Inferring strain-level mutational drivers of phage-bacteria interaction phenotypes arising during coevolutionary dynamics

Adriana Lucia-Sanz[a,^], Shengyun Peng[b,*,^], Chung Yin (Joey) Leung[c,*], Animesh Gupta[d], Justin R. Meyer[e] and Joshua S. Weitz[f,g,h,*,#]

[a] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

[b] Adobe Inc., Palo Alto, California, USA

[c] GSK, Stevenage, Herts, United Kingdom

[d] Department of Physics, University of California San Diego, La Jolla, California, USA

[e] Department of Ecology, Behavior and Evolution, University of California San Diego, La Jolla, California, USA

[f] Department of Biology, University of Maryland, College Park, MD, USA

[g] Department of Physics, University of Maryland, College Park, MD, USA

[h] University of Maryland Institute for Health Computing, North Bethesda, MD, USA

^ Equal contributions

* Former address: School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA & Institut de Biologie, École Normale Superieure, Paris, France

[#] Corresponding author, E-mail: jsweitz@umd.edu (JSW)

Abstract

The enormous diversity of bacteriophages and their bacterial hosts presents a significant challenge to predict which phages infect a focal set of bacteria. Infection is largely determined by complementary – and largely uncharacterized – genetics of adsorption, injection, cell take-over and lysis. Here we present a machine learning approach to predict phage-bacteria interactions trained on genome sequences of and phenotypic interactions amongst 51 *Escherichia coli* strains and 45 phage λ strains that coevolved in laboratory conditions for 37 days. Leveraging multiple inference strategies and without *a priori* knowledge of driver mutations, this framework predicts both who infects whom and the quantitative levels of infections across a suite of 2,295 potential interactions. We found that the most effective approach inferred interaction phenotypes from independent contributions from phage and bacteria mutations, accurately predicting 86% of interactions while reducing the relative error in the estimated strength of the infection phenotype by 40%. Feature selection revealed key phage λ and *E. coli* mutations that have a significant influence on the outcome of phage-bacteria interactions, corroborating sites previously known to affect phage λ infections, as well as identifying mutations in genes of unknown function not previously shown to influence bacterial resistance. The method's success in recapitulating strain-level infection outcomes arising during coevolutionary dynamics may also help inform generalized approaches for imputing genetic drivers of interaction phenotypes in complex communities of phage and bacteria.

Introduction

Next-generation sequencing technology has revealed widespread diversity in microbial and viral communities (Aylward et al., 2017; Munson-McGee et al., 2018; Breitbart et al., 2018; Dominguez-Huerta et al., 2022; Sunagawa et al., 2015; Nayfach et al., 2021; Sunagawa et al., 2020). In parallel, the development of analytical tools to characterize species interaction networks from co-occurrence and/or time series data has led to a better understanding of microbial community structure and function (Faust and Raes 2012; Flannick et al., 2006; Stein et al., 2013; Berry and Widder 2014; Liao et al., 2020; Jiliang et al., 2022; Tamar and Kishony 2022). In principle, it should be possible to infer microbial interaction networks directly from genotypes and environmental context (Manrubia et al., 2021). In the case of phage-bacteria interactions, phage infection of a focal bacterial strain requires adsorption to specific cell-surface receptors (e.g., protein, lipid, carbohydrate) (Neurath et al., 1986; Wang, Hofnung and Charbit 2000; Chatterjee and Rothenberg 2012; Gaborieau et al. 2024), although in many cases the specific receptor remains unknown or modulated by poorly characterized biosynthetic pathways (Tetz and Tetz 2022). However, even if a phage adsorbs to a bacterium, there are many intracellular resistance mechanisms that could assist or inactivate phage infection altogether (Zborowsky and Lindell 2019; Koonin et al., 2020; Gao and Feng 2023). Categorizing effective, extracellular adsorption and intracellular replication remains challenging. Hence, despite significant progress in linking microbial genotype to phenotype, less progress has been made with understanding the genetics of traits that influence microbial species interactions (including virus and host pairs) given the additional complication that the phenotypic output of an association may depend on

41 the joint effects of two separate genomes (Bajic et al., 2018; de Jonge et al., 2019; Buckling and
42 Rainey 2002; Elena and Lenski 2003; Koskella and Brockhurst 2014; Poullain et al., 2008; Kaltz
43 and Shykoff 2002; Beckett and Williams 2013; Weitz et al., 2013; Gurney et al., 2017).
44 The problem of understanding the genetic basis of interactions requires the development of
45 computational approaches to construct genotype-to-phenotype maps. Conventional approaches
46 try to correlate phenotypic differences with genetic variation (e.g., this is true for the broad scope
47 of work in genome-wide associated studies (Horton et al., 2014; Falush 2016; Power et al., 2017)).
48 The challenge for inferring interaction-associated phenotypes is that such interactions arise due
49 to the combination of multiple genotypes (e.g., phage and host genotypes) leading to new
50 combinatorial challenges. Initial steps towards interaction inference have been made through
51 mutation-based association approaches that have successfully uncovered combinations of virus
52 and host mutations that correlate with successful virus-host interactions (Tamar and Kishony
53 2022; MacPherson et al., 2018; Jallow et al., 2009; Scanlan et al., 2011; Borin, Lee, Lucia-Sanz et
54 al., 2023; Boeckaerts et al., 2024). Conceptually, the challenge of uncovering interaction
55 phenotypes is similar to attempts to tackle the problem of studying complex traits where gene-
56 by-gene (G x G) interactions or gene-by-environment (G x E) interactions shape phenotypes (Wei
57 et al., 2014; An et al., 2009; Gibson 2015; Gupta, Zaman et al., 2022).
58 In the case of virus-microbe systems, efforts to predict interaction phenotypes require leveraging
59 specific system features and may depend on taxonomic scales. For example, computational
60 approaches are increasingly used to predict the host range of viruses in a broad taxonomic sense,
61 e.g., leveraging tetranucleotide frequencies and other sequence-specific information (Edwards

62 et al., 2016; Dutilh et al., 2017; Roux et al., 2023, Bastien et al., 2024). However, predicting strain-
63 specific interactions remains a difficult task, particularly because taxonomic markers are known
64 to be a poor proxy for infection profiles (Sullivan et al., 2003; Kauffman et al., 2022). Recent
65 studies have shown some improvement in resolving strain-specific interaction phenotypes, e.g.,
66 by using CRISPR spacers and metagenomic data to identify recent phage infection (Roux et al.,
67 2021; Szabo et al., 2022; George and Hug 2023; Shang and Sun 2022) or by co-clustering phage
68 and bacteria mutations, respectively, amongst strains that tend to interact as a means to identify
69 associated gene or sequence differences (Kauffman et al., 2022).
70 Here, we link whole genome-wide changes in phage and bacteria with observed changes in
71 interaction phenotypes in a coevolutionary context using a machine learning inference
72 framework. We do so by leveraging emergent genotype and phenotype changes in coevolving
73 populations of *Escherichia coli* B strain REL606 and bacteriophage λ strain cI26 during a 37-day
74 experiment (Gupta, Peng et al., 2022). The key idea is to recapitulate infection phenotypes within
75 an interaction network through a hierarchical regression approach without *a priori* assumptions
76 about driver mutations or the nature of genetic interactions. In contrast, prior work on
77 microevolutionary changes in infectivity have focused on changes to genes or proteins with
78 known functions in model organisms (Gaborieau et al., 2024; Meyer et al., 2012; Lobo et al., 2009;
79 Modi et al., 2013). Such approaches depend on existing annotation of genes or mutations, and
80 thus are limited by both the quality and quantity of annotations. Our regression framework
81 predicts a substantial portion of phage-host infection phenotypes, including: i) who infects whom
82 and ii) with what efficiency. In doing so, we identify prioritized phage and bacterial mutations

83 underlying changes in infection phenotypes and reveal that additive effects of phage and host
84 mutations can be sufficient to predict interaction phenotypes. As we explain, this finding suggests
85 a route to generate testable hypotheses for genome sites underlying interactions between phage
86 and bacteria amongst closely related strains.

87 ## Results

88 **Phage-bacteria mutation profiles and cross-infection matrix**

89 We used genome sequences of 50 bacterial hosts (descended from E. *coli* B strain REL606) and
90 44 phage strains (descended from λ strain cI26) isolated at varying time points during a 37-day
91 coevolution experiment (Gupta, Peng et al., 2022). Mutation profiles of the host and phage
92 revealed many changes in their genomes, including 18 and 176 unique mutations for the host
93 and phage, respectively (Supplementary Data S1 and Data S2). The interactions between phage
94 and bacteria (including evolved strains and ancestors) were quantified via plaque assay in terms
95 of the relative value of the efficiency of plating (EOP) in a target host by a focal phage strain
96 compared to that in the sensitive ancestral host. Additional details of the EOP calculations are
97 described in Methods section "Experimental setup and data collection" (Gupta, Peng et al., 2022).
98 The interactions of all phage-bacterial pairs (including ancestors) yielded a 51 x 45 phage-bacteria
99 cross-infection matrix. A summary of the mutation profiles and the phage-bacteria cross-
100 infection matrix is shown in Fig. 1. We observed 913 successful (EOP > 0) and 1382 unsuccessful
101 (EOP = 0) phage infections of 2295 phage-bacteria pairs. The distribution of EOP values was
102 skewed, with 95% of values ranging from 0 to 1.5, including a long tail with a significant variability

103 in the observed phenotypes (Supplementary Fig. S1). The co-occurrence of mutations in different
104 genomic contexts suggested it might be feasible to infer host and phage mutations that
105 disproportionately impact the phage-bacteria interaction phenotype (Supplementary Fig. S2).

106 **Model for predicting a coevolution-induced phage-bacteria interaction network**

107 We developed a framework for predicting the combined effect of mutational profiles in phage
108 and bacteria on interaction phenotypes. The underlying framework uses a logistic regression
109 approach to determine infection outcomes – specifically predicting the presence or absence of
110 infection (referred to here as POA)– for each phage-bacteria pair in the cross-infection network
111 (see Methods section "Framework design"). Additionally, we examined how combining different
112 mutation profiles (i.e., phage and host genotypes separately or in combination) improves POA
113 inference. We tested five models in total. Three of these models use linear combinations of
114 mutation features: one for host-only mutations (H model), one for phage-only mutations (P
115 model), and one for the addition of phage and host mutations (linear model). The other two
116 models use nonlinear combinations of phage and bacteria mutations with a first-order term
117 (nonlinear model) and with a second-order term (mixed model). A comprehensive description of
118 each model is provided in the Methods section "Feature construction".

119 By comparing the performance of the models, we find that those with combinations of phage
120 and bacteria mutational features predict the original POA phenotypes significantly better than
121 the null model. In particular, we find that the linear model outperforms all the rest in the
122 validation step (P < 9.44e-5) with a mean classification accuracy of ~86% (Supplementary Fig.

123 S3a). The predicted and experimentally measured POA phenotypes are shown in Fig. 2 along with
124 the associated coefficients of the mutational features resulted from the linear, nonlinear, and
125 mixed models. The value and sign of the inferred coefficients indicate the contributions that each
126 mutational feature has on the predicted phenotype: positive coefficients increase the probability
127 of infection, and the opposite is true for negative coefficients. Notably, we observe that bacterial
128 mutations are more likely to have a negative effect (negative coefficients) due to the evolution
129 of host resistance, whereas phage mutations tend to have a positive effect (positive coefficients),
130 indicating selection for counter-defense traits that expand host range (see (Gupta, Peng et al.,
131 2022)). We show that the additive effects of phage and host mutations alone can recapitulate
132 the POA matrix without explicit inclusion of higher order interaction effects. We perform feature
133 importance analysis of the linear model (detailed in the Methods section "Final predictions and
134 feature importance analysis") which reveals 5 host mutations and 32 phage mutations that have
135 a positive effect on predicting the phage-host interaction network, compared with 7 host
136 mutations and 15 phage mutations that have a negative effect (Supplementary Fig. S4a,
137 Supplementary Data S3).

138 **Model for predicting the efficiency of phage-bacteria infections**

139 Next, we extended the prediction framework to infer the efficiency of phage-bacteria
140 interactions (EFF phenotype) as a function of phage and bacteria genotypes. EFF phenotypes
141 were defined as log-transformed EOP measurements of individual infection pairs (see Methods
142 section "Framework design" and Supplementary Fig. S5), while keeping the cross-interaction

143 network fixed (Fig. 3a). Performances of the linear regression models based on the five different
144 mutational feature sets (models H, P, linear, nonlinear, and mixed) were validated via mean
145 absolute error (MAE). Results show that models containing combinations of phage and bacteria
146 mutational features predict the original EFF phenotypes significantly better than the null model
147 (Supplementary Fig. S3b; see panels in Fig. 3b-d). In particular, the linear model has the lowest
148 validation MAE (P < 3.95e-14) with ~40% reduction of the mean error compared to the null
149 model. The coefficients associated to each mutational feature in the linear model denote the
150 relative impact on the EFF phenotype. Feature importance analysis identified 8 host mutations
151 and 25 phage mutations that increase the efficiency of phage infection (positive coefficients),
152 compared to 6 host mutations and 28 phage mutations that reduce the efficiency of phage
153 infection (negative coefficients). The full set of coefficients is listed in Supplementary Data S4,
154 and driver mutations are shown in Supplementary Fig. S4a.

155 **Characterizing the functional role of putative driver mutations of phage-bacteria interaction**

156 The prediction framework yielded dozens of phage and bacterial mutations that significantly
157 impact POA and EFF predictions (Fig. 2 and 3, respectively). In Fig. 4 we organize putatively
158 important mutations revealed by the feature analysis using the linear models associated with
159 POA (Fig. 4a, Supplementary Data S3) and EFF (Fig. 4b, Supplementary Data S4) matrices. We
160 found 3 phage mutations and 1 bacterial mutation that show a significant positive effect for the
161 POA phenotype. For phage, these mutations include 2 nonsynonymous mutations in genes *S* and
162 *J* and a synonymous mutation in gene *J.* For the bacteria we identified a nonsynonymous

163 mutation in the *ccmA* gene which encodes a subunit of an ABC transporter to the periplasm (Fig.
164 4a and Supplementary Fig. S4a). We also found 3 mutations in bacteria with a significant negative
165 effect on POA: a nonsynonymous mutation in *ompF* and two deletions Δ777bp in *insB* and
166 Δ141bp in *malT*; whereas for phage we identified a nonsynonymous mutation in *J* (Fig. 4a and
167 Supplementary Fig. S4a).
168 For the EFF phenotype, 16 mutations are predicted to have a significant effect (7 positive and 9
169 negative; Fig. 4b, Supplementary Fig. S4b and Supplementary Data S4). Of the 7 positive
170 predicted features, only 1 is bacterial, a nonsynonymous mutation in the *uup* gene which encodes
171 a UvrA-like ABC family ATPase. For phage, we identified 2 insertions, 1 deletion, and 1
172 synonymous mutation in *J* gene, another synonymous mutation in *bor* gene, and a
173 nonsynonymous mutation in the *lom* gene that should increase the efficiency of infection. Of the
174 9 features that negatively impact EFF, 1 is in the bacteria and 8 are in phage. The only bacterial
175 mutation that negatively affects the EFF was already identified by the POA model: the Δ777bp
176 deletion in *insB*. For the phage we identify 2 different intergenic mutations with significant
177 negative effects downstream of *lambdap79* gene; 3 nonsynonymous, 1 synonymous (that was
178 positive for POA) and Δ1bp deletion mutations in *J* gene and 1 intergenic mutation between *Rz*
179 and *bor* genes (Fig. 4b, Supplementary Fig. S4b).
180 Results from feature importance analysis demonstrate the ability of the machine learning
181 inference framework to identify candidate, pivotal genes involved in phage-bacteria interactions
182 across various available phenotypes while recapitulating known biology of λ phage infection. We
183 find mutations in the bacterial *malT* gene, a trans positive regulator of LamB (Debarbouille et al.,

184 1978; Blanche et al., 2013; Maynard et al., 2010; Meyer and Lenski 2020), and several mutations
185 located in the phage *J* gene region that were important for both POA and EFF phenotype
186 predictions. The *J* gene encodes the tail fiber of phage λ which is critical to the process of
187 adsorption to the host and injecting phage DNA via LamB (Ge and Wang 2024; Wang, Hofnung
188 and Charbit 2000; Werts et al., 1994; Wang et al., 1998; Maddamsetti et al., 2018). Therefore,
189 mutations in both *malT* and *J* gene region are expected to impact the phage-bacteria interaction
190 network and the relative efficiency of infection – consistent with our model predicting the
191 mutations to be important for both POA and EFF phenotypes. A nonsynonymous mutation in the
192 outer membrane porin OmpF, is the most important feature for predicting a decrease in POA,
193 but was not found to be important for predicting EFF. This mutation is shared by 10 host strains,
194 2 of which were sampled from day 28 and 8 were from day 37. These 10 host strains were super-
195 resistant, that is, they were resistant to the ancestral phage λ strain, and all the phage isolates
196 from the coevolution experiment. Previous studies on this bacterial population showed that
197 phage λ evolves to use OmpF as a second receptor after *E. coli* evolves to down-regulate LamB
198 (Meyer et al., 2012). Therefore, this OmpF mutation is expected to confer resistance to these
199 evolved phage λ strains and so affects the POA (host-range), but not the EFF (efficiency of phage
200 infection). Similar OmpF mutations have been described to provide resistance to a related phage,
201 phi21, after it similarly evolved to use OmpF (Borin, Lee, Lucia-Sanz et al., 2023). Both models
202 also identified mutations in *manY* which is an inner membrane transporter that enables phage λ
203 to inject its DNA into the cytoplasm. Mutations in this protein or others in the ManXYZ complex
204 are known to confer resistance to λ (Erni and Kocher 1987; Burmeister et al., 2021; Borin, Lee,

205 Gerbino et al., 2023), and all of them impacted negatively both POA and EFF phenotypes. Most
206 interestingly, both models were able to identify the importance of Δ777bp deletion in *insB* by an
207 IS element from *E. coli* which affects genes not previously identified to interact with phage λ
208 (Blanche et al., 2013; Maynard et al., 2010), but was recently identified to confer resistance
209 through epistasis with other resistance mutation in *malT* through an unknown mechanism
210 (Gupta, Peng et al., 2022).

211 **Discussion**

212 In this study, we developed a machine learning framework leveraging hierarchical logistic
213 regression to predict phage-bacteria interactions by linking infection phenotypes with genetic
214 mutation profiles of both phage and bacterial host. By comparing models incorporating
215 increasing layers of genotype interactions, we found that incorporating independent and additive
216 mutational effects of phage and bacteria had the highest predictive value in inferring phenotype
217 from genotype. In doing so, the framework identified gene regions already recognized in
218 mediating the efficiency of infection for bacteriophage λ and *E. coli* (Gupta, Zaman et al., 2022;
219 Meyer et al., 2012; Blanche et al., 2013; Burmeister et al., 2021) and predicted mutations that
220 confer a resistant phenotype in bacteria through epistasis with other mutations (Gupta, Zaman
221 et al., 2022). The model also identified features that were located in the phage gene *J* region,
222 including a number of synonymous mutations as well as insertions and deletions that in principle
223 should be detrimental, but have been shown to modulate host-range expansion and counter-

224 defense through recombination (Borin et al. 2021). Finally, the framework identified potentially
225 novel sites that impact phage-bacteria interaction in this bacteriophage λ and *E. coli* system.
226 Interaction inference was enabled by comparing the performance of different combinations of
227 genotype information in predicting phage-bacteria interaction phenotypes. Model performance
228 analysis revealed the additive model as the best predictor of interaction phenotype from phage
229 and bacterial genotype. In the additive model, phage and bacterial mutations act independently,
230 rather than synergistically (whether positively or negatively), to determine the infection
231 outcome. Hence complex interaction networks may be (partially) predictable based on direct
232 effects rather than relying on direct inference of complex interactive effects that are more
233 challenging to measure (Tamar and Kishony 2022). This result may be limited by sampling and
234 does not exclude the possibility that higher order gene-gene interactions affect infection
235 phenotypes. The number of phage-host mutation pairs scales as the product of the number of
236 phage and host mutations in higher order models (including both the nonlinear and mixed
237 models), but most of these mutational combinations were not observed in our set of sequenced
238 strains. As such, fitting higher order models leads to underdetermined systems even with the
239 introduction of regularization terms meant to limit the number of weak contributions from
240 mutations – whether direct or in combination. Future work would have to significantly scale-up
241 genotyped combinations of overlapping mutations in different contexts to robustly infer phage-
242 bacteria interaction mutational pairs.
243 The machine learning inference framework was able to detect the importance of previously
244 identified adaptive mutations that modify phage-host interactions and potentially novel genes

245 and mutations that modulate qualitative and quantitative features of virus-microbe interactions.
246 Although identification of novel genes is not expected to be comprehensive, we did identify
247 several nonsynonymous mutations that impact POA and EFF phenotypes and that are not directly
248 involved in phage-host surface recognition – notably, mutations in *uup*, and *ccmA* genes in
249 bacteria and *S* and *lom* genes in phage. The nonsynonymous mutation in the phage *S* gene region
250 is identified as important for predicting the presence (or absence) of infection. This gene encodes
251 a holin – a small inner membrane protein required for phage-induced host lysis (Chang, Nam and
252 Young 1995). Thus, we interpret the feature analysis to imply that a mutation in the S gene has a
253 direct impact on the lysis of host cells and resulting POA/EFF phenotypes. Similar mutations were
254 uncovered via experimental evolution to counteract a gene deletion in the host that helps
255 facilitate phage DNA replication (Gupta et al., 2020). This mutation may potentially extend the
256 infection process and allow phage more time to initiate DNA replication in the debilitated host,
257 increasing the chance of a successful infection. Another candidate mutation that shapes EFF
258 interaction phenotypes was found in the phage *lom* gene region. We note that this site was
259 previously reported to increase phage resistance through an unknown mechanism (Borin et al.,
260 2021), future work may prioritize the impact of mutations in this gene region on within-cell viral
261 dynamics. Finally, we note that despite the potential for false positives and negatives,
262 evolutionary effects including genetic hitchhiking and recombination may move adaptive
263 mutations onto different backgrounds, improving detection of driver mutations of infection.
264 In summary, we have developed a machine learning framework for predicting genotypic drivers
265 of both the qualitative and quantitative nature of host-pathogen interactions that can serve as a

266 foundation for similar analyses in other co-evolutionary contexts where strong selection
267 pressures enable the selection of virus and host mutations that modulate interaction
268 phenotypes. Moving forward, it will be essential to investigate how variations in host-phage
269 systems, ecological contexts, and genetic constraints introduce sufficient diversity during
270 coevolution to allow phenotype prediction from genotype and we acknowledge that determining
271 a minimum viable training sample size for consistent model performance would be a valuable
272 direction in future applications. Nevertheless, this framework could help prioritize research on
273 identifying novel drivers of infection, focusing efforts on mutations most likely to alter phage-
274 bacteria phenotypes. Although we applied this framework in a relatively low genetic diversity
275 context, this data-driven approach does not require *a priori* knowledge of driver genes and
276 mutations. Hence, this inference framework could be applied to other, even poorly
277 characterized, phage-bacteria systems, potentially improving understanding of interactions in
278 complex, natural systems as well as for phages that target bacterial pathogens.

279 Materials and Methods

280 Experimental setup and data collection

281 We analyzed phenotypes and genomes associated with *E. coli* B strain REL606 and phage λ strain
282 cI26, cocultured for a 37-day period. Samples were taken on checkpoint days for pairwise
283 quantitative plaque assays as described in (Gupta, Peng et al., 2022). The EOP value measures
284 the efficiency of a phage infecting a derived host strain relative to that for infecting the ancestral
285 strain. The EOP value for a phage, j, infecting a host, i, is computed as:

286 $$e_{ij} = \frac{q_{(i,j)}}{q_{(anc,j)}} \times \frac{d^{s(i,j)}}{d^{s(anc,j)}}, \quad (1)$$

287 where $q_{(i,j)}$ is the number of plaques for phage j against host i, $q_{(anc,j)}$ is the number of plaques
288 for phage j against the ancestral host strain, $s_{(i,j)}$ is the number of dilutions performed to observe
289 distinguishable and countable clear plaques for phage j against host i, $s_{(anc,j)}$ is the number of
290 dilutions performed to observe distinguishable and countable clear plaques for phage j against
291 the ancestral host strain and d is the dilution ratio which is 5 in our experiment. A positive EOP
292 value from the cross-infection plaque assay indicates a successful infection event for a given
293 phage-host pair. In contrast, a zero EOP value indicates the phage has no capacity to infect. A
294 larger EOP value from the cross-infection plaque assay indicates that the phage can infect a given
295 host more efficiently than the ancestral host strain.

296 For each phage and host samples taken from each checkpoint, the DNA extraction, library
297 preparation and sequencing experiment was performed as described in (Gupta, Peng et al.,
298 2022). Mutation profiles based on the genome sequencing data were constructed using *breseq*
299 as described in (Gupta, Peng et al., 2022). In addition to the mutations revealed by *breseq* results,
300 for both host and phage we created an artificial mutation as the indicator for the ancestral strain
301 to add the ancestral strain into the mutation profile table. For this artificial mutation, only the
302 ancestral strain is indicated to have this mutation. All other strains were shown to not have this
303 mutation in the mutation profile table.

Feature construction

For a total number of U host samples and V phage samples, we denote the EOP value for the i-th host against j-th phage as e_{ij} where $i \in [1, U]$ and $j \in [1, V]$. Let N be the total number of unique mutations observed for the host and M be the total number of unique mutations observed for the phage, the host mutation profile H is a matrix of dimension U by N, and the phage mutation profile P is a matrix of dimension V by M. Let h_{il} be an element from H, then $h_{il} = 1$ corresponds to the presence of the l-th mutation in the i-th host whereas $h_{il} = 0$ corresponds to the absence of the l-th mutation in the ith host. Similarly, let p_{jk} be an element from P, then $p_{jk} = 1$ corresponds to the presence of the k-th mutation in j-th phage whereas $p_{jk} = 0$ corresponds to the absence of the k-th mutation in the j-th phage.

Five sets of features were constructed based on the mutation profiles of the host and phage. The H-only model is constructed based on a linear combination of 'host only' mutation profiles. The H-only model, denoted as $\phi_{ij}^{(1)}$, can be represented as:

$$\phi_{ij}^{(1)} = \gamma_1 + \sum_{l=1}^{N} \alpha_l h_{il} \,, \qquad (2)$$

where γ_1 represents a scalar of the bias term and α_l is the coefficient for the l-th host mutation. γ_1 and α_l will be learned from the model. The H-only model can also be represented in matrix form as:

$$\Phi^{(1)} = \Gamma_1 + H \cdot R_\alpha, \qquad (3)$$

where Γ_1 is a U by V matrix by repeating γ_1, i.e. $\Gamma_1 = [\gamma_1]_{U \times V}$, R_α is a N by V matrix by stacking the same coefficient vector α horizontally, i.e. $[\alpha|\alpha| \cdots |\alpha|\alpha]_{N \times V}$.

The P-only model is constructed based on a linear combination of 'phage only' mutational profiles. The P-only model, denoted as $\phi_{ij}^{(2)}$, can be represented as:

$$\phi_{ij}^{(2)} = \gamma_2 + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk}, \qquad (4)$$

where γ_2 represents a scalar of the bias term and $\tilde{\alpha}_k$ is the coefficient for the k-th phage mutation. γ_2 and $\tilde{\alpha}_k$ will be learned from the model. This model can also be represented in matrix form as:

$$\Phi^{(2)} = \Gamma_2 + [P \cdot R_{\tilde{\alpha}}]^T , \qquad (5)$$

where Γ_2 is a U by V matrix by repeating γ_2 and $R_{\tilde{\alpha}}$ is a M by U matrix by stacking the same coefficient vector $\tilde{\alpha}$ horizontally, i.e. $[\tilde{\alpha}|\tilde{\alpha}|\cdots|\tilde{\alpha}|\tilde{\alpha}]_{M\times U}$.

The linear model, denoted as $\phi_{ij}^{(3)}$, utilizes a linear combination of phage and host mutational features and can be represented as:

$$\phi_{ij}^{(3)} = \gamma_3 + \sum_{l=1}^{N} \alpha_l h_{il} + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk}, \qquad (6)$$

where γ_3 represents a scalar of the bias term, α_l is the coefficient for the l-th host mutation and $\tilde{\alpha}_k$ is the coefficient for the k-th phage mutation. γ_3, α_l and $\tilde{\alpha}_k$ will be learned from the model. The linear model can also be represented in matrix form as:

$$\Phi^{(3)} = \Gamma_3 + H \cdot R_{\alpha} + [P \cdot R_{\tilde{\alpha}}]^T , \qquad (7)$$

where Γ_3 is a U by V matrix by repeating γ_3, i.e. $\Gamma_3 = [\gamma_3]_{U\times V}$, R_{α} is a N by V matrix by stacking the same coefficient vector α horizontally, i.e. $[\alpha|\alpha|\cdots|\alpha|\alpha]_{N\times V}$ and $R_{\tilde{\alpha}}$ is a M by U matrix by stacking the same coefficient vector $\tilde{\alpha}$ horizontally, i.e. $[\tilde{\alpha}|\tilde{\alpha}|\cdots|\tilde{\alpha}|\tilde{\alpha}]_{M\times U}$. The assumption for

the linear model is that the impact of mutations from both the phage and host have additive effects on the observed outcome.

The nonlinear model, denoted as $\phi_{ij}^{(4)}$, utilizes nonlinear combination of phage and host mutational features as the input and can be represented as:

$$\phi_{ij}^{(4)} = \gamma_4 + \sum_{l=1}^{N} \sum_{k=1}^{M} \beta_{lk} h_{il} p_{jk} \ , \qquad (8)$$

where γ_4 represents a scalar of the bias term, β_{lk} denotes the coefficient for the l-th host mutation and k-th phage mutation in the corresponding i-th host and j-th phage pair. γ_4 and β_{lk} will be learned from the model. This nonlinear model can also be represented in the matrix form as:

$$\Phi^{(4)} = \Gamma_4 + H \cdot B \cdot P^T \ , \qquad (9)$$

where Γ_4 is a U by V matrix by repeating γ_4, i.e. $\Gamma_4 = [\gamma_4]_{U \times V}$, B is the N by M coefficient matrix. The assumption for the nonlinear model is that the impact of the genetic mutations on the observed outcome comes from the additive effects of co-occurring phage-host mutation pairs. In other words, $h_{il} p_{jk} = 1$ only when both the host i has mutation l and phage j has mutation k.

Based on the formulation of the linear and nonlinear models, it is natural to combine both effects to get a more sophisticated input feature, by adding up both effects. The mixed model, denoted as $\phi_{ij}^{(5)}$, utilizes a mixed combination of linear and nonlinear effects of host and phage mutation features and can be represented as:

$$\phi_{ij}^{(5)} = \gamma_5 + \sum_{l=1}^{N} \alpha_l h_{il} + \sum_{k=1}^{M} \tilde{\alpha}_k p_{jk} + \sum_{l=1}^{N} \sum_{k=1}^{M} \beta_{lk} h_{il} p_{jk} \ . \qquad (10)$$

363 The matrix form of the mixed model is:

364 $$\Phi^{(5)} = \Gamma_5 + H \cdot R_\alpha + [P \cdot R_{\tilde{\alpha}}]^T + H \cdot B \cdot P^T \,, \qquad (11)$$

365 where Γ_5 is a U by V matrix by repeating γ_5, i.e. $\Gamma_5 = [\gamma_5]_{U \times V}$.

366 **Framework design**

367 We designed a framework comprised of two types of predictions. First, we designed a framework
368 that predicts the phage-host cross interaction network (i.e., host range). This model tries to find
369 the set of mutational features that best distinguish between presence (EOP > 0) and absence
370 (EOP = 0) of infection (POA) using classification models. Second, we built a framework to predict
371 the efficiency of the interaction (EFF). This model is designed to find the set of mutational
372 features that best predicts the efficiency of a phage infecting a host in those phage-host pairs
373 where the host is susceptible to the phage (EOP > 0).

374 **Model for predicting phage-host cross-interaction network (POA)**

375 In order to determine the presence or absence of a successful infection event for a phage-host
376 pair, we binarized the EOP values e_{ij} into 0 and 1, i.e.

377 $$d_{ij} = 1_{\{e_{ij} > 0\}} \,, \qquad (12)$$

378 where $d_{ij} = 0$ indicates a failure of the infection and $d_{ij} = 1$ indicates success. Here we used
379 logistic regression to model the relationship between mutation profiles and the existence of
380 successful infection in phage-host pairs, that is:

381 $$\phi_{ij}^{(\cdot)} = ln\left(\frac{d_{ij}}{1 - d_{ij}}\right). \qquad (13)$$

382 Each of the five sets of features, namely H-only, P-only, linear, nonlinear and mixed, were used
383 as the input features for the models $\phi_{ij}^{(1)}$, $\phi_{ij}^{(2)}$, $\phi_{ij}^{(3)}$, $\phi_{ij}^{(4)}$ and $\phi_{ij}^{(5)}$ respectively. In practice, we
384 used LASSO for feature selection and regularization. The penalty term parameter for LASSO was
385 determined by using 10-fold cross-validation on the training data. The prediction classification
386 error, $\frac{FalsePositives+FasleNegatives}{TestSamples}$, was used to assess the performance for this model. The mean
387 classification error was calculated by taking the mean of classification error from 200 runs.

388 **Model for predicting the efficiency of phage-bacteria infections (EFF)**

389 We applied a log transformation on the positive EOP values to normalize the distribution
390 (Supplementary Fig. S5a). For a given phage-host pair where a successful infection event is
391 present, that is $e_{ij} > 0$, we denote the natural log transformed EOP value as:

392 $$e'_{ij} = ln(e_{ij}) . \qquad (14)$$

393 Shapiro-Wilk test was performed to check the normality of the distribution of e'_{ij} (Supplementary
394 Fig. S5b).

395 Linear regression was used to model the relationship between mutation profiles and the intensity
396 of successful infections in phage-host pairs, that is:

397 $$\phi_{ij}^{(\cdot)} = e'_{ij} . \qquad (15)$$

398 Each of the five sets of features, namely H-only, P-only, linear, nonlinear and mixed, were used
399 as the input features for the models $\phi_{ij}^{(1)}$, $\phi_{ij}^{(2)}$, $\phi_{ij}^{(3)}$, $\phi_{ij}^{(4)}$ and $\phi_{ij}^{(5)}$ respectively. For the linear
400 model, we also used LASSO for feature selection and regularization. The penalty term parameter

401 for LASSO was determined by using 10-fold cross-validation on the training data. Finally, the MAE
402 was used to evaluate the performance of the model.

403 **Train-validation split and feature evaluation**

404 To assess the performance of different features for the logistic regression model, we performed
405 200 bootstrap runs to predict the existence of phage infection. Specifically, in each run the
406 training set was generated by randomly select $U \times V$ samples from the entire dataset with
407 replacement. The d_{ij} values that were not selected as training samples form the validation set.
408 As a control, for each run, a null model was built to predict the outcomes by randomly sample
409 d_{ij} values from a Bernoulli distribution $Bern(\hat{p})$ where $\hat{p}$ is the maximum likelihood estimator
410 (MLE) of the proportion of successful infection from the training set of that run. After the 200
411 runs, the training and validation prediction error were compared between pairs of the models
412 including the null model and models based on phage and host mutations only and linear,
413 nonlinear, and mixed combinations of phage and host mutational features.

414 Similarly, we also performed 200 bootstrap runs for the linear model to predict the infection
415 efficiency. Specifically, in each run the training set was generated by randomly sample e'_{ij} with
416 replacement. The size of e'_{ij} sampled as the training set in each run matches the total number of
417 the e'_{ij}. The e'_{ij} that were not selected in the training set forms the validation set. As a control,
418 for each run, a null model was built by always predicting the efficiency of infection as the mean
419 e'_{ij} of the training set for that run. After the 200 runs, the training and validation MAEs were
420 compared between pairs of the models including the null model and every feature model set.

Final predictions and feature importance analysis

After comparing the training and validation performance of models based on the different mutational sets with 200 bootstrap runs, a final model, that integrates predictions of POA and EFF was constructed. The penalty term parameter for each of the prediction frameworks was chosen as the mean of the best penalty term parameter from each of the 200 bootstrap runs. After model fitting, the predicted outcome $\tilde{d}_{ij}$ for the POA model and $\tilde{e}'_{ij}$ for the EFF model was calculated. For each step of the final models, the importance of feature was measured by the absolute value of coefficients learned from each step.

Author Contributions

Conceptualization: JSW, CYL & JRM

Methodology: ALS, SP, CYL & JSW

Investigation: ALS, SP, CYL, AG

Visualization: ALS, SP

Writing – original draft: ALS, SP & JSW

Writing – review & editing: ALS, SP, CYL, AG, JRM, JSW

Software Availability

https://github.com/aluciasanz/genotype_to_phenotype_inference_model

Acknowledgments

JSW - Army Research Office (W911NF1910384), NSF (2200269), NIH (R01 AI146592), Simons Foundation (930283), Chaires Blaise Pascal of the Île-de-France region.

JRM - Howard Hughes Medical Institute Emerging Pathogens Initiative grant 311169.

442 **References**

An, P., et al. (2009) 'The challenge of detecting epistasis (G x G interactions): Genetic Analysis Workshop 16', *Genetic Epidemiology*, 33. doi:10.1002/gepi.20474

Anderson, B., et al. (2011) 'Enumeration of bacteriophage particles: comparative analysis of the traditional plaque assay and real-time QPCR- and nanosight-based assays', *Bacteriophage*, 1: 86-93. doi:10.4161/bact.1.2.15456

Aylward, F.O., et al. (2017) 'Diel cycling and long-term persistence of viruses in the ocean's euphotic zone', *Proc Natl Acad Sci USA,* 114: 11446-51. doi:10.1073/pnas.1714821114

Bajic, D., et al. (2018) 'On the deformability of an empirical fitness landscape by microbial evolution', *Proc Natl Acad Sci USA*, 115:11286-11291. https://doi.org/10.1073/pnas.1808485115

Bastien G.E., et al. (2024) 'Virus-host interactions predictor (VHIP): Machine learning approach to resolve microbial virus-host interaction networks', *PLOS Computational Biology*, 20: e1011649. https://doi.org/10.1371/journal.pcbi.1011649

Beckett, S.J. and Williams, H.T.P. (2013) 'Coevolutionary diversification creates nested-modular structure in phage-bacteria interaction networks', *Interface Focus*, 3: 20130033. doi:10.1098/rsfs.2013.0033

Berry, D. and Widder, S. (2014) 'Deciphering microbial interactions and detecting keystone species with co-occurrence networks', *Frontiers in Microbiology*, 5:219. doi:10.3389/fmicb.2014.00219

Blanche, S., et al. (2013) 'The protein interaction network of bacteriophage lambda with its host Escherichia coli', *Journal of Virology*, 87:12745-12755. https://doi.org/10.1128/JVI.02495-13

Breitbart, M., et al. (2018) 'Phage puppet masters of the marine microbial realm', *Nature Microbiology,* 3: 754–766. https://doi.org/10.1038/s41564-018-0166-y

Boeckaerts, D., et al. (2024) 'Prediction of Klebsiella phage-host specificity at the strain level', *Nature Communications*, 15: 4355. https://doi.org/10.1038/s41467-024-48675-6

Borin, J.M., et al. (2021) 'Coevolutionary phage training leads to greater bacterial suppression and delays the evolution of phage resistance', *Proc Natl Acad Sci USA*, 118: e2104592118. https://doi.org/10.1073/pnas.2104592118

Borin, J.M., Lee, J.J., Gerbino, K., et al. (2023) 'Comparison of bacterial suppression by phage cocktails, dual-receptor generalists, and coevolutionarily trained phages', *Evolutionary Applications*, 16: 152-162. https://doi.org/10.1111/eva.13518

Borin, J.M., Lee, J.J., Lucia-Sanz, A., Gerbino, K.R., Weitz, J.S. and Meyer, J.R. (2023) 'Rapid bacteria-phage coevolution drives the emergence of multi-scale networks', *Science*, 382: 674-678. doi:10.1126/science.adi5536

Buckling, A. and Rainey, P.B. (2002) 'Antagonistic coevolution between a bacterium and a bacteriophage', *Proceedings of the Royal Society B*, 269:931-936. doi: 10.1098/rspb.2001.1945

Burmeister, A.R., et al. (2021) 'Sustained coevolution of phage lambda and Escherichia coli involves inner- as well as outer-membrane defenses and counter-defenses', *Microbiology*, 167: 001063. doi:10.1099/mic.0.001063

Chang, C.Y., et al. (1995) 'S gene expression and the timing of lysis by bacteriophage lambda', *Journal of Bacteriology*, 177: 3283-3294.

Chatterjee, S. and Rothenberg, E. (2012) 'Interaction of bacteriophage lambda with its E. coli receptor, LamB', *Viruses*, 4: 3162-3178. doi: 10.3390/v4113162

de Jonge, P.A., et al. (2019) 'Molecular and evolutionary determinants of bacteriophage host range', *Trends in Microbiology*, 27:51-63. doi: 10.1016/j.tim.2018.08.006

Debarbouille, M., et al. (1978) 'Dominant constitutive mutations in malT, the positive regulator gene of the maltose regulon in Escherichia coli', *Journal of Molecular Biology*, 124: 359-371. doi:10.1016/0022-2836(78)90304-2

Dominguez-Huerta, G., et al. (2022) ‘Diversity and ecological footprint of Global Ocean RNA viruses’, *Science,* 376: 1202-1208. doi: 10.1126/science.abn6358

Dutilh, B.E., Reyes, A., Hall, R.J. and Whiteson, K.L. (2017) 'Virus discovery by metagenomics: the (im)possibilities', *Frontiers in Microbiology*, 8:1710. doi: 10.3389/fmicb.2017.01710

Edwards, R.A., McNair, K., Faust, K., Raes, J. and Dutilh, B.E. (2016) 'Computational approaches to predict bacteriophage-host relationships', *FEMS Microbiology Reviews*, 40:258-272. doi: 10.1093/femsre/fuv048

Elena, S.F. and Lenski, R.E. (2003) 'Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation', *Nature Reviews Genetics*, 4:457-469. doi:10.1038/nrg1088

Erni, B. and Kocher, H.P. (1987) 'The mannose permease of Escherichia coli consists of three different proteins. Amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda DNA', *Journal of Biological Chemistry*, 262: 5238-5247

Falush, D. (2016) 'Bacterial genomics: Microbial GWAS coming of age', *Nature Microbiology*, 1:16059. doi:10.1038/nmicrobiol.2016.59

Faust, K. and Raes, J. (2012) 'Microbial interactions: from networks to models', *Nature Reviews Microbiology*, 10: 538-550. doi: 10.1038/nrmicro2832

Flannick, J., et al. (2006) 'Graemlin: general and robust alignment of multiple large interaction networks', *Genome Research*, 16: 1169-1181. doi: 10.1101/gr.523570

Gaborieau, B., et al. (2024) 'Prediction of strain level phage–host interactions across the *Escherichia* genus using only genomic information', *Nature Microbiology*, 9: 2847–286. https://doi.org/10.1038/s41564-024-01832-5

Gao, Z. and Feng, Y. (2023) 'Bacteriophage strategies for overcoming host antiviral immunity', *Frontiers in Microbiology*, 14: 1211793. https://doi.org/10.3389/fmicb.2023.1211793

Ge, X. and Wang, J. (2024) 'Structural mechanism of bacteriophage lambda tail's interaction with the bacterial receptor', *Nature Communications* 15: 4185 (2024). https://doi.org/10.1038/s41467-024-48686-3.

George, N.A. and Hug, L.A. (2023) 'CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts', *Scientific Reports*, 13:5611. https://doi.org/10.1038/s41598-023-32078-6.

Gibson, G. (2015) *A Primer of Human Genetics*, Sinauer Associates, an imprint of Oxford University Press.

Gupta, A., et al. (2020) 'Bacteriophage lambda overcomes a perturbation in its host–viral genetic network through mutualism and evolution of life history traits', *Evolution*, 74: 764-774. https://doi.org/10.1111/evo.13920

Gupta, A., Peng, S., et al. (2022) 'Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing', *Ecology Letters*, 25: 876-888. doi:10.1111/ele.1

Gupta, A., Zaman, L., et al. (2022) 'Host-parasite coevolution promotes innovation through deformations in fitness landscapes', *eLife*, 11: e76162. https://doi.org/10.7554/eLife.76162

Gurney, J., et al. (2017) 'Network structure and local adaptation in co-evolving bacteria-phage interactions', *Molecular Ecology*, 26: 1764-1777. doi:10.1111/mec.14008

Horton, M.W., et al. (2014) 'Genome-wide association study of *Arabidopsis thaliana* leaf microbial community', *Nature Communications*, 5: 5320. doi:10.1038/ncomms6320

Jallow, M., et al. (2009) 'Genome-wide and fine-resolution association analysis of malaria in West Africa', *Nature Genetics*, 41: 657-665. doi:10.1038/ng.388

Jiliang, H., et al. (2022) 'Emergent phases of ecological diversity and dynamics mapped in microcosms', *Science*, 378: 85-89. doi:10.1126/science.abm7841

Kaltz, O. and Shykoff, J.A. (2002) 'Within-and among-population variation in infectivity, latency and spore production in a host–pathogen system', *Journal of Evolutionary Biology*, 15: 850-860. https://doi.org/10.1046/j.1420-9101.2002.00433.x

Kauffman, K.M., et al. (2022) 'Resolving the structure of phage-bacteria interactions in the context of natural diversity', *Nature Communications*, 13:372. https://doi.org/10.1038/s41467-021-27583-z

Koonin, E.V., et al. (2020) 'Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire', *Nature Reviews Genetics*, 21: 119-131. https://doi.org/10.1038/s41576-019-0172-9

Koskella, B. and Brockhurst, M.A. (2014) 'Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities', *FEMS Microbiology Reviews*, 38:916-931. doi: 10.1111/1574-6976.12072.

Liao, C., et al. (2020) 'Modeling microbial cross-feeding at intermediate scale portrays community dynamics and species coexistence', *PLOS Computational Biology*, 16: e1008135. https://doi.org/10.1371/journal.pcbi.1008135

Lobo, F.P., et al. (2009) 'Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts', *PLoS One*, 4: e6282. doi:10.1371/journal.pone.0006282

MacPherson, A., Otto, S.P. and Nuismer, S.L. (2018) 'Keeping pace with the Red Queen: identifying the genetic basis of susceptibility to infectious disease', *Genetics*, 208: 779-789. doi:10.1534/genetics.117.300481

Maddamsetti, R., et al. (2018) 'Gain-of-function experiments with bacteriophage lambda uncover residues under diversifying selection in nature', *Evolution*, 72: 2234-2243. https://doi.org/10.1111/evo.13586

Manrubia, S., et al. (2021) 'From genotypes to organisms: state-of-the-art and perspectives of a cornerstone in evolutionary dynamics', *Physics of Life Reviews*, 38: 55-106. doi: 10.1016/j.plrev.2021.03.004

Maynard, N.D., et al. (2010) 'A forward-genetic screen and dynamic analysis of lambda phage host-dependencies reveals an extensive interaction network and a new anti-viral strategy', *PLoS Genetics*, 6: e1001017. https://doi.org/10.1371/journal.pgen.1001017

Meyer, J.R., et al. (2012) 'Repeatability and contingency in the evolution of a key innovation in phage lambda', *Science*, 335: 428-432. doi:10.1126/science.1214449

Meyer, J.R. and Lenski, R.E. (2020) 'Subtle environmental differences have cascading effects on the ecology and evolution of a model microbial community'. In: *Evolution in Action: Past, Present and Future*, Springer International Publishing.

Modi, S.R., et al. (2013) 'Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome', *Nature*, 499: 219-222. doi:10.1038/nature12212

Munson-McGee, J.H., et al. (2018) 'A virus or more in (nearly) every cell: ubiquitous networks of virus-host interactions in extreme environments', *ISME J,* 12: 1706-14. https://doi.org/10.1038/s41396-018-0071-7

Nayfach, S., et al. (2021) 'Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome', *Nature Microbiology*, 6: 960-970. https://doi.org/10.1038/s41564-021-00928-6

Neurath, A.R., et al. (1986) 'Identification and chemical synthesis of a host cell receptor binding site on hepatitis B virus', *Cell*, 46: 429-436. doi: 10.1016/0092-8674(86)90663-x

Poullain, V., et al. (2008) 'The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage', *Evolution*, 62:1-11. doi: 10.1111/j.1558-5646.2007.00

Power, R.A., Parkhill, J. and de Oliveira, T. (2017) 'Microbial genome-wide association studies: lessons from human GWAS', *Nature Reviews Genetics*, 18: 41-50. doi:10.1038/nrg.2016.132

Roux, S., Camargo, A.P., Coutinho, F.H., Dabdoub, S.M., Dutilh, B.E., et al. (2023) 'iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria', *PLOS Biology*, 21. https://doi.org/10.1371/journal.pbio.3002083

Roux, S., et al. (2021) 'IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses', *Nucleic Acids Research*, 49: 764-775. doi:10.1093/nar/gkaa946

Sambrook, J. and Russell, D.W. (2006) *The Condensed Protocols from Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Scanlan, P.D., et al. (2011) 'Genetic basis of infectivity evolution in a bacteriophage', *Molecular Ecology*, 20:981-989. doi:10.1111/j.1365-294X.2010.04903.x

Shang, J. and Sun, Y. (2022) 'CHERRY: a computational method for accurate prediction of virus-prokaryotic interactions using a graph encoder-decoder model', *Briefings in Bioinformatics*, 23. https://doi.org/10.1093/bib/bbac182

Stein, R.R., et al. (2013) 'Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota', *PLOS Computational Biology*, 9: e1003388. doi: 10.1371/journal.pcbi.1003388

Sullivan, N.J., et al. (2003) 'Accelerated vaccination for Ebola virus haemorrhagic fever in non-human primates', *Nature*, 424:681-684. doi: 10.1038/nature01876.

Sunagawa, S., et al. (2015) 'Structure and function of the global ocean microbiome', *Science*, 348, 1261359. doi:10.1126/science.1261359

Sunagawa, S., et al. (2020) 'Tara Oceans: towards global ocean ecosystems biology', *Nature Reviews Microbiology*, 18:428–445. https://doi.org/10.1038/s41579-020-0364-5

Szabo, R.E., et al. (2022) 'Historical contingencies and phage induction diversify bacterioplankton communities at the microscale', *Proc Natl Acad Sci USA*, 119: e2117748119. https://doi.org/10.1073/pnas.2117748119

Tamar, E.S. and Kishony, R. (2022) 'Multistep diversification in spatiotemporal bacterial-phage coevolution', *Nature Communications*, 13: 7971. https://doi.org/10.1038/s41467-022-35351-w

Tetz, V. and Tetz, G. (2022) 'Novel prokaryotic system employing previously unknown nucleic acids-based receptors', *Microbial Cell Factories*, 21: 202. https://doi.org/10.1186/s12934-022-01923-0

Wang, J., et al. (1998) 'Cloning of the J gene of bacteriophage lambda, expression and solubilization of the J protein: first in vitro studies on the interactions between J and LamB, its cell surface receptor', *Research in Microbiology*, 149: 611-624. doi:10.1016/s0923-2508(99)80009-6

Wang, J., Hofnung, M. and Charbit, A. (2000) 'The C-terminal portion of the tail fiber protein of bacteriophage lambda is responsible for binding to LamB, its receptor at the surface of Escherichia coli K-12', *Journal of Bacteriology*, 182: 508-512. doi:10.1128/JB.182.2.508-512.2000

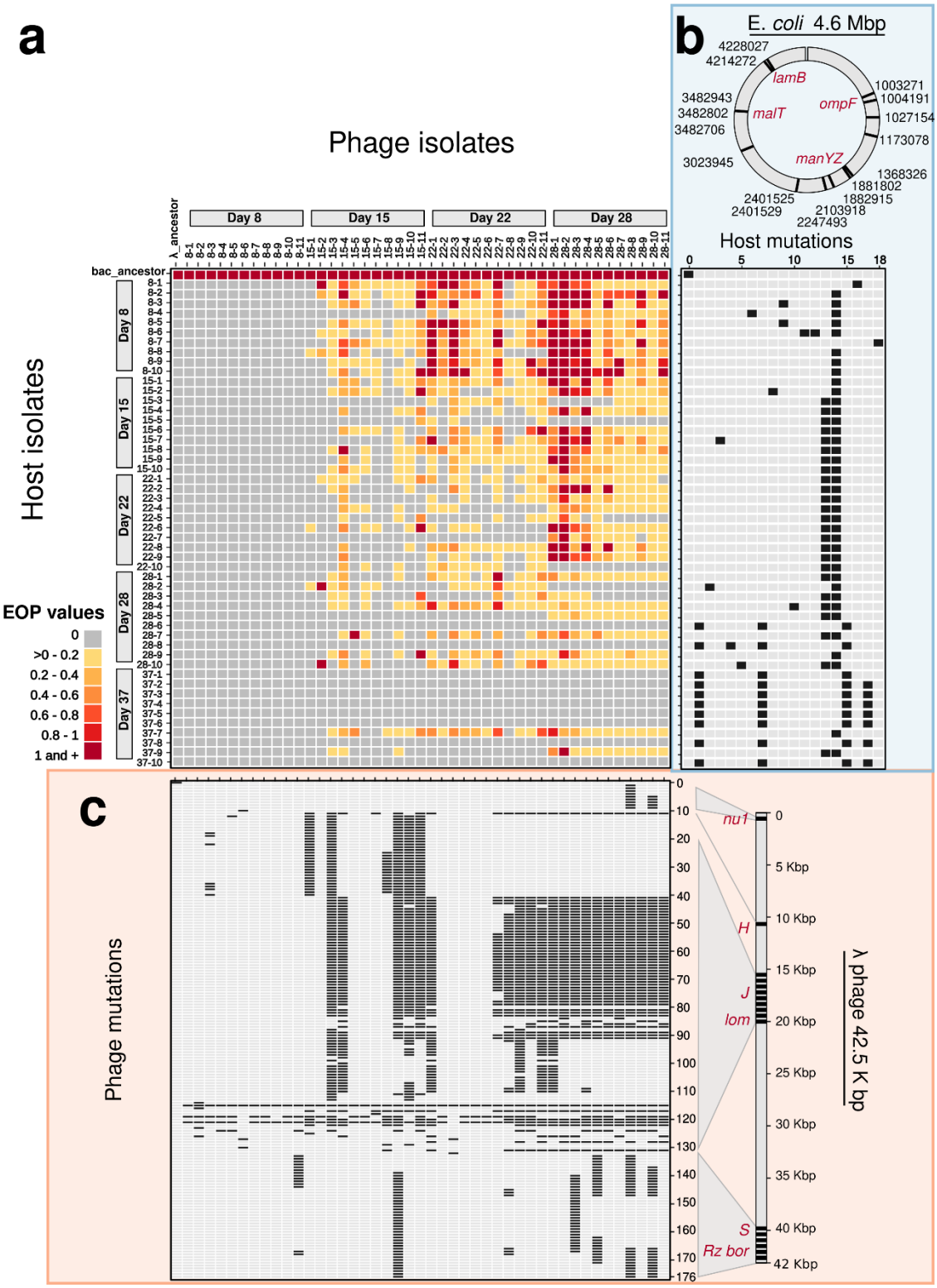
Wei, W., Hemani, G. and Haley, C.S. (2014) 'Detecting epistasis in human complex traits', *Nature Reviews Genetics*, 15: 722-733. doi:10.1038/nrg3747

Weitz, J.S., et al. (2013) 'Phage-bacteria infection networks', *Trends in Microbiology*, 21: 82-91. doi:10.1016/j.tim.2012.11.00

Werts, C., et al. (1994) 'Adsorption of bacteriophage lambda on the LamB protein of Escherichia coli K-12: point mutations in gene J of lambda responsible for extended host range', *Journal of Bacteriology*, 176: 941-947. doi:10.1128/jb.176.4.941-947.1994

Zborowsky, S. and Lindell, D. (2019) 'Resistance in marine cyanobacteria differs against specialist and generalist cyanophages', *Proc Natl Acad Sci USA*, 116: 16899-16908. https://doi.org/10.1073/pnas.1906897116

443 **Figures**


a
Phage isolates
Host isolates
Day 8
Day 15
Day 22
Day 28
Day 37
λ_ancestor
bac_ancestor
EOP values
0
>0 - 0.2
0.2 - 0.4
0.4 - 0.6
0.6 - 0.8
0.8 - 1
1 and +
b
E. coli 4.6 Mbp
lamB
ompF
malT
manYZ
Host mutations
c
Phage mutations
nu1
H
J
lom
S
Rz bor
λ phage 42.5 K bp
5 Kbp
10 Kbp
15 Kbp
20 Kbp
25 Kbp
30 Kbp
35 Kbp
40 Kbp
42 Kbp

444 **Figure 1. Phage-bacteria cross-infection matrix and mutation profiles.** (a) Cross-infection matrix,
445 including host and phage ancestor strains, and 50 bacteria (rows) and 44 phage (columns) strains
446 isolated during 37-day coevolution experiment (day of isolation indicated). Names correspond to
447 "day of isolation – number of isolate". Colored cells are EOP values of infection as in legend, grey
448 cells indicate no infection. (b-c) Mutation profiles for each isolate (positions mutated are in black
449 and in grey otherwise) for 18 (host) and 127 (phage) found mutations numbered in sequential
450 order of appearance in the corresponding genome. (b, in blue) Host isolates (rows) and mutation
451 profiles (columns) for 1 to 18 unique mutations found in nt position 1,003,271 to 4,228,027 of
452 the E. *coli* genome (c, in orange) Phage isolates (columns) and mutation profiles (rows) for 1 to
453 127 unique mutations found in nt position 175 to 42,491 of the λ phage genome. For the
454 complete list of host and phage mutations see Supplementary Data S1 and Data S2. Important
455 genes for phage-host interaction are highlighted in red and discussed in the main text.

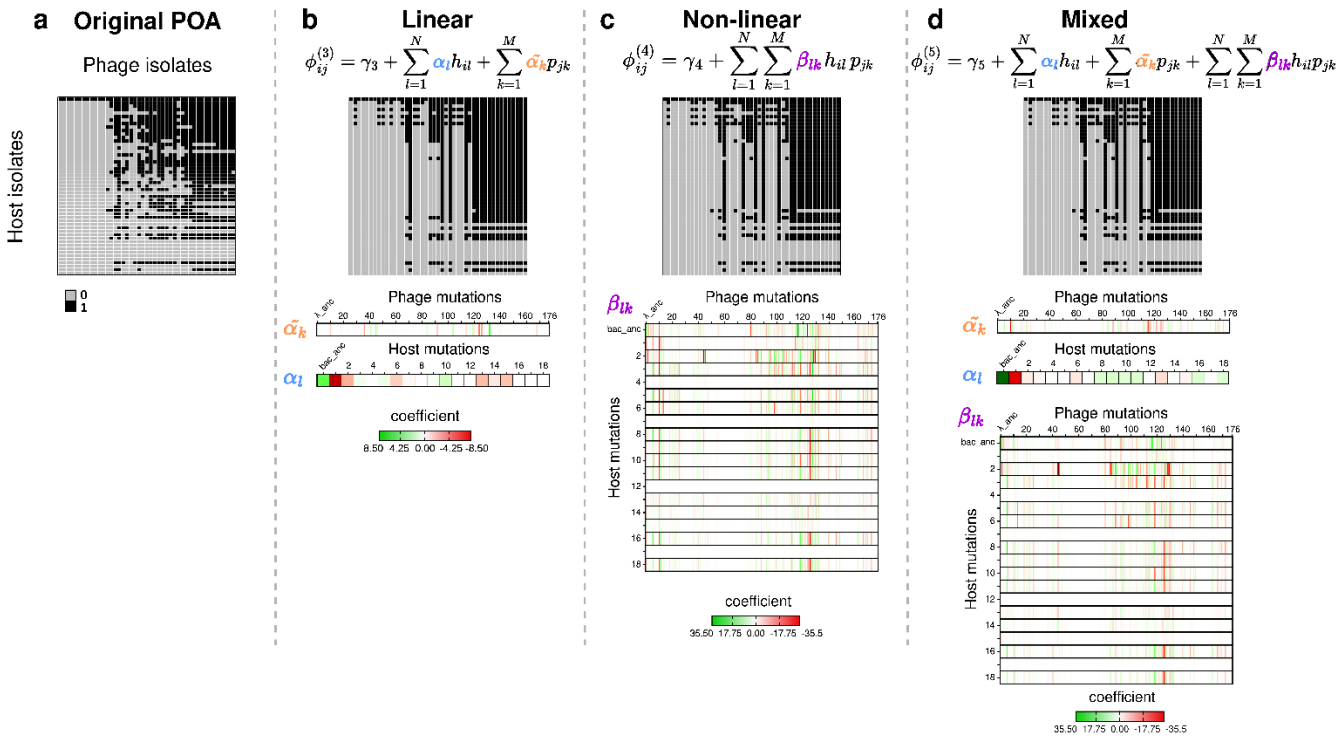


Figure 2. Model for predicting phage-host interaction network. (a) Original POA matrix showing presence (black) and absence (gray) of successful infection between phage (columns) and host (rows) isolated pairs. (b-d) Results of the different model predictions as of the POA matrices, and coefficient values for 176 phage and 18 host mutational features plus the ancestor trait using (b) a linear mutation set (equation [6]), (c) nonlinear mutation set (equation [8]) and (d) mixed combination of phage and host mutation set (for details see Methods section "feature construction"). The color of the coefficient indicates positive (green) to negative (red) effects of each mutational feature (phage: $\tilde{\alpha}_k$, host: α_l) or combination of mutations, β_{lk}.

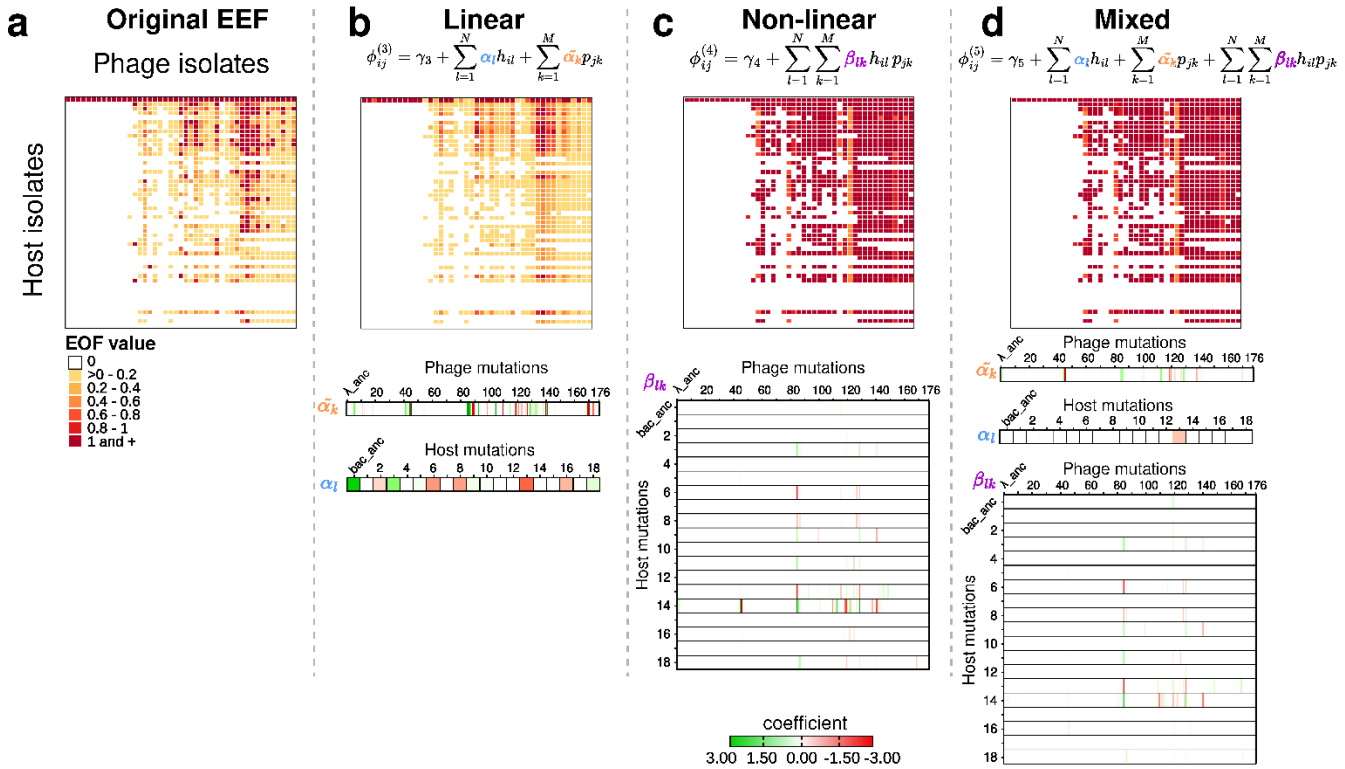


Figure 3. Model for predicting the efficiency of infection. (a) Original cross-infection matrix where colors are EOP values of infection between phage (columns) and host (rows) isolate pairs, white cells indicate no infection. (b-d) Results of the different model predictions as of the EFF matrices, and coefficient values for 176 phage and 18 host mutations plus the ancestor trait using (b) a linear mutation set (equation [6]), (c) nonlinear mutation set (equation [8]) and (d) mixed combination of phage and host mutation set (equation [10]) (for details see Methods section "feature construction"). The color of the coefficient indicates positive (green) to negative (red) effects of each mutational feature (phage: $\tilde{\alpha}_k$, host: α_l) combination of mutations, β_{lk}.

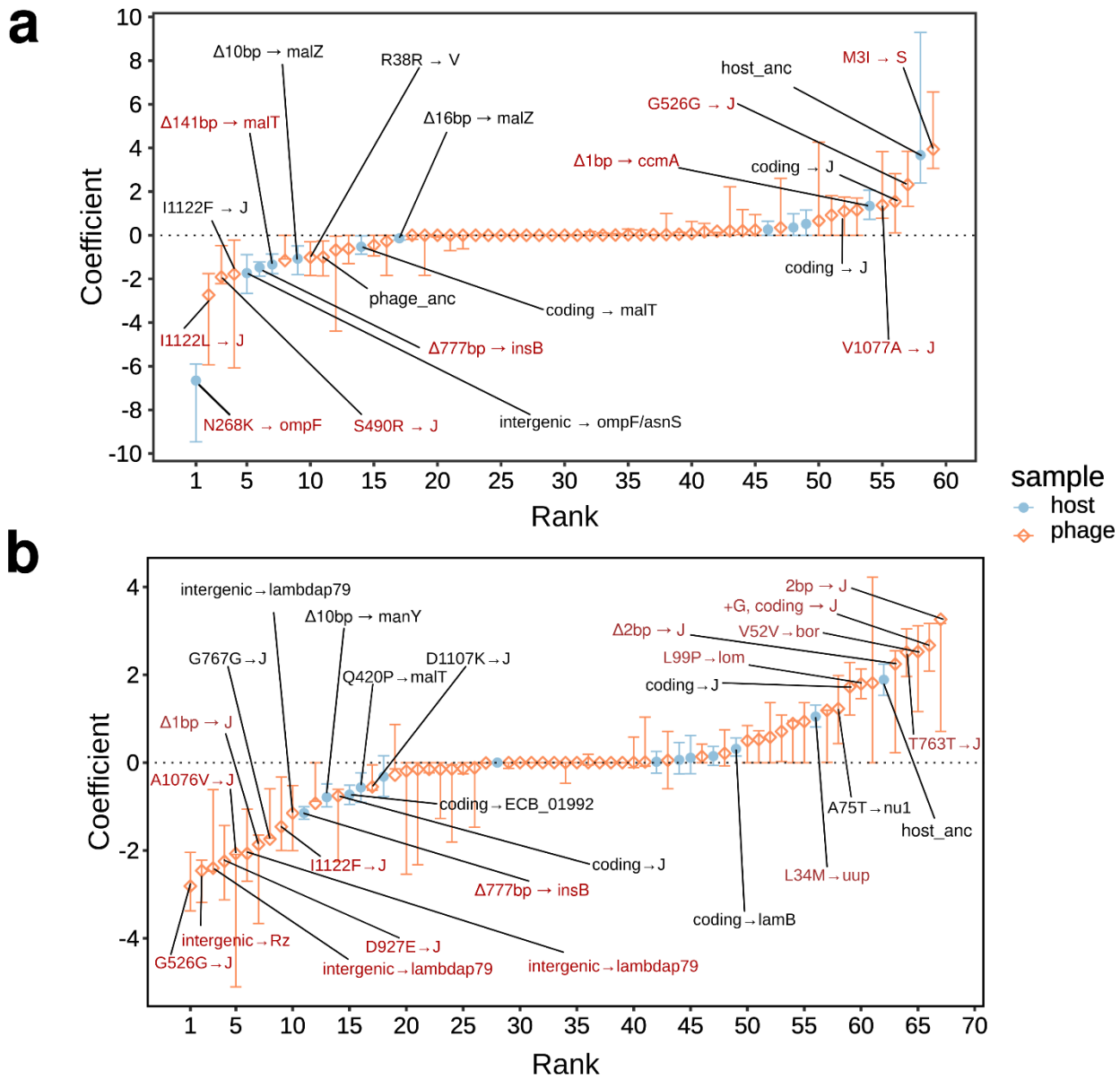


Figure 4. Rank ordered (most negative to most positive) coefficients inferred for putative mutations. The importance of mutational features was measured by the absolute value of the coefficients learned from each model. Error bars indicate 0.9 quantile. Labels indicate "mutation → gene" when the 90th quantile excludes 0. Mutations in red have the highest positive (negative) coefficients which lowest (highest) value is larger (smaller) or equal to 0 (from 200 bootstrap

478 runs) and are discussed in the main text. Important features for (a) the final model predicting
479 POA include a total of 59 non-zero coefficients, and (b) 67 non-zero coefficient values for the final
480 model predicting EFF. The complete lists of mean, maximum and minimum values of the
481 coefficients associated to mutations predicting POA and EFF are shown in Supplementary Data
482 S3 and Supplementary Data S4 respectively. For further details see Methods section "Feature
483 importance analysis"