Phylogeny-agnostic strain-level prediction of phage-host interactions from genomes

Avery J. C. Noonan[1], Lucas Moriniere[1,2], Edwin O. Rivera-López[3], Krish Patel[4], Melina Pena[1], Madeline Svab[1,5,6], Alexey Kazakov[1], Adam Deutschbauer[1,6], Edward G. Dudley[3,7], Vivek K. Mutalik[2] & Adam P. Arkin[1,2,4]

[1] Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2] California Institute for Quantitative Biosciences, University of California Berkeley, CA, USA
[3] Department of Food Science, The Pennsylvania State University, University Park, PA, USA
[4] Department of Bioengineering, University of California Berkeley, CA, USA
[5] Department of Molecular and Cell Biology, University of California Berkeley, CA, USA
[6] Department of Plant and Microbial Biology, University of California Berkeley, CA, USA
[7] *E. coli* Reference Center, The Pennsylvania State University, University Park, PA, USA

*Author to whom correspondence should be addressed; aparkin@lbl.gov or vkmutalik@lbl.gov

1 Supplementary Information

1.1 Extended Results

1.1.1 Dataset overview

The *E. coli* interaction matrix was assayed on solid medium, where susceptibility was qualitatively assessed based on the clearance achieved by three phage dilutions, whereas all other datasets were assayed in liquid medium and relied on the area under the optical density curve to quantify susceptibility. Positive interactions, or those resulting in infection, were defined as any non-zero interaction in the *E. coli* dataset and based on criteria specified by the publishing authors for *Klebsiella*, *Pseudomonas*, and *Vibrionaceae* datasets.

Phylogenetic trees based on amino acid alignments of conserved genes were generated for each dataset [71, 72]. These show host-strain diversity within individual datasets and shared phylogenetic clades between the two Klebsiella datasets, which are the only datasets covering the same host phylogenies (*Supp. Fig. 2-6*). A phage gene-sharing network representing all datasets showed that phages across datasets and host phylogenies were interspersed within phage clusters (*Supp. Fig. 7-8*) [18]. This suggested that phages with similar gene content, possibly representing structural or mechanistic similarities, are capable of infecting diverse hosts. These results suggest that gene content-based features could likely be shared across datasets and that pooling of datasets may improve model performance.

1.1.2 Workflow development and optimization

To generate phylogeny-independent numerical features representing strain and phage genomes, we used binary presence/absence matrices of either protein families, in a pangenome-like structure, or amino acid *k*-mers, allowing us to capture sequence variants and key residues. These feature generation workflows were run independently, with whole strain and phage proteomes represented as protein families or *k*-mers, or sequentially, representing only proteins belonging to predictive protein families as *k*-mers. We tested MCL and MMSeqs2 clustering algorithms for protein-family construction, including varied inflation values (the MCL argument controlling clustering stringency) and MMSeqs2 sequence identity and coverage thresholds (*Supp. Fig. 9-10*) [51, 52]. We then compared eight machine learning algorithms across four development datasets, identifying CatBoost gradient-boosted decision trees as the most consistently high-performing model (*Supp. Fig. 12-13*). We implemented recursive feature elimination (RFE) to identify predictive subsets of feature tables of up to 20,000 protein family-based features or 100,000 *k*-mer features and benchmarked this strategy against five alternative feature selection methods

(*Supp. Fig. 14*). Feature selection was chosen over dimension reduction to enable both simplified feature assignment to novel strains and biological interpretation of predictive features.

We tested whether performance was improved by filtering phylogenetically linked features prior to feature selection, decreasing the risk of overfitting and increasing the likelihood of identifying features linked to mechanism, rather than phylogeny. To accomplish this, strains and phages are clustered based on feature content (hierarchical or HDBSCAN clustering [57]) and features unique to a single cluster were removed. This typically removed less than 10% of features, while improving model generalizability. We also explored the impact of iterative feature selection and modeling across various train-test splits, selecting features based on recurrence across feature selection iterations and using an ensemble-learning approach for final predictions (*Supp. Fig. 15-19*). Train-test splitting based on strain or phage clusters was also tested, forcing models to learn features predictive of interaction in distinct phylogenetic clades. This also prevents very similar strains from being present in training and testing datasets, limiting overfitting during feature selection and model training.

Finally, we tested the impact of *k*-length on model performance for *k*-mer-based feature tables. Large feature sets with *k*-mer-based genome representations (20^k possible AA *k*-mer sequences) limited full proteome *k* values to 3-4, resulting in up to 160,000 features. Values of *k* from 3-15 were tested when filtering to predictive proteins only, which significantly decreased *k*-mer diversity enabling larger *k* values.

1.1.3 Modeling Error Analysis

First, an analysis of overall model bias revealed consistent over-prediction bias, with 58 strains showing systematic over-prediction (>20% deviation from observed infection rates) compared to only 11 under-predictors. This bias correlated negatively with strain susceptibility ($r = -0.207$, $p = 8.3 \times 10^{-5}$), being most pronounced in less susceptible strains, consistent with the positive correlation between strain susceptibility and model performance ($r = 0.310$, $p = 2.3 \times 10^{-9}$) (*Supp. Fig. 29-30*). This indicates that a larger proportion of interactions are falsely predicted to be infectious in strains with narrow susceptibility profiles. Several factors likely contribute to this behavior. First, by nature of their narrow susceptibility, the dataset includes few phages capable of infecting similar strains, preventing models from learning relevant feature sets. Whether these strains are truly only suspectable to a small set of phages, possibly through extensive defense mechanisms or distinct receptors variants, cannot be distinguished from the possibility that the tested phage set is not representative of appropriate phage diversity for these strains. This highlights the importance of including a phylogenetically broad set of phages in experimental datasets.

To explore this further, we investigated the impact of phylogenetic distribution of strains and phages on prediction accuracy. For all strain genomes, we constructed phylogenetic trees based on core-genome concatenated marker genes and examined how prediction performance varied across phylogenetic space (*Supp. Fig. 2-6*). For each strain, we calculated the mean distance to its five closest phylogenetic neighbors as a measure of phylogenetic isolation. Surprisingly, we found no consistent relationship between phylogenetic isolation and strain-level prediction performance (MCC) ($r = -0.055$, $p = 0.60$), indicating that the presence of closely related strains in the training data does not necessarily improve prediction accuracy (*Supp. Fig. 29*). This further supports our observations when modeling combined *Klebsiella* datasets, suggesting that mechanisms of phage susceptibility are not directly linked to phylogeny and are not necessarily vertically inherited. In contrast to this, a significant negative correlation was observed between phage performance and phylogenetic isolation ($r = -0.425$, $p = 1.9 \times 10^{-5}$), where proteomic equivalence was used as a metric of phage genomic similarity (*Supp. Fig. 30*). This relationship indicates that models performed less well when predicting infection by more distinct phages. This highlights the importance of phage diversity in interaction datasets and the possible performance benefit of phage phylogenetic redundancy.

To systematically quantify where models fail to recognize distinguishing genetic features dictating strain infection patterns, we calculated pairwise Jaccard similarities between strain pairs and phage pairs based on their experimental interaction profiles and compared these to model-predicted interaction similarities across 66,795 strain pairs and 4,371 phage pairs. This calculation evaluates whether strains pairs and phages pairs with similar interaction profiles are predicted to behave similarly. The result is a sensitive metric of strain- or phage-level predictive performance, where a one-to-one relationship indicates perfect accordance and a large distance from this line indicates discordance and the inability of the model to capture critical information. Given our strain-based cross-validation design, where all phages appear in both training and test sets, phage-phage similarity correlations were expectedly high ($r = 0.893$, $p < 1 \times 10^{-12}$), indicating models effectively capture phage-specific infectivity patterns (*Supp. Fig. 32C-D*). In contrast, strain-strain similarity correlations were moderate ($r = 0.387$ ($p < 1 \times 10^{-12}$) for all strains, $r = 0.493$ ($p < 1 \times 10^{-12}$) for strains infected by ≥20 phages) (*Supp. Fig. 32A-B*). The improved correlation for broadly susceptible strains matches the observation that models better capture similarity patterns among these strains. Analysis of extreme discordance cases, where shared interactions differed significantly from shared predicted interactions, revealed fundamental model failures in capturing relevant biological information. We identified strain pairs where models predicted nearly identical interaction patterns (similarity > 0.9) despite minimal biological overlap (true similarity < 0.3), exemplified by *E. coli* BL21 comparisons to *E. coli* strains ECOR22, ECOR42 and ECOR54. This indicates the model's failure to capture BL21's distinct phage susceptibility patterns, which result from its laboratory-adapted genetic background including truncated

lipopolysaccharide and altered outer membrane protein composition compared to wild-type strains. Conversely, strain pairs like *E. coli* IAI45 compared to *E. coli* strains ECOR57, ECOR59, and ECOR71 showed high biological similarity (>0.7) but low predicted similarity (<0.3), indicating failure to recognize shared mediators of infection. This information may help identify strains and phages with abnormal interaction patterns, possibly uncovering novel mediators or sequence variants of biological interest.

2 Supplementary Tables

Supplementary Table 1. Model performance across datasets when predicting infection of known strains by unseen phages.

Dataset	AUROC	MCC	Normalized AUPR	Brier score
Pseudomonas	0.925	0.682	0.869	0.115
Vibrionaceae	0.924	0.487	0.659	0.030
E. coli	0.870	0.510	0.563	0.146
Klebsiella-2	0.814	0.294	0.312	0.037
Klebsiella-1	0.687	0.178	0.102	0.088

Supplementary Table 2. Model performance across datasets when predicting interactions between unseen strain-phage pairs.

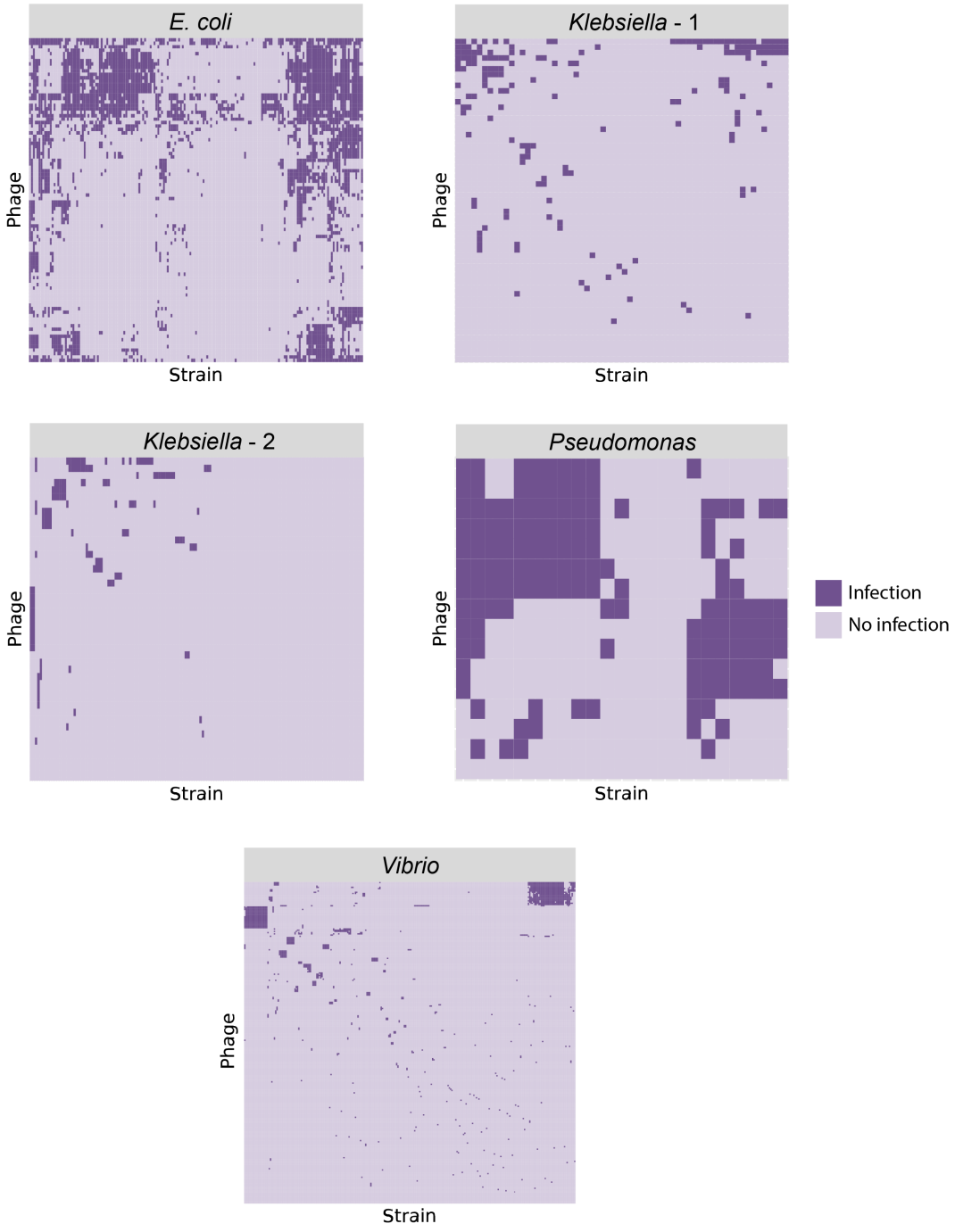
Dataset	AUROC	MCC	Normalized AUPR	Brier score
Vibrionaceae	0.920	0.436	0.515	0.025
E. coli	0.812	0.410	0.381	0.134
Pseudomonas	0.766	0.283	0.576	0.238
Klebsiella-2	0.675	0.194	0.147	0.022
Klebsiella-1	0.577	0.068	0.032	0.064

Supplementary Table 3. Phage list used in experimental validation.

Phage ID	Full Phage Name	Genome Accession	Genome Size (bp)	Interaction matrix	RB-TnSeq
T2	Escherichia phage T2	MH751506.1	163832		x
P1vir	Escherichia phage P1vir	NC_005856.1	94800		x
phi92	Escherichia phage phi92	FR775895.2	148612		x
Bas01	Escherichia phage AugustePiccard	MZ501051	50126	x	
Bas02	Escherichia phage JeanPiccard	MZ501080	47149	x	
Bas03	Escherichia phage JulesPiccard	MZ501087	47731	x	
Bas04	Escherichia phage FritzSarasin	MZ501069	51777	x	
Bas05	Escherichia phage PeterMerian	MZ501101	47775	x	x
Bas06	Escherichia phage KarlJaspers	MZ501090	51474	x	
Bas07	Escherichia phage JakobBernoulli	MZ501079	51129	x	x
Bas08	Escherichia phage DanielBernoulli	MZ501059	50731	x	
Bas09	Escherichia phage PaulSarasin	MZ501099	51285	x	x
Bas10	Escherichia phage IsaakIselin	MZ501077	49911	x	x
Bas12	Escherichia phage BrunoManser	MZ501053	49298	x	
Bas13	Escherichia phage LeonhardEuler	MZ501092	50192	x	
Bas14	Escherichia phage TheodorHerzl	MZ501107	43627	x	
Bas15	Escherichia phage PaulFeyerabend	MZ501097	44550	x	
Bas16	Escherichia phage GeorgBuechner	MZ501070	44295	x	
Bas17	Escherichia phage KarlBarth	MZ501088	43104	x	
Bas18	Escherichia phage Oekolampad	MZ501095	44882		x
Bas19	Escherichia phage ChristophMerian	MZ501057	58073	x	x
Bas20	Escherichia phage FritzHoffmann	MZ501068	59834	x	
Bas21	Escherichia phage GottfriedDienst	MZ501071	56589	x	x
Bas22	Escherichia phage KurtStettler	MZ501091	58567	x	
Bas25	Escherichia phage VogelGryff	MZ501110	58342	x	x
Bas27	Escherichia phage TrudiGerster	MZ501108	114179	x	
Bas28	Escherichia phage IrmaTschudi	MZ501076	115446	x	
Bas29	Escherichia phage SuperGirl	MZ501105	110821	x	
Bas31	Escherichia phage DaisyDussoix	MZ501058	113532	x	
Bas33	Escherichia phage HildyBeyeler	MZ501074	111607	x	
Bas34	Escherichia phage SelmaRatti	MZ501103	107080	x	

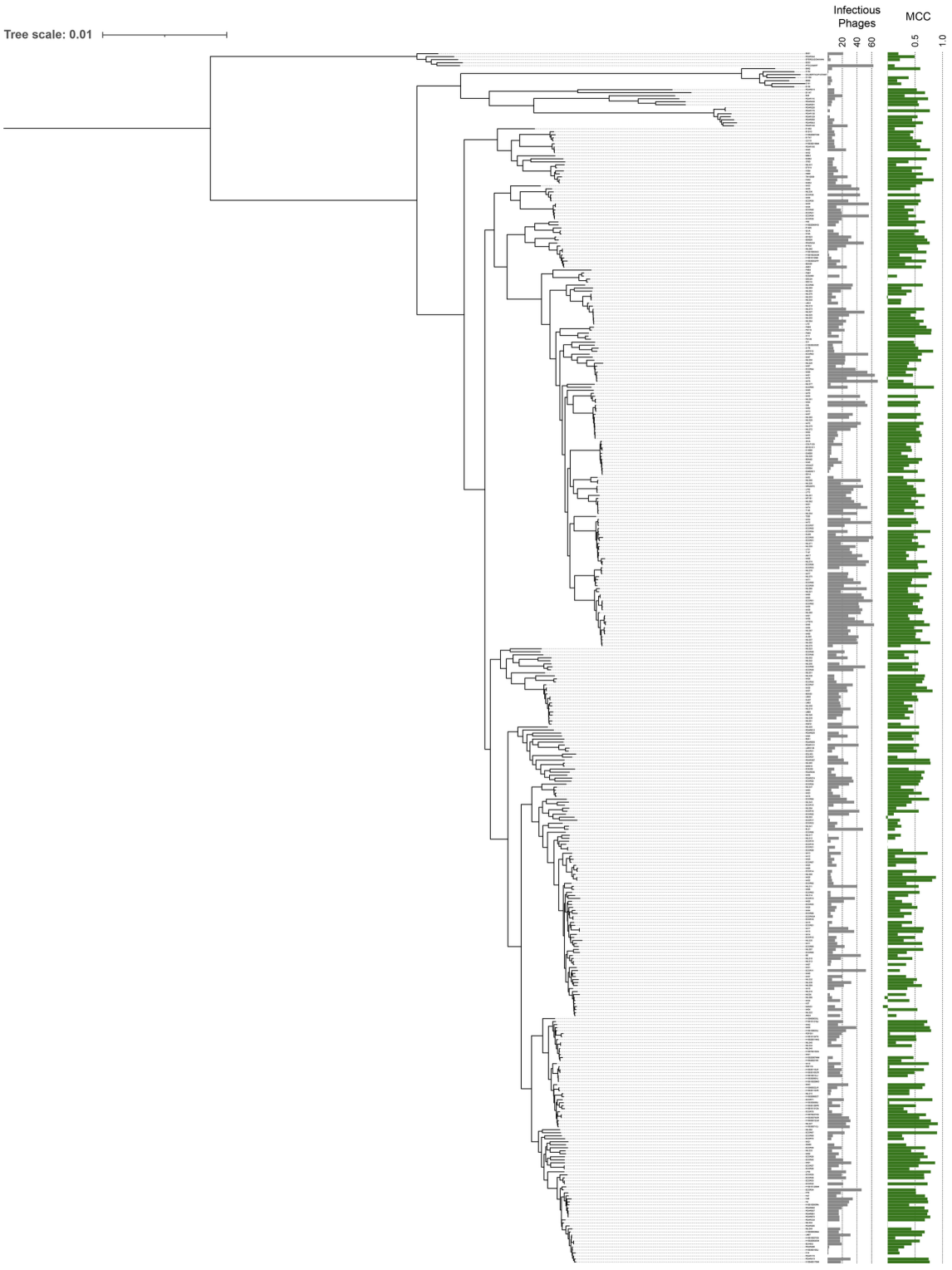
Bas35	Escherichia phage WilhelmHis	MZ501113	166861	x	
Bas36	Escherichia phage Paracelsus	MZ501096	167732	x	
Bas37	Escherichia phage KarlGJung	MZ501089	167832	x	x
Bas38	Escherichia phage AugustSocin	MZ501052	167411	x	x
Bas39	Escherichia phage FriedrichMiescher	MZ501066	169509	x	x
Bas41	Escherichia phage FriedrichZschokke	MZ501067	165574	x	
Bas42	Escherichia phage AndreasVesalius	MZ501050	168255	x	
Bas43	Escherichia phage TadeuszReichstein	MZ501106	166115	x	
Bas44	Escherichia phage AdolfPortmann	MZ501046	166773	x	
Bas45	Escherichia phage PaulHMueller	MZ501098	170053	x	x
Bas46	Escherichia phage ChristianSchoenbein	MZ501056	169031	x	
Bas47	Escherichia phage AlbertHofmann	MZ501047	168968	x	
Bas48	Escherichia phage CarlMeissner	MZ501054	137623	x	x
Bas49	Escherichia phage EmilHeitz	MZ501062	139980	x	
Bas51	Escherichia phage WalterGehring	MZ501111	140659	x	
Bas52	Escherichia phage RudolfGeigy	MZ501102	137380	x	
Bas54	Escherichia phage MaxBurger	MZ501093	136346	x	
Bas56	Escherichia phage AlexBoehm	MZ501048	135345	x	
Bas57	Escherichia phage MaxTheCat	MZ501094	135061	x	
Bas58	Escherichia phage HeinrichReichert	MZ501073	136906	x	
Bas59	Escherichia phage EduardKellenberger	MZ501061	131526	x	
Bas60	Escherichia phage PaulScherrer	MZ501100	150425	x	
Bas61	Escherichia phage EmilieFrey	MZ501063	146666	x	x
Bas63	Escherichia phage JohannRWettstein	MZ501086	87100	x	x
Bas64	Escherichia phage JeanTinguely	MZ501081	39451	x	
Bas65	Escherichia phage JacobBurckhardt	MZ501078	39451	x	
Bas67	Escherichia phage ErnstBeyeler	MZ501064	39315	x	x
Bas68	Escherichia phage CarlSpitteler	MZ501055	39466	x	x
Bas69	Escherichia phage AlfredRasser	MZ501049	70849	x	x

3 Supplementary Figures

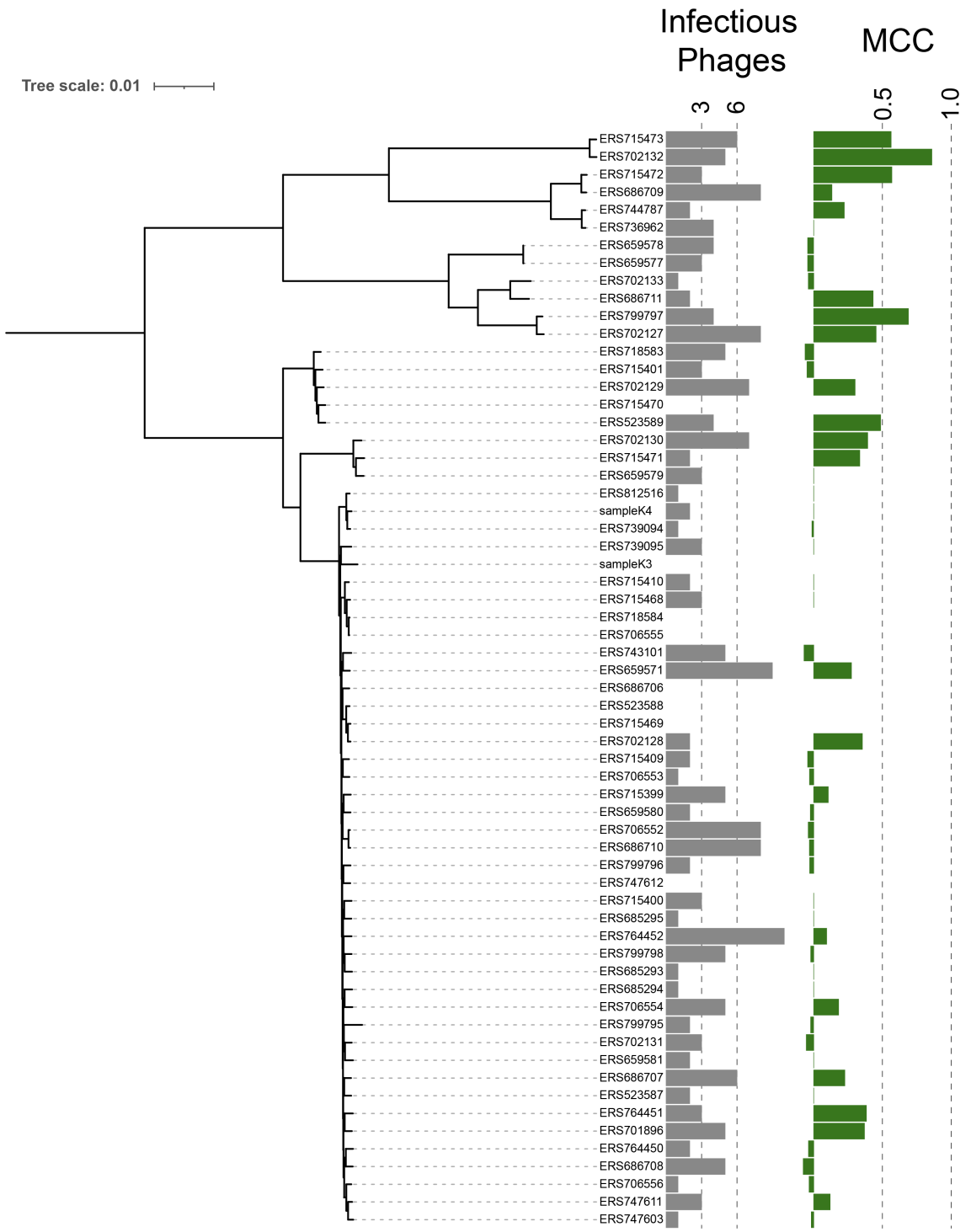


Supplementary Figure 1. Phage-host interaction matrices. The interaction matrices for five published datasets. Clusters of strains and phages indicate conserved infection patterns within these groups. Dark purple represents interactions resulting in infection and light purple indicates interaction that did not result in infection.

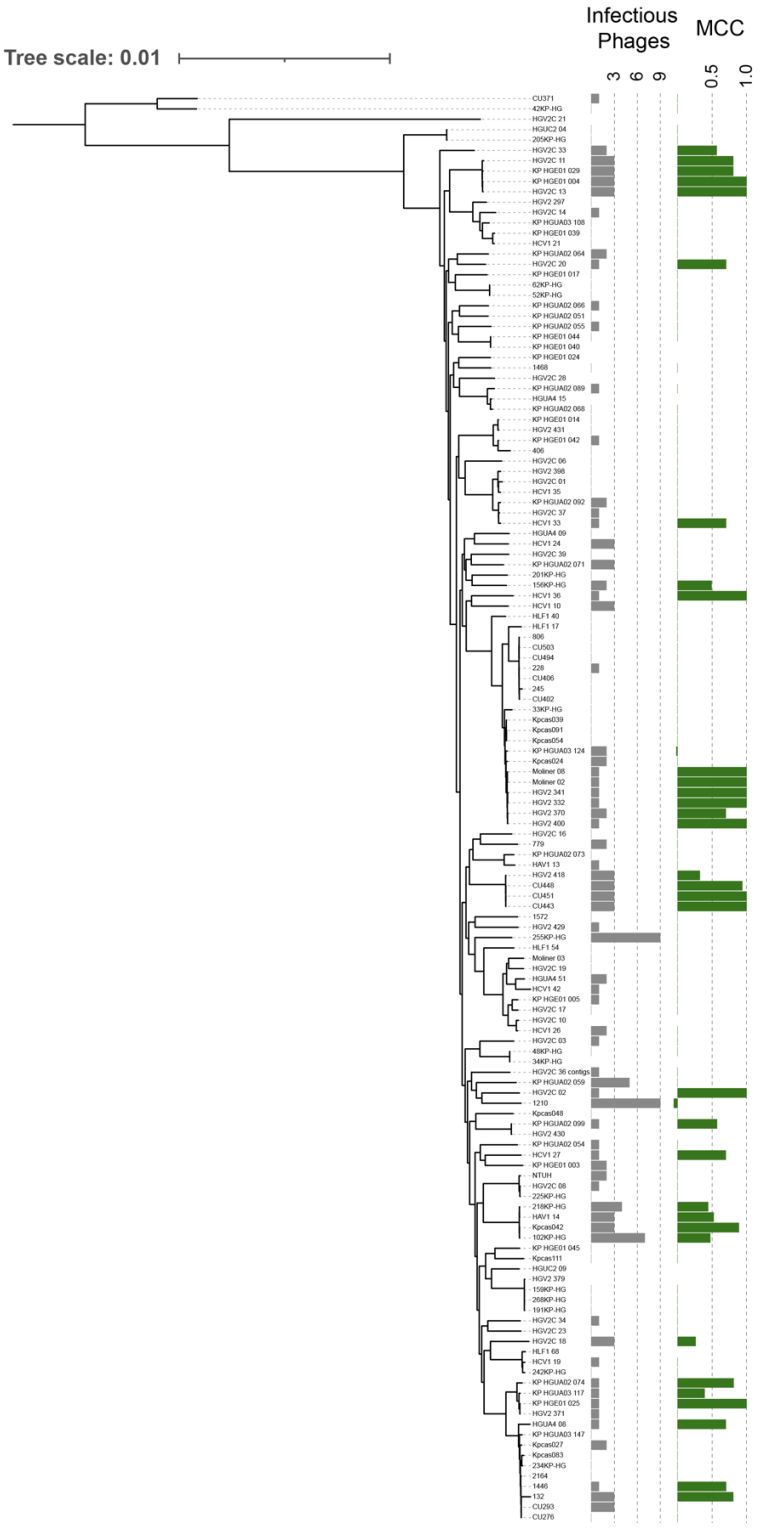
Supplementary Figure 2. Klebsiella dataset tree. Phylogenetic tree based on core-genome concatenated marker genes alignments. Dark orange indicates mixed *Klebsiella* spp. dataset and light orange indicates *Klebsiella pneumoniae* dataset.



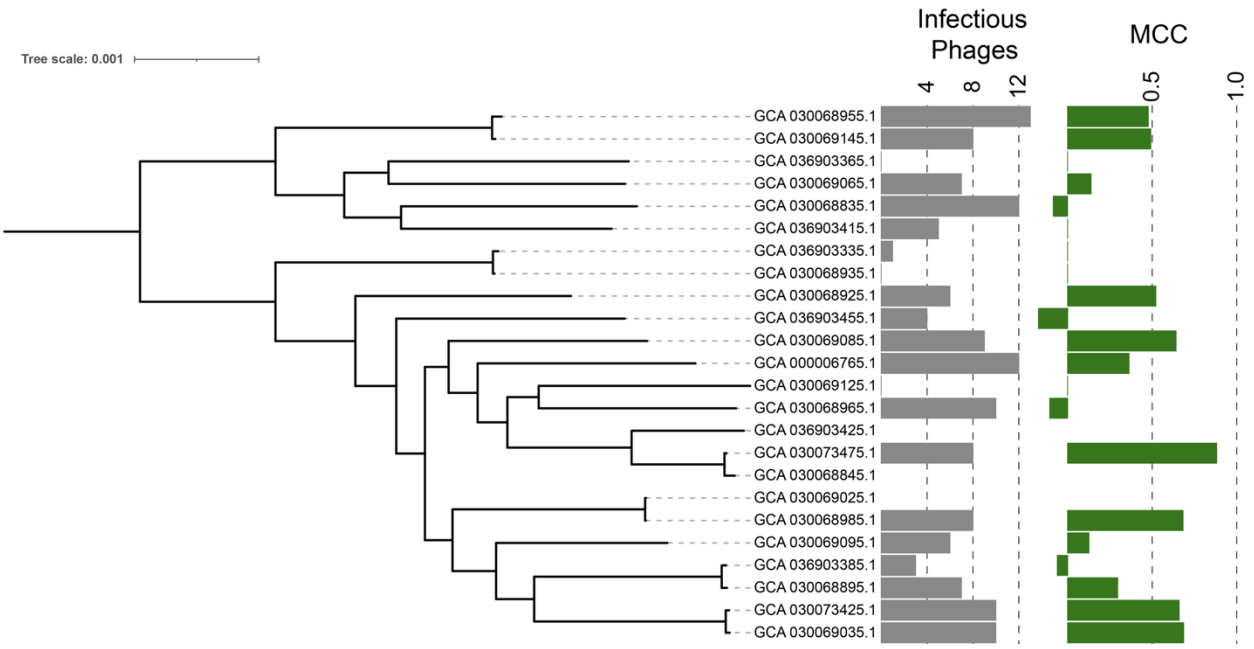
Supplementary Figure 3. E. coli dataset tree. Phylogenetic tree based on core-genome concatenated marker genes alignments. Grey bars indicate number of infectious phages out of 94-phage set. Green bars indicate prediction performance in 20-fold cross-validation for models with strain-based train-test splitting (unseen strains).



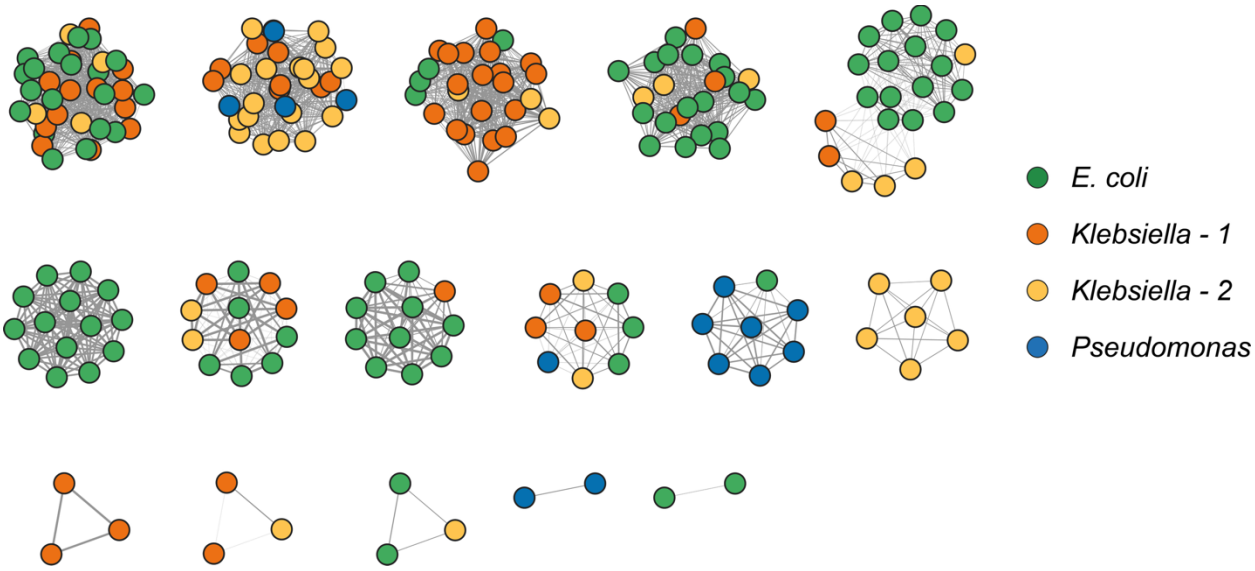
Supplementary Figure 4. Klebsiella-1 dataset tree Phylogenetic tree based on core-genome concatenated marker genes alignments. Grey bars indicate number of infectious phages out of 59-phage set. Green bars indicate prediction performance in 20-fold cross-validation for models with strain-based train-test splitting (unseen strains).



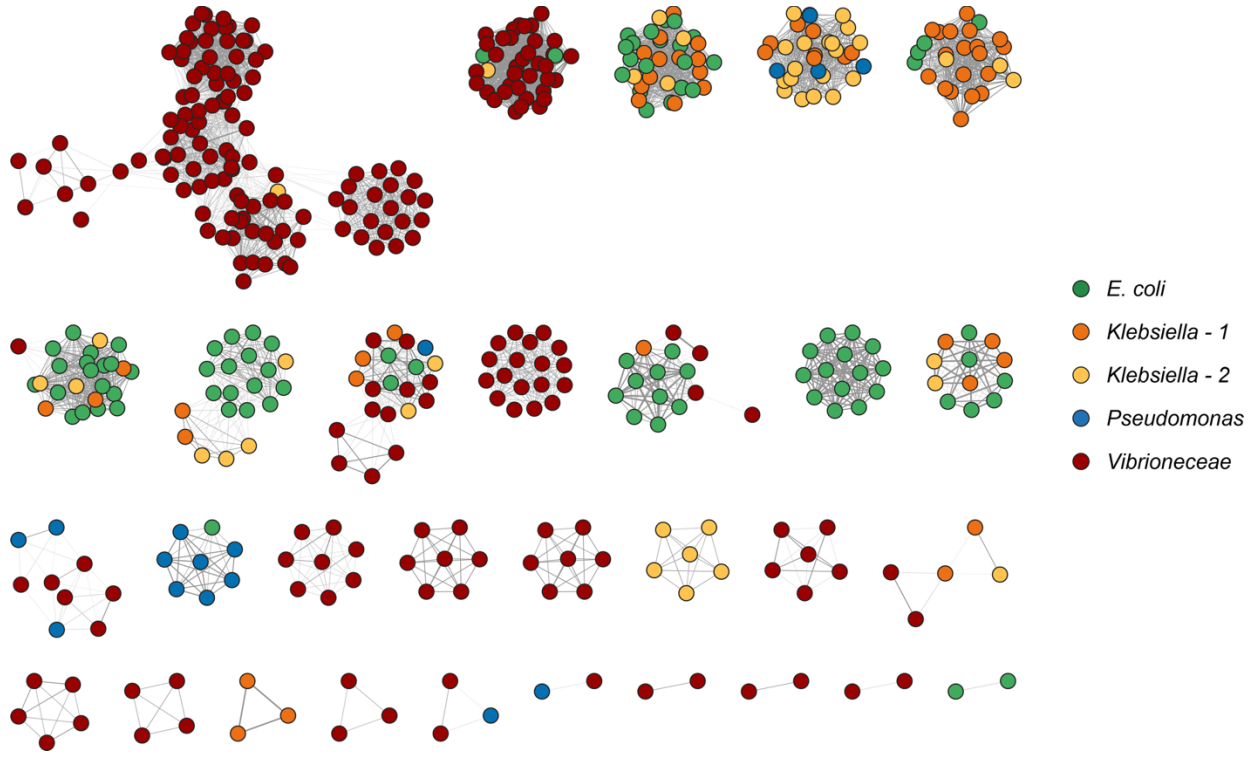
Supplementary Figure 5. Klebsiella-2 dataset tree. Phylogenetic tree based on core-genome concatenated marker genes alignments. Grey bars indicate number of infectious phages out of 46-phage set. Green bars indicate prediction performance in 20-fold cross-validation for models with strain-based train-test splitting (unseen strains).



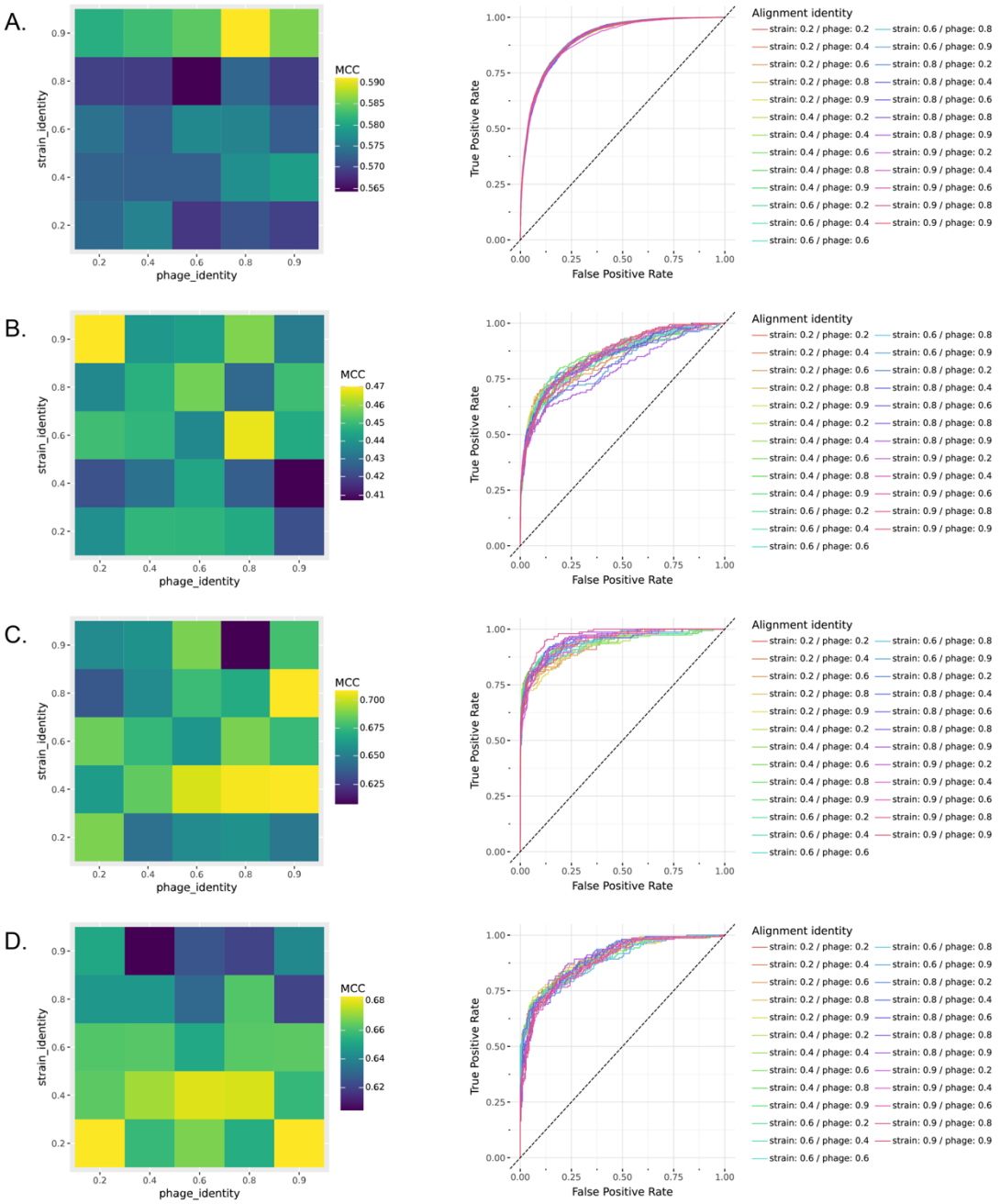
Supplementary Figure 6. Pseudomonas aeruginosa dataset tree. Phylogenetic tree based on core-genome concatenated marker genes alignments. Grey bars indicate number of infectious phages out of 19-phage set. Green bars indicate prediction performance in 20-fold cross-validation for models with strain-based train-test splitting (unseen strains).



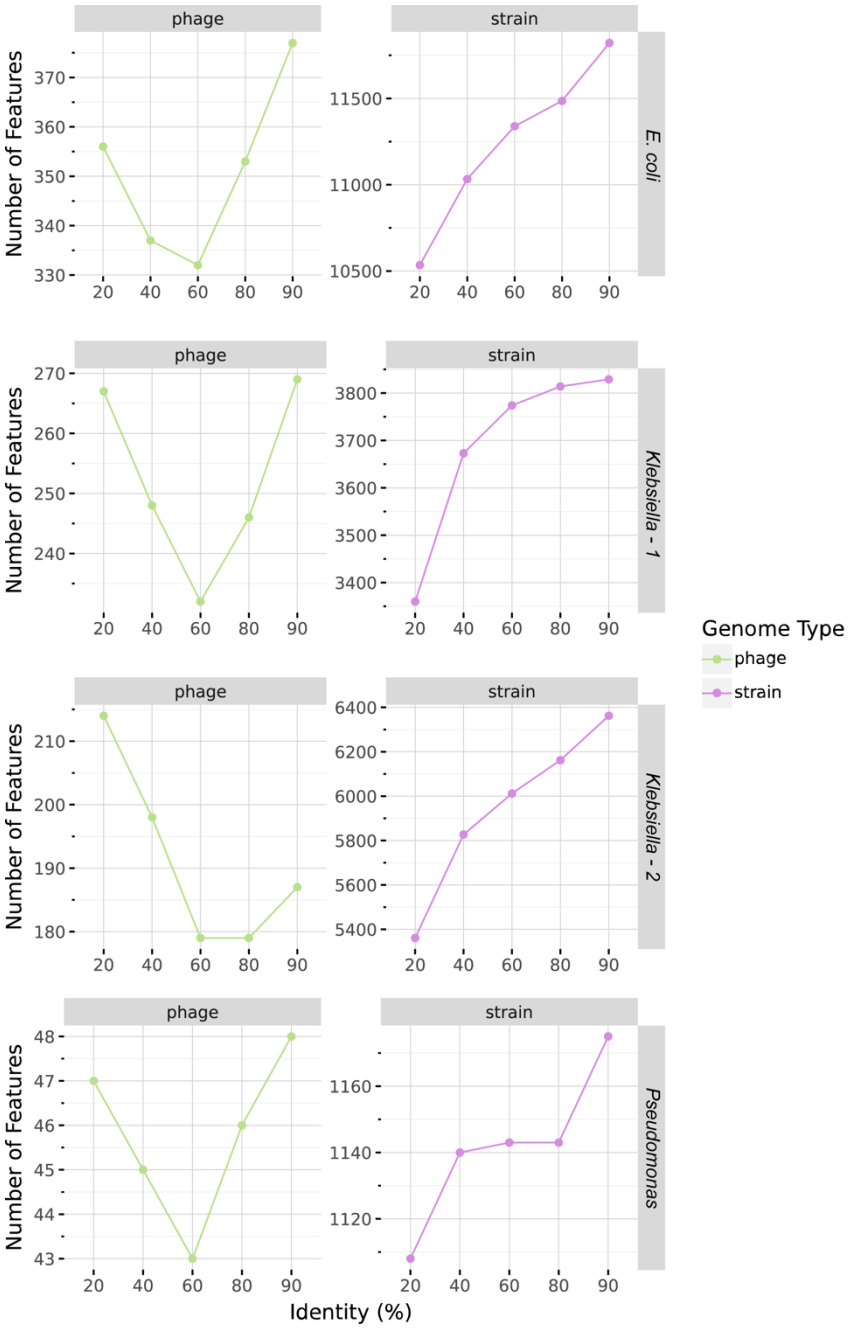
Supplementary Figure 7. Phage gene sharing network for development datasets. Gene sharing network shows distribution of phages across datasets (color). Edges indicate vConTACT2 gene sharing similarity metric.



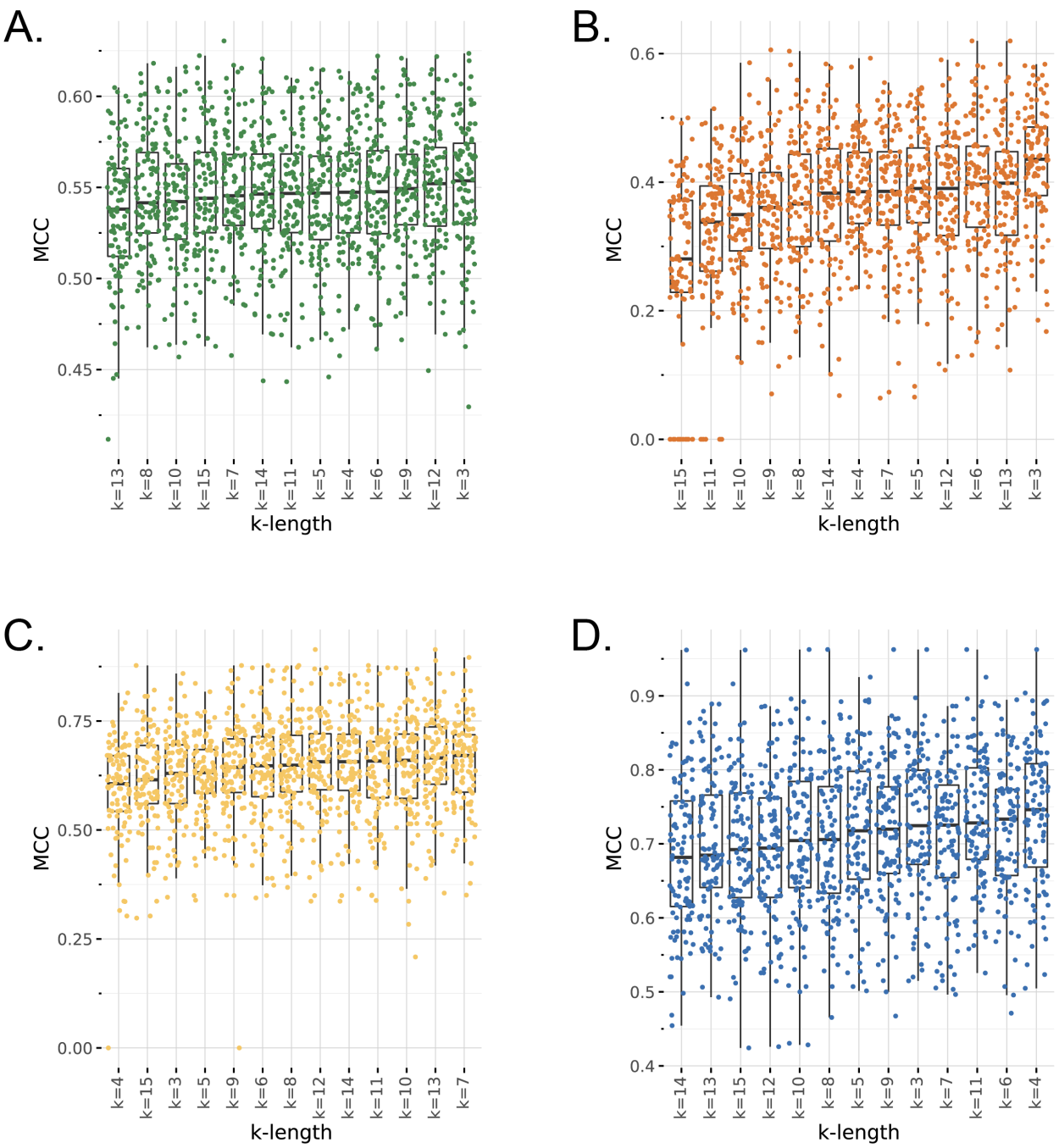
Supplementary Figure 8. Phage gene sharing network for public datasets. Gene sharing network shows distribution of phages across datasets (color). Edges indicate vConTACT2 gene sharing similarity metric.



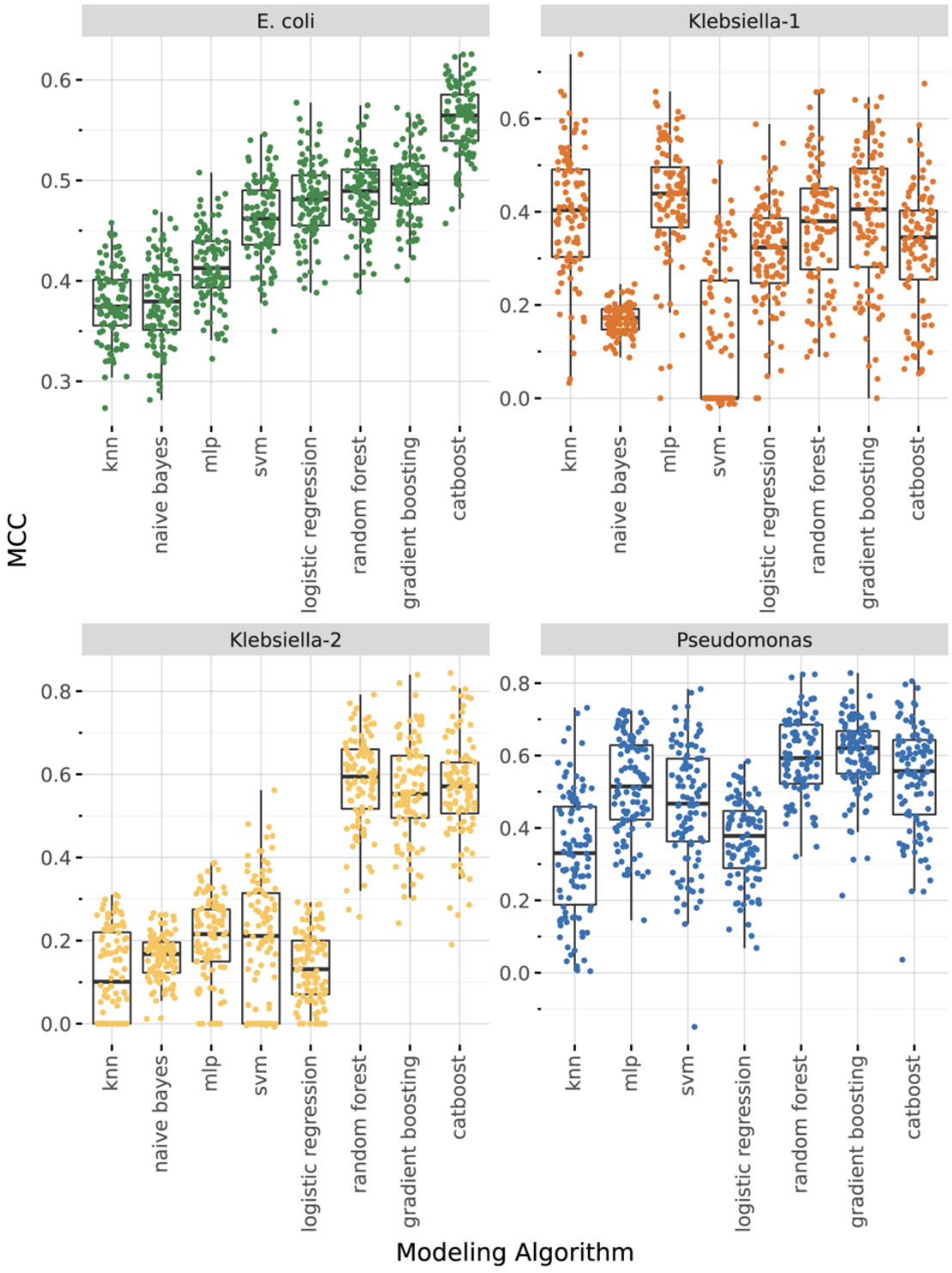
Supplementary Figure 9. Impact of MMSeqs2 identity thresholds on model performance. Heatmaps show model performance (MCC) across 25 MMSeqs2 identity threshold combinations. Receiver operating characteristic curves should model performance (AUROC) across 25 MMSeqs2 identity threshold combinations across *E. coli* (A), *Klebsiella*-1 (B), *Klebsiella*-2 (C), and *Pseudomonas* (D) datasets



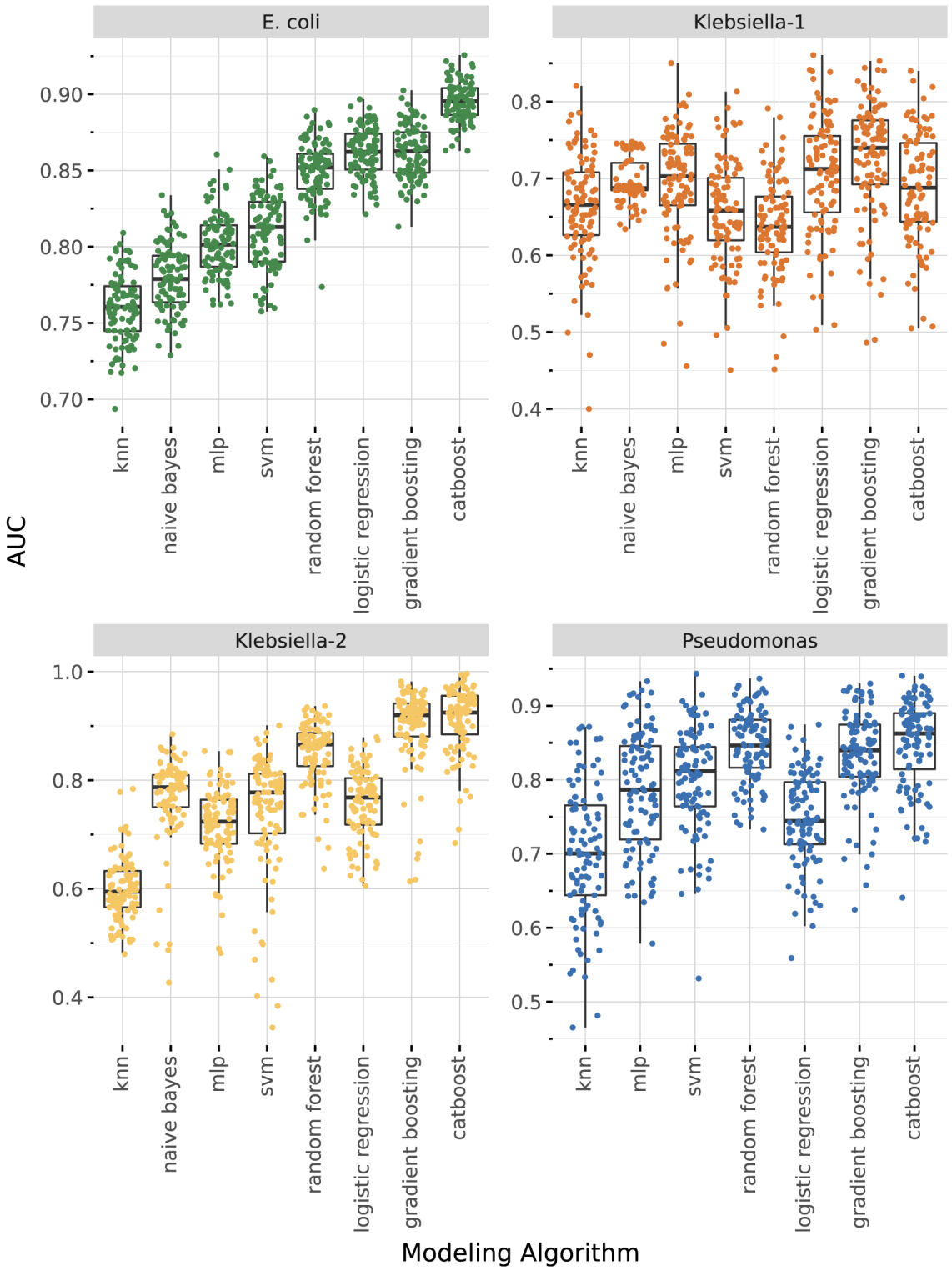
Supplementary Figure 10. Feature counts for MMSeqs2 identity thresholds. Lines show the impact of MMSeqs2 identity threshold on the total number of phage (green) and strain (purple) protein family features across development datasets.



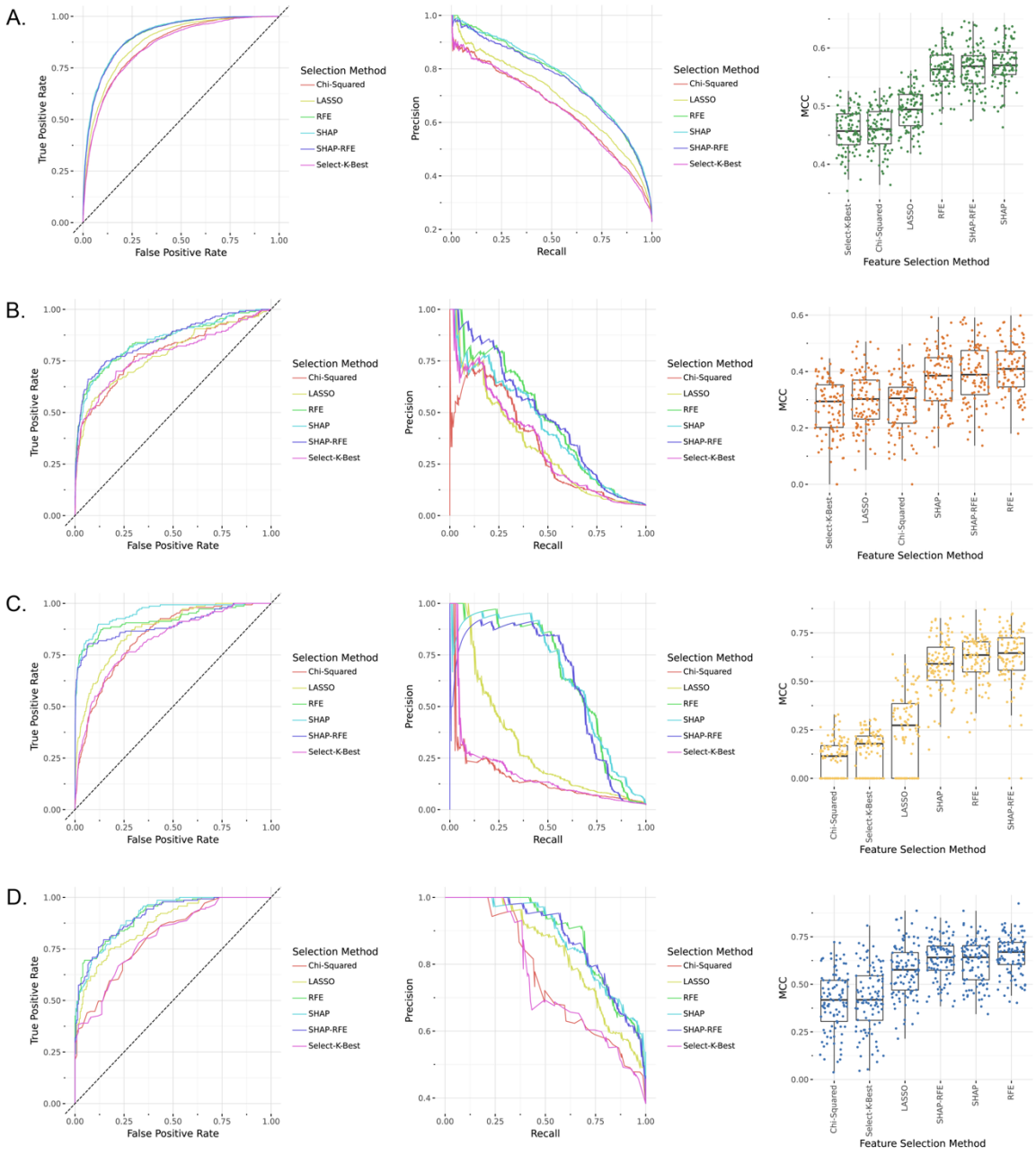
Supplementary Figure 11. The impact of k-length in E. coli (A), Klebsiella-1 (B), Klebsiella-2 (C), and Pseudomonas (D) datasets.



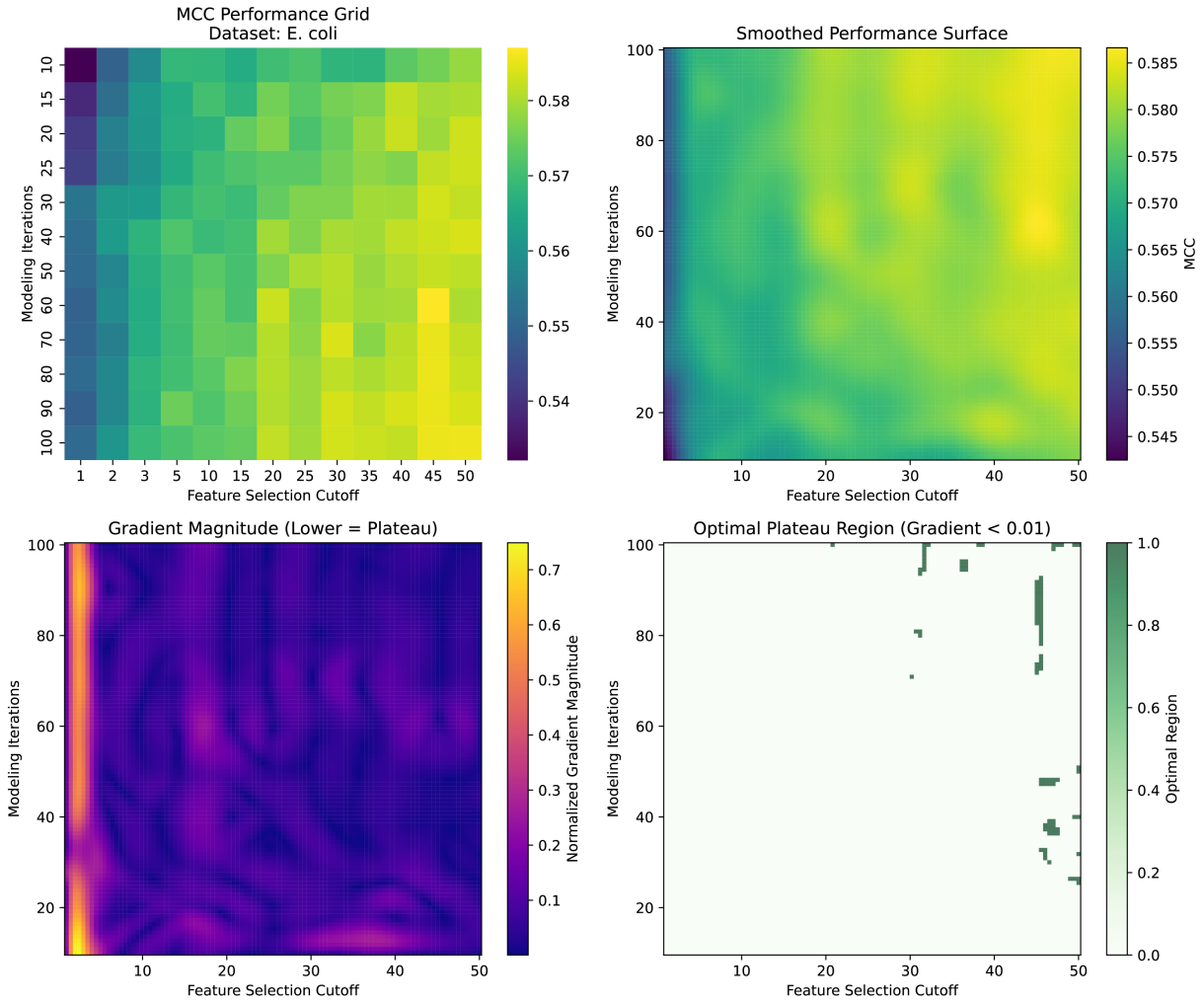
Supplementary Figure 12. Modeling algorithm comparison. Boxplots show performance (MCC) of 100 models used for ensemble learning approach across modeling algorithms and development datasets.



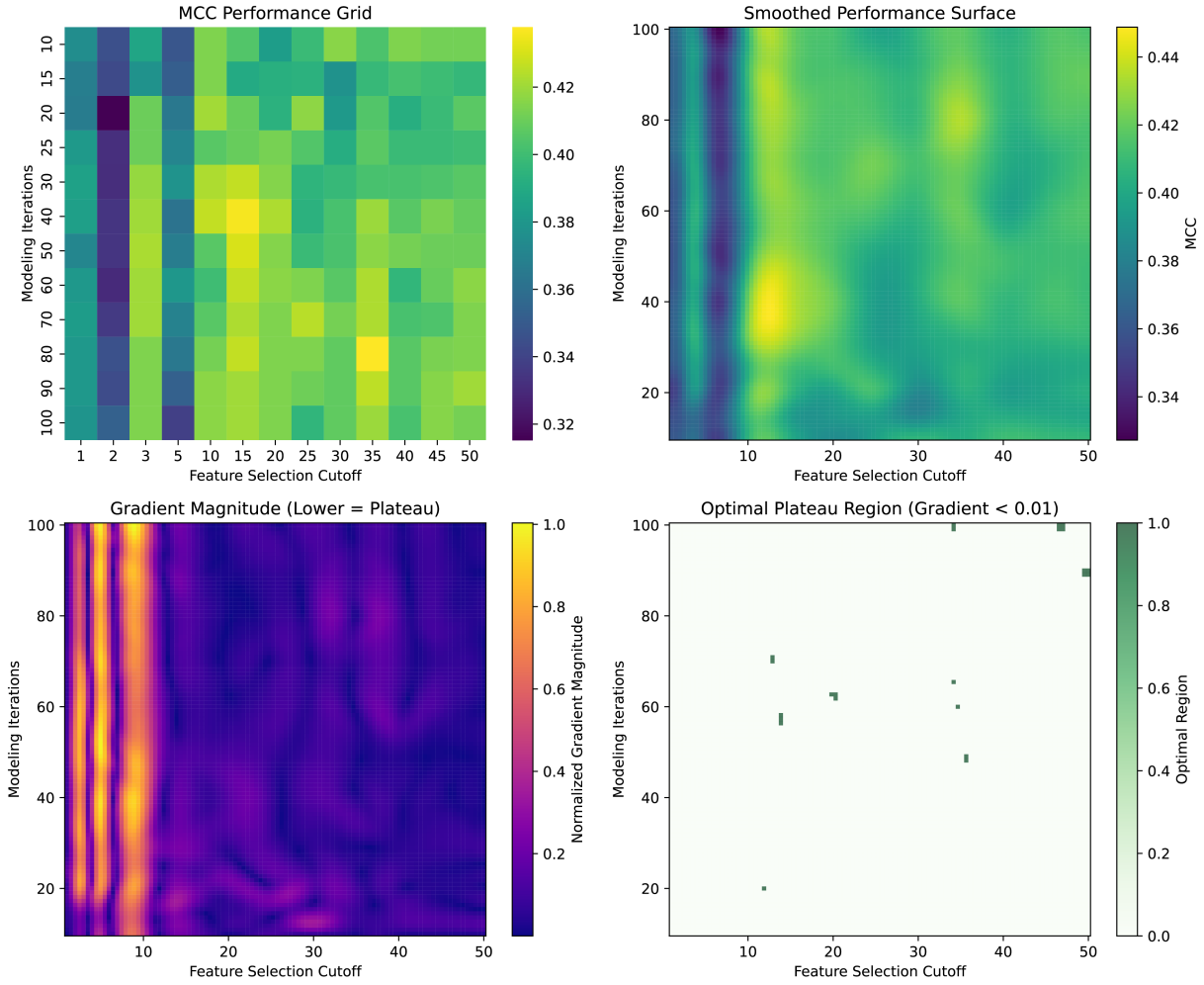
Supplementary Figure 13. Modeling algorithm comparison. Boxplots show performance (AUROC) of 100 models used for ensemble learning approach across modeling algorithms and development datasets.



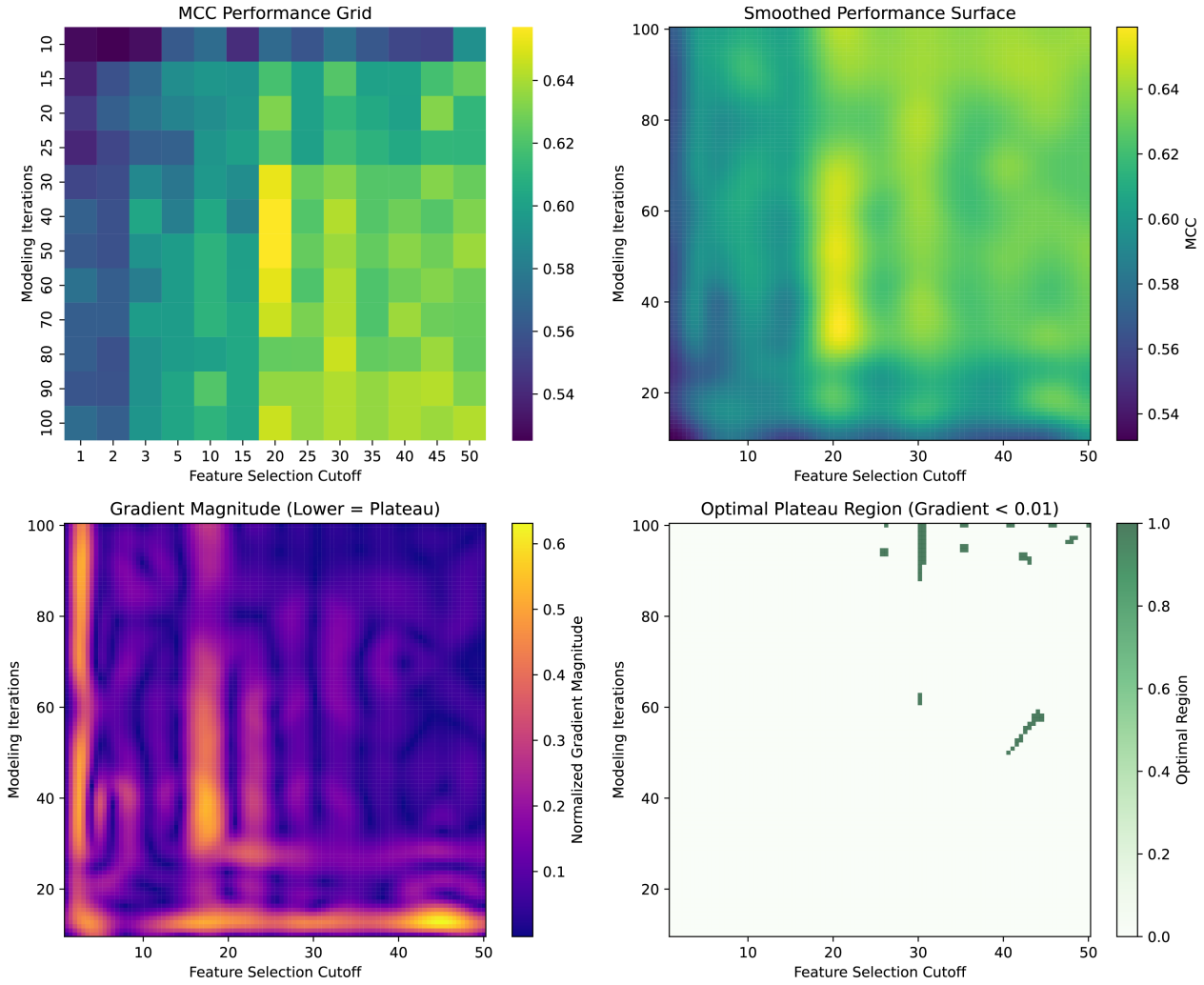
Supplementary Figure 14. Impact of feature selection algorithm on model performance. ROC curves, precision-recall curves and the distribution of model performances comparing 6 feature selection algorithms across *E. coli* (A), *Klebsiella*-1 (B), *Klebsiella*-2 (C), and *Pseudomonas* (D) datasets.



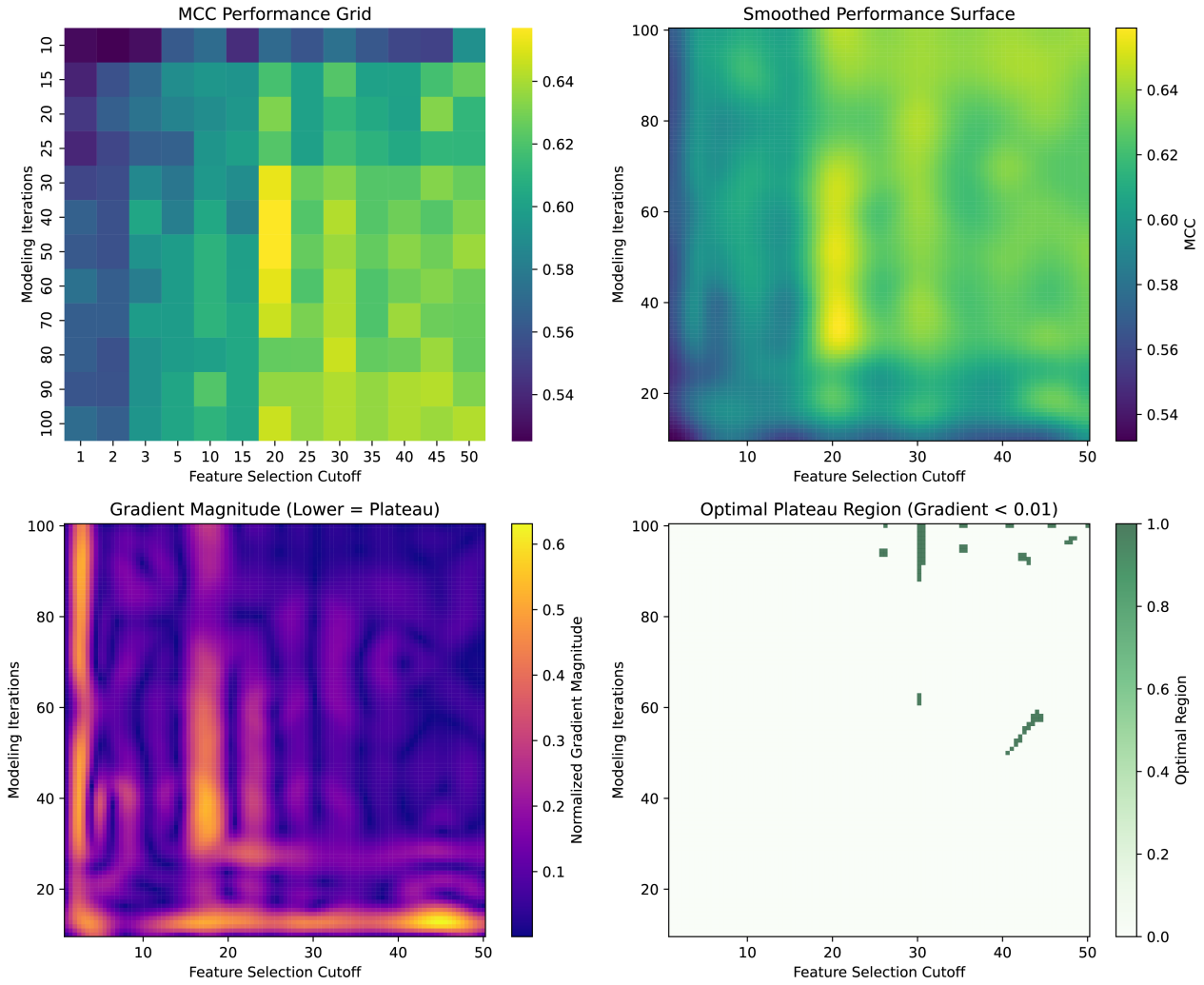
Supplementary Figure 15. Impact of ensemble features selection and modeling approach iterations in E. coli dataset.



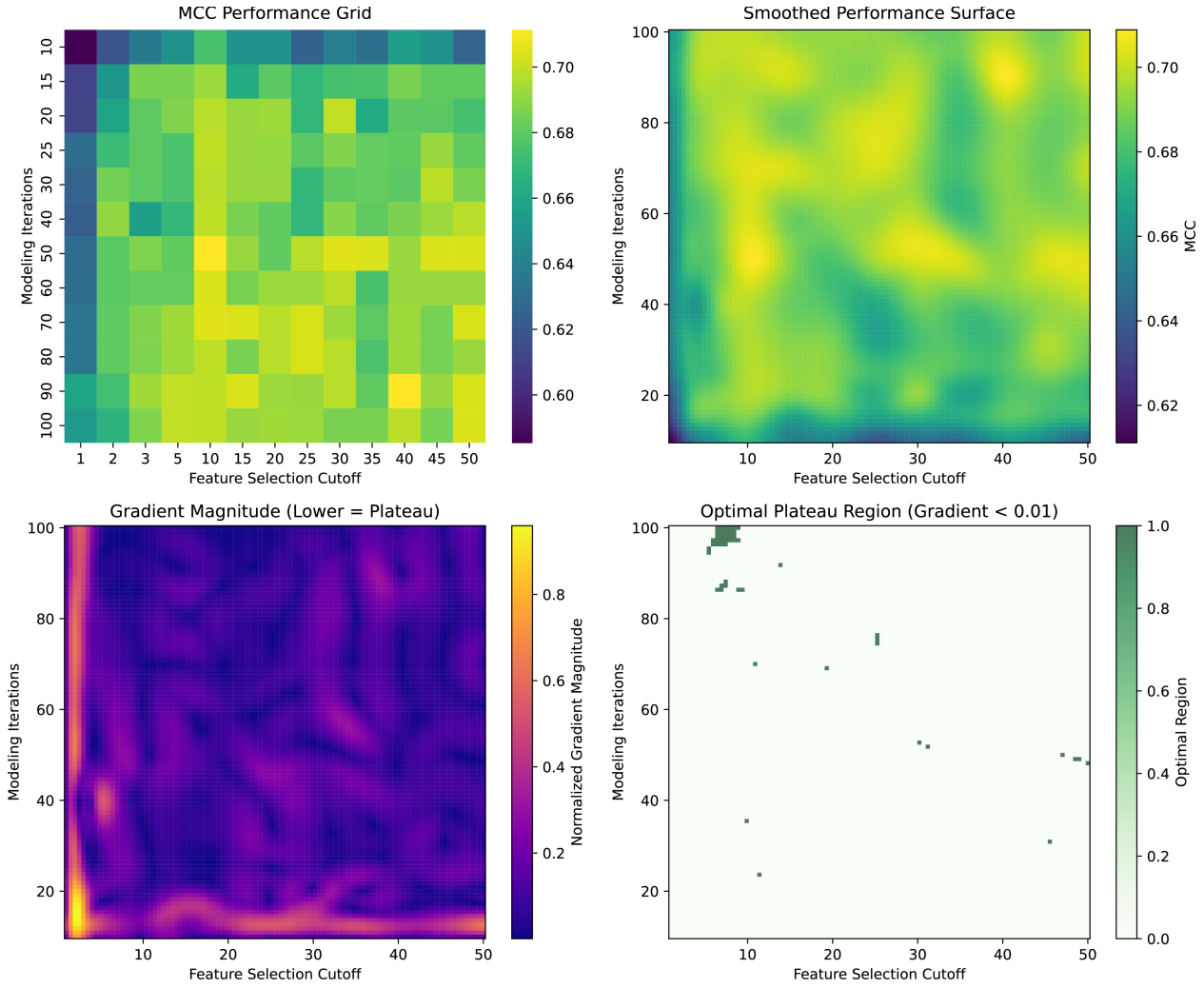
Supplementary Figure 16. Impact of ensemble features selection and modeling approach iterations in Klebsiella-1 dataset.



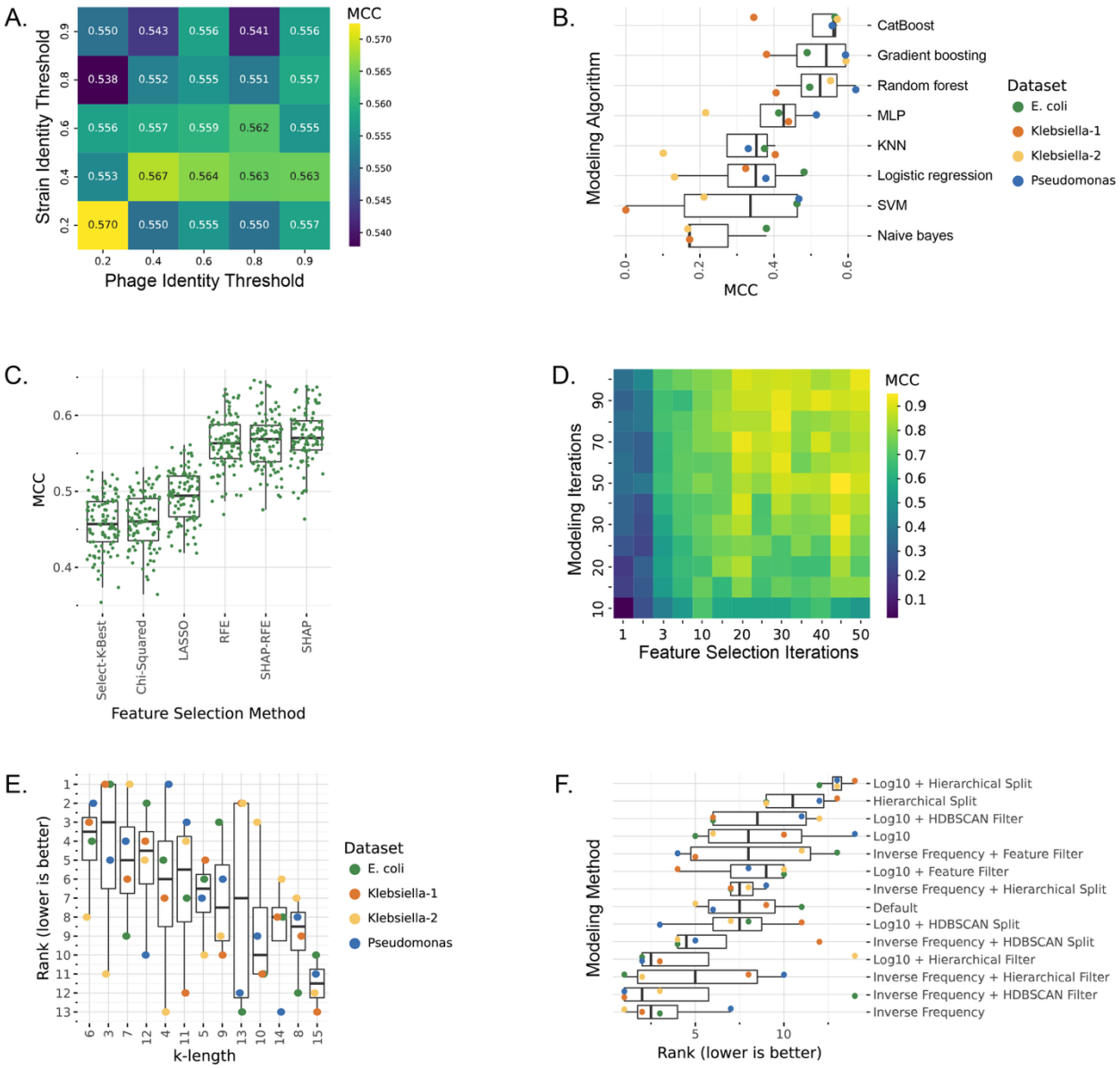
Supplementary Figure 17. Impact of ensemble features selection and modeling approach iterations in Klebsiella-2 dataset.



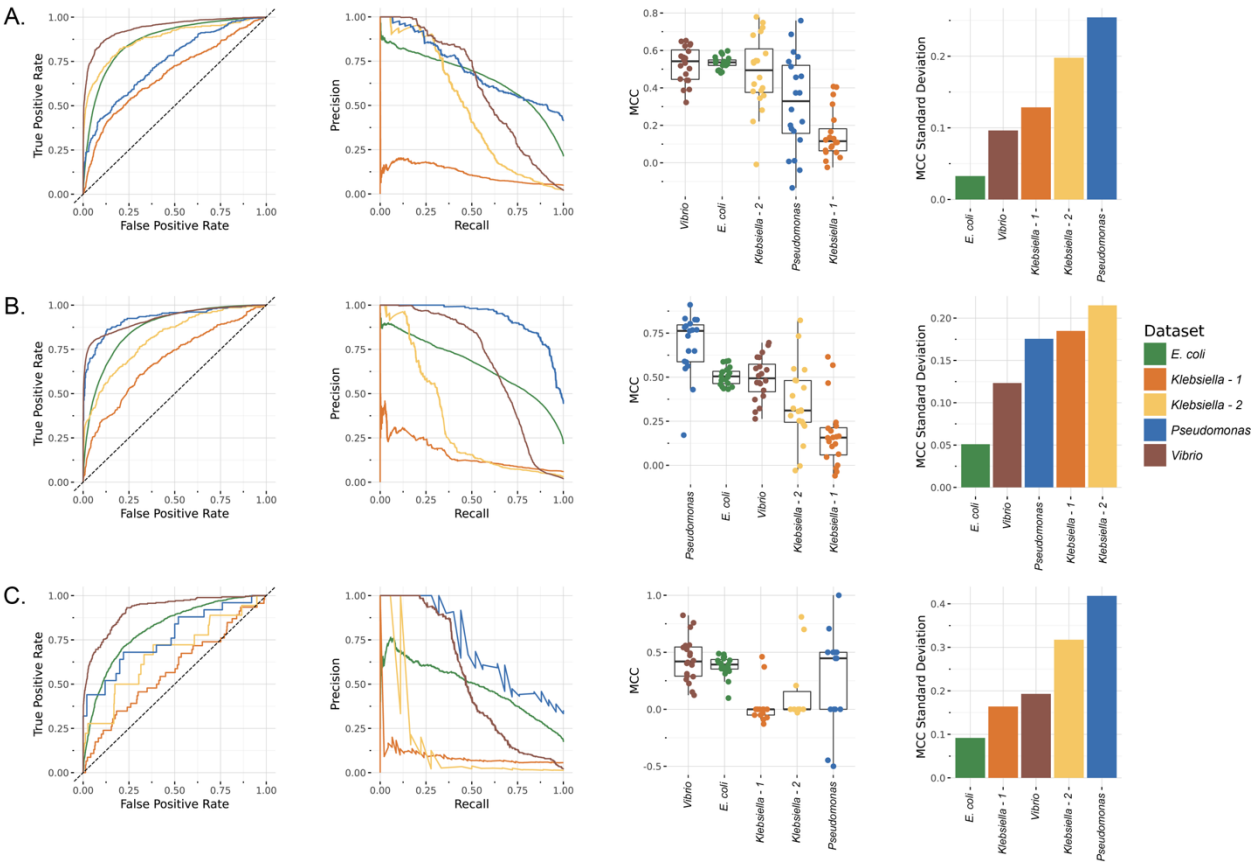
Supplementary Figure 18. Impact of ensemble features selection and modeling approach iterations in Pseudomonas dataset.



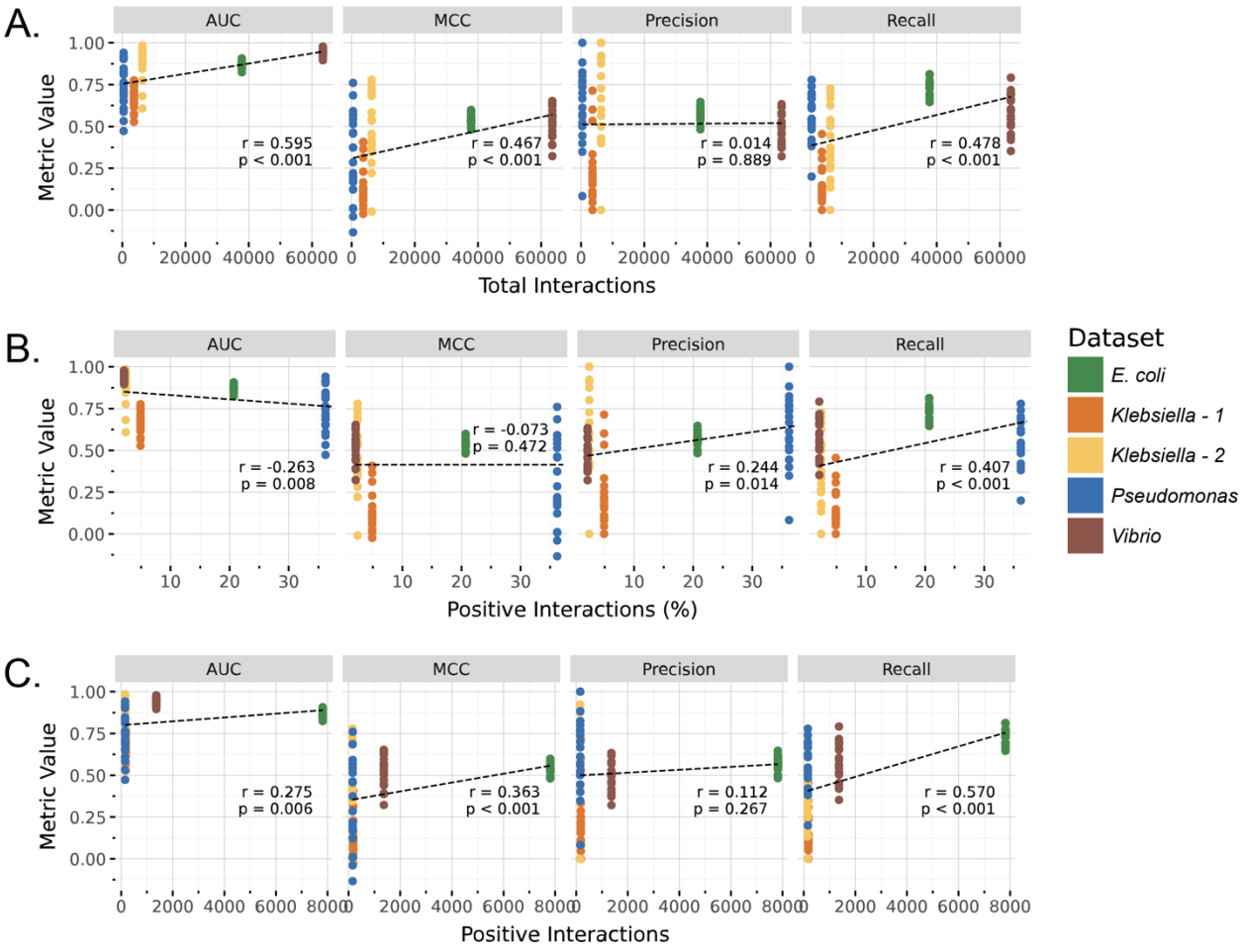
Supplementary Figure 19. Combined analysis of the impact of ensemble features selection and modeling approach iterations in all development datasets.



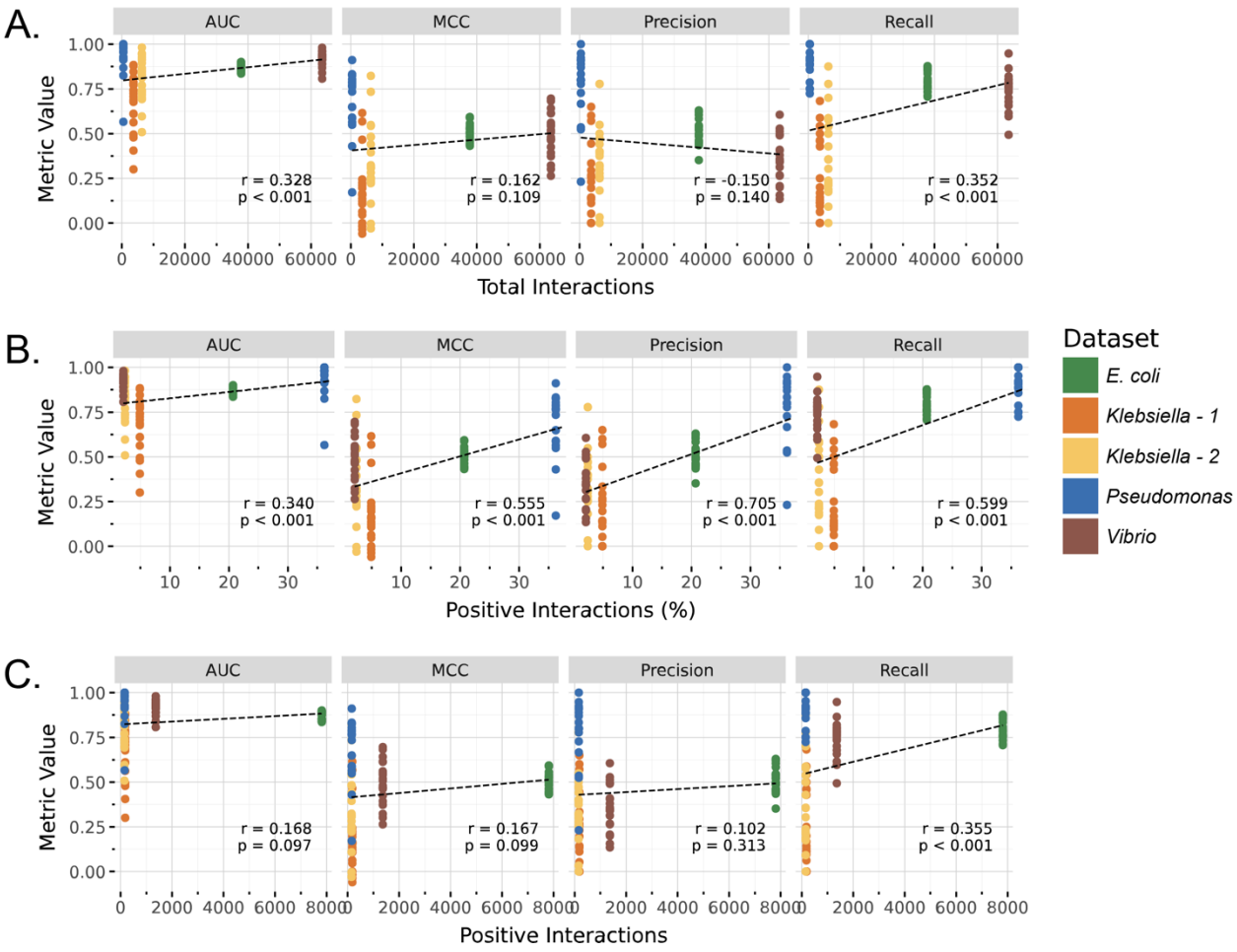
Supplementary Figure 20. Modeling workflow optimization. A) Mixed effect model of MMSeqs2 identity thresholds on model performance (MCC). B) Impact of modeling algorithm on model performance across datasets (MCC). C) Impact of feature selection method on training model performance (MCC) in the *E. coli* dataset. D) Impact of feature selection and modeling iterations on model performance (MCC) in the *E. coli* dataset. E) Rank order impact of *k* length on model performance (MCC) of unseen strains across datasets, based on 20-fold cross-validation F) Rank order impact of hyperparameters on model performance (MCC) of unseen strains across datasets, based on 20-fold cross-validation.



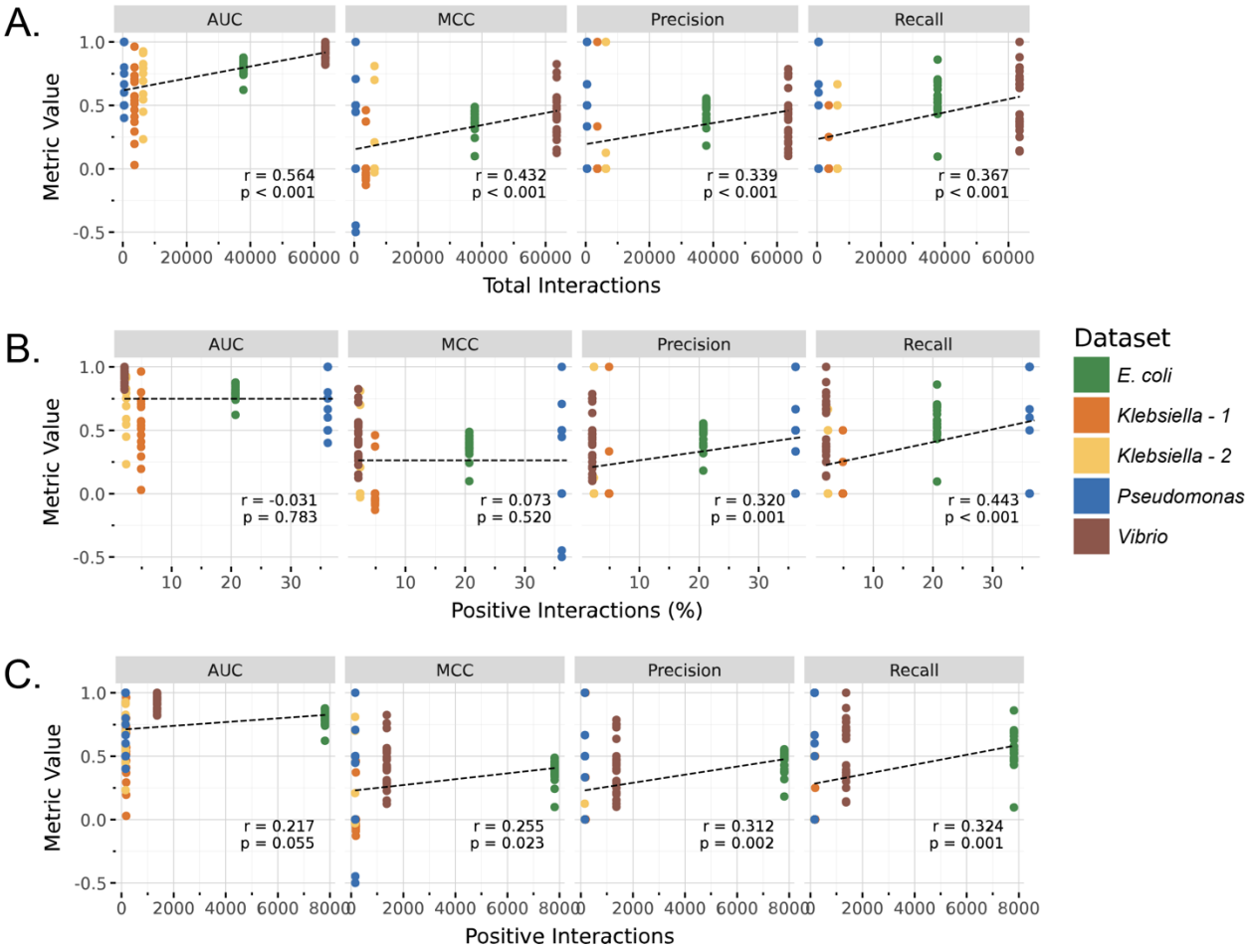
Supplementary Figure 21. Model performance across datasets. Receiver operating characteristic (ROC) curves show prediction performance variability across datasets, when predicting phages infecting previously unseen strains (A), hosts for previously unseen phages (B), and interactions between previously unseen strain-phage pairs (C).



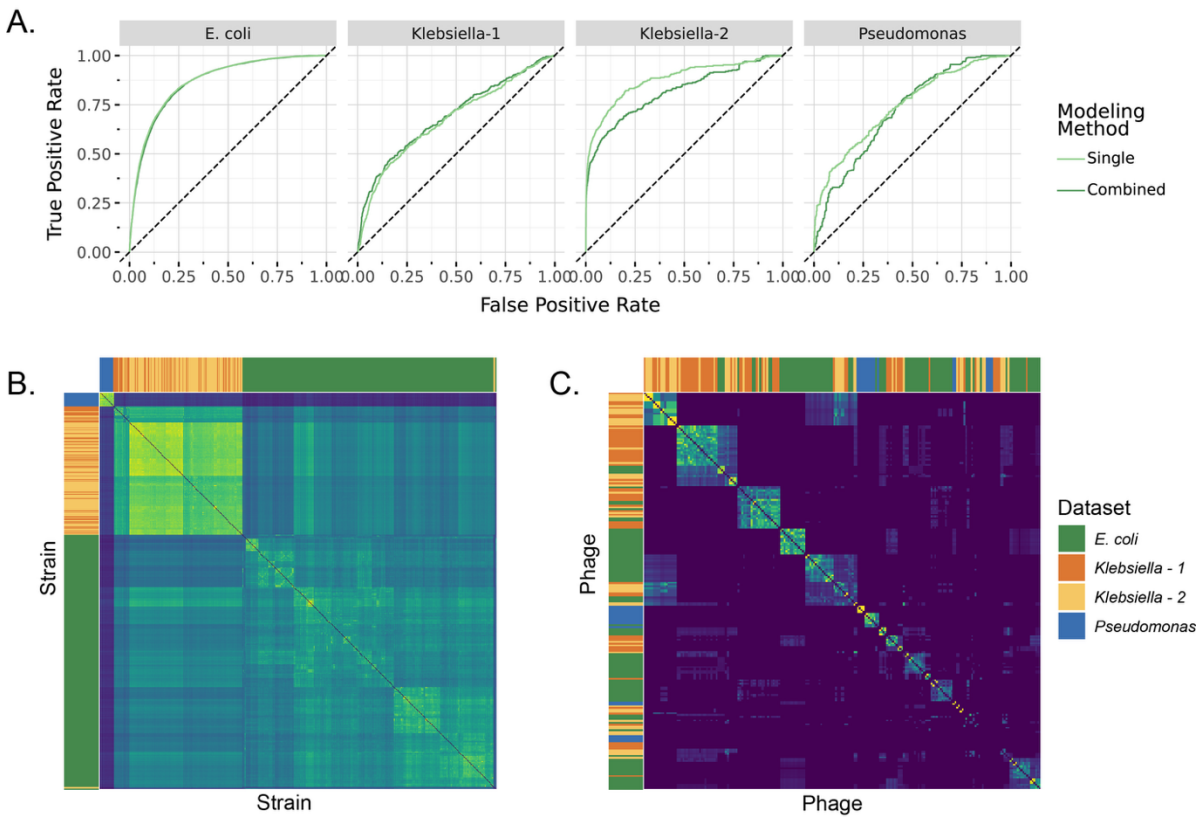
Supplementary Figure 22. Relationship between model performance and dataset characteristics when predicting phages infecting previously unseen strains.



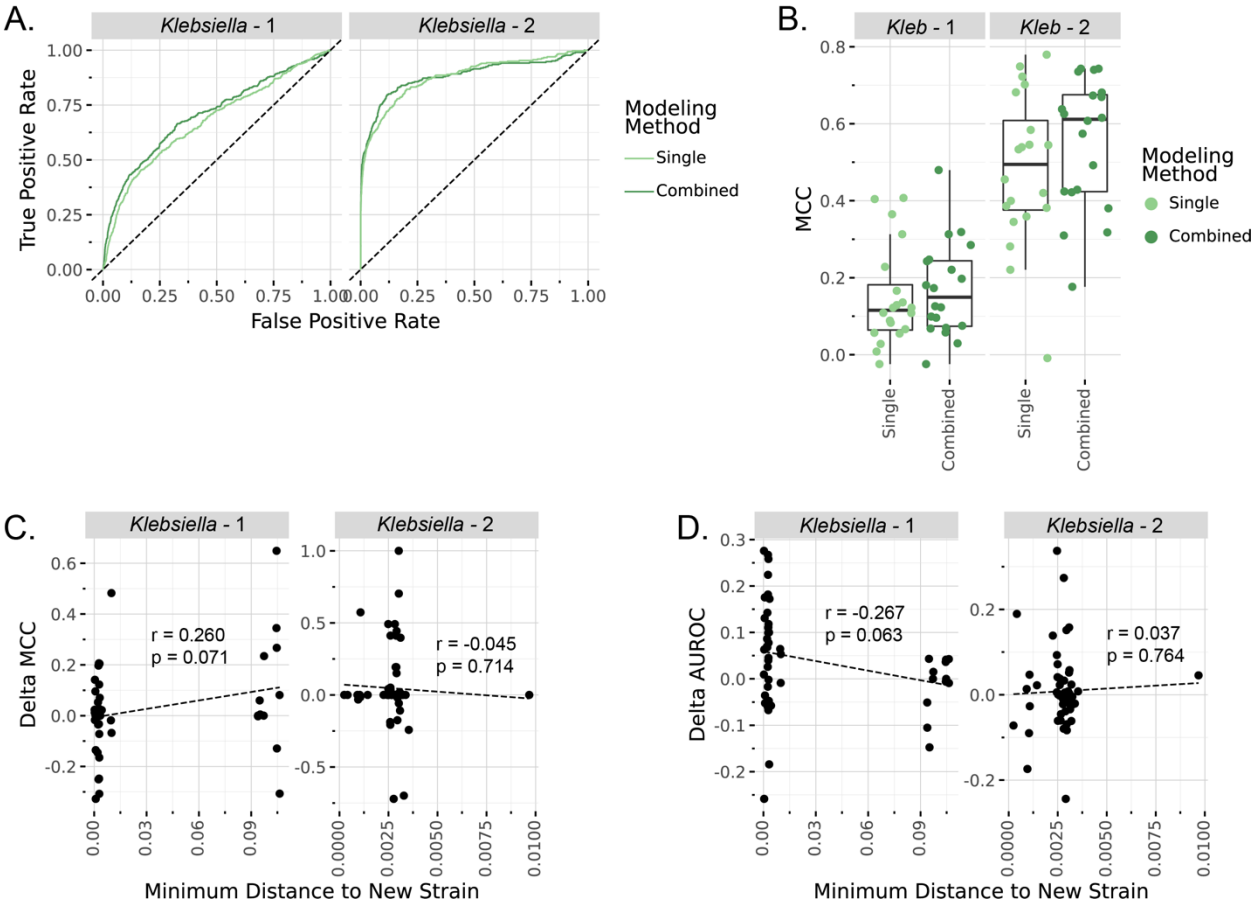
Supplementary Figure 23. Relationship between model performance and dataset characteristics when predicting hosts for previously unseen phages.



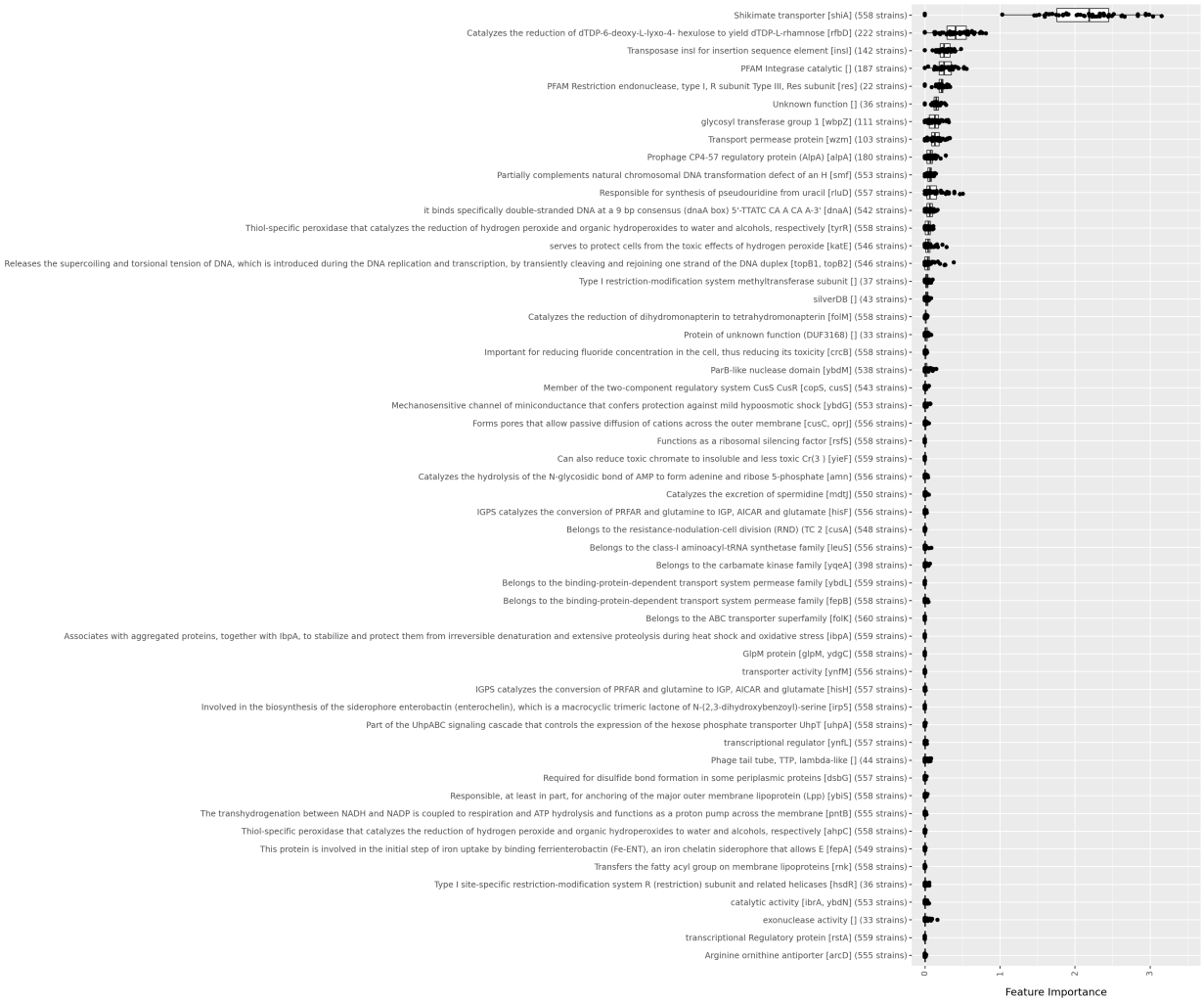
Supplementary Figure 24. Relationship between model performance and dataset characteristics when predicting interactions between previously unseen strain-phage pairs.



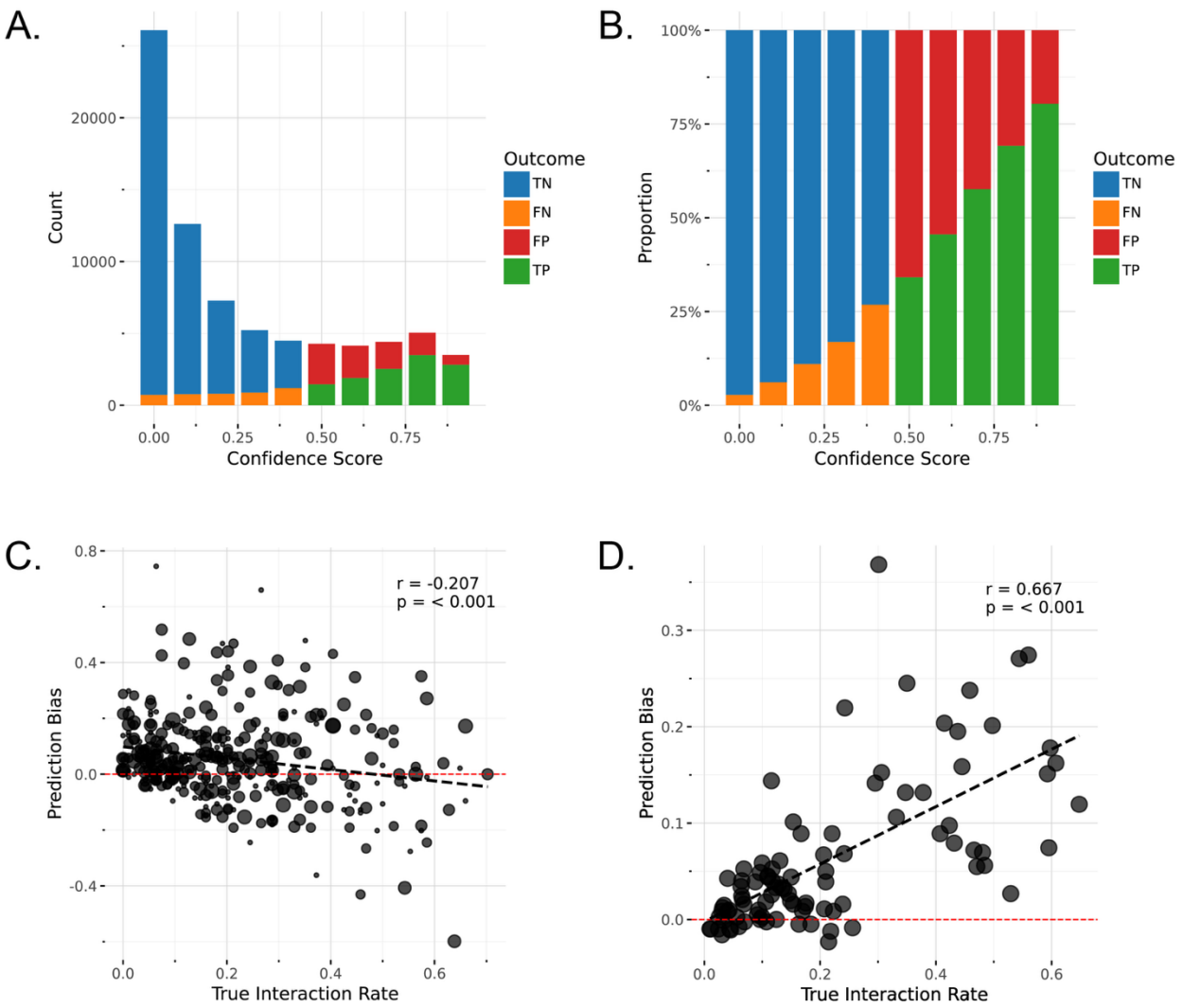
Supplementary Figure 25. Combining datasets across genera. A) Models trained by combining all datasets (dark green) showed increased performance in the *Klebsiella* datasets, and no change or a slightly decrease in performance in *E. coli* and *Pseudomonas* datasets, in comparison to models trained on single datasets (light green). Displayed results are from 20-fold cross-validation. B) Jaccard distances between strains, based on predictive feature content, shows show distinct clusters by genus. Colored bars show strain dataset. C) Jaccard distances between phages, based on predictive feature content, show more modularity and overlap between genera. Colored bars show phage dataset.



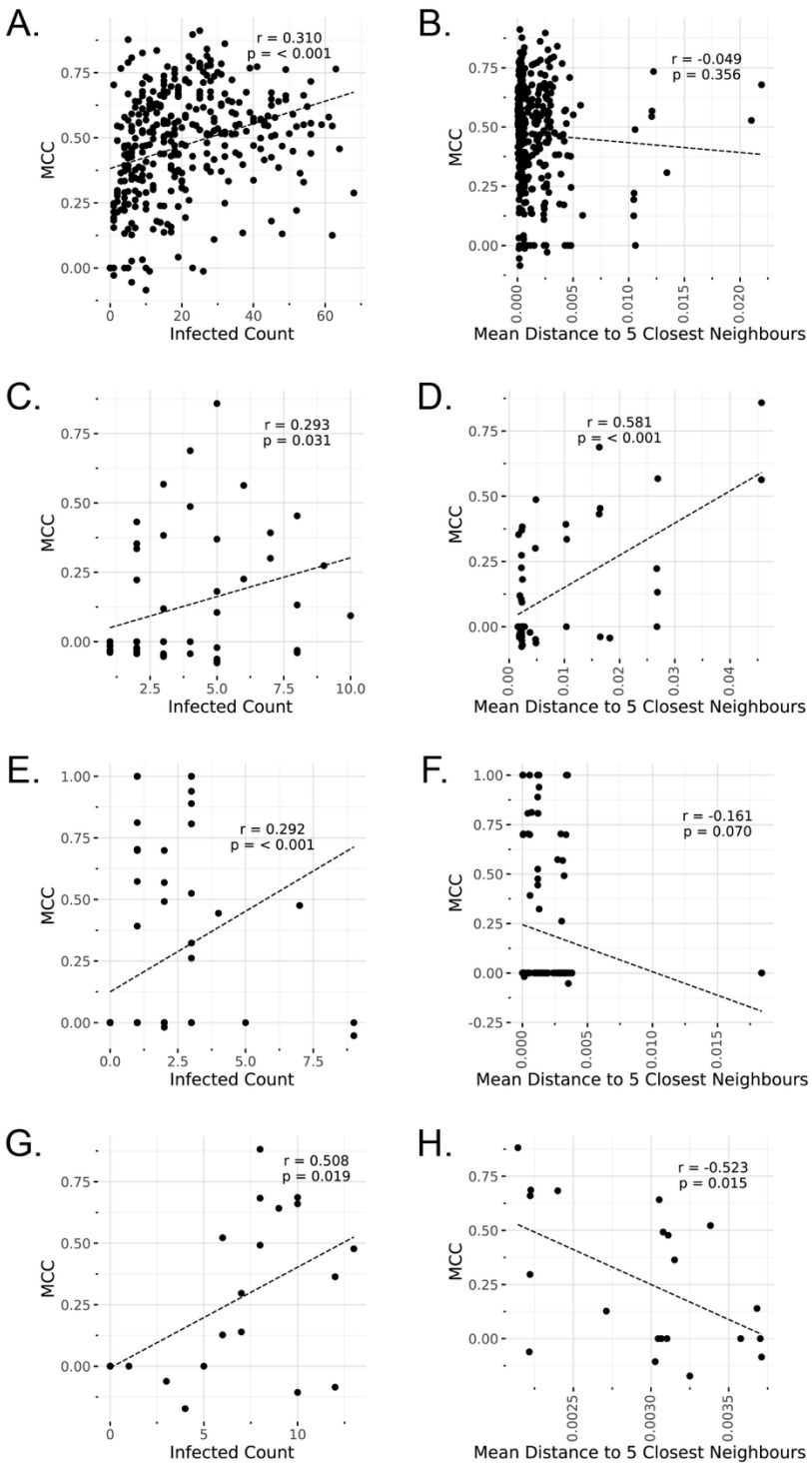
Supplementary Figure 26. Combining datasets within the Klebsiella genus. A) Models trained by combining both *Klebsiella* datasets (dark green) showed increased performance across both datasets, in comparison to models trained on single datasets (light green). Displayed AUROC curves are from 20-fold cross-validation. B) Boxplots show performance of individual rounds of cross validation when combining both *Klebsiella* datasets (dark green) in comparison to models trained on single datasets (light green). A significant increase in performance was observed in the *Klebsiella*-2 dataset (p = 0.024). C) The change in MCC of individual *Klebsiella* strains was compared to the phylogenetic distance to the nearest strain from the added dataset. D) The change in AUROC of individual *Klebsiella* strains was compared to the phylogenetic distance to the nearest strain from the added dataset.



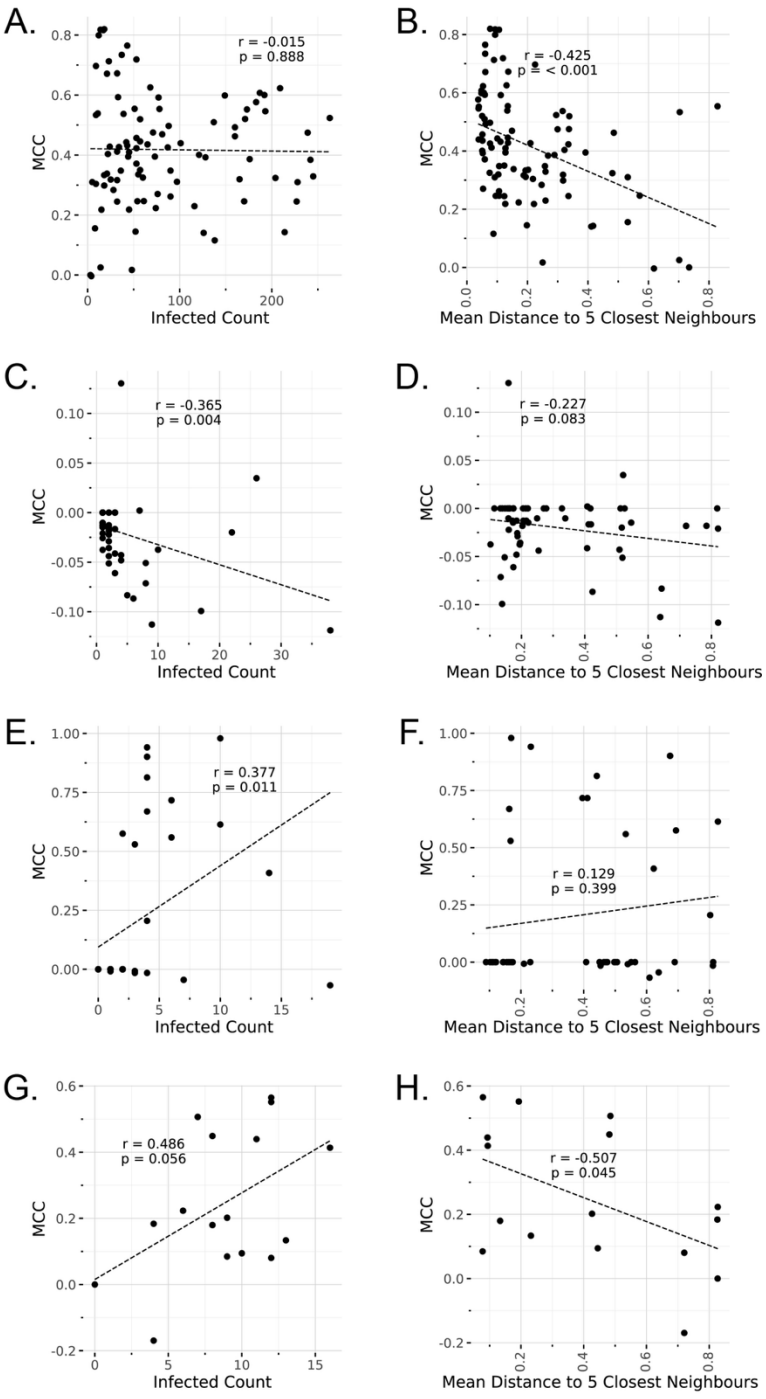
Supplementary Figure 27. Annotations of predictive features shared across datasets. Plot shows feature importance of 54 predictive features that were shared across *E. coli*, *Klebsiella*, and *Pseudomonas* datasets. Y-axis shows the most common annotation of genes associated with each predictive features and number of strains encoding that features. X-axis shows feature importance across 50 models used in ensemble learning.



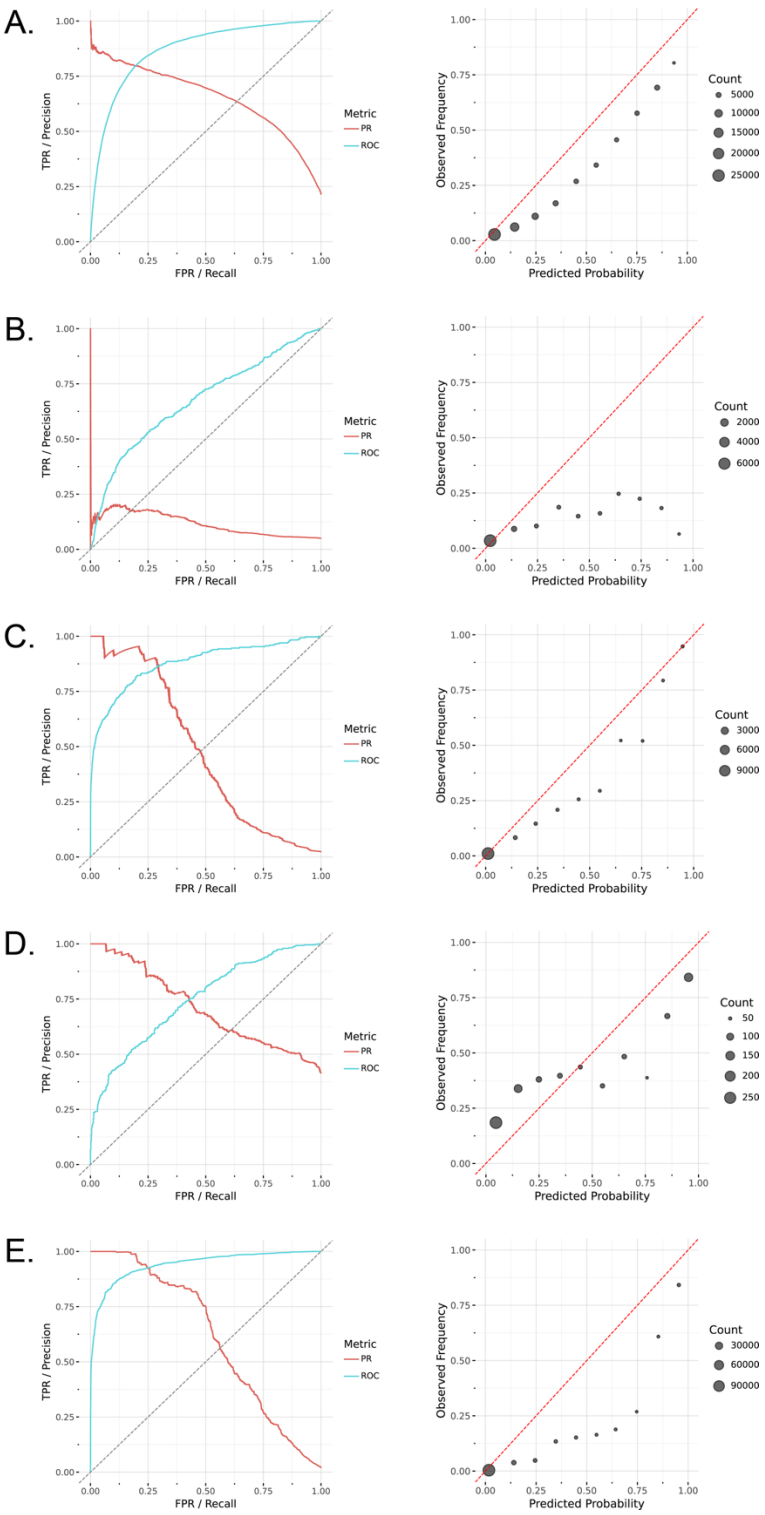
Supplementary Figure 28. E. coli dataset prediction bias overview. The distribution of prediction confidences across 20-fold cross-validation leaving out 10% of strains from the *E. coli* dataset. The majority of predictions were of no-infection, consistent with distributions observed in the experimental dataset (A). The proportion of true-positives (TP) and true-negatives (TN) increases as confidence scores move away from 0.5 classification threshold (B). Prediction bias compared to true infection rates show a negative correlation in strains (C) and a positive correlation in phages (D).



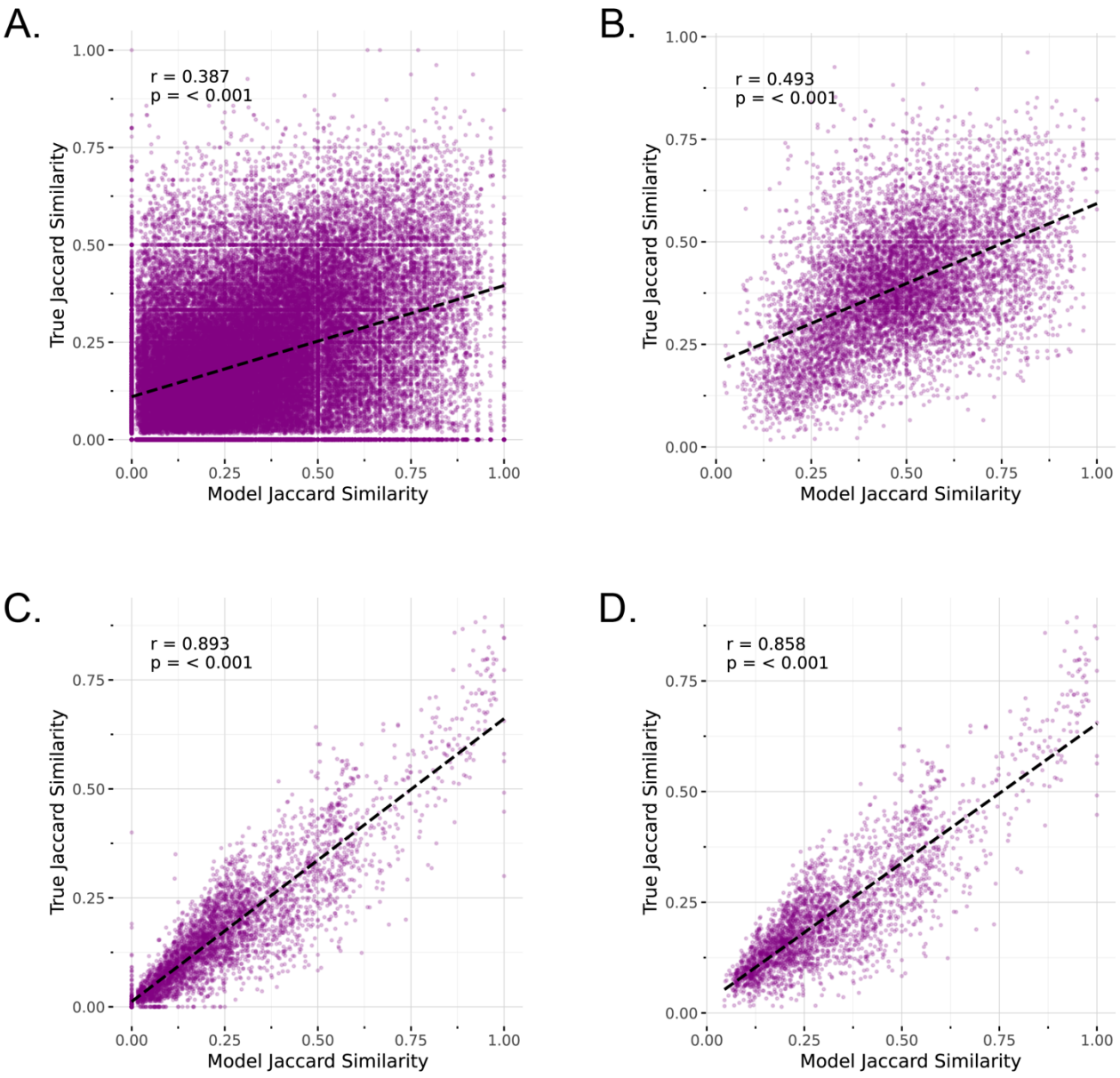
Supplementary Figure 29. Impact of strain characteristics on prediction performance. The relationship between strain susceptibility (number of infecting phages) and strain-level MCC (left) and phylogenetic isolation based on concatenated marker gene alignments and strain-level MCC (right) were evaluated for the *E. coli* (A / B), *Klebiella*-1 (C / D), *Klebiella*-2 (E / F), and *Pseudomonas* (G / H) datasets.



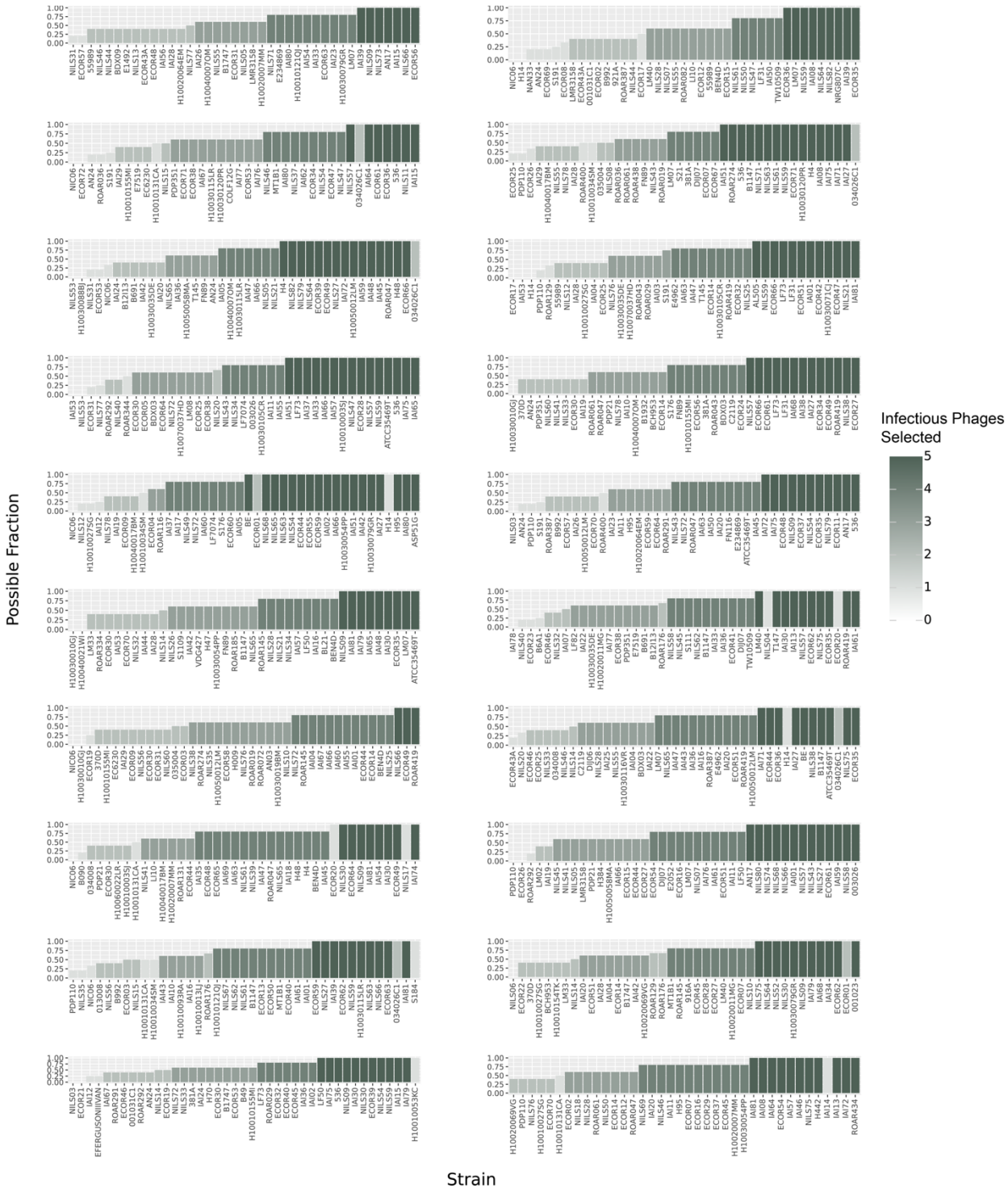
Supplementary Figure 30. Impact of phage characteristics on prediction performance. The relationship between phage infectivity (number of infected strains) and phage-level MCC (left) and phylogenetic isolation based on proteomic equivalence (PEG) and phage -level MCC (right) were evaluated for the *E. coli* (A / B), *Klebiella*-1 (C / D), *Klebiella*-2 (E / F), and *Pseudomonas* (G / H) datasets.



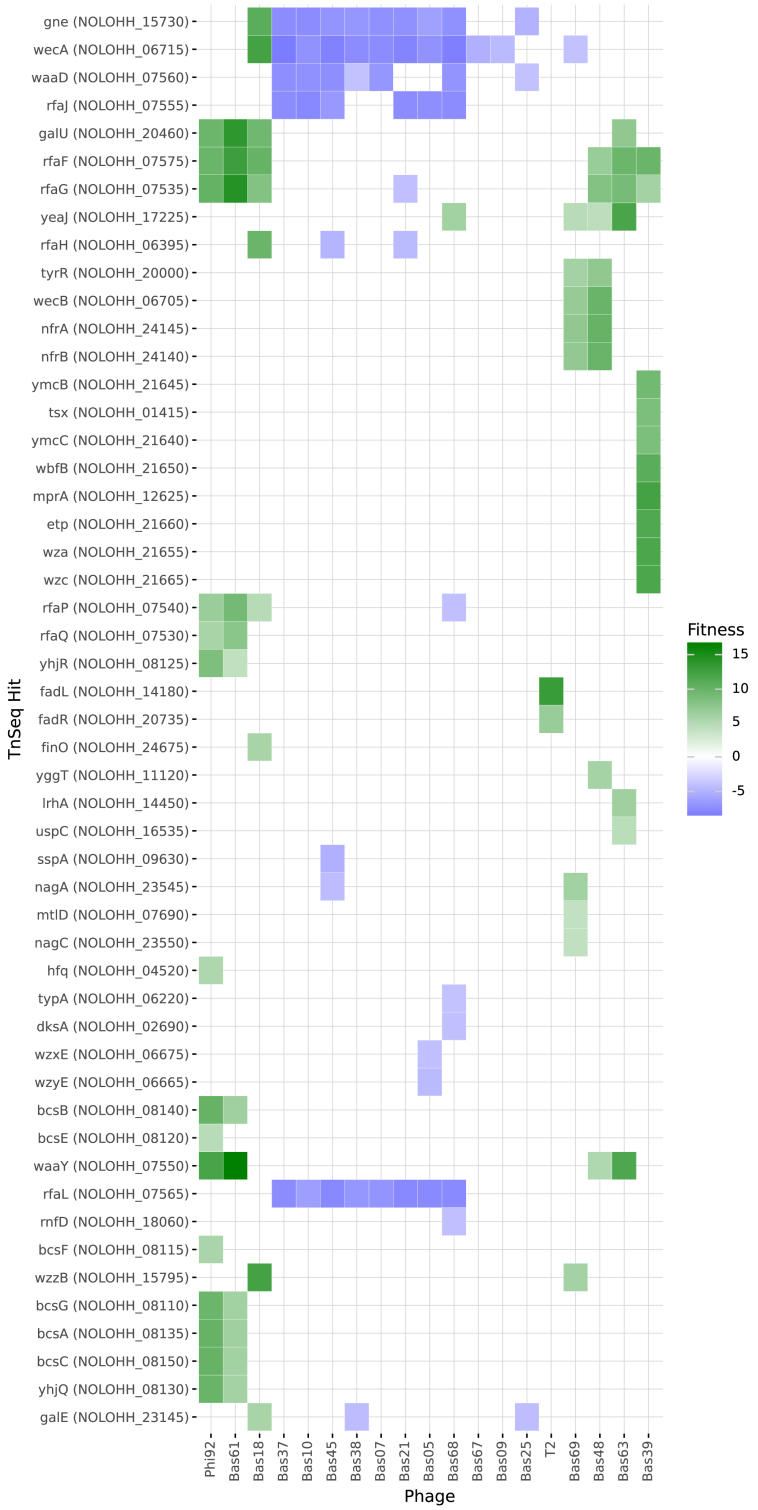
Supplementary Figure 31. Predictive model performance and calibration curves. Column 1 shows ROC (blue) and precision-recall (red) curves. Column 2 shows calibration curves comparing predicted vs. observed infection rates for the *E. coli* (A), *Klebiella*-1 (B), *Klebiella*-2 (C), *Pseudomonas* (D), and *Vibrionaceae* (E) datasets.



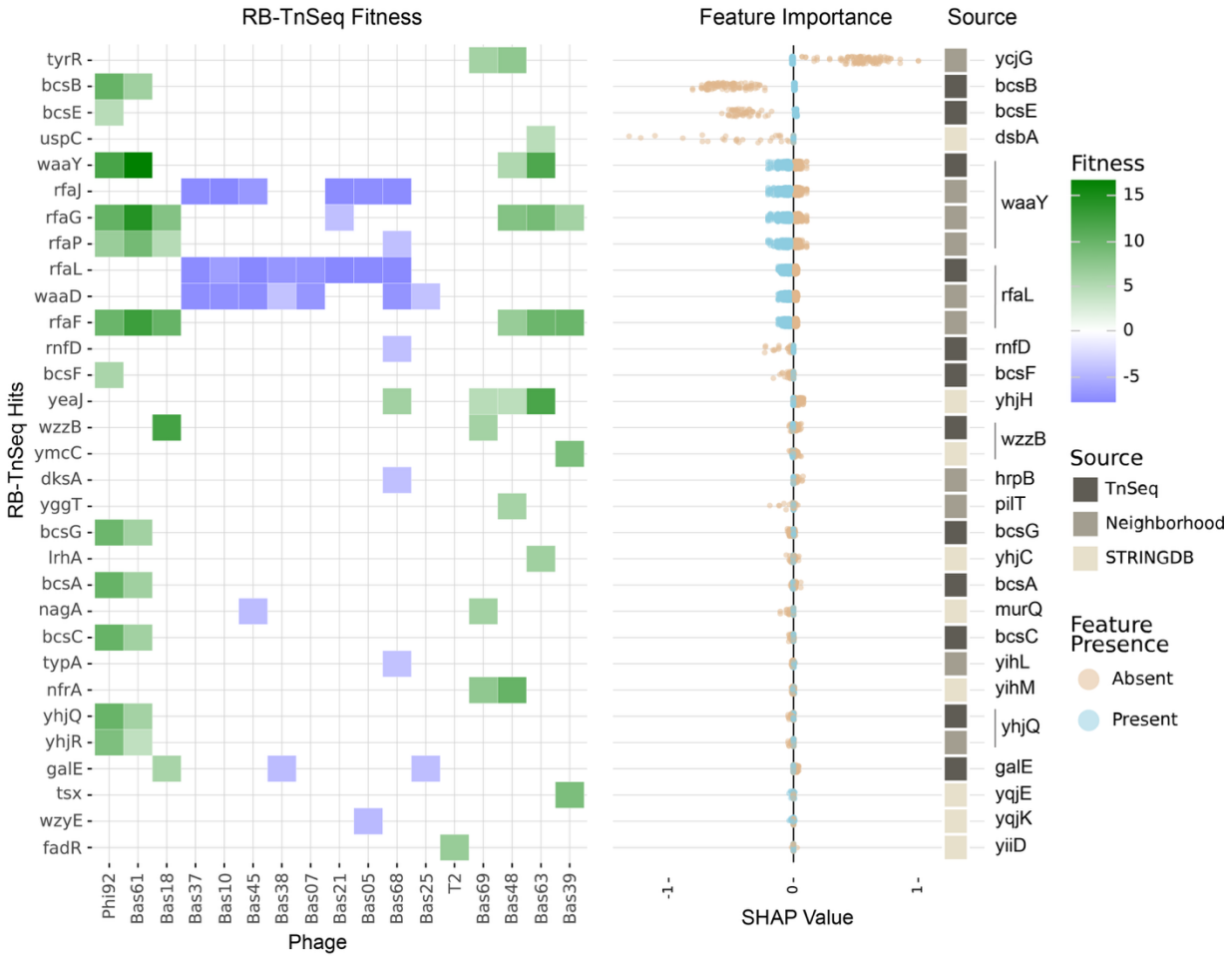
Supplementary Figure 32. Comparing experimental and predicted interaction profiles. Jaccard similarities between strain pairs and phage pairs based on their experimental interaction profiles and compared these to model-predicted interaction similarities across 66,795 strain pairs (A) and 4,371 phage pairs (C), and between broadly susceptible strains (infected by >= 20 phages) (B) and broadly infectious phages (infects >= 20 strains) (D).



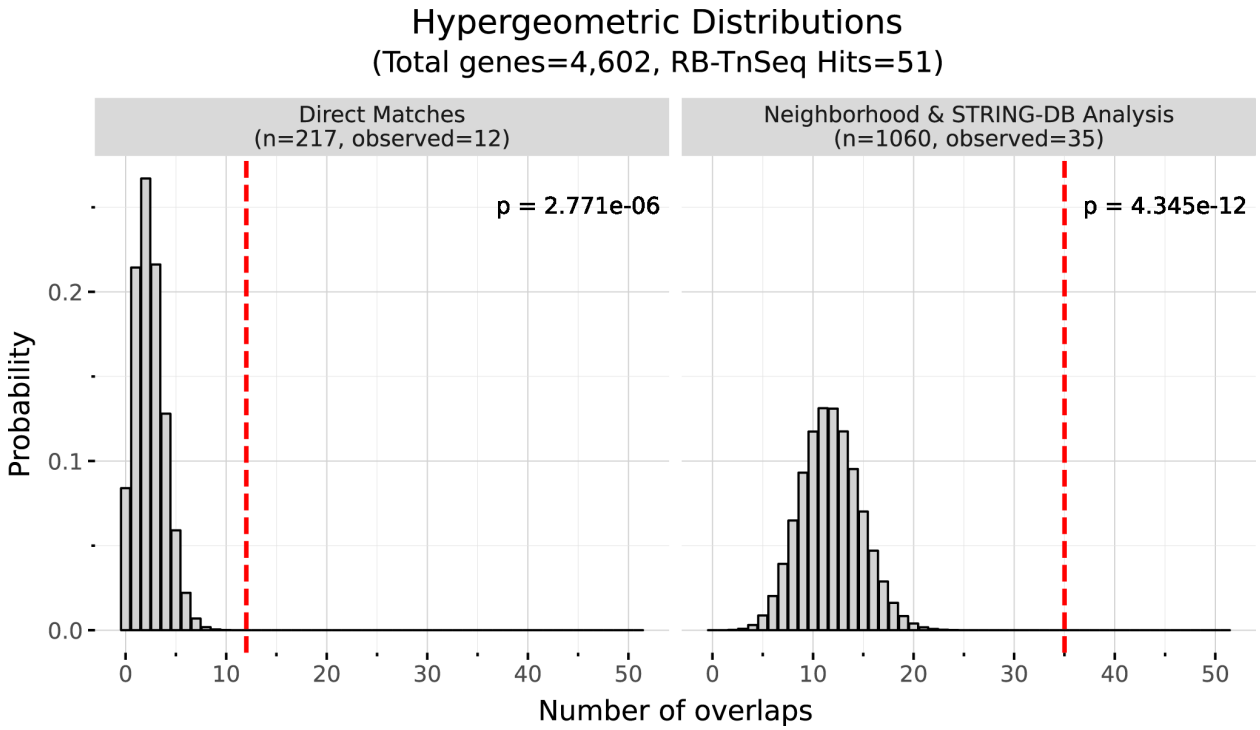
Supplementary Figure 33. Phage cocktail design in E. coli over 20-fold cross-validation. Bars show phage cocktail design success rate for all validation strains in 20-fold cross-validation. Bar height represents the fraction of possible phages selected in 5-phage cocktail designs, showing success rate regardless of how many infectious phages were present in the dataset. Bar color represents the total number of infectious phages selected in 5-phage cocktail.



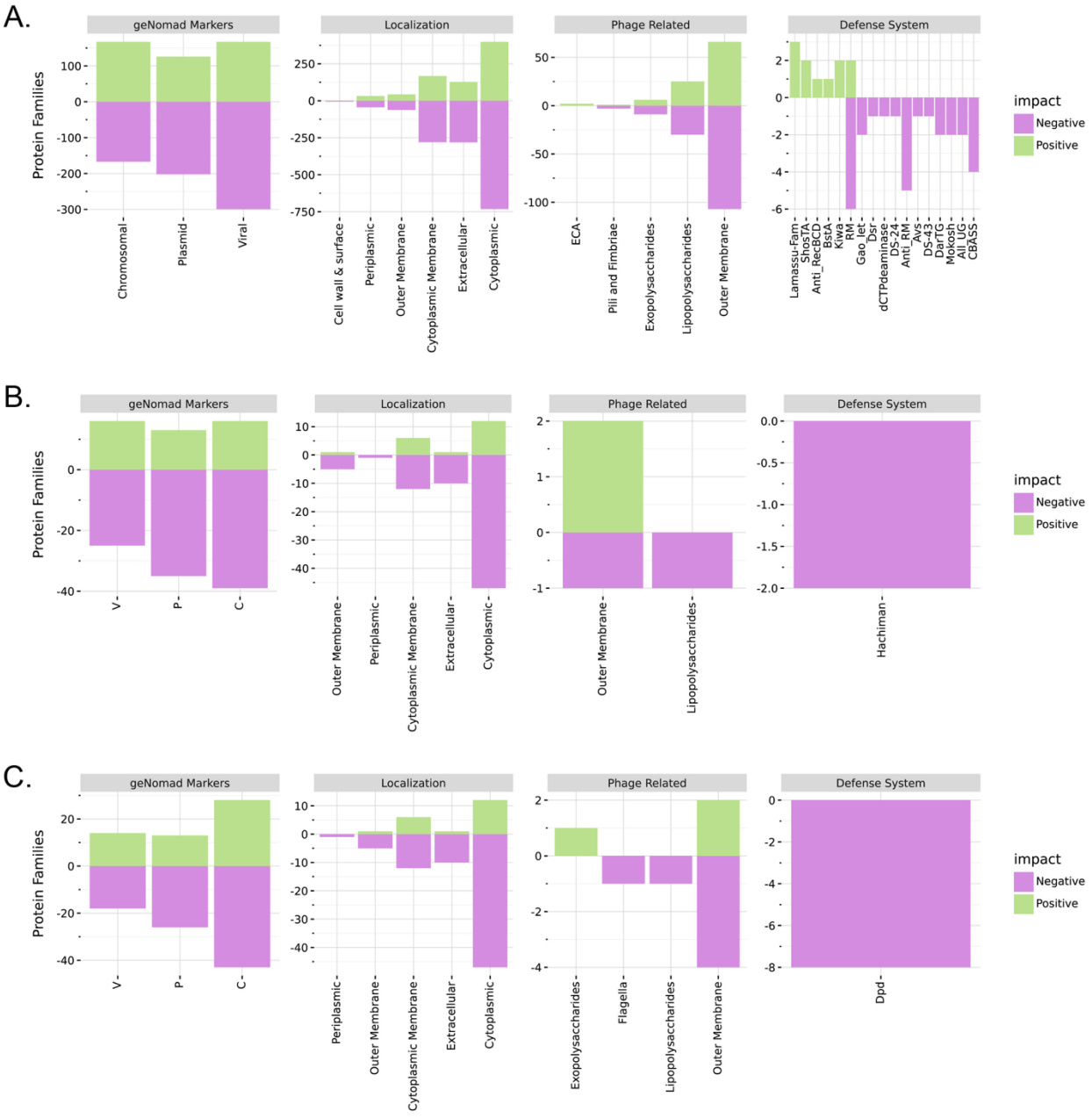
Supplementary Figure 34. RB-TnSeq hits in E. coli ECOR27 identified across 19 phages. Heatmap shows *E. coli* ECOR27 genes (X-axis) with strong fitness effects when submitted to one of 19 phages (Y-axis). Color represents fitness scores with positive in green and negative in blue.



Supplementary Figure 35. Mapping RB-TnSeq hits from E. coli ECOR27 to predictive features. The heatmap shows fitness values associated with all RB-TnSeq hits with links predictive features in *E. coli* ECOR27. The X-axis shows tested phages and the Y-axis gene names of RB-TnSeq hits ordered by Shapley Additive exPlanations (SHAP) feature importances. SHAP values indicate whether presence of a given feature (blue) or absence (orange) is associated with increased likelihood of infection (SHAP value > 0) or decreased likelihood of infection (SHAP value < 0). The colored bar shows relationship between RB-TnSeq hits and predictive features, including whether the predictive feature in directly linked to the gene in question (dark brown), linked through a neighborhood analysis (medium brown), or linked through STRING-DB (light brown). Gene names on the far right show the annotation associated with the predictive feature.



Supplementary Figure 36. Enrichment of RB-TnSeq hits in predictive features. Hypergeometric distributions show the expected overlap in predictive and RB-TnSeqs given random sampling. Dashed lines show observed overlap and associated p-values.



Supplementary Figure 37. Classification of predictive features based on relevance to known mediators of phage-host interactions. Proteins associated with predictive features were annotated and classified into having a positive (green) or negative (purple) impact of phage-host interactions across *E. coli* (A), *Klebsiella*-1 (B), and *Klebsiella*-2 (C) datasets.