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

Abstract

Bacteriophages offer promising alternatives to antibiotics for treating drug-resistant infections and engineering microbiomes, but applications are limited by inability to select phages infecting specific strains. Selecting suitable phages requires either one-to-one experimental assays or strain-level predictions of phage-host interactions. Existing computational approaches either predict host taxonomy at broad ranks unsuitable for strain-level targeting, or require species-specific mechanistic knowledge limiting generalizability. Here, we present a phylogeny-agnostic machine learning framework predicting strain-level phage-host interactions across diverse bacterial genera from genome sequences alone. Systematically optimizing the workflow over 13.2 million training runs across five datasets (128,357 interactions, 1,058 strains, 560 phages), we achieved performance matching species-specific methods (AUROC 0.67-0.94) while eliminating phylogenetic constraints. Comprehensive feature engineering identifies biologically interpretable genetic determinants while minimizing overfitting in sparse, imbalanced datasets. Experimental validation through 1,328 novel interactions confirmed generalizability (AUROC 0.84), while genome-wide RB-TnSeq screens verified that 68.8% of experimentally identified infection mediators were captured computationally, including receptors and cell wall biosynthesis pathways. Model-guided cocktail design achieved 80.7-97.0% strain coverage with five phages, and up to a 6.2-fold improvement in one-shot phage selection over promiscuity-based selection. This platform enables rational phage therapy design and precision microbiome engineering with applications in combating antimicrobial resistance across clinical, agricultural, and industrial contexts.

Keywords: *Bacteriophage-host interactions, Machine Learning, Phenotype-Genotype Prediction, High-throughput Screening*

Introduction

Bacteriophage host specificity is the byproduct of an evolutionary arms race, where phages continuously respond to evolving host receptors, cell wall structures and defense mechanisms [1-7]. In the context of phage therapies, this specific, multivalent 'recognition' system enables the targeting of individual bacterial strains while minimizing collateral impact on surrounding microbiota. This offers significant advantages over traditional broad-spectrum antibiotics in the treatment of bacterial infections and contamination [8, 9], representing a promising frontier in microbiome engineering, with applications spanning therapeutics, agriculture, and industrial microbiology. Phages could provide an alternative treatment option for antibiotic-resistant infections, which the World Health Organization has identified as one of the most pressing global health threats [10]. They may also be effective tools for basic science, providing precision targeting for microbiome editing to dissect mechanisms of microbial ecological interaction and function *in situ* [11].

The clinical and biotechnological application of phages depends on our ability to accurately select phages capable of infecting specific bacterial strains [8, 9, 12, 13]. Although their narrow host-specificity relative to antibiotics can reduce the impact of treatment on beneficial commensal bacteria, this also presents a challenge in identifying suitable therapeutic phages for specific targets [14]. Recent phage-therapy clinical trials and examples of compassionate use rely on either broad-spectrum phage cocktail formulations or experimental testing of phages on isolated pathogenic strains, commonly through plaquing assays [14]. As the size of phage banks increase to capture the breadth of bacterial hosts, it becomes less feasible to test phage susceptibility experimentally for a new target pathogen or contaminant [14, 15]. In situations where pathogens are not susceptible to broad host-range cocktails, predictive models offer strategies to quickly identify candidate phages across available phage banks from host genetic information [12, 13], enabling clinicians to focus efforts on high-probability phages. However, because of the complex, multifactorial mediators of phage-host interactions, robust predictive modeling of these interactions at strain-level requires high-quality interaction datasets capturing a range of representative host and phage phylogenies and their associated phage infection mechanisms [13, 16, 17].

Several tools have been developed to assign host taxonomies to bacteriophage genomes often derived from metagenomic sequencing data [18-23], representing by far the largest resource of phage genome sequences [24, 25]. These host classification tools can achieve species-level

resolution by identifying shared genomic sequences such as CRISPR spacers or by evaluating amino acid and nucleotide frequencies, showing strong performance at family and order-level classifications [22, 26-38]. This allows inference of general characteristics of phage lifestyle and host range, providing valuable scientific insights and expanding our understanding of phage biology and their impact on microbial community structure. However, reliance on databases of potential host strains means they do not support the prediction of individual phage-host interactions required for targeting specific bacterial strains [16, 17, 19, 39].

Predicting phage-host interactions at single-strain granularity requires datasets of confirmed one-to-one interactions of sufficient size and with appropriate variation of genetic components that document both successful and unsuccessful infections: data not provided by metagenomic sequencing approaches [16, 17]. These datasets are time-consuming to generate, often relying on plaquing assays or microplate-based liquid assays, as well as isolation and sequencing of individual strain and phage genomes [9, 16, 17, 40]. Results of these experiments have revealed that interaction datasets can be highly imbalanced, with only a small fraction of interactions resulting in infection, and significant variability in susceptibility even within a single host species [40, 41]. With as little as 2-5% of interactions resulting in infection in published *Klebsiella* and *Vibrionaceae* interaction matrices [40, 42, 43], this further increases the scale of experimental efforts required to generate sufficient representations of infectious interactions for model generalizability. Although, droplet and Hi-C-based strategies may increase throughput of phage-host interactions assays in the future [44, 45], experimental challenges currently impact the availability and breadth of relevant datasets, representing a critical limitation to our capacity to accurately model these relationships.

Individual labs often generate only a limited amount of data about a target host and phage, while focusing on specific host phylogenies [6, 16, 40, 42, 43, 46, 47]. Consequently, there are few published datasets of sufficient size for predictive modeling, and relatively few tools have been developed to specifically predict interactions between unrelated strain-phage pairs [13, 16, 17], none of which support phylogeny-independent prediction of interactions between novel strains and phages. These data limitations manifest in several ways that complicate predictive modeling. First, aggregating datasets across laboratories to increase statistical power introduces challenges related to experimental heterogeneity and differences in assay methodologies. Second, while leveraging transferable knowledge across phylogenetically diverse hosts and phages could expand model applicability by identifying conserved interaction mechanisms, this approach has yet to be extensively explored and must contend with substantial lineage-specific variation in genetic architecture. Third, even within well-studied species, uneven phylogenetic sampling could limit prediction accuracy for underrepresented strain lineages. It would therefore be advantageous to establish predictive modeling workflows that can leverage and combine relevant datasets generated across the research community, supporting phylogenetically diverse hosts and phages,

and the prediction of interactions in novel strains and phages. This approach would increase statistical power while enabling the learning of transferable knowledge across host and phage phylogenies, potentially capturing shared mechanisms of phage infection and resistance.

Machine-learning (ML)-based predictive modeling of phage-host interactions requires the representation of training data as numerical feature tables associated with a phenotype. Converting complex genomic data into numerical values results in high-dimensionality feature tables, presenting a risk of overfitting when training on small and imbalanced phenotypic datasets [13, 48]. These are well-documented challenges associated with whole-genome microbial phenotype prediction, where phenotypic sparsity, the identification of suitable genomic representation space, high-dimensionality of datasets, and challenges in distinguishing phenotypically causal and associated genetic features limit biological interpretation and model generalizability [48, 49]. Many existing tools for strain-level phage-host interaction prediction address these challenges by reducing feature-set dimensionality by selecting genetic features known to mediate these interactions [16, 17]. Although effective, this strategy limits the applicability of these workflows across diverse phage and host phylogenies and hinders the identification of novel features mediating their interactions [16, 17]. Some approaches also employ dimension-reduction steps that impede biological interpretability [16]. In the context of phage-host interaction prediction, many existing workflows rely on phage-specific models, which restricts predictions to whether unseen strains are infected by an existing set of phages [13, 16]. Collectively, these constraints limit the applicability of existing models in combatting diverse emerging pathogens and contaminants.

To address these limitations, we leveraged 5 published phage-host interaction datasets totaling 128,357 interactions, across 1058 strains and 560 phages [16, 40, 42, 43, 46] to develop a predictive modeling framework that incorporates versatile feature assignment, comprehensive feature selection, and ensemble-learning-based modeling, parameterized for the characteristics of phage-host interaction data and other sparse microbial phenotypic data. Our approach represents strain and phage genomes by generating a pangenome-like presence-absence matrix of protein or amino acid (AA) *k*-mer content. Systematic feature engineering and selection identifies predictive features while minimizing the risk of overfitting associated with high dimensionality and imbalanced datasets. This strategy can be applied to any phage and host phylogenies without requiring mechanistic knowledge of a given phage-host interaction. Simplified feature assignment also enables prediction of interactions with previously unseen bacterial strains and phages, demonstrating its potential for real-world applications in phage therapy and microbiome engineering.

In parallel, this strategy generates predictive features that can be associated with individual strain or phage genomes and linked to specific genetic elements. The filtering of phylogenetically-linked

genomic features minimizes the likelihood of identifying signatures of phylogenetic clades rather than functionally relevant mechanisms, enabling biological interpretation of model predictions. The presented modeling workflow identified known mediators of phage-host interactions, including host receptors, defense systems, phage receptor-binding proteins, and counter-defense mechanisms, while also suggesting potential novel determinants of phage susceptibility.

Here, we show that our approach can match or improve upon the performance of previous models trained on published datasets, without prior assumptions of important genetic features; demonstrate predictive capacity on new hosts; demonstrate how the models can guide phage cocktails design for previously unseen strains; and reveal predictive features that highlight proteins responsible for key specificity mechanism between phage and host, which could serve as targets for future engineering. Finally, we validated model performance experimentally by building a 1400-interaction phage-host interaction matrix and demonstrate the biological relevance of predictive features by performing RB-TnSeq genome-wide fitness assays in presence of panel of phages [7, 50]. Combined, this integrated approach provides a versatile framework to predict phage-host interactions at the strain level, with potential applications in precision medicine, microbiome manipulation, and biotechnological applications.

Results

Dataset overview

To develop a phylogeny-agnostic phage-host interaction prediction workflow, we identified five previously published phage-host interaction datasets [16, 40, 42, 43, 46]. The datasets needed to include both phage-host interaction matrices, where interactions were reported as, or can be converted to, binary phenotypes (infection or no-infection), and the genome sequences of associated phages and host strains (*Figure 1A*). These totaled 128,357 interactions, across 1058 strains and 560 phages (Table 1) (*Supp. Fig. 1-8*) (*Supp. Materials 1.1.1*). Four datasets were used for workflow development: an *E. coli* dataset (with a subset used for workflow development) (*Figure 1B*) [16], a mixed *Klebsiella* spp. dataset [42], a *Klebsiella pneumoniae* dataset [17, 43], and a *Pseudomonas aeruginosa* dataset (*Supp. Fig. 1*) [46]. The fifth, a mixed *Vibrionaceae* dataset (242 *Vibrio*, 8 *Enterovibrio*, and 6 *Shewanella* strains), was excluded from the workflow optimization process because of computational resource requirements but was reserved for validation [40].

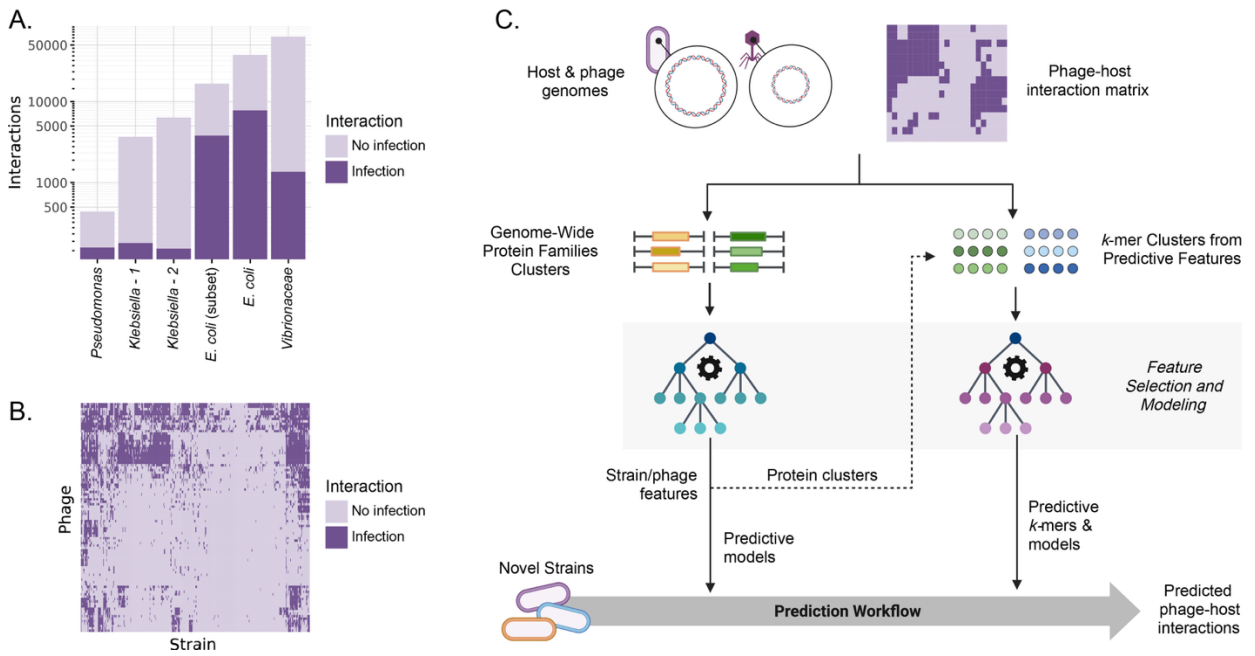

Figure 1. Phage-host interaction dataset and modeling workflow overview. A) Interaction datasets range from 437 to 63,488 total interactions, with proportions of infectious interactions ranging from 2.1% to 36.2%. B) The interaction matrix from the *E. coli* dataset includes 402 *E. coli* strains and 94 phages, with 20.7% of interactions resulting in infection. Clusters of strains and phages indicate conserved infection patterns within these groups. Dark purple represents interactions resulting in infection and light purple indicates interactions that did not result in infection. C) The predictive modeling workflow uses phage and strain genomes and phage-host interaction matrices to generate protein-family- or *k*-mer-based feature tables for feature selection and predictive modeling. Amino acid sequences from predictive proteins can also be used to generate *k*-mer features for further modeling. Protein-family and *k*-mer based features can be assigned to new strains and phages to predict novel interactions or can be used to explore biological mechanisms mediating phage-host interactions.

The four datasets used for workflow development totaled 27,081 interactions, including 4,304 positive interactions, and range from 23 strains by 19 phages in the *Pseudomonas* dataset to 177 strains by 94 phages in the *E. coli* dataset subset (*Figure 1B*). The percentage of infectious (positive) interactions ranged from 2.4% to 36.2% (Table 1).

199 *Table 1. Phage-host interaction dataset overview.*

Dataset ID	Host Taxonomy	Host Strains	Phages	Total Interactions	Positive Interactions	Positive Interactions (%)	Ref.
E. coli (subset)	*E. coli*	402 (177)	94	37788 (16638)	7813 (3813)	20.7 (22.9)	[16]
Klebsiella-1	*Klebsiella pneumoniae*	62	59	3658	180	4.9	[42]
Klebsiella-2	*Klebsiella*	138	46	6348	153	2.4	[43]
Pseudomonas	*Pseudomonas*	23	19	437	158	36.2	[46]
Vibrionaceae	*Vibrio* / *Enterovibrio* / *Shewanella*	256	248	63488	1364	2.1	[40]

200

201 *Workflow development and optimization*

202 Our objective was to develop a modeling workflow for strain-level prediction of phage-host
203 interactions applicable to phage therapy design, industrial biotechnology, and biological research.
204 The workflow needed to meet three key criteria: 1) predict interactions between individual strain-
205 phage pairs; 2) be independent of host and phage phylogeny without requiring prior knowledge
206 of genetic determinants of phage-host interaction; 3) enable predictions for novel strains or
207 phages based on genomic information alone, without model re-training. These criteria distinguish
208 our workflow from existing strategies that often assign host taxonomies to phages from an
209 existing host database, often at higher taxonomic ranks, preventing these models from being used
210 to evaluate specific phage-host interactions [18, 22, 27-38]. Modeling strategies that do predict
211 one-to-one phage-host interactions use prior mechanistic knowledge, limiting applicability to
212 single species with known mediators of phage infection [16, 17], do not directly address feature
213 assignment and prediction for new strains, impacting their use in targeting new hosts strains [16],
214 or generate phage-specific models, enabling only the prediction of infection by an existing set of
215 phages [13, 16]. Additionally, we considered secondary criteria, including biological
216 interpretability of predictive features and the applicability of resulting prediction to phage
217 selection and phage cocktail design.

218 To meet these criteria, we performed an extensive modeling parameter optimization experiment,
219 evaluating over 200 parameters settings and training more than 13.2 million individual predictive
220 models. This enabled optimization of multiple modeling processes, including feature engineering
221 and selection, predictive model training, and feature assignment to novel strains and phages.
222 Specifically, workflow optimization involved testing genomic representations including protein-
223 family clustering strategies (*Supp. Fig. 9-10*) and amino acid (AA) *k*-mer lengths (*Supp. Fig. 11*)
224 [51, 52], modeling algorithms (*Supp. Fig. 12-13*) [53, 54], feature selection algorithms (*Supp. Fig.*
225 *14*) [53, 55], ensemble-learning approaches (*Supp. Fig. 15-19*), feature table filtering and

engineering strategies, and training strategies (*Supp. Fig. 20*) [53], such as class weights and clustering strategies for train-test splits [53, 56, 57] (*Supp. Materials 1.1.2*). To quantify the overall importance of each methodological decision, we calculated the performance delta (Δ) between the worst- and best-performing configuration within each parameter. This showed that modeling algorithm (Δ MCC = 0.365), training strategies (Δ MCC = 0.203), and feature selection method (Δ MCC = 0.184) had the largest impact on performance, while genomic representation parameters, including protein-family clustering thresholds (Δ MCC = 0.072) and *k*-mer length (Δ MCC = 0.020) showed limited effects (*Table 2* / *Supp. Fig. 20*).

Table 2. Impact of modeling parameters on predictive performance.

Parameter	Δ MCC	Δ AUROC	Configurations per dataset
Modeling Algorithm	0.365	0.148	8
Training Strategies (class weights / feature filtering)	0.203	0.110	14
Feature Selection Algorithm	0.184	0.086	5
Modeling / Feature Selection Iterations	0.124	0.069	156
MMSeqs2 Clustering Threshold	0.072	0.031	25
k length	0.066	0.020	13

To select a final parameter set, we employed rank-order analysis across datasets, with performance differences assessed using Friedman tests and pairwise Mann-Whitney U comparisons. This identified configurations that consistently maximized Matthews Correlation Coefficient (MCC) across phylogenetically diverse bacterial genera. The optimized workflow uses MMSeqs2-based protein family clustering (sequence identity 0.4, coverage 0.8), CatBoost gradient-boosted decision tree modeling, 25 rounds of RFE-based feature selection, 50 modeling iterations for ensemble-learning, and a *k*-mer length of 4 in the standalone *k*-mer workflow and 6 when combining with predictive protein families. To address dataset imbalance (2.1 %-36.2% positive interactions) and minimize overfitting, we implemented inverse-frequency-based phage-specific class weights, hierarchical clustering of strains (20 clusters) during train-test split, and feature filtering that maintains only features present across multiple strain clusters, removing features with strong phylogenetic linkage (*Supp. Fig. 20*). These optimization decisions produced a robust modeling pipeline to handle the high degree of sparsity, imbalance, and feature table dimensionality characteristic of phage-host interaction and genomic datasets (*Figure 1A-B* / *Supp. Fig. 21*). We then systematically evaluated model performance and biological interpretability across the five published datasets.

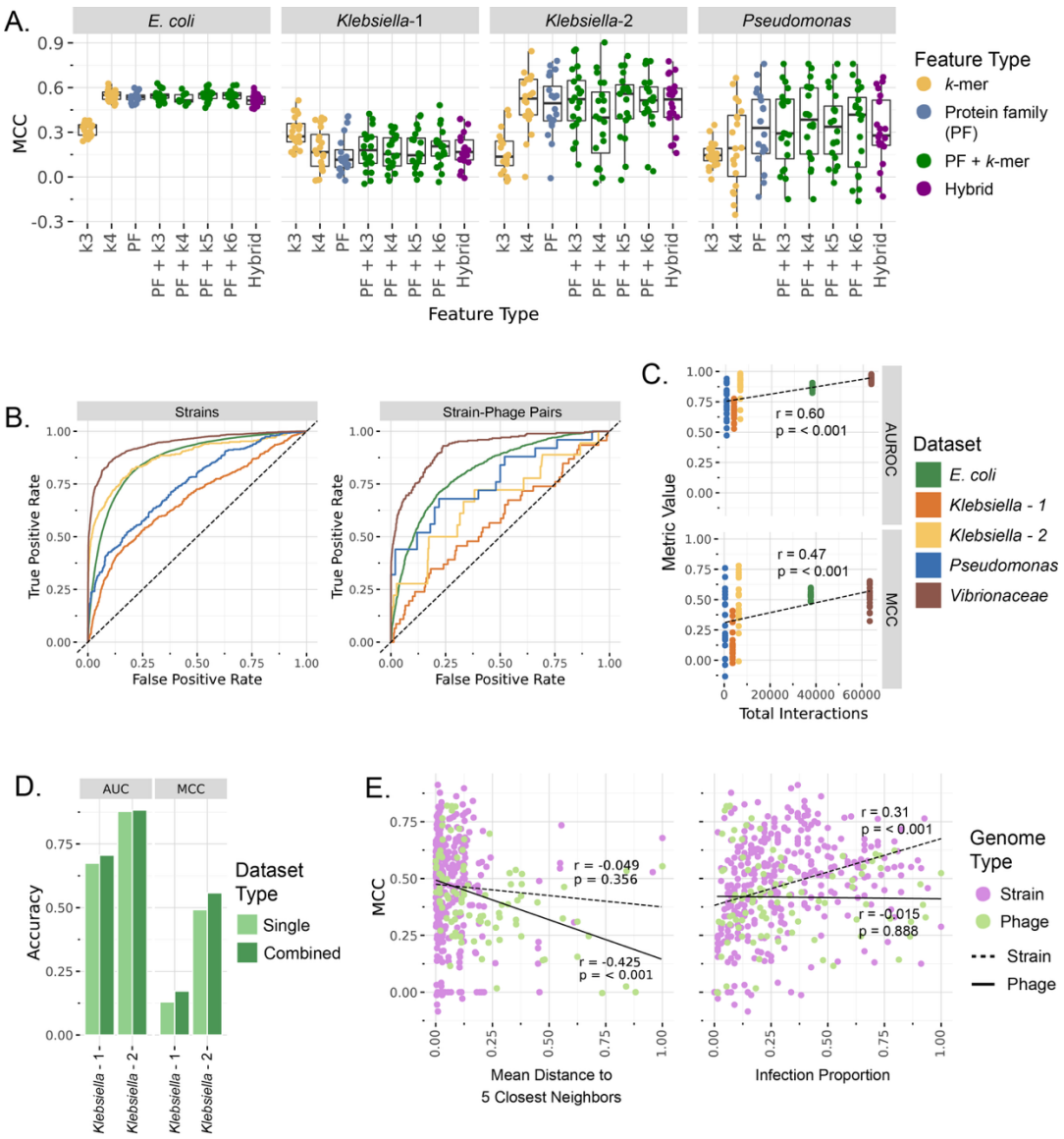

Figure 2. Phage-host interaction prediction performance in 20-fold cross-validation. A) Performance of models (MCC) across genome representations (*k*-mer: yellow / protein family (PF): blue / protein families with *k*-mers from predictive proteins: green / hybrid with protein families for strain and *k*-mers for phage: purple) show no significant differences with $k > 3$. B) Receiver operating characteristic (ROC) curves show prediction performance variability across datasets, when predicting phages infecting previously unseen strains ("Strain") and interactions between previously unseen strain-phage pairs ("Strain-Phage Pairs"). Colors represent datasets. The dashed line indicates performance of random predictions (AUROC=0.5) C) Dataset size correlates with prediction performance. D) Combining datasets (dark green) leads to improved performance across *Klebsiella* datasets, in comparison to single-dataset models (light green). E-1) The relationship between phylogenetic isolation and strain / phage level performance shows no correlation between strain (purple) performance and phylogenetic isolation (dashed line) and a significant negative correlation between phage (green) performance and phylogenetic isolation (solid line). E-2) A significant positive correlation was observed between strain (purple) performance and susceptibility (dashed line) and no correlation between phage (green) performance and host range (solid line).

Model performance

When evaluating performance, we tested three model configurations that reflect possible applications in the laboratory or the clinic. Each of these dictate train-test splitting paradigms during feature selection and training, and were tested in 20-fold nested cross-validation experiments. The first configuration represents the most likely application, where users want to select a phage from an existing phage set ("phage bank") that infects a novel strain, such as in the context of phage therapies or microbiome engineering. For cross-validation of these models, 10% of strains are excluded from the entire feature selection and modeling workflow and features are assigned to these strains for prediction with resulting models. During feature selection and training, strains are split across train and test datasets, whereas phages are maintained. The second mirrors this strategy, leaving out 10% of phages from the modeling workflow and predicting infection of a characterized set of strains by previously unseen phages. This reflects a scenario where researchers want to select a suitable host-strain for a target phage. Here phages are split across train and test datasets, with strains maintained. The final cross-validation workflow leaves out 10% of strains and 10% of phages, enabling predictions in completely unseen strain-phage pairs. For all validation sets, features were assigned to held-out genomes using similarity search against training protein clusters (MMSeqs2, sequence identity ≥0.4, coverage ≥0.8), ensuring feature assignment reflected realistic inference conditions. Model performance varied across datasets for all model configurations (*Supp. Fig. 21*)., but with consistent relationships between dataset size and model performance (*Supp. Fig. 22-24*).

First, we evaluated the performance impact of genomic representation strategies by comparing 20-fold nested cross-validation performance of protein-family and *k*-mer-based models using the primary model configuration. Here we observed no significant differences in performance between models, suggesting that all representations captured sufficient biological information for predicting phage-host interactions (*Figure 2A*). Although *k*-mer-based representations may provide interesting insights into specific functional residues and sequence variants, they often identified *k*-mers that are shared across many genes, complicating biological interpretation with no performance improvement. For this reason, we focus on protein family-based models moving forward, but believe that alternative representations could provide value in other contexts, such as with small genomes or when a specific subset of relevant genes has already been identified, and should be considered when training predictive models.

When predicting phages infecting novel strains, simulating the process of designing phage therapies, area under the receiver operating characteristic curve (AUROC) ranges from 0.67 to 0.94, Matthews Correlation Coefficient (MCC) from 0.13 to 0.54, and normalized area under the precision-recall curve (AUPR) of 0.07 to 0.60 (*Figure 2B - Strains / Supp. Fig. 21A / Table 2*). For *E. coli*, our models achieved an AUROC of 0.87, closely matching the performance reported by the

original study (AUROC = 0.86) despite fundamental differences in approach. Notably, the published method required phage-specific models and host features based on known *E. coli* mediators of phage infection [16], while our approach achieved comparable performance without these constraints, enabling application to novel phage-host systems. Similarly, when predicting hosts for novel phages, AUROC ranged from 0.69 to 0.93, MCC from 0.18 to 0.68, and normalized AUPR of 0.10 to 0.87 (*Supp. Fig. 21B / Supp. Table 1*). The final model configuration, predicting interaction between unseen strain-phage pairs, was most sensitive to dataset size, yielding AUROC values of 0.58 to 0.92, MCC of 0.07 to 0.44, and normalized AUPR of 0.03 to 0.58 (*Figure 2B - Strain-Phage Pairs / Supp. Fig 21C / Supp. Table 2*).

Across all cross-validation experiments, we observed significant positive correlations between model performance and dataset characteristics, including both total and infectious interaction counts (*Figure 2C / Supp. Fig 22-24*). When predicting phages infecting unseen hosts, performance increased with dataset size, showing Pearson correlation coefficients of $r = 0.60$ for AUROC and $r = 0.47$ for MCC (both $p < 1\times10^{-6}$). Positive interaction counts also correlated with performance, with $r = 0.28$ for AUROC ($p = 0.0057$) and $r = 0.36$ for MCC ($p = 2.0\times10^{-4}$). By contrast, the proportion of positive interactions showed a marginally significant negative correlation with AUROC ($r = -0.26$, $p = 0.0081$) and no correlation with MCC ($r = -0.20$, $p = 0.472$). This slight decrease in performance with an increased proportion of infectious interactions likely reflect characteristics of our specific datasets, including the high performance of the large *Vibrionaceae* dataset (63,488 interactions; AUROC = 0.94 and MCC = 0.54) with only 2.2% positive interactions and the variable performance of the small *Pseudomonas* dataset (437 interactions; AUROC = 0.75 and MCC = 0.33) with 36.2% positive interactions (*Supp. Fig 22*). Overall, these results suggest that a sufficient total interaction count can mitigate the negative performance impact of dataset imbalance. Standard deviation in MCC and AUROC across rounds of cross-validation also decreased with dataset size, showing increasing stability and reliability in larger datasets (*Supp. Fig 21*).

Table 3. Model performance across datasets when predicting infection of unseen strains.

Dataset	AUROC	MCC	Normalized AUPR	Brier score
Vibrionaceae	0.941	0.533	0.602	0.017
Klebsiella-2	0.878	0.492	0.464	0.017
E. coli	0.869	0.536	0.562	0.125
Pseudomonas	0.745	0.326	0.504	0.221
Klebsiella-1	0.674	0.130	0.069	0.061

333 We next evaluated the impact of training models on combined datasets both across
334 phylogenetically diverse hosts and within single host genera. The cross-genus combination
335 included *E. coli*, *Klebsiella*, and *Pseudomonas* datasets (*Supp. Fig. 25*), while within-genus
336 combination merged the two *Klebsiella* datasets (*Supp. Fig. 26*). Combining *Klebsiella* datasets led
337 to slight but significant performance improvements in the *Klebsiella*-2 dataset (p = 0.024) and a
338 non-significant improvement in the *Klebsiella*-1 dataset (p = 0.398) compared to single-dataset
339 models. This represents MCC increases of 0.04 and 0.07 over single-datasets models, despite the
340 relatively small 10,006 interaction combined dataset (*Figure 2D* / *Supp. Fig. 26A-B*). In these
341 combined *Klebsiella* models, there was no significant correlation between the strain-level
342 increase in predictive performance and the proximity of the closest relative added by the
343 additional data, suggesting that this performance improvement cannot be explained by the
344 addition of highly similar strains (*Supp. Fig. 26C-D*). This indicates a more complex relationship
345 between genotype and phage-host interaction phenotype than phylogenetic relatedness. The
346 AUROC of 0.812 of this combined model also nearly matches the performance of models reported
347 by a previous study (AUROC = 0.818), but whose method is heavily dependent on *Klebsiella*-
348 specific knowledge using only *K*-locus genes and phage tail fibers [17]. This again indicates our
349 workflow can provide comparable performance to alternative workflows without requiring
350 phylogeny-specific knowledge of the mediators of phage-host interactions. Unlike within-genus
351 combinations, merging cross-genus datasets did not significantly improve performance of any
352 dataset (*Supp. Fig. 25A*). This response likely relates to the sharing of predictive features across
353 datasets and the requirement for a minimum set of shared features for accurate predictions.
354 Jaccard distance analysis based on feature content revealed clear separations between host
355 phylogenies (*Supp. Fig. 25B*), with more overlap observed between phages (*Supp. Fig. 25C*),
356 suggesting that sufficient feature overlap is required for performance gains through dataset
357 combination. The fraction of predictive features that were shared across genera were often
358 associated with either core microbial metabolic processes or mobile genetic elements (MGEs),
359 suggesting horizontal transmission of phage defense systems or previous cross-genera phage
360 infection (*Supp. Fig. 27*). These shared features may represent interesting conserved mechanisms
361 of phage infection and defense, but their apparent limited impact of predictive performance
362 indicates that phylogeny specific genetic features may play a more critical role.

363

364 *Modeling Error Analysis*

365 Although these models performed well across phylogenetically distinct datasets, understanding
366 the sources and patterns of prediction errors will help identify model limitations, directing both
367 their downstream application and strategies for model improvement. To evaluate prediction
368 error, we conducted systematic error analysis focusing primarily on the *E. coli* dataset, examining

bias patterns, the relationship between phylogenetic distribution and performance, and interaction pattern fidelity (*Supp. Materials - 1.1.3*).

Overall, this analysis showed a general over-prediction bias, specifically in strains with narrow susceptibility profiles ($r = -0.207$, $p = 8.3 \times 10^{-5}$) (*Supp. Fig. 28*). This is consistent with a positive correlation observed between strain susceptibility and model performance ($r = 0.310$, $p = 2.3 \times 10^{-9}$), indicating that models perform better on more susceptible strains (*Supp. Fig. 29*). 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 often 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 ($r = -0.425$, $p = 1.9 \times 10^{-5}$) (*Supp. Fig. 30*). This relationship indicates that models performed less well when predicting infection by more distantly-related phages, highlighting the importance of phage diversity in interaction datasets and the possible performance benefit of phage phylogenetic redundancy. This could also be a limitation of the modeling dataset that included 402 *E. coli* strains, in comparison to only 94 phages.

Despite these limitations, high-confidence errors (false-negatives with predicted likelihood of infection of 0-0.1 and false-positives with predicted likelihood of infection of 0.9-1) were rare, totalling only 1.5% of predictions (*Supp. Fig. 28*). This suggests most failures occur in uncertain boundary cases rather than confident misclassifications. Calibration curves showing predicted vs. observed infection rates (*Supp. Fig. 31*) and Brier scores (*Table 1* / *Supp. Table 1-2*) support this observation, with scores under 0.15 for all datasets and model configurations except the 437-interaction *Pseudomonas* dataset. In all but the low-performing *Klebsiella*-1 dataset, high-confidence predictions are suitable targets for experimental validation, with observed infection rates between 80-95% when model prediction confidence ≥ 0.9. High AUROC values further demonstrate that models successfully rank phages by infection probability (*Figure 2B / Supp Fig. 21*). Together, these results demonstrate that this workflow identifies predictive genetic features from whole genomes and that prediction confidence can effectively guide phage selection for engineering and therapeutic applications.

Phage Selection and Phage Cocktail Design

Applying predictive models to microbiome engineering or phage therapy involves selecting phages targeting a specific host strain based on model predictions. As models generate

404 continuous probability scores (0-1) representing the likelihood of phage infection in each host,
405 these probabilities must be translated into actionable phage selection strategies. In the context
406 of phage therapies, combining multiple phages into "phage cocktails" simultaneously increases
407 infection probability through multiple independent mechanisms while reducing the likelihood
408 that target bacteria possess or can evolve resistance to all cocktail components [13, 15, 39]. We
409 evaluated model-based phage selection strategies and phage cocktail design by conducting *in*
410 *silico* experiments where models selected phage sets to target unseen strains from our cross-
411 validation holdouts (*Supp. Fig. 33*).

412 Previous studies have demonstrated that effective phage cocktail design requires balancing both
413 phage infectivity and genetic diversity, selecting phages with complementary host interaction
414 mechanisms to maximize strain coverage while minimizing resistance development [13, 15, 16,
415 39]. This combination of phages simultaneously helps ensure targeting by at least one phage,
416 while mitigating the likelihood of escape by providing multiple selective pressures. Based on
417 these principles, we assessed three phage cocktail design strategies incorporating predicted
418 infection probability and phage mechanistic diversity: (1) ranking-based selection of the top *n*
419 phages by predicted infection probability, (2) diversity-based selection using hierarchical
420 clustering of phages followed by selection of the highest-ranked phage from each of *n* clusters,
421 and (3) unsupervised density-based clustering using HDBSCAN to identify natural phage
422 groupings and selecting top-ranked phages from the *n* highest-scoring clusters (*Figure 3A-B*) [57].
423 For clustering approaches, phages were grouped based on their predictive feature profiles,
424 ensuring cocktails contained phages with diverse genetic determinants of host interaction.

425 Phage selection and cocktail performance was measured as the proportion of target strains for
426 which at least one selected phage achieved infection. This does not directly test whether selected
427 phages will decrease the likelihood of escape when used in combination, but allows us to evaluate
428 whether we can mitigate model uncertainty and increase the probability of selecting at least one
429 infectious phage. We tested single phage selections alongside 3-phage and 5-phage cocktails,
430 comparing performance against baseline cocktails composed of the most promiscuous phages
431 (highest infection rates across training strains). Performance was evaluated across 20-fold cross-
432 validation, maintaining the same strain holdout strategy used for individual interaction
433 predictions. Model-guided cocktail design outperformed promiscuity-based selection in the three
434 largest datasets (*E. coli*, *Klebsiella*-2, and *Vibrionaceae* datasets) across 1-, 3-, and 5-phage
435 cocktails (*Figure 3C*). For one-shot phage selection, models identified infectious phages for 66.9%
436 of *E. coli* strains, 66.7% of *Klebsiella* strains, and 88.3% of *Vibrionaceae* strains, equating to 1.1-
437 fold, 3.4-fold, and 6.2-fold increase over promiscuity-based cocktails, respectively. For 5-phage
438 cocktails, models selected infectious phages for 97.0%, 80.7%, and 94.5% of *E. coli*, *Klebsiella*, and
439 *Vibrionaceae* strains, respectively, showing consistently high performance. Here, HDBSCAN
440 clustering-based cocktail design consistently outperformed ranking-based selection in these high-

performing datasets, demonstrating that selecting genetically diverse phages with high infection probability can increase the likelihood of identifying suitable phages. These results demonstrate that predictive models can be effectively integrated into a phage selection and cocktail design workflow, and that models can identify predictive features that meaningfully improve cocktail design beyond simple promiscuity-based heuristics.

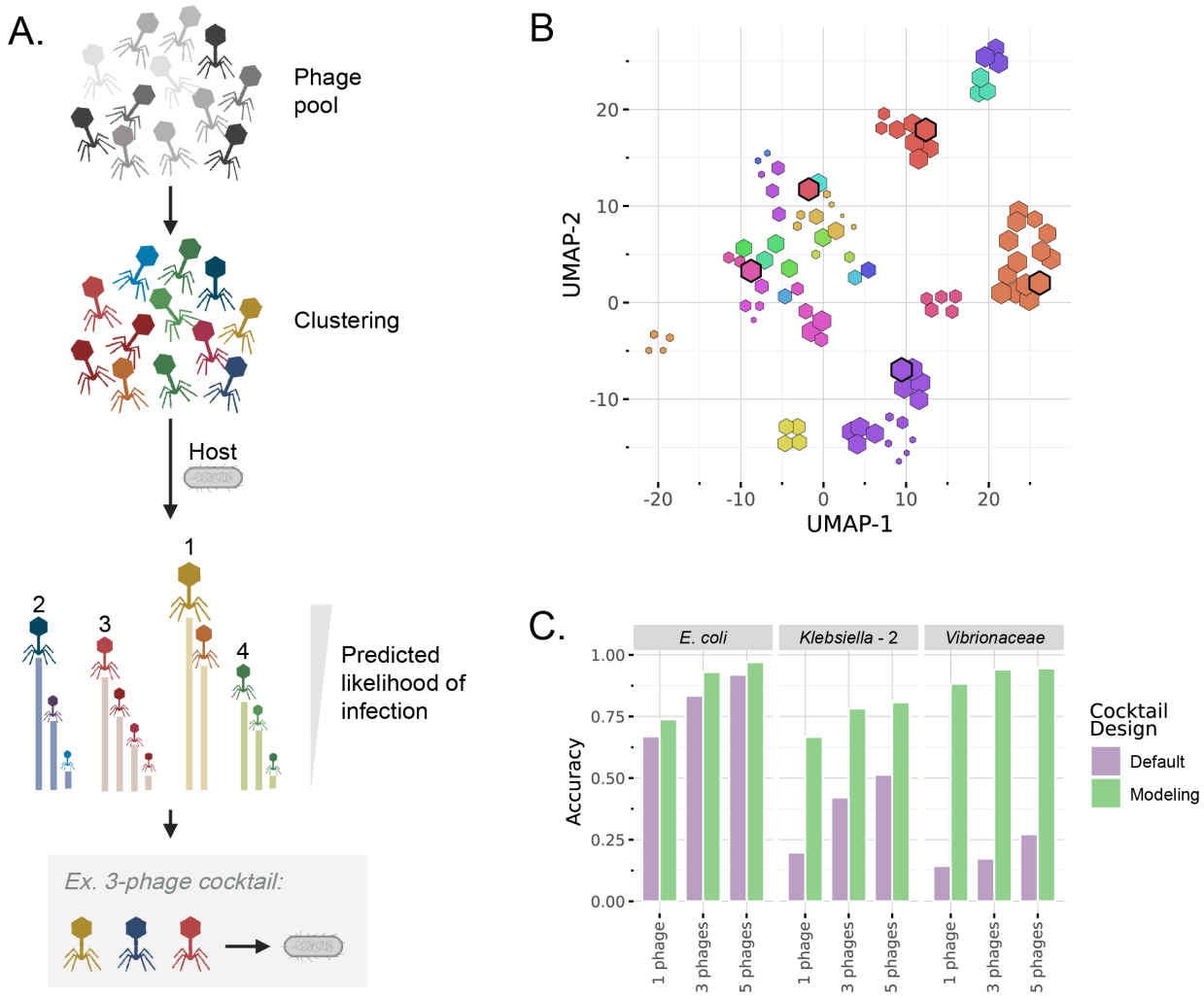

Figure 3. Phage-host cocktail design workflow and in-silico validation. A) The phage-cocktail design workflow involves HDBSCAN clustering of phages based on predictive feature content, ranking of phages within each cluster based on model predictions and selecting the top performing phages from *n* clusters. B) UMAP projection shows the distribution of phages sized by their predicted likelihood of infecting *E. coli* ECOR13. Colors represent HDBSCAN phage clusters and selected phages are highlighted in black. C) Phage cocktail based on model predictions (green) significantly outperformed cocktails based on phage promiscuity (purple) across the 3 largest and highest-performing datasets.

456 *Experimental Validation*

457 Although cross-validation experiments showed strong performances across multiple datasets, for
458 this modeling workflow to be broadly applicable by researchers, clinicians, and engineers we need
459 to confirm that it is identifying generalizable genetic features that can predict phage-host
460 interaction in multiple experimental contexts. To validate model generalizability and the biological
461 relevance of predictive genetic features, we generated two complementary high-throughput
462 experimental datasets. The first is a 1,400-interaction matrix testing 56 novel phages (*Supp. Table*
463 *3*) against 25 *E. coli* strains present in the modeling dataset. The second was a genome-wide RB-
464 TnSeq screen testing the importance of 3,804 gene knockouts in *E. coli* ECOR27 to infection by a
465 panel of 21 phages (*Supp. Table 3*).

466 The interaction matrix allowed us to directly test predictive performance of the *E. coli* dataset
467 models on experimental data generated in a distinct laboratory environment. As we did not have
468 access to any of the phages used in the published dataset, we could not test the primary model
469 configuration selecting phages infecting unseen strains. We instead selected 25 *E. coli* strains
470 belonging to the ECOR collection that were included in the published dataset (*E. coli* ECOR13 -
471 ECOR38) and predicted infection of these strains by a set of 56 previously unseen phages
472 belonging to the BASEL phage collection [5, 16, 58-60]. Spotting assays were then performed with
473 these phages on all 25 *E. coli* strains, resulting in 1,400 tested interactions of which 72 (5.1%) did
474 not show a clear phenotype (*Figure 4A*). We then compared experimental plaquing phenotypes
475 of the remaining 1,328 interactions to predicted infections, resulting in an AUROC of 0.84 (*Figure*
476 *4B*). This is slightly lower than the cross-validation performance (AUROC = 0.87) but shows the
477 capacity of the models to generalize to unseen phages and to predict phage-host interactions in
478 experimental data generated in a distinct laboratory setting with different experimental
479 protocols.

480 To validate the biological relevance of the predictive features, we performed an RB-TnSeq genetic
481 screening experiment in *E. coli* ECOR27, which was present in both the modeling dataset and our
482 interaction matrix described above. In these experiments, pooled genome-wide mutant libraries
483 are subjected to phages, allowing us to identify host factors associated with phage infection
484 through the increased fitness of their mutants (*Supp. Fig. 34*) [7]. Although these high-throughput
485 genetic screening experiments are known to enrich for phage receptors, whose regulation has
486 the most significant impact on host fitness [7, 44], this comparison revealed that our workflow
487 successfully identified multiple screening hits, related regulators, or genes within the same
488 operon or gene cluster as predictive features. The *E. coli* ECOR27 library was screened against 21
489 phages, identifying 51 genes as high scoring hits for phage infection in this host (*Supp. Fig. 34*).
490 We then compared these RB-TnSeq hits to predictive features, ranked based on SHapley Additive
491 exPlanations (SHAP) values as a metric of model feature importance. Of the 51 genes identified

through RB-TnSeq, 12 mapped directly to predictive features, 10 fell within a 3-gene window of a predictive feature (4 directly adjacent), and 13 were functionally related to predictive features through STRING-DB [61], totaling 35 predictive genes with links to RB-TnSeq hits (68.8%) (*Figure 4C* / *Supp. Fig. 35*). With 155 features assigned to *E. coli* ECOR27 corresponding to 217 genes (1060 after neighborhood and STRING-BD analysis), this represents significant enrichment in mechanistically important genes, with direct matches 5.0-fold above random expectation ($p = 2.8 \times 10^{-6}$) and a 3.0-fold enrichment for the neighborhood and STRING-DB extended set ($p = 4.3 \times 10^{-12}$) (*Supp. Fig. 37*). We considered gene neighborhood and STRING-DB relationships because functionally related genes are often co-inherited or co-maintained within genomes [62-64] making models incapable of distinguishing them during feature selection as these have equal predictive power. Although multiple genes in single gene clusters often all showed fitness phenotypes in RB-TnSeq data, we observed that the predictive models would frequently only identify a single member of these pathways, including regulators for porins *tsx*, *fadL*, and *nfrA*, as well as single genes in LPS, O-antigen, and capsule biosynthesis clusters (*Figure 4C*). This pattern demonstrates that the models may capture pathway- or operon-level regulatory mechanisms rather than just direct gene effects. Our comparison to RB-TnSeq hits demonstrated that the predictive modeling workflow was particularly effective at identifying features associated with cell-wall biosynthesis pathways, including LPS, O-antigen, cellulose, and capsule biosynthesis, whereas porins were often represented by regulators or adjacent genes. We also compared annotations of predictive features to published RB-TnSeq, DubSeq, and CRISPRi experiments, which identify host genes affecting phage infection through transposon mutagenesis, overexpression, and inactivation screens, respectively [7]. This revealed multiple additional screening hits or related genes that our workflow successfully identified as predictive features. These included *aes*, *bcsA* (*bcsA* / *bcsE* / *bcsF* / *bcsQ*), *ycgF* (*bluR* anti-repressor), *uvrY* (*csrA* regulator), *cyaA*, *dsbB* (*dsbA*), *fhuA*, *ydiV* (negative regulator of *flhD*), *gadC*, *glmS*, *macB* (*macA*), *malT* (*malX* / *malQ*), *mdtB* (*mdtB* / *mdtF*), *nfrA*, *ompC*, *ompF*, *ompT*, *ptsG*, *rpiR* (negative regulator of *rpiB*), *cytR* (negative regulator of *tsx*), *waaY*, *yhbX*, and *yphG*. Of these, multiple are porins known to be the primary phage receptors (*fhuA, nfrA*, *ompC*, *ompF*, *ompT*) or important contributors to *E. coli* outer-membrane structure (*bcsA*, *waaY*).

These complementary experimental validations demonstrate that our approach generalizes effectively to novel datasets while capturing biologically meaningful determinants of phage infection, providing confidence that computationally identified features represent genuine mediators of phage-host interactions and that resulting models could be valuable tools in phage cocktail design and targeted microbiome engineering.

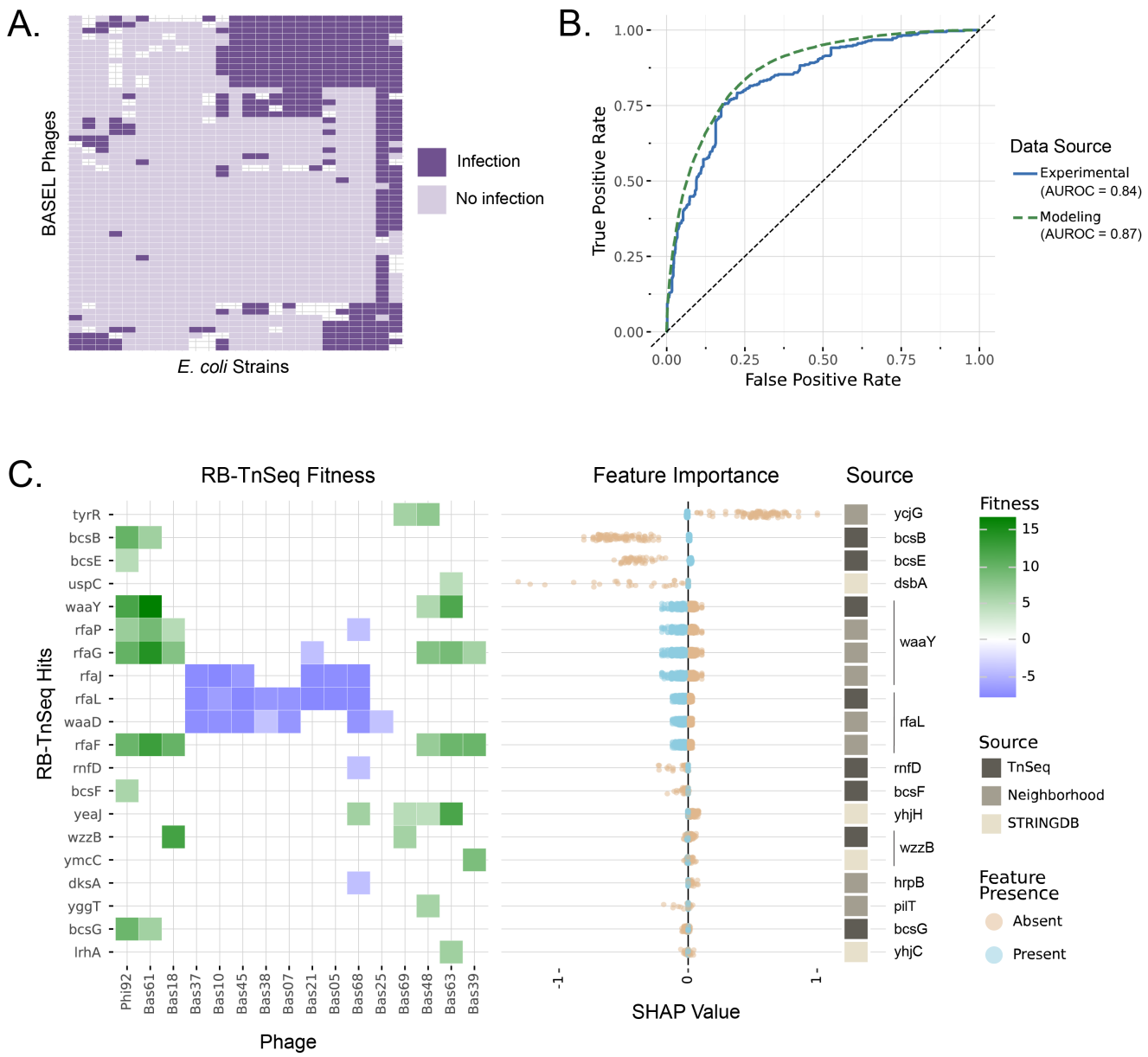

Figure 4. Experimental validation of model performance and predictive features. A) A 1400-interaction one-to-one phage-host interactions matrix showing results of spotting assays. Dark purple shows interactions resulting in infection and light purple shows interaction not resulting in infection. Missing values indicate interactions where phenotypes were unclear. B) ROC curve showing performance of hosts predictions for previously unseen BASEL phages (blue) in comparison to the 20-fold cross-validation performance (dashed green). C) The heatmap shows fitness values associated with 20 RB-TnSeq hits linked to top 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.

542 *Biological Interpretation*

543 A key objective of this modeling workflow was to enable biological interpretation of predictions
544 by identification of specific genetic determinants of phage-host interactions. Protein family
545 features allow direct mapping of predictive features to genes and gene variants that may mediate
546 infection. Having demonstrated through RB-TnSeq experiments that computationally identified
547 features correspond to functionally important genes, we sought to understand the mechanistic
548 basis of these predictions and the relationship between these features and known mediators of
549 phage-host interaction across datasets.

550 We leveraged SHAP values to quantify feature importance across datasets and classify model
551 features based on their predicted impact on infection [55, 65]. Features were categorized as
552 having a "positive" impact when their presence increased infection probability or a "negative"
553 impact when their presence decreased infection probability. To systematically interpret these
554 findings, we used established bioinformatics workflows to classify host proteins associated with
555 predictive features based on genomic origin (viral, plasmid, or chromosomal) [26], cellular
556 localization (cytoplasm, membrane, or extracellular) [66], relationship to known phage-host
557 mechanisms, and classification as defense or anti-defense systems [67] (*Supp. Fig. 37*). Across the
558 four development datasets, we did not observe an enrichment of features with either "positive"
559 or "negative" impacts across any feature functional classifications, apart from defense and anti-
560 defense systems where defense systems typically had negative impacts on infection while anti-
561 defense systems had positive impacts.

562 Analysis of the top 25 most predictive features in the *E. coli* dataset revealed various functional
563 categories associated with known phage-host interaction mechanisms (*Figure 5A*). Three of these
564 features were directly associated with RB-TnSeq data in *E. coli* ECOR27, relating to cellulose, O-
565 antigen, and LPS biosynthesis. Of the unvalidated features, the two that related directly to cell
566 wall biosynthesis showed varied effects, with presence of the O-antigen biosynthesis gene *rfbB*
567 increasing infection probability and the absence of the peptidoglycan biosynthesis gene *agaR*
568 decreasing infection probability. Two features associated with genes of viral origin could either
569 promote or inhibit infection, potentially reflecting superinfection exclusion mechanisms versus
570 indicators of host susceptibility. Seven features were related to mobile genetic element and two
571 with components of RM systems, showing varying impacts on infection likelihood. A total of eight
572 features showed no clear relationship to known phage-host interaction mechanisms, potentially
573 representing novel mediators requiring further investigation.

574 To explore possible phage-infection mechanisms associated with the top two predictive features,
575 we employed the same STRING-DB and gene neighborhood analysis used to link predictive
576 features to RB-TnSeq hits. We found that *goaG* (also referred to as *puuE*) and *gabT* are associated
577 with putrescine catabolism (*Figure 5B*). Putrescine has been shown to prevent bacteriophage

replication in *Pseudomonas aeruginosa* and may act as a host signaling molecule during phage infection [68], indicating that this phenotype may be conserved in *E. coli*. A gene neighborhood analysis showed that *ybdK*, which is encoded in 392 of 402 strains, is located adjacent to the outer membrane protein *fepA*, a known phage receptor (*Figure 5C*)[69]. Although *fepA* is still present in strains missing *ybdK*, this may indicate changes in regulation that impact receptor expression. Although these phenotypes would require experimental validation, these analyses show that modeling features are potentially valuable tools in hypothesis generation and the discovery of novel mediators of phage infection.

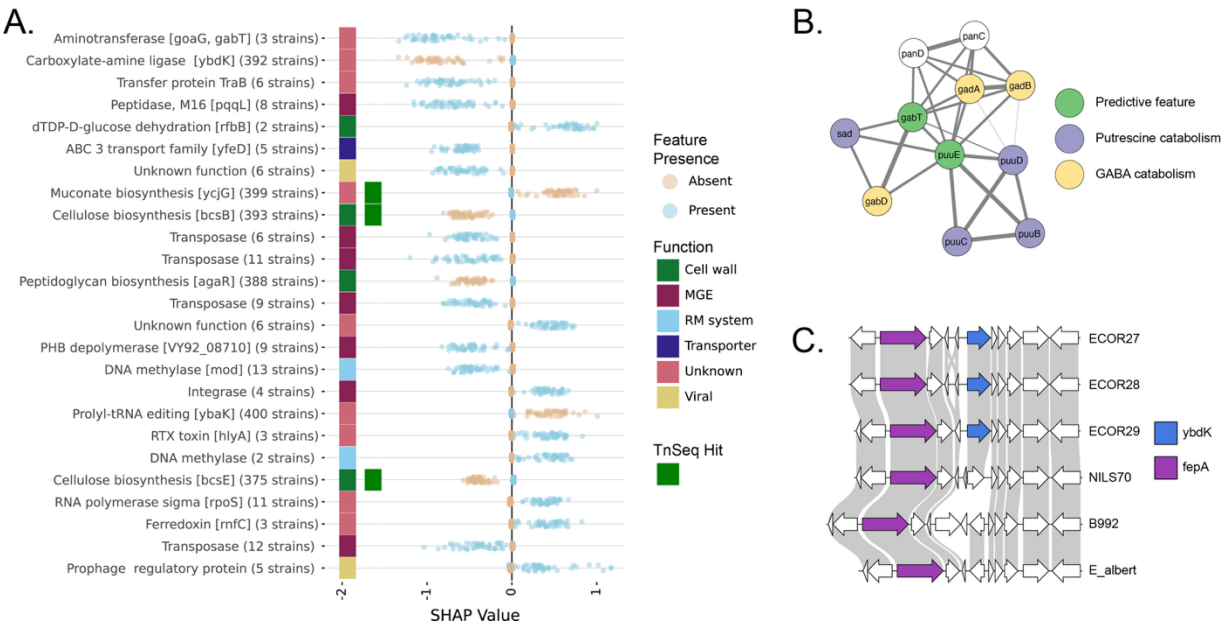

Figure 5. Biological interpretation from predictive features. A) Feature annotations ordered by Shapley Additive exPlanations (SHAP) feature importances. SHAP values indicate whether presence of a given feature (blue) or absence (orange) is associated with increased likelihood of infection (SHAP value > 0) or decreased likelihood of infection (SHAP value < 0). The colored bar shows functional categories associated with predictive features or whether the feature is associated with an RB-TnSeq hit in *E. coli* ECOR27. B) The STRING-DB network associated with genes *goaG* (*puuE*) and *gabT*. C) Synteny plot showing *fepA* adjacent to predictive feature *ybdK* in 3 strains with and 3 strains without *ybdK*.

Discussion

Recent studies have advanced machine learning approaches for predicting bacteriophage host range, exploring various genomic representation strategies and modeling algorithms [13, 16, 17, 70]. However, most have remained restricted to single genera, fixed receptor mechanisms, or taxonomic classification tasks, and lack experimental or genetic validation of predicted determinants.

In contrast, our framework predicts binary infection outcomes between arbitrary phage–host pairs across multiple genera without requiring prior knowledge of receptor biology or species-specific mechanisms. Through recursive feature elimination, cluster-aware train-test splitting, and per-phage class weighting, we optimized model configurations to produce a phylogeny-agnostic workflow that generalizes across five independent datasets. Direct performance comparison demonstrated our approach matched existing methods (AUROC 0.87 vs. 0.86 for *E. coli* and AUROC 0.812 vs. 0.818 for *Klebsiella*) while removing major constraints: the published methods required features based on known infection mediators, whereas our framework achieved comparable performance across diverse bacterial genera without these limitations. Crucially, we demonstrate external generalization to 56 unseen phages from the independent BASEL collection (AUROC 0.84) and mechanistic concordance with genome-wide RB-TnSeq assays, in which 68.8% of genetically validated host-determinant genes align with computationally identified predictive features.

The workflow's key innovation lies in converting complex whole-genome data into numerical feature representations that capture sufficient biological information for accurate prediction without mechanistic assumptions, while minimizing overfitting through dynamic feature filtering and selection. Extensive workflow optimization revealed that algorithmic choices in feature selection and machine learning implementation had greater impact on performance than genomic representation strategies, indicating that extracting relevant signals from high-dimensionality data represents the primary computational challenge. Integration with cocktail design strategies translated these probabilistic predictions into actionable therapeutic selection, achieving 80.7-97.0% strain coverage using 5-phage cocktails and up to 6.2-fold improvements in one-shot phage selection over promiscuity-based methods.

Biological interpretation of predictive features revealed both established and potentially novel genetic determinants of phage susceptibility while providing mechanistic insights into infection processes. RB-TnSeq experiments directly validated the functional relevance of computationally identified features, with 68.8% of RB-TnSeq hits mapping to predicted determinants of phage susceptibility, including established receptors (*tsx*, *nfrA*) and cell wall biosynthesis pathways. Importantly, the modeling approach frequently identified single regulatory genes within broader

pathways rather than all pathway components, suggesting the framework could capture pathway-level regulatory mechanisms rather than just direct gene effects.

Performance correlations with dataset characteristics highlight strategies for collaborative model development while revealing current limitations. Larger datasets consistently improved performance, with the requirement for substantial positive interaction data currently constraining application to a limited set of host systems. However, successful combination of *Klebsiella* datasets from different research groups demonstrated viable strategies for distributed dataset generation, suggesting that standardized experimental protocols could enable community-driven model improvement. The surprising absence of correlation between strain phylogenetic relatedness and prediction accuracy within single phylogenetic groups challenges conventional assumptions about evolutionary distance and phenotypic similarity, suggesting that horizontally transferred elements may be primary determinants of susceptibility patterns. The lack of performance improvement when combining phylogenetically-diverse datasets suggests that this protein-family-based strategy may not be suitable for capturing conserved mediators of phage infection. In the future, representations that capture structural information, such as protein language model embeddings, may improve cross-genus transfer learning.

This framework establishes computational infrastructure for rational phage therapy design and precision microbiome engineering applications. The demonstrated ability to achieve near-complete strain coverage (>94% of the largest datasets) across phylogenetic clades addresses key requirements for therapeutic efficacy and resistance management. Future developments should prioritize active learning approaches that leverage model confidence metrics to strategically guide experimental design. Our analysis revealed a strong correlation between model confidence and prediction reliability, suggesting that uncertain predictions (those near the 0.5 classification threshold) could identify high-value experiments for model improvement. Such adaptive experimental design could substantially reduce the scale of interaction screening required while maximizing information gain. Additionally, cross-dataset generalization represents a particularly promising direction, enabling models trained on well-characterized bacterial species to accelerate predictive capabilities for related pathogens lacking comprehensive datasets. Integration with high-throughput screening methodologies and emerging synthetic biology tools for phage engineering could further extend the framework's applicability. As antimicrobial resistance continues to escalate globally, this computational approach provides a scalable platform for evidence-based therapeutic design, potentially accelerating the translation of phage therapy from experimental treatment to standard clinical practice.

Methods

Dataset Collection and Preprocessing

Five phage-host interaction datasets with binary phenotype data and corresponding genomes were obtained from published sources [16, 17, 40, 42, 43, 46]. Four datasets were used for model development: *E. coli* subset (16,638 interactions, 177 strains × 94 phages) [16], mixed *Klebsiella* spp. (3,658 interactions, 62 strains × 59 phages) [42], *Klebsiella pneumoniae* (6,348 interactions, 138 strains × 46 phages) [43], and *Pseudomonas* (437 interactions, 23 strains × 19 phages) [46]. The mixed *Vibrionaceae* dataset (63,488 interactions, 256 strains x 248 phages)[40] and full *E. coli* dataset (37,788 interactions, 402 strains x 94 phages)[16] were reserved for validation. Development datasets totaled 27,081 interactions including 4,304 positive interactions (2.4%-36.2% positive rate across datasets). The *E. coli* interaction matrix was generated by assaying on solid medium, with susceptibility qualitatively assessed based on the clearance achieved by three phage dilutions and positive interactions defined as any non-zero interaction. All other datasets were generated by assaying in liquid medium and relied on the area under the optical density curve to quantify susceptibility, with positive interactions defined based on thresholds set by publishing authors for other datasets.

Phylogenetic Analysis

Bacterial strain phylogenies were inferred from core genome alignments using PPanGGOLiN (v2.0.3) [71]. Pangenome analysis identified persistent (core) genes across all strains in each dataset, and multiple sequence alignments of concatenated amino acid sequences from these genes were generated. Maximum-likelihood phylogenetic trees were constructed using FastTree (v2.1.11) with default parameters [72]. Pairwise patristic distances (sum of branch lengths between terminal nodes) were calculated for all strain pairs using BioPython's Phylo module, generating complete distance matrices for each dataset.

Phage evolutionary distances were calculated using PhamClust protein family similarity scores (PEQ values), which quantify shared gene content between phage pairs [73]. PEQ similarity scores were converted to distance metrics as $d(i,j) = 1 - PEQ(i,j)$, where $d(i,j)$ represents the distance between phages i and j. This approach captures both sequence similarity and gene content divergence. Phage gene-sharing networks were additionally constructed using vConTACT2 (v0.11.3) to visualize evolutionary relationships and identify phage clusters based on shared protein content across datasets [18].

Phylogenetic isolation was quantified for each strain or phage as the mean patristic distance to its 5 nearest phylogenetic neighbors, excluding self-distance. This metric captures local phylogenetic context rather than global tree position and provides a measure of how

evolutionarily distinct each entity is from its closest relatives. Correlations between phylogenetic isolation and prediction performance (MCC) were assessed using Pearson correlation coefficients with two-tailed significance tests (α = 0.05).

Protein Family Construction

Protein sequences from all genomes were clustered using MMSeqs2 (v. 15.6f452) with optimized parameters [51]. MMSeqs2 databases were created using the `createdb` command for all input FASTA files. Protein clustering was performed using the `cluster` command with sequence identity threshold 0.4, coverage requirement 0.8, and sensitivity 7.5. The complete clustering workflow used the following MMSeqs2 commands:

```
mmseqs createdb [input_files] [database] -v 3
mmseqs cluster [database] [cluster_output] [tmp_dir] -c 0.8 --min-seq-id
0.4 -s 7.5 --threads [n] -v 3
mmseqs createtsv [database] [database] [cluster_output] [clusters.tsv] -
-threads [n] -v 3
```

Clusters present in only a single genome were removed to focus on shared protein families. Binary presence-absence matrices were constructed with genomes as rows and protein clusters as columns. Clusters with identical presence-absence patterns were consolidated using hash-based grouping to reduce redundancy.

Parameter optimization tested pairwise combinations of MMSeqs2 identity thresholds (0.2, 0.4, 0.6, 0.8, 0.9) with fixed coverage of 0.8 across 25 parameter combinations using mixed-effects modeling. Parameter optimization was performed within ensemble-learning runs, with performance metrics calculated for each threshold combination across multiple modeling iterations. Model predictions were aggregated using median confidence scores across iterations before calculating performance metrics for each parameter combination. The model formula `performance_metric ~ C(strain_identity) * C(phage_identity)` with dataset as random effect was used to validate parameter selection. Results showed strain and phage identity thresholds explained 0.0% to 3.35% of Matthews Correlation Coefficient (MCC) variance across datasets, confirming minimal impact of parameter choice. When standard fitting failed due to convergence issues, optimization used the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method, a quasi-Newton optimization algorithm that provides improved convergence for complex mixed-effects models [74].

An alternative clustering approach using DIAMOND (version 2.0+) and Markov Clustering (MCL) was evaluated for comparison [52, 75]. This workflow employed DIAMOND blastp for homology search with sensitive mode and e-value threshold of 0.001, followed by MCL clustering with

pairwise testing of inflation parameter values of 1.4, 2.0, 4.0, and 6.0 for phage and strain genomes. DIAMOND results were converted to ABC format for MCL input, and resulting clusters were processed to generate equivalent presence-absence matrices. While this approach resulted in comparable model performance, MMSeqs2 was selected for final analyses due to superior computational efficiency.

K-mer Feature Construction

Amino acid *k*-mers of *length k* were extracted from all proteins using sliding window approaches. *K*-mers present in only a single gene were filtered to improve signal-to-noise ratio. Binary presence-absence matrices were generated for unique *k*-mers either across complete proteomes or within protein families. *K*-mer features were assigned to genomes based on exact subsequence matching or protein family membership and exact subsequence matching, respectively. Hash-based consolidation was applied to *k*-mers with identical occurrence patterns across genomes. Alternative k-values (3-15) were tested showing minimal performance impact, but with *k*=6 results in the most consistently high performing models in training data. A value of *k*=4 was used for complete proteome feature tables as increasing above this value led to untenably large feature tables (20^k theoretical maximum).

For novel genome feature assignment, *k*-mer matching required presence of exact *k*-mer subsequences within assigned protein sequences. A permissive threshold of 0.001 was used for *k*-mer assignment, allowing features to be assigned when ≥0.1% of *k*-mers were present in the assigned proteins.

Machine Learning Algorithm Selection and Optimization

Eight machine learning algorithms were systematically compared across all development datasets: CatBoost gradient-boosted decision trees (CatBoost) [54], random forest, *k*-nearest neighbor (KNN), logistic regression, multilayer perceptron (MLP), naïve Bayes, support vector machines (SVM), and gradient boosting [53]. Algorithm comparison employed comprehensive grid search hyperparameter optimization with algorithm-specific parameter ranges tested across multiple iterations. Grid search parameters included:

- **CatBoost**: iterations (500, 1000), learning_rate (0.05, 0.1), depth (4, 6)
- **Random Forest**: n_estimators (100, 200), max_depth (None, 10, 20)
- **Logistic Regression**: C (0.1, 1, 10), solver ('liblinear', 'lbfgs')
- **SVM**: C (0.1, 1, 10), kernel ('linear', 'rbf')
- **Gradient Boosting**: n_estimators (100, 200), learning_rate (0.01, 0.1), max_depth (3, 5)

- **KNN**: n_neighbors (3, 5, 7), weights ('uniform', 'distance')
- **MLP**: hidden_layer_sizes ((50,), (100,), (50, 50)), activation ('tanh', 'relu')

Model evaluation employed strain-based train-test splitting (80% strains for training). Performance was evaluated using multiple metrics (MCC, AUROC, accuracy, precision, recall, F1-score) with best models selected based on Matthews Correlation Coefficient to account for dataset imbalance. Statistical significance was assessed using Mann-Whitney U tests for pairwise algorithm comparisons. Model predictions were aggregated using median confidence scores across 100 modeling iterations, with 0.5 threshold for binary classification.

CatBoost gradient-boosted decision trees demonstrated the most consistent high-performance across datasets, achieving equal or superior MCC scores compared to alternative algorithms while also minimizing computational resource requirements.

Phage-specific class weights were implemented to address dataset imbalance with three available methods: 'log10' and 'inverse_frequency'. For each phage, positive and negative interaction counts were calculated separately, and weights were computed as follows:

- Log10 method: `weight_1 = max(1.0, log10(neg_count / (pos_count + smoothing)) + 1)`
- Inverse frequency method: `weight_1 = total_count / (pos_count + smoothing)`

where smoothing = 1.0 to prevent extreme weight ratios when positive counts are low. Negative class weights were set to 1.0 for all phages, while positive class weights were adjusted based on the phage-specific positive/negative ratio. This approach accounts for the varying interaction rates across different phages in the dataset.

Ensemble predictions were generated using median values across 50 modeling iterations to improve prediction stability and reduce overfitting.

Feature Selection Methodology

Six feature selection methods were systematically compared: Recursive Feature Elimination (RFE), SHAP-based selection, SHAP-RFE hybrid, LASSO, chi-squared, and SelectKBest. RFE demonstrated optimal performance in 3 of 5 datasets with significantly lower computational requirements and was therefore selected for final analyses. Performance was evaluated across multiple datasets using Mann-Whitney U tests for pairwise method comparisons.

In the final modeling workflow RFE was implemented using CatBoost as the base estimator with the following configuration:

- n_features_to_select: automatically determined or specified

- step_size: `max(1, int((total_features - n_features_to_select) / 10))`

RFE was performed for 25 rounds with different train-test splits, with 10 feature elimination steps per round. Features were ranked by occurrence frequency across feature selection iterations, with final feature tables generated using occurrence cut-offs. Final models used features present across the maximum number of iterations, including no more than total_features/20 features in the final selection. This feature occurrence tracking across iterations enabled identification of stable, predictive features while reducing overfitting to specific data splits. Final models used features selected based on optimal occurrence thresholds determined by model performance.

Dataset Balancing and Overfitting Prevention

Dataset imbalance (2.1%-36.2% positive interactions) was addressed through systematic comparison of 14 parameterization strategies across three methodological dimensions: clustering methods for train-test splitting (HDBSCAN, hierarchical, or random splitting), phage-specific class weighting approaches (log10 or inverse-frequency), and feature filtering strategies (cluster-based, train-test presence, or no filtering). Strategy performance was evaluated using twenty-fold cross-validation with 10% of strains systematically withheld from the entire feature selection and modeling workflow.

The optimized approach combined: (1) hierarchical clustering of strains into 20 clusters using Ward linkage and Euclidean distance (2) inverse-frequency class weights computed per phage, and (3) feature filtering requiring presence across ≥2 clusters to exclude features unique to single strain groupings. This strategy provided optimal balance between model generalizability and performance while addressing dataset-specific imbalance and strain heterogeneity.

Cross-Validation and Model Evaluation

Model generalization was evaluated using three distinct model configurations designed to assess different aspects of biological transferability. Twenty-fold cross-validation was performed for each strategy, with 10% of samples systematically withheld from the entire feature selection and modeling workflow to ensure unbiased evaluation. Features for validation genomes were assigned using similarity search (MMSeqs2, sequence identity ≥0.4, coverage ≥0.8) against training protein clusters to simulate realistic inference conditions.

Strain-based cross-validation evaluated generalization to novel bacterial hosts by training models on 90% of strains (using all available phages) and validating predictions on the remaining 10% of strains. This approach tests the model's ability to predict phage susceptibility for previously unseen bacterial genotypes. The phage-based configuration assessed generalization to novel

phages by training on all strains with 90% of phages and validating on interactions between all strains and the held-out 10% of phages. This strategy evaluates model performance when predicting host ranges for unseen phages. Strain and phage-based cross-validation examined the most challenging scenario of complete novelty by training exclusively on interactions between 90% of strains and 90% of phages (81% of total interactions), then validating on interactions between the held-out 10% of strains and 10% of phages. This approach tests generalization to entirely novel strain-phage combinations representing real-world discovery scenarios.

Statistical comparison of modeling parameters using cross-validation employed rank-order analysis with median performance metrics across the 20 iterations per parameters set. Pairwise differences were assessed using Mann-Whitney U tests, and overall parameterization differences were evaluated using the Friedman test for non-parametric comparison of ranks across multiple datasets.

Model performance was quantified using multiple complementary metrics implemented in scikit-learn (v1.5.2) to account for dataset imbalance and provide comprehensive evaluation. [53] Predicted probabilities were converted to binary predictions using a classification threshold of 0.5 for calculation of Matthews Correlation Coefficient (MCC), accuracy, precision, recall, and F1-score. MCC was selected as the primary performance metric because it provides a balanced measure accounting for all four confusion matrix categories (true/false positives and negatives), ranging from -1 (complete disagreement) to +1 (perfect prediction), with particular suitability for imbalanced datasets.

Threshold-independent metrics included area under the receiver operating characteristic curve (AUROC), which measures the model's ability to rank positive instances higher than negative instances with values from 0.5 (random) to 1.0 (perfect), and area under the precision-recall curve (AUPR), which quantifies precision-recall trade-offs and is particularly informative for imbalanced datasets. AUPR values were normalized relative to random baseline performance as AUPR_normalized = (AUPR - baseline) / (1.0 - baseline), where baseline equals the proportion of positive interactions. This normalization enables direct comparison across datasets with different imbalance ratios, with values ranging from 0 (random) to 1 (perfect).

Prediction calibration was assessed using Brier score (described in Model Calibration Analysis section) and calibration curves comparing predicted probabilities to observed infection frequencies.

Feature Assignment Protocol for Novel Genomes

Novel genome proteins were queried against reference protein clusters using MMSeqs2 similarity search with optimized parameters: sensitivity 7.5, coverage ≥0.8, and sequence identity threshold ≥0.4. The search was performed using:

```
mmseqs search query_db target_db result_db tmp_dir -c 0.8 --min-seq-id
0.4 -s 7.5 --threads [n] -v 3
```

Each protein was assigned to the best-matching cluster above the identity threshold, with duplicate protein IDs resolved by prefixing with genome identifiers when detected. Cluster assignments were then mapped to specific features using the feature mapping established during model training, which defined the relationship between protein clusters and selected features from the reference dataset. Features were assigned to a strain if a single protein family from a given feature was present. Binary feature vectors were constructed indicating protein family presence/absence for each genome based on these cluster-to-feature mappings.

K-mer features were assigned based on amino acid sequence content, requiring exact *k*-mer subsequence matches within assigned protein sequences. A permissive threshold of 0.001 was used for *k*-mer assignment, allowing features to be assigned when ≥0.1% of *k*-mers within a protein family were present in the assigned proteins. This protocol ensured feature assignment reflected realistic conditions where novel genomes lack complete sequence identity to reference databases while maintaining biological relevance through sequence similarity requirements.

Model Error Analysis

To identify systematic failures in phage-host interaction prediction and prioritize targets for biological investigation, we developed a comprehensive error analysis framework consisting of five complementary approaches.

High-Confidence Misclassification Analysis

High-confidence misclassifications were identified as predictions where $C_{abs} > 0.8$ and the prediction disagreed with ground truth. These represent cases where the model is most confident but incorrect, indicating systematic feature misinterpretation rather than uncertain boundary cases. The threshold of 0.8 corresponds to model confidence scores <0.1 or >0.9, representing the top quintile of model certainty.

Systematic Bias Detection

Entities exhibiting consistent over- or under-prediction were identified using prediction bias thresholds of ±0.2, corresponding to systematic errors of >20 percentage points between

predicted and observed interaction rates. This threshold captures biologically meaningful differences while excluding noise from natural prediction variance. Entities were classified as over-predictors ($Bias_i > 0.2$) or under-predictors $Bias_i < 0.2$.

Interaction Pattern Analysis Using Jaccard Similarity

To identify cases where the model fails to distinguish biologically distinct interaction patterns, we implemented pairwise comparisons using Jaccard similarities. For each strain or phage pair (i, j), we calculated model prediction similarity $J_{pred}(i,j) = \frac{|P_i \cap P_j|}{|P_i \cup P_j|}$ and true interaction similarity $J_{true}(i,j) = \frac{|Y_i \cap Y_j|}{|Y_i \cup Y_j|}$, where P_i and Y_i represent binary prediction and interaction vectors for entity i across all tested partners.

This approach directly compares interaction overlap patterns rather than summary statistics, capturing the complete interaction landscape and enabling direct comparison of predicted outcome to experimentally determined interaction patterns. Analysis was conducted independently for both strain-level and phage-level perspectives.

Extreme discordance cases were identified as high-priority targets for biological investigation: Type I cases (true Jaccard similarity < 0.3 with model similarity > 0.9) represent cases where the model incorrectly predicts nearly identical interaction patterns for biologically distinct entities, while Type II cases (true Jaccard similarity > 0.7 with model similarity < 0.3) represent cases where the model fails to recognize highly similar interaction patterns.

Individual Strain Performance Analysis

Individual strain performance was quantified using standard binary classification metrics (MCC, AUROC, accuracy, precision, recall) calculated from each strain's complete set of phage interaction predictions. AUROC calculations required at least two distinct classes; strains with exclusively positive or negative interactions were assigned AUROC = NaN. The relationship between strain infectivity (total number of infecting phages) and prediction performance was assessed using Pearson correlation analysis to determine whether models perform better on promiscuous versus restrictive strains.

Confidence Distribution Analysis

Model confidence scores were analyzed across different prediction outcomes to identify systematic uncertainty patterns. Predictions were categorized by confusion matrix outcomes (TP, FP, TN, FN) and confidence distributions were visualized using kernel density estimation. Confidence scores were binned into 0.1-width intervals to assess model calibration and identify confidence ranges associated with specific error types.

Model Calibration Analysis

Model calibration was assessed to evaluate the reliability of predicted probabilities and quantify prediction accuracy. Calibration curves were generated by binning predictions into 10 decile intervals (0.0-0.1, 0.1-0.2, ..., 0.9-1.0) and calculating the mean predicted probability and observed infection frequency within each bin. Perfect calibration, where predicted probabilities match observed frequencies, is represented by the diagonal line y = x. Deviations from this line indicate over- or under-confidence in predictions.

The Brier score was calculated to provide a single metric of prediction accuracy that combines both calibration and resolution:

$$BS = \left(\frac{1}{N}\right)\sum(f_i - o_i)^2$$

where f_i is the predicted probability for interaction i, o_i is the observed binary outcome (0 or 1), and N is the total number of predictions. Brier scores range from 0 (perfect accuracy) to 1 (complete inaccuracy), with scores <0.25 generally indicating useful predictions. Brier scores were calculated using scikit-learn's brier_score_loss function across all predictions for each dataset and model configuration.

Phage Cocktail Design Strategies

Three phage cocktail design strategies were compared using two phage ranking approaches (model-predicted confidence of infection and phage promiscuity in training dataset), creating a 3×2 factorial experimental design to optimize cocktail efficiency while accounting for phage diversity.

Ranking-based selection chose the top *n* phages by ranking score without diversity constraints, representing a purely score-driven approach. Hierarchical clustering-based selection employed Ward linkage and Euclidean distance based on phage feature profiles, clustering into *n* clusters and selecting the highest-ranked phage from each. HDBSCAN clustering-based selection used

964 density-based clustering with minimum cluster size of 2, handling noise points by assigning
965 unique cluster identifiers, then selected the top-ranked phage from *n* cluster. Clustering-based
966 cocktail design methods were designed to ensure broad-spectrum coverage while preventing
967 selection of highly similar phages.

968 Each selection strategy was evaluated using two independent ranking methods: prediction-based
969 ranking using median confidence scores from machine learning models across modeling
970 iterations, and historical success-based ranking (phage promiscuity) using empirical infection
971 rates calculated as the fraction of strains infected in training data. This dual ranking approach
972 enabled comparison between model-dependent and model-independent phage selection
973 criteria.

974 Performance was evaluated as the fraction of target strains achieving ≥1 predicted interactions
975 per cocktail. Cocktail sizes of 1, 3, and 5 phages were evaluated across all six strategy-ranking
976 combinations using twenty-fold cross-validation, enabling direct evaluation of diversity-based
977 versus ranking-only approaches and prediction-driven versus empirically-guided phage selection
978 under realistic deployment scenarios.

979

980 *Phage Acquisition, Amplification and Titration*

981 BASEL phage aliquots were kindly provided by A. Harms (Biozentrum, University of Basel,
982 Switzerland) and phage phi92 was provided by Petr Lieman (Department of Biochemistry &
983 Molecular Biology, University of Texas Medical Branch, USA), while the remaining phages were
984 from our lab stocks [7]. Phages were amplified from single plaques in 35 mL LB broth cultures of
985 their respective hosts supplemented with 10 mM $CaCl_2$. Incubation was carried out at 37°C for up
986 to 4h or until complete clearance was observed. Phage lysates were then centrifuged for 10 min
987 at 20,913 x g to pellet cell debris and the supernatants were filtered on 0.22-µm Millex®
988 polyethersulfone syringe filters (Millipore-Sigma). Phage titers were determined by serially-
989 diluting lysates in SM buffer (Teknova) + 10 mM $CaCl_2$ and spotting 2 µL droplets on lawns of the
990 appropriate bacterial hosts following the standard double-agar overlay method. High-titer phage
991 lysates were kept at 4°C and considered stable over the duration of the experiments.

992

993 *Experimental Validation of Model Performance*

994 56 BASEL collection phages were used for experimental validation on 25 *E. coli* strains from the
995 ECOR collection present in the training dataset (ECOR13 - ECOR38) [59, 60]. Phage suspensions
996 were diluted in a 96-well deep-well plate to normalize titers to 10^8 – 10^9 PFU/mL, achieving a
997 multiplicity of infection (MOI) range of 1-10 in the experiment. Using a Rainin MicroPro300 semi-

998 automated pipettor (Mettler-Toledo Rainin), 2 µL droplets of all 56 phages were then spotted on
999 bacterial lawns of the tested strains prepared in rectangular Nunc™ Omnitray™ single-well plate
1000 (Thermo Scientific). Technical replicates were conducted by spotting 3 plates for each bacterial
1001 host. Phage clearance was manually recorded after an overnight incubation at 37°C and was
1002 scored as a binary infection phenotype (infection/no-infection). These interaction phenotypes
1003 were then compared to model predictions to evaluate model performance and generalizability.

1004

1005 *RB-TnSeq Validation of Predictive Features*

1006 The RB-TnSeq library in *E. coli* ECOR27 was built following published protocols by conjugating the
1007 pHLL250 mariner transposon vector library (AMD290) from donor strain *E. coli WM3064* [50, 76,
1008 77]. An aliquot of the barcoded donor library was thawed and allowed to recover at 37°C for 4h
1009 in 60 mL of LB supplemented with diaminopimelic acid (DAP) at 300 µM and carbenicillin at 50
1010 µg/mL. The pellet was subsequently washed by centrifuging 5 min at 7,155 x g and resuspending
1011 in an equal volume of LB + DAP. Recipient strain ECOR27 was grown overnight at 37°C in 10 mL of
1012 LB. Mating was achieved by mixing the donor and recipient cultures at a 1:1 ratio, centrifuging 5
1013 min at 7,155 x g, resuspending and spreading the pellet on a LB + DAP plate, and incubating for
1014 5h at room temperature. The mixture was scraped and resuspended in LB before being plated on
1015 LB plates supplemented with kanamycin at 25 µg/mL to select for transconjugants. After an
1016 overnight incubation at 37°C, the resulting colonies were scraped and resuspended in 25 mL of
1017 LB. The mutant suspension was diluted to an OD_{600} of 0.5 in 50 mL of LB + Kan^{25} and allowed to
1018 grow for 4h at 37°C. Glycerol was then added at 15% final concentration and 1 mL aliquots were
1019 prepared in cryovials and frozen at -80°C for later use. Genomic DNA was extracted from one
1020 aliquot with the DNeasy Blood and Tissue Kit (QIAgen). Barcode mapping to transposon insertion
1021 sites was carried out using a two-step PCR method [78], revealing that the library contained
1022 157,637 uniquely barcoded covering 4,032 protein-coding genes out of 4,677 total, with a mean
1023 coverage of 25 unique mutants per gene.

1024 Phage competitive fitness assays were conducted in a high-throughput manner in a 48-well
1025 culture plate. An aliquot of the library was inoculated into 25 mL of LB + Kan^{25} and grown under
1026 agitation for 4 h at 37°C to reach an OD_{600} of 0.4. Three samples of 1 mL each were immediately
1027 retrieved, centrifuged for 5 min at 15,900 x g and temporarily frozen at -20°C to serve as time-
1028 zero reference for BarSeq. The mutant library was diluted to an OD_{600} of 0.04 in 2X LB + Kan^{25} (~
1029 4×10^7 CFU/mL) and 350 µL was transferred to each well of a sterile 48-well culture microplate.
1030 The same volume of phage suspensions normalized at 10^9 PFU/mL was added to the wells. The
1031 plate was then sealed with a Breathe-Easy® membrane and incubated at 37ºC with orbital shaking
1032 for 16h in a BioTek 800 TS plate reader (Agilent Technologies). Surviving mutants were collected
1033 by retrieving the content of the wells and centrifuging for 5 min at 15,900 x g. Genomic DNAs

1034 were extracted using the QIAamp 96 DNA QIAcube HT Kit (QIAgen) following manufacturer's
1035 instructions. BarSeq PCRs and Illumina sequencing were conducted as described previously [7].
1036 Genes displaying a log2 fold-change above or equal to 4 (16-fold increase in mutant abundance
1037 compared to time zero) and a t-statistic [50] absolute value above or equal to 5 were considered
1038 to be significantly enriched in the final pool and thus critical for phage infection in this host strain.
1039 Similarly, genes showing a log2 fold-change below or equal to -4 (16-fold decrease in mutant
1040 abundance) and a t-test absolute value above or equal to 5 were classified as low-fitness genes
1041 for which disruption increases the host phage susceptibility.

1042

1043 *Biological Interpretation*

1044 Shapley Additive exPlanations (SHAP) values were calculated for all features across all samples
1045 using TreeExplainer with approximate=True to handle large feature spaces efficiently [55].
1046 Features were classified as having a positive impact (presence increased infection probability) or
1047 negative impact (presence decreased infection probability) based on mean SHAP values across all
1048 predictions.

1049 Functional annotation was performed using multiple complementary computational approaches.
1050 Defense system classification employed Defense-Finder to identify bacterial defense mechanisms
1051 and anti-defense systems relevant to phage-bacteria interactions [67]. Functional annotation
1052 utilized eggNOG-mapper for comprehensive protein family assignment and Gene Ontology (GO)
1053 classification, creating high-confidence and low-confidence categories for phage-host interaction
1054 mechanisms including lipopolysaccharides, outer membrane proteins, flagella, pili, and
1055 enterobacterial common antigen [79]. Viral protein clustering and taxonomic classification were
1056 performed using vContact2 with Diamond similarity search and MCL clustering to identify viral
1057 protein families and evolutionary relationships [18, 52, 75]. Provirus detection and genomic
1058 context analysis employed geNomad to classify genomic components as viral, bacterial, or
1059 plasmid-derived, enabling identification of horizontally transferred elements and prophage
1060 regions [26]. Protein localization was predicted using DeepLocPro, to classify predictive proteins
1061 into cellular localizations "Cell wall & surface", "Periplasmic", "Outer membrane", "Cytoplasmic
1062 membrane", "Extracellular", and "Cytoplasmic".

1063 Predictive features were validated against known mediators of phage-host interactions
1064 characterized through RB-TnSeq, DubSeq, and CRISPRi experiments, which identify host genes
1065 affecting phage infection through transposon mutagenesis, overexpression, and inactivation
1066 screens [7]. This comparison revealed that the workflow successfully identified multiple screening
1067 hits, related regulators, or genes within the same operon or gene cluster as predictive features.
1068 This validation demonstrates that the approach captures biologically relevant determinants of

phage infection, suggesting features with unknown relevance to phage infection may indicate novel mediators.

Gene Neighborhood and Protein-Protein Interaction Network Analysis

To identify functional relationships between RB-TnSeq hits and computationally identified predictive features, genomic neighborhoods were analyzed based on chromosomal proximity. Genes were sorted by scaffold identifier and genomic start position within each strain. For each predictive feature, neighboring genes were identified as those within a 3-gene window (±3 genes) on the same scaffold. Distances were calculated both as the number of intervening genes and as genomic distance in base pairs. This approach accounts for operonic organization and co-regulated gene clusters that may share functional roles in phage-host interactions.

For visualization of gene neighborhoods, extended regions spanning 5 genes upstream and 5 genes downstream of genes of interest were extracted from sorted genomic coordinates. Synteny plots visualizing gene arrangement and conservation across strains were generated using clinker through the CAGECAT (Comparative Gene Cluster Analysis Toolbox) web server (https://cagecat.bioinformatics.nl/) [80].

Functional relationships between proteins were assessed using STRING-DB (v12.0) protein-protein interaction networks [61]. Custom interaction networks for *E. coli* strains BW25113 were generated by uploading complete proteomes to the STRING web application, enabling network construction using strain-specific locus identifiers. Network confidence scores were calculated from individual evidence channels (neighborhood, fusion, cooccurrence, homology, co-expression, experimental evidence, and database annotations) using STRING's Bayesian integration framework with prior probability of 0.041:

For each evidence channel: score_adjusted = (score - prior) / (1 - prior)

Combined score: 1 - ∏(1 - score_adjusted_i) × (1 - prior) + prior

Protein interactions were filtered to retain high-confidence associations (combined score ≥ 600). To link RB-TnSeq hits with predictive features, networks were restricted to include only interactions where at least one partner was a computationally identified predictive feature. Features were considered functionally related to RB-TnSeq hits if they showed direct STRING interactions (combined score ≥ 600) with genes identified in genetic screens.

Statistical Analysis

Statistical comparisons throughout employed non-parametric tests appropriate for the distribution of performance metrics. For comparisons of multiple configurations across datasets, each configuration was evaluated using 20-fold cross-validation, and configurations were ranked by mean performance on each dataset independently. The Friedman test assessed overall significance of rank differences across datasets. Pairwise comparisons of configurations used two-tailed Mann-Whitney U tests on performance scores, with effect sizes quantified using Common Language Effect Size (CLES = $U/(n_1 \times n_2)$) and Cliff's delta to characterize the magnitude of observed differences. Correlations between dataset characteristics and model performance were assessed using Pearson correlation coefficients with two-tailed significance tests. Statistical significance was defined as $p < 0.05$.

Parameter optimization involved systematic evaluation of modeling configurations through multiple training iterations. Final parameter selection was based on rank-ordered performance across datasets (assessed via Friedman test), integrated with multiple independent validation criteria: (1) consistency across phylogenetically diverse datasets, (2) stability across 20-fold cross-validation, and (3) biological interpretability of predictions. This multi-criteria approach, combined with independent experimental validation (1,328 novel interactions and RB-TnSeq screening), provides robust evidence for selected parameters.

Software Implementation and Compute Resources

Machine learning modeling was performed using the GenoPHI package available on GitHub (https://github.com/Noonanav/GenoPHI). Downstream analyses were conducted in Jupyter notebooks also available in the GitHub repository. The following software versions were used:

Core Bioinformatics Tools:

- MMSeqs2: 15.6f452
- DefenseFinder: 2.0.0
- eggNOG-mapper: 2.1.12 (with Diamond 2.1.9, eggNOG DB 5.0.2)
- vConTACT2: 0.11.3
- geNomad: 1.11.0
- PPanGGOLiN: 2.0.3
- Prodigal: 2.6.3
- Prodigal-gv: 2.11.0
- Pharokka: 1.7.3
- FastTree: 2.1.11

- DeepLocPro: 1.0.0

Python Environment:

- Python: 3.12.5
- pandas: 2.2.2, numpy: 1.26.4, scipy: 1.14.1
- scikit-learn: 1.5.2, CatBoost: 1.2.7, SHAP: 0.46.0, umap: 0.5.6, hdbscan: 0.8.33
- matplotlib: 3.9.2, seaborn: 0.13.2, plotnine: 0.13.6
- biopython: 1.84

Machine learning modeling was performed on the Lawrencium high-performance computing cluster at Lawrence Berkeley National Laboratory. Jobs were executed on lr7 partition nodes, each equipped with dual Intel Xeon Gold 6330 28-core processors (56 cores per node) and 256-512 GB memory, interconnected via HDR InfiniBand. Memory allocations ranged from 40-240 GB RAM with 8-32 CPU threads depending on dataset size and computational requirements.

Downstream analysis of results was conducted on local computational servers equipped with AMD processors (64-core AMD Opteron 6376 and 64-core AMD EPYC systems) for data processing, visualization, and statistical analysis using Jupyter notebooks and custom Python scripts.

Acknowledgments

The authors thank members of the BRaVE Phage Foundry team, the Arkin lab, Simon Roux, Morgan N. Price, and Aeron T. Hammack for feedback and discussion, and Alexander Harms and Petr Lieman for providing phages used in experimental validation.

Author Contributions

A.J.C.N.: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing—Original Draft, Writing—Review & Editing
L.M.: Conceptualization, Methodology, Investigation, Validation, Writing—Review & Editing
E.O.R.: Investigation, Validation, Writing—Review & Editing
K.P.: Software
M.P.: Investigation, Validation
M.S.: Investigation, Validation
A.K.: Software, Data Curation
A.D.: Validation, Writing—Review & Editing, Supervision
E.G.D.: Writing—Review & Editing, Supervision
V.K.M.: Conceptualization, Methodology, Writing—Review & Editing, Supervision
A.P.A.: Conceptualization, Methodology, Writing—Review & Editing, Supervision

Data and Code Availability

All data and code are available through related publications and the GenoPHI GitHub repository at https://github.com/Noonanav/GenoPHI, including published interaction matrices, modeling workflows in a pip-installable Python package, analysis scripts, RB-TnSeq data, and the validation phage-host interaction matrix.

Funding Statement

This was completed as part of the BRaVE Phage Foundry at Lawrence Berkeley National Laboratory which is supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231. This work was also supported by the National Science Foundation (NSF) of the United States under grant award No. 2220735 (EDGE CMT: Predicting bacteriophage susceptibility from *Escherichia coli* genotype).

1185 **Competing Interests**

1186 A.P.A. is a shareholder in and advisor to Nutcracker Therapeutics. The authors declare no other
1187 competing interests.

References

1. Costa, A.R., et al., *Bacteriophage genomes encode both broad and specific counter-defense repertoires to overcome bacterial defense systems.* Cell Host & Microbe, 2025. **33**(7): p. 1161-1172.e5.
2. Scanlan, P.D., et al., *Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations.* Mol Biol Evol, 2015. **32**(6): p. 1425-35.
3. Borin, J.M., et al., *Coevolutionary phage training leads to greater bacterial suppression and delays the evolution of phage resistance.* Proceedings of the National Academy of Sciences, 2021. **118**(23): p. e2104592118.
4. Nobrega, F.L., et al., *Targeting mechanisms of tailed bacteriophages.* Nature Reviews Microbiology, 2018. **16**(12): p. 760-773.
5. Maffei, E., et al., *Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection.* PLOS Biology, 2021. **19**(11): p. e3001424.
6. Moller Abraham, G., et al., *Genes Influencing Phage Host Range in Staphylococcus aureus on a Species-Wide Scale.* mSphere, 2021. **6**(1): p. 10.1128/msphere.01263-20.
7. Mutalik, V.K., et al., *High-throughput mapping of the phage resistance landscape in E. coli.* PLOS Biology, 2020. **18**(10): p. e3000877.
8. Kortright, K.E., et al., *Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria.* Cell Host & Microbe, 2019. **25**(2): p. 219-232.
9. Strathdee, S.A., et al., *Phage therapy: From biological mechanisms to future directions.* Cell, 2023. **186**(1): p. 17-31.
10. World Health, O., *Antimicrobial resistance.* 2023.
11. Brödel, A.K., et al., *In situ targeted base editing of bacteria in the mouse gut.* Nature, 2024. **632**(8026): p. 877-884.
12. Nie, W., et al., *Advances in phage–host interaction prediction: in silico method enhances the development of phage therapies.* Briefings in Bioinformatics, 2024. **25**(3): p. bbae117.
13. Keith, M., et al., *Predictive phage therapy for Escherichia coli urinary tract infections: Cocktail selection for therapy based on machine learning models.* Proceedings of the National Academy of Sciences, 2024. **121**(12): p. e2313574121.
14. Sawa, T., K. Moriyama, and M. Kinoshita, *Current status of bacteriophage therapy for severe bacterial infections.* Journal of Intensive Care, 2024. **12**(1): p. 44.
15. Kim, M.K., et al., *A blueprint for broadly effective bacteriophage-antibiotic cocktails against bacterial infections.* Nature Communications, 2024. **15**(1): p. 9987.
16. Gaborieau, B., et al., *Prediction of strain level phage–host interactions across the Escherichia genus using only genomic information.* Nature Microbiology, 2024. **9**(11): p. 2847-2861.
17. Boeckaerts, D., et al., *Prediction of Klebsiella phage-host specificity at the strain level.* Nature Communications, 2024. **15**(1): p. 4355.
18. Bin Jang, H., et al., *Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks.* Nature Biotechnology, 2019. **37**(6): p. 632-639.
19. Coclet, C. and S. Roux, *Global overview and major challenges of host prediction methods for uncultivated phages.* Current Opinion in Virology, 2021. **49**: p. 117-126.

20. Guo, J., et al., *VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses.* Microbiome, 2021. **9**(1): p. 37.
21. Ho, S.F.S., et al., *Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data.* Microbiome, 2023. **11**(1): p. 84.
22. Roux, S., et al., *iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria.* PLOS Biology, 2023. **21**(4): p. e3002083.
23. Schackart, K.E., et al., *Evaluation of computational phage detection tools for metagenomic datasets.* Frontiers in Microbiology, 2023. **Volume 14 - 2023**.
24. Camargo, A.P., et al., *IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata.* Nucleic Acids Research, 2023. **51**(D1): p. D733-D743.
25. Roux, S. and C. Coclet, *Viromics approaches for the study of viral diversity and ecology in microbiomes.* Nature Reviews Genetics, 2025.
26. Camargo, A.P., et al., *Identification of mobile genetic elements with geNomad.* Nature Biotechnology, 2024. **42**(8): p. 1303-1312.
27. Gonzales, M.E.M., J.C. Ureta, and A.M.S. Shrestha, *Protein embeddings improve phage-host interaction prediction.* PLOS ONE, 2023. **18**(7): p. e0289030.
28. Gonzales, M.E.M., J.C. Ureta, and A.M.S. Shrestha, *PHIStruct: improving phage–host interaction prediction at low sequence similarity settings using structure-aware protein embeddings.* Bioinformatics, 2025. **41**(1): p. btaf016.
29. Pan, J., et al., *Predicting phage-host interaction via hyperbolic Poincaré graph embedding and large-scale protein language technique.* iScience, 2025. **28**(1): p. 111647.
30. Baláž, A., et al. *PHERI—Phage Host ExploRation Pipeline*. Microorganisms, 2023. **11**, DOI: 10.3390/microorganisms11061398.
31. Qiu, J., et al., *PB-LKS: a python package for predicting phage–bacteria interaction through local K-mer strategy.* Briefings in Bioinformatics, 2024. **25**(2): p. bbae010.
32. Galiez, C., et al., *WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs.* Bioinformatics, 2017. **33**(19): p. 3113-3114.
33. Shang, J. and Y. Sun, *Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning.* BMC Biology, 2021. **19**(1): p. 250.
34. Amgarten, D., et al., *vHULK, a New Tool for Bacteriophage Host Prediction Based on Annotated Genomic Features and Neural Networks.* Phage (New Rochelle), 2022. **3**(4): p. 204-212.
35. Li, M. and W. Zhang, *PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion.* Briefings in Bioinformatics, 2022. **23**(1): p. bbab348.
36. Coutinho, F.H., et al., *RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content.* Patterns, 2021. **2**(7): p. 100274.
37. Villarroel, J., et al. *HostPhinder: A Phage Host Prediction Tool*. Viruses, 2016. **8**, DOI: 10.3390/v8050116.
38. Boeckaerts, D., et al., *Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins.* Scientific Reports, 2021. **11**(1): p. 1467.

39. Schooley Robert, T., et al., *Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection.* Antimicrobial Agents and Chemotherapy, 2017. **61**(10): p. 10.1128/aac.00954-17.
40. Kauffman, K.M., et al., *Resolving the structure of phage–bacteria interactions in the context of natural diversity.* Nature Communications, 2022. **13**(1): p. 372.
41. Flores, C.O., et al., *Statistical structure of host-phage interactions.* Proc Natl Acad Sci U S A, 2011. **108**(28): p. E288-97.
42. Ferriol-González, C., et al., *Targeted phage hunting to specific Klebsiella pneumoniae clinical isolates is an efficient antibiotic resistance and infection control strategy.* Microbiology Spectrum, 2024. **12**(10): p. e00254-24.
43. Beamud, B., et al., *Genetic determinants of host tropism in Klebsiella phages.* Cell Rep, 2023. **42**(2): p. 112048.
44. Song, F., et al., *Droplet-based Random Barcode Transposon-site Sequencing (Droplet RB-TnSeq) to characterize phage-host interactions.* bioRxiv, 2025: p. 2025.09.29.679331.
45. Bignaud, A., et al., *Phages with a broad host range are common across ecosystems.* Nature Microbiology, 2025. **10**(10): p. 2537-2549.
46. Costa, A.R., et al., *Accumulation of defense systems in phage-resistant strains of Pseudomonas aeruginosa.* Science Advances, 2024. **10**(8): p. eadj0341.
47. Göller, P.C., et al., *Multi-species host range of staphylococcal phages isolated from wastewater.* Nature Communications, 2021. **12**(1): p. 6965.
48. James, T., et al., *Whole-genome phenotype prediction with machine learning: open problems in bacterial genomics.* Bioinformatics, 2025. **41**(7): p. btaf206.
49. Lees John, A., et al., *Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions.* mBio, 2020. **11**(4): p. 10.1128/mbio.01344-20.
50. Wetmore Kelly, M., et al., *Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons.* mBio, 2015. **6**(3): p. 10.1128/mbio.00306-15.
51. Steinegger, M. and J. Söding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nature Biotechnology, 2017. **35**(11): p. 1026-1028.
52. Van Dongen, S., *Graph Clustering Via a Discrete Uncoupling Process.* SIAM Journal on Matrix Analysis and Applications, 2008. **30**(1): p. 121-141.
53. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2871.
54. Prokhorenkova, L., et al., *CatBoost: unbiased boosting with categorical features*, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, Curran Associates Inc.: Montréal, Canada. p. 6639–6649.
55. Shapley, L.S., *Contributions to the Theory of Games, Volume II - A Value for n-Person Games*, H.W. Kuhn and A.W. Tucker, Editors. 1953, Princeton University Press. p. 307-318.
56. Campello, R.J.G.B., D. Moulavi, and J. Sander. *Density-Based Clustering Based on Hierarchical Density Estimates*. in *Advances in Knowledge Discovery and Data Mining*. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg.

57. McInnes, L., J. Healy, and S. Astels, *hdbscan: Hierarchical density based clustering.* Journal of Open Source Software, 2017. **2**(11): p. 205.
58. Humolli, D., et al., *Completing the BASEL phage collection to unlock hidden diversity for systematic exploration of phage–host interactions.* PLOS Biology, 2025. **23**(4): p. e3003063.
59. Ochman, H. and R.K. Selander, *Standard reference strains of Escherichia coli from natural populations.* J Bacteriol, 1984. **157**(2): p. 690-3.
60. Patel Isha, R., et al., *Draft Genome Sequences of the Escherichia coli Reference (ECOR) Collection.* Microbiology Resource Announcements, 2018. **7**(14): p. 10.1128/mra.01133-18.
61. Szklarczyk, D., et al., *The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.* Nucleic Acids Res, 2023. **51**(D1): p. D638-d646.
62. Kupczok, A., et al., *Co-transfer of functionally interdependent genes contributes to genome mosaicism in lambdoid phages.* Microb Genom, 2022. **8**(11).
63. Osbourn, A.E. and B. Field, *Operons.* Cell Mol Life Sci, 2009. **66**(23): p. 3755-75.
64. Hall, R.J., et al., *Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility.* Microbial Genomics, 2021. **7**(9).
65. Lundberg, S.M., et al., *From local explanations to global understanding with explainable AI for trees.* Nature Machine Intelligence, 2020. **2**(1): p. 56-67.
66. Moreno, J., et al., *Predicting the subcellular location of prokaryotic proteins with DeepLocPro.* Bioinformatics, 2024. **40**(12): p. btae677.
67. Tesson, F., et al., *Systematic and quantitative view of the antiviral arsenal of prokaryotes.* Nature Communications, 2022. **13**(1): p. 2561.
68. de Mattos, C.D., et al., *Polyamines and linear DNA mediate bacterial threat assessment of bacteriophage infection.* Proc Natl Acad Sci U S A, 2023. **120**(9): p. e2216430120.
69. Rabsch, W., et al., *FepA- and TonB-Dependent Bacteriophage H8: Receptor Binding and Genomic Sequence.* Journal of Bacteriology, 2007. **189**(15): p. 5658-5674.
70. Lucia-Sanz, A., et al., *Inferring strain-level mutational drivers of phage-bacteria interaction phenotypes arising during coevolutionary dynamics.* Virus Evolution, 2024. **10**(1): p. veae104.
71. Gautreau, G., et al., *PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph.* PLOS Computational Biology, 2020. **16**(3): p. e1007732.
72. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
73. Gauthier Christian, H. and F. Hatfull Graham, *PhamClust: a phage genome clustering tool using proteomic equivalence.* mSystems, 2023. **8**(5): p. e00443-23.
74. Liu, D.C. and J. Nocedal, *On the limited memory BFGS method for large scale optimization.* Mathematical Programming, 1989. **45**(1): p. 503-528.
75. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature Methods, 2015. **12**(1): p. 59-60.
76. Liu, H., et al., *Magic Pools: Parallel Assessment of Transposon Delivery Vectors in Bacteria.* mSystems, 2018. **3**(1): p. 10.1128/msystems.00143-17.

1360 77. Adler, B.A., et al., *The genetic basis of phage susceptibility, cross-resistance and host-range*
1361 *in Salmonella.* Microbiology (Reading), 2021. **167**(12).
1362 78. Rubin, B.E., et al., *Species- and site-specific genome editing in complex bacterial*
1363 *communities.* Nature Microbiology, 2022. **7**(1): p. 34-47.
1364 79. Cantalapiedra, C.P., et al., *eggNOG-mapper v2: Functional Annotation, Orthology*
1365 *Assignments, and Domain Prediction at the Metagenomic Scale.* Molecular Biology and
1366 Evolution, 2021. **38**(12): p. 5825-5829.
1367 80. Gilchrist, C.L.M. and Y.-H. Chooi, *clinker & clustermap.js: automatic generation of*
1368 *gene cluster comparison figures.* Bioinformatics, 2021. **37**(16): p. 2473-2475.

1369