ANOVA for performance parameters by effects of feature selection methods and ML models on of *Arabidopsis–Botrytis* infection.

Table S7 Results ANOVA for predicted disease class error of *Arabidopsis thaliana–Sclerotinia sclerotiorum* infection.

Table S8 Results ANOVA for predicted disease class error of *Arabidopsis thaliana–Pseudomonas syringae* infection.

Table S9 Chi-square results for GWAS gene set with each of the feature selection sets.

Table S10 Chi-square results for Bjornson identified gene set with each of the feature selection sets.

Table S11 Gene ontology biological function enrichment for NFD 3000 feature selection set.

Table S12 Gene ontology biological function enrichment for Betweenness 500 feature selection set.

Table S13 Gene ontology biological function enrichment for Degree 500 feature selection set.

Table S14 Gene ontology biological function enrichment for RF 100 feature selection set.

Table S15 Gene ontology biological function enrichment for FDC 3000 feature selection set.

Table S16 High-level GO biological function proportional rank across all seven feature selection sets.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.