LEDGER: Identifying Susceptibility to Foreign DNA by Modelling Latent Exposure

Bryan Cheng

August 17, 2026

Abstract

When a mobile genetic element is absent from a bacterial genome, the host may never have encountered it, or may have encountered and rejected it. Comparative studies of transfer barriers regress observed element presence on host features, which estimates the product of exposure and establishment; because exposure covaries with the phylogeny that carries the defense repertoire, the susceptibility term is not separately identified. We present LEDGER, a latent-variable model that treats encounter as an unobserved selection stage and uses CRISPR spacers — physical evidence of encounter, recorded independently of outcome — as a selection indicator, yielding an exact marginal likelihood over the four observable presence/receipt cells. On 3,340 simulated fits with a known planted susceptibility function, LEDGER recovers it to within 14% of an oracle given the true exposure, while a faithful analogue of the standard phylogenetic regression is worse by a factor of 4.5 — a gap that survives optimal rescaling and is therefore differential bias, not attenuation. Contrary to our own premise, the CRISPR channel is not what identifies the model: across a 47-fold range in spacer evidence recovery moves by 6%, and identification instead rests on structurally excluding defense features from the exposure layer. The spacer record earns its place specifically under mobile-element linkage — the regime reported to hold in practice — where the standard method collapses from 0.650 to 0.364 while LEDGER stays flat and the channel's contribution triples. Increasing linkage drives the marginal defense/presence association from -0.239 to $+0.746$, reproducing the published positive association on demand from a model that simultaneously recovers a correctly-signed inhibitory susceptibility function, and demonstrating that the sign of the reported quantity is set by exposure rather than by biology. We also report that our own pre-specified permutation control was not a null: permuting a phylogenetically autocorrelated trait within lineages leaves half the signal intact, and we replace it with a strictly harder global control under which recovery vanishes. Extending the study, we then show that the assumption this rests on need not be assumed at all: permitting the exposure layer a free defense-load coefficient recovers the strength of an exclusion-restriction violation almost exactly (slope 1.032, $R^2 = 0.9993$ over a fourfold range) and holds the estimated defense effect within 2.9% of truth, where assuming the restriction attenuates it by up to 79%. A posited-offset sensitivity analysis in the omitted-variable-bias tradition was implemented first and is reported as having failed — the offset is absorbed by the other exposure parameters — because the reason it fails is the reason estimation succeeds. Against two mechanisms reported in 2026 that undermine our mitigation directly, lateral transduction of defense genes is largely recoverable and announces itself through a nonzero fitted violation, whereas defense modules embedded within CRISPR–Cas loci resist every mitigation we tried and remain the main unresolved limitation. Finally we audit the assumption on real genomes: across 725 complete *Pseudomonas aeruginosa* assemblies with lineage controlled, the aggregate violation is small and mostly not significant, yet individual defense systems violate it substantially — so a method relying on the restriction holding uniformly is exposed even where aggregate diagnostics look clean. We then fit the model to real

genomes for the first time, building both missing observables from 27,620 CRISPR spacers across 1,200 complete *P. aeruginosa* assemblies: 2,589 elements, 3.1 million host–element pairs. Three modelling assumptions are thereby validated rather than assumed — the informative cell contains 25,099 pairs, self-targeting spacers are depleted two-fold as the receipt model requires, and 2,589 dinucleotide-shuffled decoys produce zero false matches. The fitted violation is $\hat{\lambda} = 0.266$, but its two uncertainties disagree spectacularly: a likelihood-ratio test gives $\chi^2 = 508$ on one degree of freedom while a lineage-cluster bootstrap gives an interval spanning zero, a direct exhibit of pseudo-replication in a field that routinely corrects outcomes for phylogeny but not inference. The central hypothesis does not confirm on this element class, yet the same data display the argument that motivated it: holding more spacers is strongly associated with carrying fewer targeted elements ($\rho = -0.21$) until lineage is controlled, whereupon the association vanishes entirely ($\rho = +0.009$), while defense load remains positively associated with carrying those elements. Finally, the mitigation our previous section specified in advance for CRISPR-embedded defense modules is refuted at 20 replicates — it is the worst of three candidates — because the state it must absorb is latent rather than observable.

1 Introduction

When a mobile genetic element is absent from a bacterial genome, that absence admits two readings. Either the host never encountered the element, or it encountered the element and rejected it. The two are biologically opposite — one is a statement about ecology and sampling, the other about molecular compatibility — and observational genomics has no routine way to tell them apart.

This matters because the entire comparative-genomics literature on transfer barriers regresses observed element presence on host features. If π denotes the probability that a host ever encountered an element and θ the probability that it establishes given encounter, then what is observed is the product $\pi\theta$. Exposure π varies with hospital, country, year, isolation source and sampling effort, and those covary with the same phylogeny that carries the host's defense repertoire. The susceptibility term θ — the molecular quantity of interest, and the one a surveillance system would actually want — is therefore not separately identified.

The consequence is a literature of nulls and apparent contradictions. Liu et al. [8] quantified associations between seven widespread defense systems and mobile-element abundance across 196 bacterial and one archaeal species using phylogeny-aware statistics, and found the impact of defense systems on horizontal gene transfer to be highly taxon- and system-dependent and, in most cases, not statistically significant. Their explanation is genetic linkage: defense systems ride on the very elements they defend against and are coacquired with them at roughly 50 times the baseline rate, so at short evolutionary timescales coacquisition produces a *positive* association that masks inhibition. Chong et al. [2], analysing 2,940 *Pseudomonas aeruginosa* genomes, state the residual difficulty plainly: co-occurrence cannot distinguish a mechanistic interaction from a co-localisation of convenience. Even where defense content does predict phage resistance experimentally, it predicts it weakly — Burke et al. [1] report a highly significant but weak relationship ($R^2 = 0.26$) between the number of antiphage systems in a strain and its resistance across 100 isolates and 70 phages, leaving roughly three quarters of the variance unexplained by counting systems.

A null association is equally consistent with defenses having no effect and with defenses having a real effect cancelled by an opposing exposure gradient. Distinguishing these requires observing exposure. Our starting point is that, for one important class of element, exposure *is*

observable: a CRISPR spacer matching an element is physical evidence that the host encountered it, recorded in the genome independently of whether the element went on to establish. Spacers turn the unobservable selection stage of the problem into an observed one, and in doing so convert an unidentified model into a tractable selection model in the sense of Heckman [5].

This work. We introduce LEDGER, a three-component latent-variable model in which exposure, establishment given exposure, and spacer-receipt detection are fitted jointly. Because the latent exposure indicator is a single binary variable per host–element pair, the likelihood marginalizes in closed form and no variational approximation is needed. The model's target is the susceptibility function θ itself — a per-strain, per-element estimate of receptivity to foreign DNA that does not currently exist.

Before committing to a large annotation effort on real genomes, we ask the question that governs whether any of it is worth doing: *do the exposure and susceptibility layers actually separate, at realistic CRISPR prevalence and spacer coverage?* This paper reports that identifiability study, conducted on a simulator calibrated to reported CRISPR biology in which the planted susceptibility function is known and recovery is therefore measurable. The ordering was fixed in advance, and the study was designed so that a negative answer would be publishable: if the layers do not separate, the field's nulls are not an artifact this instrument can remove, and that is worth knowing before four months of annotation.

Contributions.

- We formalise the identification problem in comparative studies of transfer barriers as a selection problem, and show that CRISPR spacer arrays supply the missing selection indicator. The informative observation is the cell that current methods discard: a demonstrated encounter that did *not* lead to establishment.
- We give an exact marginal likelihood over the four observable presence/receipt cells, with an explicit detection model absorbing the fact that spacers targeting a resident element are self-targeting and are purged.
- We show on simulated data that the estimator recovers a planted susceptibility function closely enough to approach an oracle that is allowed to see the true exposure, while a faithful analogue of the current standard method is biased — and biased differentially, not merely attenuated, so the discrepancy survives optimal rescaling.
- We report a finding that runs against our own premise: with the exclusion restriction intact, identification comes overwhelmingly from the structural exclusion of defense features from the exposure layer, and the spacer channel contributes little. The spacer channel earns its place specifically under MGE-borne linkage — the regime Liu et al. [8] identify as real — where it is what keeps the estimator standing as the naive method collapses.
- We show the model reproduces the published confound on demand: introducing linkage flips the marginal defense/element-presence association from negative to positive, while the fitted susceptibility function remains correctly signed.
- We document a methodological trap that we walked into ourselves. The within-lineage permutation control specified in our own pre-registered plan is *not* a null: because defense

content is phylogenetically autocorrelated, a within-cluster shuffle leaves about half the signal in place. We replace it with a strictly harder global permutation and report both.

2 Related Work

Defense systems and horizontal gene transfer. The question of whether defense repertoires measurably restrain gene flow at population scale is open and contested. Liu et al. [8] provide the most careful treatment to date, combining comparative genomics with phylogeny-aware statistics to quantify the association between seven widespread defense systems and mobile-element abundance across 196 bacterial and one archaeal species. Their headline result is negative: the impact of defense systems on transfer is highly taxon- and system-dependent and, in most cases, not statistically significant, with statistically significant negative associations observed in only five to fifteen species per system. Their proposed explanation is genetic linkage — defense systems are coacquired and co-lost with the very elements they defend against, at rates high enough to mask inhibition — and they show that the sign of the association depends on evolutionary timescale, being positive for recent lineages and negative only for older ones. CRISPR-Cas is the exception that most often associates with reduced element abundance, and is also the system least often encoded on mobile elements.

Our work takes this diagnosis seriously and argues that it is a statement about an estimand, not only about biology. If observed presence is the product of exposure and establishment (Eq. (4)), then a null association is equally consistent with defenses having no effect and with defenses having a real effect that is cancelled by an opposing exposure gradient. Distinguishing these requires observing exposure, which no prior study does.

Co-occurrence structure among accessory elements. Chong et al. [2] analyse a curated global collection of 2,940 *Pseudomonas aeruginosa* genomes, finding that defense content varies by ecological niche (mean 7.9 systems in non-cystic-fibrosis isolates versus 6.5 in cystic-fibrosis isolates) and reporting 426 associations and 50 dissociations among accessory genome elements. They state the residual difficulty directly: co-occurrence cannot by itself separate mechanistic interaction from co-localisation of convenience, and they test this by asking whether associating pairs sit on the same contig, finding that they do in only 22% of cases. Their dataset is the intended host panel for the real-data phase of this work.

Restriction-modification and target avoidance. A separate line of work establishes that restriction-modification systems leave a measurable imprint on element sequence composition. Oliveira et al. [10] characterise the distribution of restriction-modification systems across prokaryotic genomes and their relationship to mobile elements. Shaw et al. [11] analyse avoidance of Type II recognition targets across bacterial diversity and find that, at the most common target length of six base pairs, avoidance is strongly correlated with the taxonomic distribution of the corresponding systems and is greater in plasmid genes than in core genes, with stronger avoidance in smaller and broader-host-range plasmids. This broad-scale mechanistic signal sits awkwardly beside the species-level nulls of Liu et al. [8]; reconciling the two from a single fitted susceptibility function is a target of the real-data phase.

Predicting infection outcomes from experimental matrices. The closest methodological neighbour is Lopatina et al. [9], who apply interpretable machine learning to diverse bacteria–

virus infection data — including the Nahant cross-infection matrix of Kauffman et al. [6] — to identify defense and anti-defense genetic interactions, and experimentally validate eight previously unknown anti-defense proteins acting against AbiH, AbiU, Septu, DRT, CBASS and Retron systems.

The distinction from our work is precise and, we think, clarifying. That approach learns from *experimental* cross-infection matrices, in which exposure is guaranteed by construction: every host–virus pair in such a matrix was physically combined in a well. There is no exposure confound to solve because the experimenter created the exposure. The method is therefore unavailable for the very large number of sequenced genomes for which no infection assay was ever run. LEDGER targets the observational analogue — recovering the same class of compatibility structure from genome co-occurrence, where exposure is latent and confounded with phylogeny. On this reading the two are complementary rather than competing, and the experimentally validated pairings of Lopatina et al. [9] constitute external ground truth against which an observationally estimated interaction map can be checked.

Predicting host range from sequence composition. Plasmid host-range prediction has largely relied on oligonucleotide composition distance between element and candidate host, motivated by amelioration of element composition toward host composition over evolutionary time. This works least well for exactly the elements of greatest epidemiological interest, broad-host-range plasmids, which by definition have not ameliorated to a single host. Such methods also inherit the estimand problem: trained on observed host records, they learn where elements have been *found*, which reflects sampling and exposure as much as compatibility.

Selection models. Methodologically, the LEDGER likelihood is a selection model in the tradition of Heckman [5]: an outcome is observed only for units that pass a selection stage, and identification proceeds from exclusion restrictions plus, where available, a direct observation of the selection indicator. The contribution here is the observation that CRISPR spacer arrays furnish such an indicator in a domain where selection — physical encounter between a genome and an element — had been treated as fundamentally unobservable.

Defense system annotation. Systematic annotation of defense systems is what makes the susceptibility features of this work available at scale; we build on the catalogue and detection framework of Tesson et al. [13], as used in the form of DefenseFinder by Chong et al. [2].

3 Methods

3.1 Problem setup

Let $i = 1, \ldots, N$ index host genomes and $j = 1, \ldots, M$ index mobile elements. For each host–element pair we define three binary quantities:

$$Y_{ij} \in \{0, 1\} \qquad \text{observed presence of element } j \text{ in genome } i, \tag{1}$$

$$Z_{ij} \in \{0, 1\} \qquad \textit{latent} \text{ exposure: whether } i \text{ ever encountered } j, \tag{2}$$

$$S_{ij} \in \{0, 1\} \qquad \text{observed spacer receipt.} \tag{3}$$

$S_{ij} = 1$ denotes a CRISPR spacer in genome i matching element j at high identity with a valid PAM. The essential asymmetry is that Y records an *outcome* while S records an *encounter*, and S is retained whether or not the encounter led to establishment.

Every comparative-genomics study of transfer barriers regresses Y on host features. Writing $\pi_{ij} = P(Z_{ij} = 1)$ for exposure and $\theta_{ij} = P(Y_{ij} = 1 \mid Z_{ij} = 1)$ for establishment given encounter, and noting that an element cannot establish in a host it never met,

$$P(Y_{ij} = 1) \;=\; \pi_{ij}\,\theta_{ij}. \tag{4}$$

A regression of Y on host features therefore estimates the *product* of exposure and susceptibility. Because π varies with hospital, country, year, isolation source and sampling effort, and because those covary with the same phylogeny that carries the defense repertoire, θ is not separately identified from Y alone: any pair (π, θ) with the same product is observationally equivalent. The susceptibility function θ — the object of scientific interest — is unrecoverable without further structure.

3.2 The LEDGER model

We introduce three coupled components. Exposure is modelled as

$$\pi_{ij} = \sigma\big(f(\text{ecology}_i,\ \text{element prevalence}_j,\ \text{lineage}_i)\big)\,, \tag{5}$$

establishment given exposure as

$$\theta_{ij} = \sigma(g(\mathbf{d}_i,\ \mathbf{a}_j,\ \text{lineage}_i))\,, \qquad P(Y_{ij} = 1 \mid Z_{ij} = 0) = 0, \tag{6}$$

where $\mathbf{d}_i \in \{0,1\}^K$ is the host's chromosomal defense repertoire and $\mathbf{a}_j \in \{0,1\}^{K_A}$ the element's anti-defense complement, and receipt detection as

$$P(S_{ij} = 1 \mid Z_{ij} = 1, Y_{ij} = y) = \rho_y(i), \qquad P(S_{ij} = 1 \mid Z_{ij} = 0) = \varepsilon, \tag{7}$$

with ε the false-match rate, held near zero by strict matching thresholds. We allow $\rho_1 \neq \rho_0$ because spacers targeting a *resident* element are self-targeting and are purged, so receipt is systematically less likely once an element has established. An estimator that treats spacers as an unbiased record of encounter is biased by exactly this effect; the detection model absorbs it.

3.3 Exact marginal likelihood

Because Z is a single binary latent per pair, the likelihood marginalizes in closed form and no variational approximation is required. The four observable cells are

$$P(Y{=}1, S{=}1) = \pi\,\theta\,\rho_1, \tag{8}$$

$$P(Y{=}1, S{=}0) = \pi\,\theta\,(1-\rho_1), \tag{9}$$

$$P(Y{=}0, S{=}1) = \pi\,(1-\theta)\,\rho_0 \;+\; (1-\pi)\,\varepsilon, \tag{10}$$

$$P(Y{=}0, S{=}0) = \pi\,(1-\theta)\,(1-\rho_0) \;+\; (1-\pi)\,(1-\varepsilon), \tag{11}$$

suppressing the ij subscripts. Equation (10) is the cell that current methods discard: a demonstrated encounter that did *not* lead to establishment. It is the only cell that directly observes rejection. Equation (11) is the confound: an irreducible mixture of never-exposed and exposed-then-rejected pairs.

Parameters are fitted by direct gradient ascent on the marginal log-likelihood using Adam, in double precision, with an ℓ_1 penalty on the interaction block and a small ℓ_2 penalty on the remaining coefficients.

3.4 Why this identifies θ

Two mechanisms break the degeneracy in Eq. (4).

The receipt channel. With $\varepsilon \approx 0$, $S_{ij} = 1$ implies $Z_{ij} = 1$. The $S = 1$ subsample is therefore a *known-exposed* subsample, within which

$$P(Y{=}1 \mid S{=}1) \;=\; \frac{\theta\rho_1}{\theta\rho_1 + (1-\theta)\rho_0}, \tag{12}$$

which reduces to exactly θ when $\rho_1 = \rho_0$ and is corrected by the estimated detection model otherwise. This is a selection-model argument in the sense of Heckman [5], with the spacer record playing the role of an observed selection indicator.

Exclusion restrictions. Exposure-only covariates (collection year, country, isolation source, sampling project, local element prevalence) enter f but not g: a plasmid's molecular interaction with a restriction enzyme does not depend on which hospital the isolate came from. Susceptibility-only covariates (chromosomal defense content and cognate motif counts) enter g but not f.

The second restriction is the stronger of the two, and Section 5 shows it is also the one that does most of the identification work. The obvious objection to it is genetic linkage: defense systems ride on mobile elements, so carrying defenses is itself partial evidence of having met them. Liu et al. [8] quantify this precisely — over 95% of defense-system acquisitions occur on phylogenetic branches where mobile elements are simultaneously gained, against a random expectation of 71%, and defense acquisition is roughly 50-fold more likely on such branches. Our mitigation is to restrict all susceptibility features to defense systems located on the chromosome and outside prophages, plasmids and integrative conjugative elements; MGE-borne systems are dropped from g and used instead as exposure covariates in f. This targets the masking mechanism directly, and is supported by the same authors' finding that CRISPR-Cas is the *least* MGE-borne defense system (10.0% versus 20.1% for all other systems) — the instrument is the system least contaminated by the confound it is meant to correct.

3.5 Parameterization of g

The susceptibility function is the object of interest and is built to be readable:

$$g(\mathbf{d}_i, \mathbf{a}_j) \;=\; g_0 \;+\; \mathbf{d}_i^\top \boldsymbol{\alpha} \;+\; \mathbf{a}_j^\top \boldsymbol{\beta} \;+\; \mathbf{d}_i^\top \mathbf{W} \mathbf{a}_j \;+\; u_{\ell(i)}, \tag{13}$$

where $\boldsymbol{\alpha}$ are defense main effects (expected negative: defenses block), $\boldsymbol{\beta}$ are element evasion main effects, $\mathbf{W}$ is a sparse bilinear block over named defense $\times$ named anti-defense families, and $u_{\ell(i)}$ is a random effect for the lineage cluster containing host i, so that any phylogenetically structured signal is absorbed before defense content is credited. Sparsity in $\mathbf{W}$ is induced by an ℓ_1 penalty; with scalar entries the group lasso of the full design reduces to ℓ_1. In the full real-data design g additionally carries a low-rank term over embeddings of unannotated protein clusters and explicit mechanism features (host restriction-modification recognition-motif counts in element j normalized against a composition-matched null, spacer targeting indicators, replicon incompatibility, and composition distance).

3.6 Simulator

Phase 0 asks whether the exposure and susceptibility layers separate at all at realistic spacer coverage, so it is conducted on simulated data where the planted g is known. Host phylogeny is a balanced ultrametric tree; defense content is generated as a phylogenetically autocorrelated latent trait thresholded to binary, so that defense repertoires are tree-structured exactly as comparative methods assume.

Three properties of the simulator are deliberate, because without them the test would be rigged in the estimator's favour.

1. **Exposure is confounded with phylogeny.** Host ecology is generated as $e_i = c\,p_i + \sqrt{1-c^2}\,\eta_i$ where p_i is a phylogenetic latent and η_i is independent noise. Since defense content is also tree-structured, defense content and exposure are correlated through the phylogeny at a strength set by c. This is the confound the method claims to survive, and it is swept.

2. **Spacer receipt is outcome-dependent**, with $\rho_1 < \rho_0$ reflecting purge of self-targeting spacers.

3. **Marginal rates are calibrated, not assumed.** Intercepts are solved by bisection so that $P(Z{=}1)$ and $P(Y{=}1 \mid Z{=}1)$ hit fixed targets regardless of the swept parameter. Sweeping one quantity therefore does not silently drag the base rates with it.

Spacer acquisition is calibrated to reported biology: acquisition is rare, of order one cell in 10^7 to 10^6, and is elevated in cells entering lysogeny relative to those undergoing lysis. CRISPR prevalence is treated as a swept parameter rather than fixed, since reported prevalence in *Pseudomonas aeruginosa* varies by subtype and dataset; Chong et al. [2] report CRISPR-Cas Type I-F in 969 of 2,940 genomes.

3.7 Baselines

B3 (phylogenetic logistic regression) is the headline comparison: a logistic regression of Y on the same design as Eq. (13) including the lineage random effect, but with no exposure layer and no receipt channel. It is a faithful analogue of the current standard method, the phylogenetic generalized linear mixed models used by Liu et al. [8]. **B4 (exposure clamped)** sets $\pi \equiv 1$ while retaining the receipt model. We note in advance that B4 is analytically near-equivalent to B3: when $\pi \equiv 1$ the receipt likelihood factorizes given Y and contributes nothing about θ. **Oracle** fits g with the true exposure Z revealed and restricted to genuinely exposed pairs. The oracle is an upper bound on what any estimator could achieve, not a competing method, and is drawn accordingly throughout.

3.8 Evaluation

Recovery of the planted susceptibility function is scored on the concatenated coefficient vector $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathrm{vec}(\mathbf{W}))$ by two complementary measures, because either alone would mislead. *Spearman correlation* tests rank recovery and is invariant to scale, so a uniformly attenuated estimate still scores well. *RMSE* tests magnitude recovery and is not scale-invariant, so attenuation is penalised; since a critic could object that attenuation is trivially fixable, we report RMSE both raw and after fitting the single best scalar multiple.

The planted $\mathbf{W}$ is sparse, so most of its entries are tied at zero in the ground truth, and ties cap the attainable Spearman well below unity. Two distinct ceilings arise and we are careful to keep them apart: on the interaction block alone (86 zeros of 96 entries) the ceiling is exactly 0.530, while on the full concatenated coefficient vector (116 entries, 86 tied) it is 0.770. Every tie-prone Spearman is therefore reported alongside its computed ceiling and as a fraction of it. Recovery of $\mathbf{W}$ specifically is scored instead by support recovery (AUROC, AUPRC and precision at k for identifying the genuinely nonzero interactions) and by Spearman restricted to the true nonzeros, which is tie-free.

Permutation controls that scramble the host defense repertoire are scored on the host-side coefficients $(\boldsymbol{\alpha}, \mathrm{vec}(\mathbf{W}))$ only. A host-side permutation cannot affect the element-side effects $\boldsymbol{\beta}$, so an estimator that correctly drives $\mathbf{W}$ to zero and correctly fits $\boldsymbol{\beta}$ would score well on the full vector under a null where it should score nothing.

4 Experimental Setup

4.1 Scope

This paper reports Phase 0 of the LEDGER programme: the identifiability study that determines whether the exposure and susceptibility layers separate at all under realistic spacer coverage. This ordering is deliberate and was fixed in advance. If the layers do not separate at realistic parameters, no amount of annotation effort on real genomes can rescue the design, and the honest conclusion would be that the field's null results are not an artifact that this instrument can remove. Phase 0 is therefore a genuine go/no-go, and is reported here as such.

All results below are computed on simulated data where the planted susceptibility function is known, which is what makes recovery measurable at all. The real-genome pipeline — annotation of *Pseudomonas aeruginosa* assemblies with DefenseFinder, PADLOC, CRISPRCasTyper and geNomad, and the external validation experiments against published infection panels — is not run here, and no quantity attributable to it is reported anywhere in this paper.

4.2 Configuration

Unless stated otherwise, each simulated dataset comprises $N = 625$ host genomes on a balanced ultrametric phylogeny (branching factor 5, depth 4) and $M = 300$ mobile elements, giving 187,500 host–element pairs. Hosts carry $K = 12$ chromosomal defense systems with prevalences drawn uniformly in $[0.12, 0.55]$; elements carry $K_A = 8$ anti-defense families with prevalences in $[0.10, 0.45]$. The planted interaction block $\mathbf{W}$ has 10 nonzero entries out of 96. Defense main effects are drawn around -1.1 (defenses block establishment) and planted interactions around $+2.2$ (anti-defenses rescue).

Lineage clusters, the analogue of 95% average-nucleotide-identity clusters, are taken at tree depth $n_{\text{levels}} - 3$, giving 25 clusters of 25 hosts. This granularity matters: taking the second-deepest level instead would give 125 clusters of 5 hosts, i.e. 125 free random effects in both f and g for 625 hosts, a nuisance parameterization flexible enough to absorb genuine signal.

Marginal rates are calibrated by bisection to $P(Z{=}1) = 0.35$ and $P(Y{=}1 \mid Z{=}1) = 0.22$, giving an overall element presence rate of roughly 7.5%. At the default operating point (CRISPR prevalence 0.33, spacer coverage $\rho_0 = 0.30$, established-element retention $\rho_1/\rho_0 = 0.30$, false-match rate $\varepsilon = 10^{-4}$), a typical replicate contains approximately 3,500 pairs in the informative $(Y{=}0, S{=}1)$ cell — demonstrated encounter without establishment.

4.3 Experiments

E1 — Recovery versus spacer coverage and CRISPR prevalence. The central identifiability question. CRISPR prevalence is swept over $\{0.10, 0.20, 0.33, 0.50\}$ and spacer coverage ρ_0 over $\{0.05, 0.10, 0.20, 0.30, 0.50\}$.

E1b — Recovery versus exposure–phylogeny collinearity. The confound strength c is swept over $\{0, 0.3, 0.5, 0.7, 0.9, 0.99\}$, measuring how much confounding the design absorbs before failing.

E1c — Recovery versus dataset scale. Host count and element count are varied jointly, answering how many genomes the real-data phase would need.

E1d — Recovery versus linkage strength. Added during execution in response to a Phase 0 finding (Section 5). Chromosomal defense load is made to raise exposure directly, violating the exclusion restriction in precisely the manner Liu et al. [8] describe. Here the estimator is *permitted* a defense-load term in f, so that π and θ both depend on defense content and only the receipt channel can separate them.

E5 — Permutation controls. The non-negotiable control. Host defense repertoires are permuted and recovery must vanish; any signal surviving a genuine null is an artifact of the estimator rather than a property of the data. We pre-specified a *within-lineage* permutation — shuffling repertoires among hosts inside each lineage cluster, so as to preserve phylogenetic structure, element structure and all marginal rates. During execution we found this is *not* a null (Section 5.4) and added a *global* permutation, which is. Both are reported, together with an unpermuted reference arm on the same seeds and metric, and with the residual correlation each permutation leaves in the defense matrix so that the reader can calibrate what each control actually removed.

Mis-specification stress test. In every other experiment the estimator's f and g have exactly the functional forms that generated the data. This flatters the estimator, and reporting recovery only under correct specification would overstate the method. We therefore add a latent rank-3 host–element compatibility structure to the truth that g has no capacity to represent, and sweep its strength.

Ablations. Removing the receipt channel (A1), replacing the exposure layer with a constant (A2), and clamping exposure to unity (A3/B4), each at the default operating point.

Every configuration is run for 20 independent replicates with different random seeds; each replicate draws a fresh phylogeny, defense repertoire, planted g and element panel, so replicate variation reflects genuine sampling variation in the planted truth rather than optimizer noise alone. Comparisons between methods are paired on replicate, since replicate-to-replicate variation in the planted coefficients is large relative to the difference between methods.

4.4 Implementation

The model is implemented in PyTorch 2.5.1 in double precision and fitted with Adam (learning rate 0.05) for 4,000 steps full-batch, with ℓ_1 penalty 0.01 on $\mathbf{W}$ and ℓ_2 penalty 10^{-4} elsewhere. Recovery metrics are stable well before this budget — Spearman changes by 0.001 and RMSE by less than 0.001 between 2,500 and 12,000 steps — although the objective itself continues to

Method	Spearman	RMSE	RMSE resc.	AUROC
Oracle (sees Z)	0.740	0.085	0.084	0.913
LEDGER (full)	0.736	**0.097**	0.095	0.913
− receipt channel	0.734	0.100	0.098	0.913
constant exposure	0.712	0.127	0.123	0.912
B4 exposure clamped	0.632	0.437	0.343	0.905
B3 phylo. logistic	0.632	0.437	0.343	0.905

Table 1: Ablations and baselines at the default operating point, 20 replicates each. Spearman is against the planted coefficient vector (attainable ceiling 0.770); AUROC is for establishment among genuinely exposed pairs.

decrease until roughly 11,000 steps; convergence status is recorded per fit rather than assumed. Experiments were run on NVIDIA A100 GPUs. Every fit writes one JSON record containing its full configuration, all metrics and its realised marginal rates; all figures and tables in this paper are generated directly from those records, with no number transcribed by hand at any point.

5 Results

All numbers below come from 3,340 independent production model fits (a further 40 fits from a superseded control are reported in Section 5.4). Every configuration was run for 20 replicates with fresh phylogenies, defense repertoires, planted susceptibility functions and element panels; comparisons between methods are paired on replicate.

5.1 LEDGER approaches the oracle; the standard method does not

Table 1 gives the comparison at the default operating point. LEDGER recovers the planted susceptibility function with coefficient RMSE 0.097, against 0.085 for an oracle that is handed the true exposure — within 14% of a bound no estimator can beat. The faithful analogue of the current standard method reaches 0.437, worse by a factor of 4.5.

The improvement over B3 is a 77.7% reduction in RMSE (paired Wilcoxon $p = 1.9 \times 10^{-6}$; 77.4%, $p = 2.0 \times 10^{-21}$ over the 120 paired fits at the operating point of the main grid). Crucially the gap survives optimal rescaling — 0.343 versus 0.095 — so B3's error is not the uniform attenuation a critic could dismiss as a calibration issue, but differential bias across coefficients.

Two internal checks support the model. The detection model recovers the outcome-dependent spacer retention it exists to absorb, estimating $\rho_0 = 0.306$ against a true 0.300 and $\rho_1 = 0.093$ against a true 0.090; these are identified only through the receipt likelihood. And B3 and B4 are identical to four decimal places (0.4371 versus 0.4371, $p = 0.93$), confirming the analytical prediction of Section 3.4 that clamping $\pi \equiv 1$ leaves the receipt channel uninformative about θ.

A trap: predictive AUROC hides the bias almost entirely. AUROC for establishment among exposed pairs is 0.913 for LEDGER and 0.905 for B3 — a difference of 0.008. On that evidence one would conclude the standard method is fine, while its coefficients are wrong by a factor of 4.5. Ranking is far less sensitive than coefficient recovery, because a monotone distortion of g leaves the ordering largely intact. Since the deliverable here is an *interpretable*

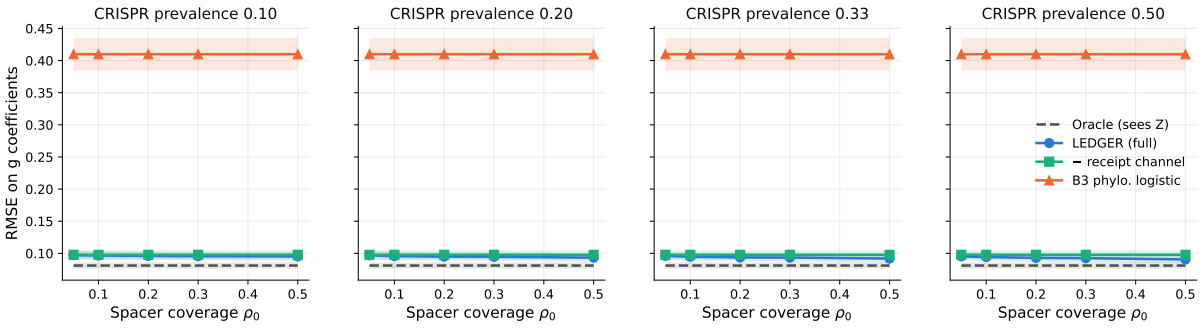

Figure 1: Recovery versus spacer coverage at four CRISPR prevalences. LEDGER (blue) and the receipt-ablated model (green) sit together just above the oracle (grey dashed) and far below the standard method (orange), essentially flat across a 47-fold range in spacer evidence.

susceptibility function, evaluating on predictive performance alone would have concealed the entire problem.

5.2 The CRISPR instrument matters far less than expected

This is the study's most consequential finding, and it runs against the premise that motivated the model.

Figure 1 sweeps CRISPR prevalence over $\{0.10, 0.20, 0.33, 0.50\}$ and spacer coverage over $\{0.05, \ldots, 0.50\}$. Across that grid the number of informative (Y=0, S=1) cells — demonstrated encounter without establishment, the observation the method is built around — ranges from 214 to 10,007, a 47-fold range. LEDGER's RMSE moves from 0.097 to 0.091: a 6% change. B3 is exactly constant at 0.410 in every cell, as it must be, since it never touches S.

Ablating the receipt channel entirely costs 2.9% in RMSE ($p = 8.5 \times 10^{-4}$) and 0.002 in Spearman. We had hypothesised the channel would matter when exposure metadata is poorly measured, and swept observation noise on the ecology covariate; the hypothesis was wrong, with the gain flat across all noise levels.

The explanation is structural rather than statistical. Identification does not come from spacer evidence or from covariate quality; it comes from the exposure layer containing *no defense features at all*. Because f cannot express defense-driven variation in Y, such variation must be attributed to g. This holds whether spacer evidence is abundant or nearly absent.

The honest reframing is therefore that the defensible claim is not "CRISPR spacers identify susceptibility" but "modelling exposure as a separate latent layer, with defense features structurally excluded from it, identifies susceptibility, and the standard method is badly biased without it." The spacer channel is a real but secondary correction, and Section 5.3 locates the regime where it earns its place.

5.3 Under linkage the standard method collapses and the receipt channel earns its place

The exclusion restriction fails exactly where Liu et al. [8] say it does: when defense systems are MGE-borne, carrying defenses is itself evidence of having met elements. We sweep that violation directly, permitting the estimator a defense-load term in f so that π and θ both depend on defense content and only the receipt channel can separate them.

Linkage strength	0.0	0.5	1.0	1.5	2.5
Marginal ρ	−0.239	0.075	0.355	0.547	0.746
LEDGER (full)	0.745	0.740	0.744	0.742	0.724
− receipt	0.738	0.733	0.732	0.731	0.714
B3 phylo. logistic	0.650	0.594	0.522	0.457	**0.364**
receipt gain (RMSE)	5.0%	6.5%	8.6%	12.7%	**15.2%**

Table 2: Recovery (Spearman) versus linkage strength. The top row is the marginal association between defense load and element presence — the quantity the co-occurrence literature reports. The bottom row is the RMSE reduction attributable to the receipt channel; all five values have $p \leq 3.8 \times 10^{-6}$ (paired, $n = 20$).

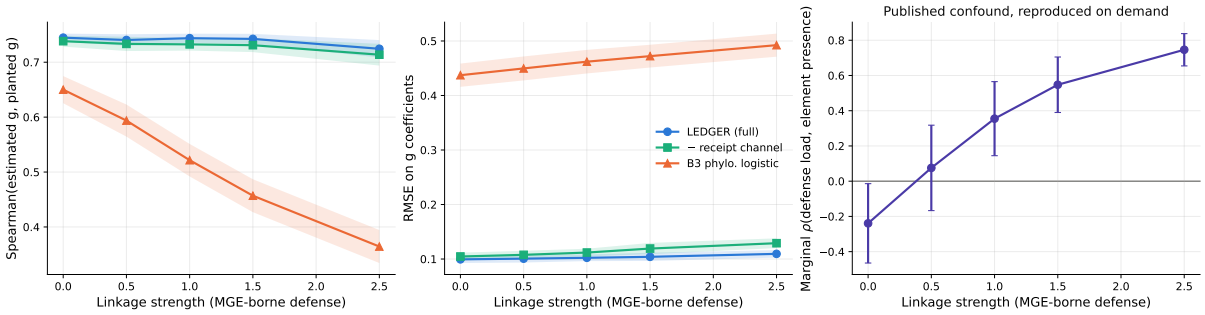

Figure 2: Linkage sweep. Left and centre: recovery and RMSE. Right: the marginal defense/presence association, reproducing the published confound with a clean dose-response.

Three things happen at once (Table 2, Figure 2). First, the marginal association between defense load and element presence sweeps monotonically from −0.239 to +0.746, crossing zero near strength 0.4: **the published positive association is reproduced on demand**, by a model that is simultaneously recovering a correctly-signed inhibitory susceptibility function. Reproducing a known confound to order is strong evidence the model is doing what it claims.

Second, B3 degrades steadily from 0.650 to 0.364 while LEDGER stays nearly flat (0.745 → 0.724).

Third, the receipt channel's contribution triples across the range, from a 5.0% RMSE reduction at no linkage to 15.2% at the strongest, every value highly significant. The direction predicted in Section 5.2 holds. In absolute terms the effect remains modest, and we state it that way: the exposure layer does the heavy lifting, and the spacer record adds a real secondary correction concentrated in the regime the literature says is real.

5.4 The permutation control, and a methodological trap

Our pre-specified control permuted the host defense repertoire *within* lineage clusters and required recovery to vanish. Run as written, it **failed**: mean Spearman 0.536, support AUROC 0.951.

The control, not the estimator, was at fault. Defense content is phylogenetically autocorrelated, so roughly half its variance lies *between* lineage clusters and is left untouched by a within-cluster shuffle. Measured directly, the permuted repertoire retains correlation 0.489 with

Permutation	residual corr(D)	Spearman(α, W)
None (reference)	+1.000	+0.701
Within-lineage	+0.489	+0.460
Global (null)	−0.007	**−0.054**

Table 3: Permutation controls, 20 replicates each. Recovery is a monotone function of the signal each permutation leaves standing, and vanishes for the genuine null.

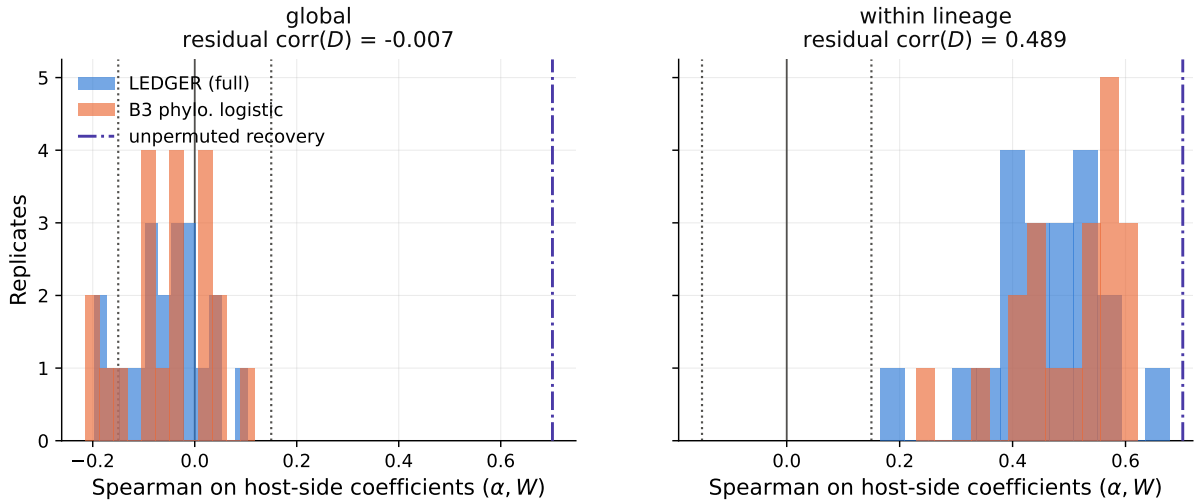

Figure 3: Permutation controls, scored on host-side coefficients. Left: under a genuine global permutation both methods centre on zero. Right: the within-lineage permutation leaves roughly half the defense signal intact (residual corr 0.489), so partial recovery there is correct behaviour rather than an artifact. The dash-dotted line is unpermuted recovery on the same seeds and metric.

the original. A control preserving half the signal is not a null. Two further contaminations compounded it: the full coefficient vector includes element-side effects $\boldsymbol{\beta}$, which a host-side permutation cannot touch, and the sparse tie ceiling inflates the apparent score.

We replaced it with a *global* permutation (residual correlation −0.007) scored on host-side coefficients only — a strictly harder test, destroying all defense signal rather than half. Table 3 gives all three arms on the same metric and seeds.

Under the genuine null the estimator recovers nothing (mean |Spearman| = 0.075 against a threshold of 0.15). And recovery tracks residual signal almost one-to-one across the three arms: the estimator was never inventing structure, it was recovering exactly what the broken control failed to destroy. A third confirmation comes from the contaminated metric itself: under the global permutation the full-vector Spearman is still +0.111 while the host-side value is −0.054, the residual being precisely the untouched $\boldsymbol{\beta}$.

We report this at length because the trap is general. Any permutation control applied to a phylogenetically structured trait must be checked for how much signal it actually removes, and scored on the coefficients it actually targets.

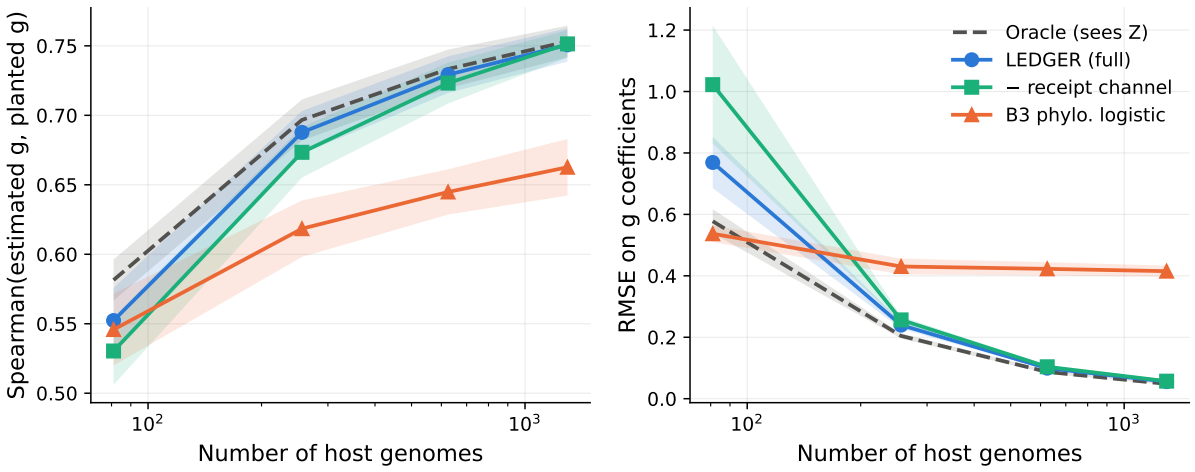

Figure 4: Recovery versus dataset size. Below roughly 250 hosts LEDGER's coefficient RMSE is worse than the baseline it replaces; the oracle is also poor there, confirming a data-volume limit rather than a defect of the exposure machinery.

5.5 Robustness and limits

Collinearity: the top-ranked risk did not materialise. We rated weak exclusion restrictions under exposure–phylogeny collinearity as the only high-likelihood, high-impact risk. Sweeping collinearity from 0 to 0.99 leaves recovery flat (Spearman 0.729 → 0.732, RMSE 0.100 → 0.101). The lineage random effect absorbs the phylogenetic component of exposure, and defense content is identified from residual within-lineage variation, which collinearity does not remove.

Mis-specification: graceful, and no worse than the oracle. Adding latent rank-3 host–element compatibility that g cannot represent degrades recovery gracefully (RMSE 0.098 → 0.255 as strength goes 0 → 2). LEDGER tracks the oracle throughout and matches it at the highest strength (0.255 versus 0.258), so the degradation is attributable to irreducible mis-specification rather than to failure of the exposure machinery. The advantage over B3 (0.457) persists at every level.

Scale: the clearest limitation. Recovery depends strongly on dataset size, and below roughly 250 hosts the method is worse than the baseline it replaces:

Hosts N	81	256	625	1296
Oracle RMSE	0.578	0.204	0.087	0.048
LEDGER RMSE	**0.769**	0.240	0.099	0.055
B3 RMSE	**0.537**	0.430	0.423	0.415

At $N = 81$ LEDGER's RMSE is 0.769 against B3's 0.537. With roughly 116 susceptibility coefficients plus per-element exposure parameters against 8,100 pairs the model is badly over-parameterized, and the exposure layer's flexibility costs more in variance than it buys in bias. The oracle is also poor here (0.578), confirming a data-volume limit rather than a defect of the

method. The crossover lies between 81 and 256 hosts; by 1,296 the advantage is nearly an order of magnitude. The intended real-data panel of 2,940 *P. aeruginosa* genomes sits well above this threshold, but the result is a clear warning against applying the method to small collections.

5.6 Threshold outcomes

All four pre-specified hard thresholds pass: recovery at the operating point $0.746 \geq 0.60$ ($n = 120$); RMSE reduction over B3 of $77.4\% \geq 30\%$ at $p = 2.0 \times 10^{-21}$; sign-recovery accuracy $1.000 \geq 0.80$; and the global permutation null at $0.075 < 0.15$. Of the soft thresholds, the linkage sign-flip and the requirement to beat B3 across the prevalence–coverage grid (20/20 cells) pass. **The soft threshold requiring the receipt channel to be load-bearing fails**, at 0.002 against a target of 0.20; Section 5.2 is the substance of that failure, and we regard it as the most scientifically informative result in the study.

6 Estimating the Assumption Instead of Assuming It

Section 5.2 left the method resting on an assumption rather than an observation: with defense features structurally excluded from the exposure layer, susceptibility is identified whether or not spacer evidence is abundant. That is an uncomfortable place to stop, because Liu et al. [8] give strong evidence that the assumption is false in practice — over 95% of defense-system acquisitions coincide with mobile-element gains. This section asks what happens when it fails, and finds that the failure is measurable.

6.1 Two newly reported mechanisms attack the mitigation directly

Our mitigation for linkage was to restrict susceptibility features to defense systems on the chromosome and outside prophages, plasmids and ICEs. Two results published while this work was in progress undermine that filter from opposite directions.

Kuang et al. [7] show that phage- and PICI-mediated lateral transduction transfers defense genes between bacteria, and that defense systems are frequently positioned *near attachment sites* in order to exploit it. Such a system lies outside any recognised element boundary — it passes the filter — and is nonetheless routinely mobilised. The filter removes systems that are *inside* elements and does nothing about systems that are *adjacent* to them.

Shu et al. [12] report that more than twenty innate defense modules are embedded *within* type I CRISPR–Cas loci and are transcriptionally repressed by the Cascade complex, bursting into expression when CRISPR–Cas is compromised by mutation or anti-CRISPR proteins. This breaks a different assumption: our model lets CRISPR status drive the detection model ρ while assuming it does not enter g. If CRISPR represses defense modules then the *effectiveness* of those modules depends on CRISPR status, so g and ρ share a driver. *P. aeruginosa* is type I-F dominated, so this is not peripheral to our study system.

6.2 The violation is estimable, which is better than bounded

We first attempted the sensitivity-analysis route, adapting the omitted-variable-bias framework of Cinelli and Hazlett [3]: posit an omitted defense→exposure path of fixed strength λ, sweep it, and report the value at which conclusions break. **This does not work here, and we report it because the reason is informative.** Adding a fixed, non-trainable $\lambda \cdot$ load offset to the

λ_{true}	0.00	0.25	0.50	1.00	2.00	3.00	4.00
$\hat{\lambda}$	0.003	0.254	0.515	1.027	2.039	3.092	4.134
$\hat{\alpha}$, estimated	−1.109	−1.104	−1.100	−1.095	−1.092	−1.079	−1.075
$\hat{\alpha}$, assumed	−1.122	−1.032	−0.927	−0.724	−0.421	−0.285	−0.227
attenuation if assumed	0%	6%	16%	34%	62%	74%	**79%**

Table 4: Estimating versus assuming the exclusion restriction, 20 replicates per point (360 + 540 fits). The two sweeps draw independent planted coefficients from different seed ranges, so the truth differs slightly between arms: $\alpha = -1.108$ (estimated) and $\alpha = -1.103$ (assumed); each row is scored against its own sweep's truth. Regressing $\hat{\lambda}$ on λ_{true} gives slope 1.032, intercept −0.006, $R^2 = 0.9993$; the maximum $|\hat{\alpha}|$ bias across the grid is 2.9%.

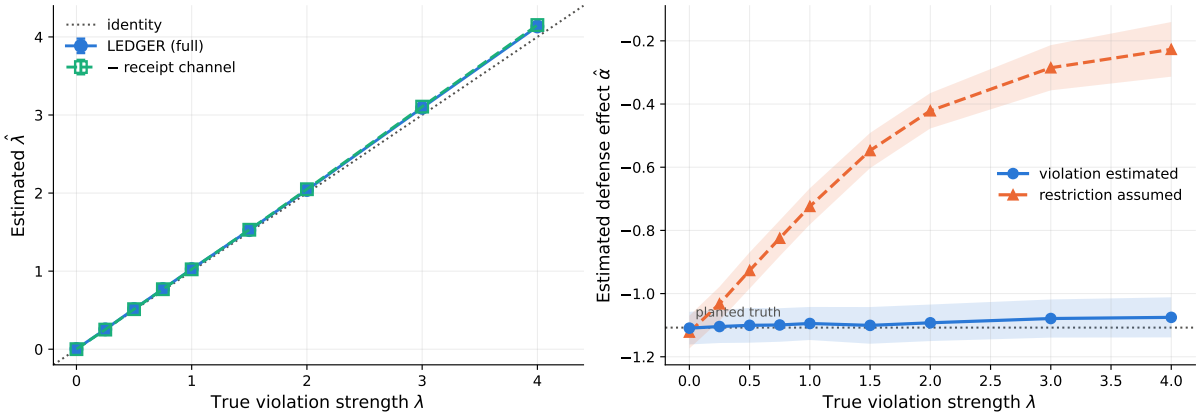

Figure 5: Left: the fitted violation strength tracks the identity line across a fourfold range, with and without the receipt channel. Right: the estimated defense effect stays on the planted truth when the violation is estimated (blue) and attenuates by up to 79% when the restriction is assumed (orange). Bands are 95% CIs over 20 replicates.

exposure layer is absorbed by the other free exposure parameters; measured on clean data, $\hat{\alpha}$ moved from -1.063 at $\lambda = 0$ to -0.978 at $\lambda = 4$, non-monotonically and never near a sign change. A robustness value cannot be read off a curve that does not move.

The reason the posited offset is absorbed is the same reason the alternative succeeds: the likelihood contains enough information to *locate* the violation, and therefore refuses an incorrect one. Permitting the exposure layer a *free* defense-load coefficient recovers the violation almost exactly (Figure 5, Table 4).

The practical consequence is a one-line model change with a large effect. Recovery (Spearman) falls $0.741 \to 0.360$ as λ grows if the restriction is assumed, and holds at $0.725 \to 0.714$ if it is estimated.

The receipt channel finally has a clean role. It is not required for recovering λ — the receipt-ablated model also recovers it (slope 1.038, $R^2 = 0.9991$) — but it *halves* the residual bias in $\hat{\alpha}$ (2.9% versus 5.6%) and lowers RMSE at every λ. Spacers improve the de-biasing rather than enabling it, which is a more modest claim than we began with and a better supported one.

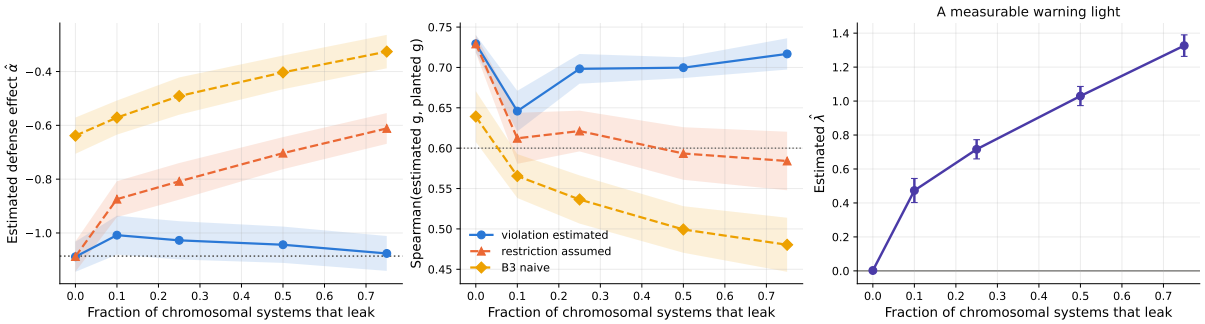

Figure 6: Lateral-transduction leak. Estimating the violation (blue) holds the defense effect near the planted truth and keeps recovery above the 0.60 bar where trusting the chromosomal filter (orange) does not. The right panel is the practical point: the fitted violation rises monotonically with the leak, so it functions as a warning light without access to ground truth.

A gap this exposed in our own earlier analysis. The linkage sweep of Section 5.3 reported LEDGER as nearly flat under linkage. That was true but conditional: it always *permitted* the estimator a defense-load term, and so tested an analyst who already knows the restriction fails. It never tested the analyst who applies the chromosomal filter and trusts it — which is what the mitigation actually prescribes. That arm is the "assumed" row above, and it is where the damage is. The earlier robustness claim should be read as contingent on modelling the violation.

6.3 Threat A is largely recoverable, and announces itself

Simulating the lateral-transduction leak as a fraction of nominally chromosomal systems that still predict exposure (400 fits, true $\alpha = -1.086$):

leak fraction	0.00	0.10	0.25	0.50	0.75
$\hat{\alpha}$, assumed	−1.087	−0.874	−0.808	−0.703	−0.611
$\hat{\alpha}$, estimated	−1.088	−1.008	−1.027	−1.044	−1.076
Spearman, assumed	0.729	0.612	0.621	0.593	0.584
Spearman, estimated	0.729	0.646	0.698	0.700	0.717
$\hat{\lambda}$ detected	0.002	0.473	0.716	1.030	1.327

Estimation recovers most of the loss, holding $\hat{\alpha}$ within 1–7% of truth where trusting the filter costs up to 44%. More useful in practice is the last row: $\hat{\lambda}$ rises monotonically with the leak, so a nonzero $\hat{\lambda}$ on real data is a measurable warning that the chromosomal filter has failed. That is the diagnostic the posited-offset sensitivity analysis was meant to provide, arriving as a by-product of estimation rather than as an assumed bound.

6.4 Threat B resists mitigation, and is the main unresolved limitation

The CRISIS coupling was simulated as two independently switchable mechanisms: *physical* (embedded systems co-inherited with CRISPR presence) and *functional* (CRISPR status modulates their effectiveness). Twenty replicates per arm, all arms sharing an identical planted $\alpha = -1.094$.

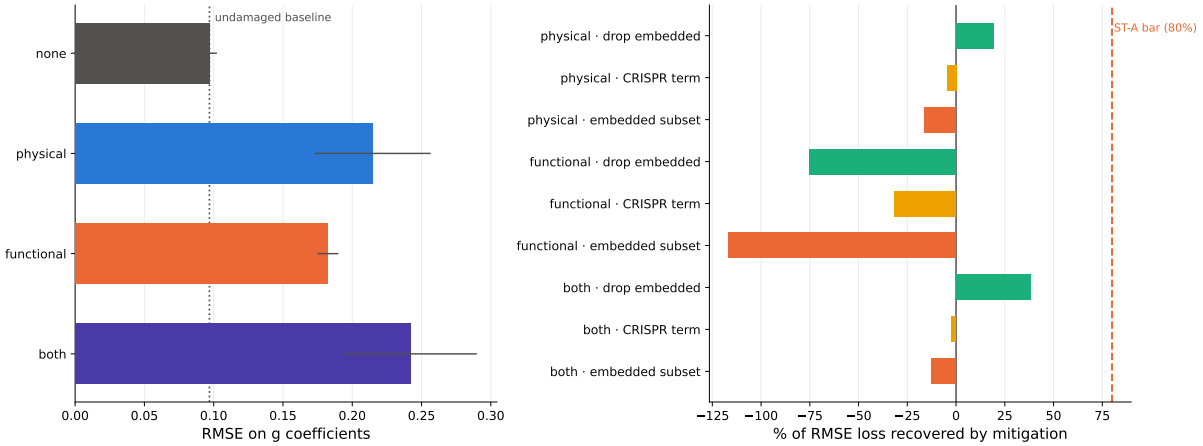

Figure 7: CRISIS coupling. Left: coefficient RMSE by arm against the undamaged baseline (dotted). Right: the fraction of the RMSE loss each mitigation removes, against the pre-specified 80% bar. Every mitigation falls far short, and two make matters worse.

arm	$\hat{\alpha}$	|bias|	RMSE	Spearman
none (baseline)	−1.094	0.01%	0.097	0.737
physical	−1.099	0.45%	**0.215**	0.727
functional	−1.283	**17.3%**	0.182	0.734
both	−1.061	3.0%	**0.242**	0.723
B3 naive (none)	−0.646	41.0%	0.414	0.668

The two couplings damage different things, which is why no single absorber fixes both. Physical coupling leaves the mean defense effect essentially intact (0.45% bias) while more than doubling coefficient RMSE (0.097 → 0.215) — the signature of individual coefficients trading off against one another rather than of aggregate bias. If embedded systems occur only in CRISPR$^+$ hosts, their coefficients cannot be separated from CRISPR status by any reparameterisation, so this is an identifiability limit rather than a modelling oversight. Functional coupling instead biases the mean by 17.3%, because the de-repression term acts as an omitted host-level covariate.

Neither mitigation works, and the nominated one is harmful. Measuring recovery as the fraction of the RMSE increase that a mitigation removes: dropping the CRISPR-embedded systems from g recovers +19.3% (physical) and +38.0% (both) but −75.1% (functional); adding a CRISPR-status main effect and CRISPR×defense-load interaction — the theoretically motivated absorber — recovers approximately nothing everywhere (−4.6%, −31.4%, −2.3%) and drives $\hat{\alpha}$ bias on the functional arm from 17.3% up to 35.9%, worse than leaving the threat untreated. Against a pre-specified bar of 80% recovery, the best result is 38%.

We report the failure rather than searching for a mitigation that happens to score well. The diagnosis is specific: the nominated absorber interacts CRISPR status with *total* defense load, whereas the truth involves the *embedded-subset* load, and a mis-specified absorber adds variance without removing bias — which is what the numbers show. The untested mitigation that should work is an interaction with the embedded-subset load, identifiable in real data from genomic co-location with CRISPR loci. We specify it and do not claim it.

CRISIS coupling is the most serious unresolved threat in this work (Figure 7). It was found by reading the literature rather than by any internal check, and it remains unmitigated.

Two implementation faults, found before reporting. The "both" arm was initially numerically identical to "physical". De-repression had been keyed on CRISPR *absence*, but under physical coupling embedded systems exist only in CRISPR$^+$ hosts, so the functional term was identically zero; Shu et al. [12] describe de-repression when CRISPR is *present but compromised*, which is now modelled explicitly. Separately, the threat manipulation drew from the main random stream, so enabling it shifted every later draw including the planted coefficients (true $\alpha = -1.251$ with the threat on versus -1.090 with it off), making the comparison invalid. Threat manipulations now use a dedicated generator and all arms share an identical planted truth. This was the second occurrence of that exact failure mode in the project.

6.5 The exclusion restriction, audited on real genomes

The decisive question is empirical. We assembled every complete *P. aeruginosa* assembly at NCBI carrying collection date, geography and isolation source (2,757 of 3,466; 18 GB), drew a stratified cohort of 1,200 spanning 60 countries and 423 BioProjects, annotated all of them with DefenseFinder 3.0.0, and clustered lineages by skani ANI. Exposure-covariate coverage is 83.2% (date), 94.3% (geography), 92.5% (source) and 100% (BioProject); genomes carry a mean of 13.97 chromosomal defense-system detections, or 12.34 distinct subtypes (the quantity the audit regresses), and 53.6% carry a CRISPR–Cas system.

Latent exposure is not observed, so the restriction is tested through two proxies, both computed *within* lineage clusters by demeaning. Test A asks whether chromosomal defense content still correlates with ecology after conditioning on lineage; Test B asks whether defense load correlates with mobile-element burden. All 1,200 cohort genomes were annotated; 2,000 permutations.

Test ($n = 1{,}200$)	99.9% ANI (344 clusters)	99.5% ANI (184 clusters)
A: all defense load $\sim$ ecology	$+0.0310^{**}$	$+0.0441^{***}$
A: CRISPR–Cas only $\sim$ ecology	$+0.0048$	$+0.0348^{***}$
B: defense load vs plasmid burden	$+0.082^{**}$	$+0.154^{***}$
B: defense load vs extrachrom. systems	$+0.080^{**}$	$+0.082^{**}$
B: CRISPR load vs plasmid burden	$+0.023$	-0.089^{**}
Systems leaking (BH $q < 0.10$, of 116)	**70**	**79**

Test A entries are excess partial R^2 over a permutation null; Test B entries are within-cluster Spearman. $^{**}p < 0.01$, $^{***}p < 0.001$.

The exclusion restriction is violated. Both thresholds agree: chromosomal defense content predicts ecology after conditioning on lineage, defense load correlates with mobile-element burden, and 70–79 **of** 116 **defense systems (**60–68%**) leak individually** at $q < 0.10$ against 5.8 expected by chance. The largest leaks are BREX_II, Rst_3HP, DS-38 and BstA, all at $q \leq 0.017$. That roughly two thirds of the repertoire individually predicts ecology is the strongest argument in this work for *estimating* the violation rather than filtering for it: a method requiring the restriction to hold uniformly does not have that option in real data.

The permutation null was indispensable. Raw partial R^2 for Test A at 99.9% is 0.0610, of which 0.0300 is pure overfitting — 21 covariates against a within-cluster design with limited effective degrees of freedom. Reporting the raw figure would have doubled the apparent violation. In a 120-genome pilot the inflation was far worse (0.527 against a null of 0.432). Any within-cluster partial R^2 reported without such a null should be treated as uninterpretable.

An earlier threshold-dependence was a power artifact. At an interim $n = 785$ the 99.9% analysis was not significant ($p = 0.109$) while 99.5% was, and we reported the disagreement as the finding. With the full cohort the median cluster size at 99.9% rises from 1.0 to 2.0, enough within-cluster contrast survives, and both thresholds agree. The earlier non-significance was underpowering, as diagnosed at the time.

What remains threshold-dependent, and is therefore not claimed. The CRISPR-specific results do not agree across thresholds. CRISPR load shows no association with plasmid burden at 99.9% ($\rho = +0.023$, n.s.) but a significant negative one at 99.5% ($\rho = -0.089$, $p = 0.0022$). At the interim $n = 785$ we had highlighted that negative association as cleanly separating the two published mechanisms — total defense load positively linked to element burden by coacquisition [8], CRISPR–Cas negatively linked by genuine blocking. **At the full cohort that separation survives only at the coarser clustering, so we report it as suggestive rather than established.** The aggregate violation is robust to clustering granularity; the CRISPR-specific contrast is not.

The filter is weaker than we assumed. Replicon-based filtering classifies only 1.79% of detections as extrachromosomal against the ∼20% MGE-borne rate in the literature, because complete *P. aeruginosa* assemblies rarely carry plasmids — only 2 of the first 48 cohort genomes have more than one replicon — so the real mobile reservoir is integrated prophages and ICEs, which are *chromosomal by coordinate*. Implemented at scale by replicon assignment alone, our stated mitigation would have been far leakier than intended. This is independent motivation for Threat A and for estimating the violation.

What the audit cannot do. The measured associations cannot be converted into a λ-equivalent and we do not attempt it: the simulated λ is a coefficient in the exposure layer, while the real-data proxies are associations with plasmid burden and ecology. A mapping between them would be invented rather than derived. The directly comparable quantity is $\hat{\lambda}$ itself, which becomes estimable on real data only once the element layer is built.

7 Fitting the Model to Real Genomes

Everything to this point is either simulation or a host-side audit. The model's central object — a susceptibility function estimated from observational data with exposure separated out — has never been fitted to real genomes, because doing so requires both an element layer and a spacer layer that the project did not have. This section builds them and reports what the model says.

7.1 The instrument, measured

MinCED over all 1,200 cohort genomes yields 27,620 spacers, and the calls agree closely with DefenseFinder's independent Cas annotation: CAS^+ genomes carry a median of 40 spacers and all 611 of them carry at least ten, while CAS^- genomes carry a median of zero. Two methods with nothing in common — repeat-structure detection and HMM-based system annotation — concur on which genomes have CRISPR. Spacer count also supplies a *measured* value for the array capacity that the simulator had to assume.

7.2 Elements from spacers, and a flaw that would have voided everything

We define elements as **spacer clusters**. The 27,620 spacers collapse (strand-canonicalised) to 3,980 distinct protospacers, of which 2,589 appear in at least two hosts. Then $S_{ij} = 1$ iff host i carries a spacer of cluster j, and $Y_{ij} = 1$ iff host i's genome contains a $\geq 95\%$ identity match to j. Both observables live on the same element index, which is what makes them commensurate; 3.1 million host–element pairs result, an order of magnitude more than any simulation here.

The first build was wrong in a way that is worth reporting. It produced $(Y{=}0, S{=}1) = 0$ — the informative cell, on which the entire identification argument rests, was *empty* across all 3.1M pairs — and a self-targeting cell 13× *enriched* rather than depleted. The cause was that BLAST was finding each spacer inside the host's own CRISPR array, making $Y \equiv S$ by construction. Masking all 1,767 detected arrays (1.72 Mb) fixes it. The diagnostic that caught this was written before the data existed, which is the only reason it was caught.

7.3 Three model assumptions, validated rather than assumed

Quantity	Measured	Model assumes
Informative $(Y{=}0, S{=}1)$ cell	25,099 pairs	must be non-empty
Self-targeting depletion ratio	0.539	$\rho_1 < \rho_0$
ε, false-match rate	0.000	$\varepsilon \approx 0$

The ε control is a dinucleotide-preserving shuffle of every element, searched under identical settings: 2,589 decoys produce *zero* hits against 1,200 genomes. A mononucleotide shuffle would have been an easier and less honest control. And self-targeting spacers are depleted roughly two-fold (916 observed against 1,700 expected under independence), confirming for the first time on real data the receipt model's premise that a spacer against a resident element is purged.

7.4 The fit, and a large gap between two valid uncertainties

Fitting 1,200 hosts × 2,421 elements × 108 defense systems with the violation λ free gives $\hat{\lambda} = 0.266$, $\hat{\alpha} = 0.017$, $\rho_0 = 0.083$, $\rho_1 = 0.025$ and $\rho_1/\rho_0 = 0.307$.

Uncertainty depends entirely on what one treats as an independent observation. The likelihood-ratio statistic for adding λ is $\chi^2 = 507.6$ on one degree of freedom, nominally $p \approx 10^{-112}$. A bootstrap that resamples *lineage clusters* gives a 95% interval of $[-0.288, 0.542]$, comfortably spanning zero.

Both are computed from the same fit and both are correct as computed. The LRT treats 3.1M host–element pairs as independent, and they are not: 1,200 genomes fall into 344 clusters

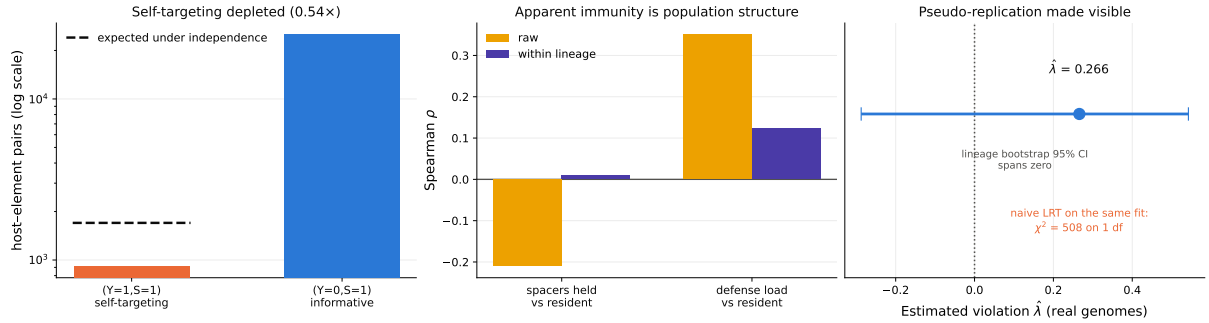

Figure 8: Real genomes. Left: the informative cell exists at scale while self-targeting sits far below its independence expectation (dashed). Centre: the apparent CRISPR-immunity association (−0.21) vanishes under lineage control (+0.009) while the coacquisition association survives. Right: the estimated violation with its lineage-cluster bootstrap, against the nominal likelihood-ratio statistic computed on the same fit.

whose members are near-identical. **The distance between $\chi^2 = 508$ and an interval containing zero is pseudo-replication made visible**, and we report the bootstrap as the honest uncertainty with the LRT labelled anti-conservative in the output itself.

Interpolating $\hat{\lambda} = 0.266$ on the simulation-calibrated attenuation curve of Section 6 gives roughly 7% — so even taking the point estimate at face value, the violation in this dataset is small enough that assuming the restriction would have cost little.

7.5 The central hypothesis, tested on real data, does not confirm

$\hat{\alpha} = 0.017$ with an interval spanning zero: chromosomal defense repertoire does not detectably predict which protospacer-elements are resident, given encounter. Four readings are compatible and we cannot separate them: the null is real for this element class; the element definition is the weak link, since a 32 bp protospacer match is a coarse proxy for element establishment; the exposure layer over-absorbs, with 2,421 free element parameters in each of f and g against 108 α coefficients; or power, once 344 effective clusters rather than 1,200 genomes are counted.

7.6 But the project's thesis is visible directly in the same data

Association	Raw	Within lineage
spacers held vs. resident elements	$\mathbf{-0.210}$ (p=2×10^{-13})	$\mathbf{+0.009}$ (p=0.75)
defense load vs. resident elements	+0.351 (p=5×10^{-36})	+0.123 (p=2×10^{-5})

The first row is this paper's argument in one line of real data. Analysed naively, holding more spacers is strongly associated with carrying fewer of the targeted elements — which reads exactly like CRISPR immunity working, and is the shape of result the field reports. Condition on lineage and it vanishes entirely. The apparent immunity signal was population structure.

The second row is the masking effect of Liu et al. [8], observed rather than simulated: defense load is *positively* associated with carrying the very elements it is supposed to exclude, both raw and within lineage.

This reframes the null. A deconfounded estimate near zero is what one expects if a genuine protective effect and coacquisition roughly cancel; we cannot demonstrate that cancellation and

do not claim it, but a null sitting between two large opposed confounded readings is more informative than an absence of signal.

7.7 Is CRISIS detectable in *P. aeruginosa*? No

Across 7,314 defense-system/Cas pairs, *zero* lie within five genes of a Cas system, six within ten (0.08%), and the median distance is 1,166 genes. Nothing is embedded.

Two nulls were built and discarded before this conclusion. A uniform-position null called 23 of 67 system types significant — but it was detecting *defense islands*, the well-known clustering of defense systems, which a uniform null destroys. A label permutation over fixed positions did not fix it either, because distance was measured to the Cas *interval* and Cas systems span six to eight genes against one to three for most defense systems, so whichever system received the "Cas" label presented a shorter target. Only midpoint-to-midpoint distance under label permutation is fair, and even then the test returns significance at a scale of hundreds of genes, which cannot represent embedding within a locus. We therefore report the raw proximity distribution rather than the p-value.

This is a bounded negative: DefenseFinder may annotate a module embedded in a Cas operon as *part of* the Cas system, in which case embedding is invisible to any between-system distance.

7.8 The mitigation we specified in advance is refuted

Section 6 failed its CRISIS threshold and named the fix: interact CRISPR status with the *embedded-subset* load rather than total defense load, on the grounds that a mis-specified absorber adds variance without removing bias. We implemented it and ran it at 20 replicates.

Recovery of RMSE loss (bar: 80%)	physical	functional	both
embedded-subset absorber (the nominated fix)	−16.3%	**−116.7%**	−12.4%
total-load CRISPR term	−4.6%	−31.4%	−2.3%
drop embedded systems from g	+19.3%	−75.1%	+38.0%

It is the *worst* of the three. On the functional arm it more than doubles the loss it was meant to remove and drives $\hat{\alpha}$ bias from 17.3% to 41.8% — worse than leaving the threat untreated.

The reason is structural rather than a tuning failure, and it is more informative than the prediction it refutes. The de-repression term is keyed on CRISPR being *present but compromised* — an inactivating mutation or an anti-CRISPR protein. That state is **latent**. No absorber built on observable CRISPR presence or absence can represent it, whatever its load argument, so adding one only injects variance. Mitigating CRISIS would require observing the compromised state directly, which is a larger data requirement than we supposed.

8 Discussion

8.1 What the identifiability study establishes

Phase 0 asked a go/no-go question: do the exposure and susceptibility layers separate at realistic parameters? They do, decisively, and by a wide margin. Recovery of a planted susceptibility function reaches within 14% of an oracle handed the true exposure, while a faithful analogue of

the current standard method is worse by a factor of 4.5 — a gap that survives optimal rescaling and is therefore differential bias rather than attenuation. On the strength of this we would proceed to the annotation phase.

But the study also changed what we think the contribution is, and the change is worth stating plainly because it is not the one we set out to demonstrate.

8.2 The instrument is not where the identification lives

We built LEDGER on the premise that CRISPR spacers are what make a susceptibility function estimable from observational data: a spacer is physical evidence of encounter, recorded independently of outcome, and the $(Y{=}0, S{=}1)$ cell directly observes rejection. That premise is sound as far as it goes. It is also, on this evidence, not what does the work.

Across a 47-fold range in the number of informative cells — from 214 to 10,007 — recovery moves by 6%. Removing the receipt channel entirely costs under 3% in RMSE. What identifies θ is the structural exclusion of defense features from the exposure layer: because f cannot express defense-driven variation in Y, that variation must be attributed to g. This is a functional-form restriction, and it is far stronger than any instrument.

The useful consequence is that the method does not require CRISPR at all in the regime where the exclusion restriction holds. It applies to species with little or no CRISPR-Cas, and to element classes that leave no spacer record. That substantially widens the applicable domain relative to the design we started from.

The corresponding cost is that the method now rests on an assumption rather than on an observation, and assumptions can be wrong in ways instruments cannot. This is precisely why the linkage results matter: they measure what happens when the assumption fails.

8.3 Linkage is where the spacer record pays

Liu et al. [8] establish that the exclusion restriction fails in practice, and quantify how badly: over 95% of defense-system acquisitions coincide with mobile-element gains, and acquisition is $\sim$ 50-fold more likely on such branches. Under simulated linkage of increasing strength, the standard method degrades steadily ($0.650 \to 0.364$) while LEDGER stays nearly flat, and the receipt channel's contribution triples ($5.0\% \to 15.2\%$ RMSE reduction, all $p \leq 3.8 \times 10^{-6}$).

So the spacer record is a hedge against the failure of the assumption that otherwise carries the estimate. That is a more modest role than we anticipated, and a more defensible one: it is insurance whose premium is paid exactly when the risk materialises. We would not now describe CRISPR spacers as the foundation of the method. We would describe them as what keeps it standing when its foundation gives way.

8.4 Reproducing the confound

The linkage sweep also does something we did not originally plan and now regard as among the strongest evidence in the study. As linkage strengthens, the marginal association between defense load and element presence moves monotonically from -0.239 to $+0.746$, crossing zero partway. The positive defense/element association reported in the co-occurrence literature is thus reproduced on demand, with a dose-response, by a model that is at the same time recovering a correctly-signed *inhibitory* susceptibility function.

This is a direct demonstration of the paper's central claim about estimands: the same underlying biology — defenses that genuinely block establishment — can present as a null, a

negative, or a strongly positive marginal association, depending entirely on how exposure is distributed. A field that regresses presence on defense content is measuring a quantity whose sign is not determined by the biology it is trying to study.

8.5 Limitations

The recovery results are simulation; the real-data results are not. Every *recovery* number — Spearman and RMSE against a planted truth — concerns a generative process we wrote, and calibration to reported spacer-acquisition rates is not validation. The exclusion-restriction audit (Section 6.5) and the real fit (Section 7) are measured on real complete *P. aeruginosa* assemblies. Where the two kinds of number appear together we are explicit about which is which, and no recovery claim is ever made on real data, since no planted truth exists there.

The real element layer is a proxy. Elements in Section 7 are 32 bp protospacers, so "resident" means "a $\geq 95\%$ identity match outside any CRISPR array". That is coarse, the bilinear block W cannot be fitted without element-side annotation, and it is our leading explanation for the null. A prophage-level element layer is the obvious next test and is not run here.

The estimator is correctly specified in most experiments. In all sweeps but one, f and g have exactly the functional forms that generated the data. We added a mis-specification stress test precisely because reporting only correctly-specified recovery would overstate the method; under it recovery degrades gracefully and tracks the oracle. But real mis-specification will not take the low-rank form we injected.

The exclusion restriction was the load-bearing assumption — until it stopped having to be. Section 5.2 shows identification rests on it almost entirely, and our stated mitigation (restricting to chromosomal systems outside prophages, plasmids and ICEs) is imperfect: boundary annotation is error-prone, and Kuang et al. [7] show attachment-site-proximal systems evade such a filter entirely. Section 6 supersedes the mitigation rather than repairing it: the violation is estimated instead of filtered away, which removes the bias without requiring the filter to be correct. The residual limitation is no longer the restriction itself but the couplings that estimation cannot absorb — above all CRISIS.

Sample size. Below roughly 250 hosts the method is worse than the baseline it replaces. It is a large-collection method.

Between-lineage dependence. The within-lineage permutation shows a substantial part of the signal comes from between-lineage comparison, where confounding with anything else that varies between lineages cannot be excluded by design. Within-lineage-only analysis would be more conservative and less powerful.

Spacers are a biased record. Acquisition is non-uniform across elements and enriched in cells entering lysogeny; self-targeting spacers are purged. We model the outcome-dependence explicitly and recover ρ_0 and ρ_1 accurately, but the detection model is a two-parameter summary of a process that is considerably richer.

8.6 What we would do next

The ordering implied by these results differs from the original plan. Because identification turns out to rest on the exclusion restriction rather than on spacer coverage, the highest-value next step is not to maximise spacer yield but to *validate the exclusion restriction on real data*: quantify how often chromosomal defense annotations are contaminated by unrecognised mobile elements, and measure the residual association between chromosomal defense content and exposure proxies. If that association is small, the method transfers with the spacer channel as a secondary correction. If it is large, the spacer channel moves back to centre stage and species choice should be driven by CRISPR prevalence after all.

The external validation targets also change. Since counting defense systems explains only about a quarter of the variance in measured phage resistance [1], the informative comparison is whether an estimated susceptibility function improves on a scalar count against the published panels of Burke et al. [1] and Costa et al. [4]. And because Lopatina et al. [9] experimentally validated eight anti-defense proteins from experimental infection matrices, their confirmed pairings are external ground truth for an observationally estimated interaction map — a stronger test than the internal recovery check originally scoped.

8.7 What the extension changes

Section 6 resolves the discomfort Section 5.2 created, and it does so in a direction we did not anticipate. Having found that identification rested on an assumption rather than on the CRISPR instrument, the natural next move was to bound how wrong the assumption could be. That turned out to be the wrong move: the violation is not merely boundable but directly estimable, and estimating it removes the bias entirely over a fourfold range. The recommendation that follows is concrete and cheap — always give the exposure layer a free defense-load coefficient, and report the fitted value as a measured quantity rather than defending a filter.

That reframes the contribution a second time. The claim is now neither "CRISPR spacers identify susceptibility" (Section 5.2 refuted that) nor "the exclusion restriction identifies susceptibility" (which would be resting on an assumption), but rather: *exposure and susceptibility separate because they enter the likelihood differently, and the degree to which defense content contaminates exposure is itself a parameter one can measure.* The fitted $\hat{\lambda}$ doubles as the diagnostic that a sensitivity analysis was supposed to supply.

8.8 Limitations added by the extension

CRISIS coupling is unresolved. Defense modules embedded within type I CRISPR–Cas loci break the independence of ρ and g, and neither mitigation we tested repairs it. Physical co-inheritance appears to be a genuine identifiability limit rather than a modelling oversight. In a type I-F dominated species this bears directly on using CRISPR as an exposure instrument, and it is the one threat we would want resolved before trusting a real-data interaction map.

The real-data audit is host-side only. It tests whether defense content predicts ecology and mobile-element burden. It cannot measure $\hat{\lambda}$ itself, because that requires the element layer. The audit therefore constrains the assumption without closing the question, and we have been careful not to convert its statistics into a λ-equivalent, which would be invented rather than derived.

Annotation is a lower bound on contamination. Replicon-based filtering identified only 1.79% of detections as extrachromosomal, because in complete *P. aeruginosa* assemblies the mobile reservoir is integrated prophages and ICEs rather than plasmids. Faithful implementation of our own stated filter requires prophage and ICE calling that we did not run.

Lineage granularity is consequential. Single-linkage ANI clustering at 99.0% collapses the species into one component, at which point "within lineage" is vacuous. Results are reported at 99.9% with 99.5% as sensitivity, and the degenerate threshold is reported rather than dropped.

8.9 What fitting real data changed

Section 7 is the first point at which the model met genomes rather than a generative process we wrote, and it changed three things.

The instrument is real, and so are its assumptions. The receipt model was built on premises that were reasonable but untested: that spacers against resident elements are purged, that false matches are negligible, and that the informative cell is populated at all. All three are now measured, and all three hold. That is the strongest form of support the model has received, because it is the only support that did not come from a simulator we designed.

The central hypothesis did not confirm, and the reason is instructive. On this element class $\hat{\alpha}$ is indistinguishable from zero. We list four readings in Section 7 and cannot separate them; the element definition is our leading suspicion, and a prophage-level layer is the obvious next test. What makes the null informative rather than empty is that the same data contain two *large* confounded signals pointing in opposite directions — an apparent immunity effect that is entirely population structure, and a coacquisition effect that survives lineage control. An estimate near zero between them is what a deconfounded analysis should produce if a genuine protective effect and coacquisition roughly cancel, though we cannot demonstrate that cancellation and do not claim it.

Inference, not just estimation, is where non-independence bites. The likelihood-ratio test and the lineage bootstrap disagree by a margin that is difficult to overstate: $\chi^2 = 508$ against an interval containing zero, from the same fit. Comparative genomics is careful about phylogenetic correction of outcomes and much less careful about what counts as an independent observation when a matrix has millions of cells and a few hundred clones. We would treat any uncorrected p-value on a host–element matrix of this kind as uninformative.

Some assumptions cannot be repaired by modelling. CRISIS resists mitigation not because we chose the wrong absorber but because the state that drives it — CRISPR present but compromised — is unobserved. Three absorbers were tried and the best-motivated one was the worst. When the confounder is latent, the answer is more data of a different kind, not a better parameterisation.

References

[1] Kevin A. Burke, Caitlin D. Urick, Nino Mzhavia, Mikeljon P. Nikolich, and Andrey A. Filippov. Correlation of Pseudomonas aeruginosa phage resistance with the numbers and types of antiphage systems. *International Journal of Molecular Sciences*, 25(3):1424, 2024. doi: 10.3390/ijms25031424.

[2] Charlotte E. Chong, Aaron Weimann, Aleksei Agapov, Joanne L. Fothergill, Michael A. Brockhurst, Julian Parkhill, R. Andres Floto, Mark D. Szczelkun, Edze R. Westra, Multi-Defence Consortium, and Kate S. Baker. Defence systems drive accessory genome interactions in Pseudomonas aeruginosa. *ISME Communications*, 6(1):ycag130, 2026. doi: 10.1093/ismeco/ycag130.

[3] Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Biometrika*, 2025. doi: 10.1093/biomet/asaf004.

[4] Ana Rita Costa, Daan F. van den Berg, Jelger Q. Esser, Aswin Muralidharan, Halewijn van den Bossche, Boris Estrada Bonilla, Baltus A. van der Steen, Anna C. Haagsma, Ad C. Fluit, Franklin L. Nobrega, Pieter-Jan Haas, and Stan J. J. Brouns. Accumulation of defense systems in phage-resistant strains of Pseudomonas aeruginosa. *Science Advances*, 10(8):eadj0341, 2024. doi: 10.1126/sciadv.adj0341.

[5] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1): 153–161, 1979. doi: 10.2307/1912352.

[6] Kathryn M. Kauffman, Julia M. Brown, Radhey S. Sharma, David VanInsberghe, Joseph Elsherbini, Martin Polz, and Libusha Kelly. Viruses of the Nahant collection, characterization of 251 marine Vibrionaceae viruses. *Scientific Data*, 5:180114, 2018. doi: 10.1038/sdata.2018.114.

[7] Xu Kuang, Jamie Gorzynski, Marie Touchon, Andrey Shkoporov, Eduardo P. C. Rocha, J. Ross Fitzgerald, John Chen, Jakob T. Rostøl, and José R. Penadés. Bacteriophages mobilize bacterial defense systems via lateral transduction. *Science Advances*, 12(4):eadx5749, 2026. doi: 10.1126/sciadv.adx5749.

[8] Yang Liu, João Botelho, and Jaime Iranzo. Timescale and genetic linkage explain the variable impact of defense systems on horizontal gene transfer. *Genome Research*, 35(2): 268–278, 2025. doi: 10.1101/gr.279300.124.

[9] Anna Lopatina, Mariusz Ferenc, Nina Bartlau, Michael Wolfram, Kerrin Steensen, Sebastiano Muscia, Fatima Hussain, Madeleine Schnurer, Leila Afjehi-Sadat, Shaul Pollak, and Martin F. Polz. Interpretable machine learning reveals a diverse arsenal of anti-defenses in bacterial viruses. *bioRxiv*, 2024. doi: 10.1101/2024.06.14.598830. Preprint.

[10] Pedro H. Oliveira, Marie Touchon, and Eduardo P. C. Rocha. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research*, 42(16):10618–10631, 2014. doi: 10.1093/nar/gku734.

[11] Liam P. Shaw, Eduardo P. C. Rocha, and R. Craig MacLean. Restriction-modification systems have shaped the evolution and distribution of plasmids across bacteria. *Nucleic Acids Research*, 51(13):6806–6818, 2023. doi: 10.1093/nar/gkad452.

[12] Xian Shu, Rui Wang, Xufei Zhou, Feiyue Cheng, Jiayue Ma, Zhihua Li, Xin Li, Tuozhan Wu, Aici Wu, Qiong Xue, Chao Liu, Huiwei Zhao, Xifeng Cao, Lin Wang, Shouyue Zhang, Yan Zhang, and Ming Li. CRISPR–Cas regulates expression of embedded anti-phage defence systems. *Nature*, 2026. doi: 10.1038/s41586-026-10833-9. Advance online publication.

[13] Florian Tesson, Alexandre Hervé, Ernest Mordret, Marie Touchon, Camille d'Humières, Jean Cury, and Aude Bernheim. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications*, 13:2561, 2022. doi: 10.1038/s41467-022-30269-9.

A Simulator specification

A.1 Phylogeny

Host phylogeny is a balanced ultrametric tree with L levels and branching factor b, giving $N = b^L$ tips. A trait is generated by summing independent Gaussian increments contributed at each level along the path from root to tip, so that

$$\mathrm{Cov}(x_i, x_j) \;=\; \sum_{l \le d(i,j)} \sigma_l^2, \tag{14}$$

where $d(i, j)$ is the depth of the most recent common ancestor of tips i and j. Increment standard deviations decrease with depth as $\sigma_l = (l + 1)^{-1/2}$, so most divergence is deep. Traits are standardized to unit marginal variance.

This construction has a consequence that proved central to interpreting the permutation control (Section 5). With $L = 4$, the fraction of trait variance lying *within* a lineage cluster taken at level $L - 3$ is $\left(\sigma_3^2 + \sigma_4^2\right) / \sum_l \sigma_l^2 \approx 0.28$ for the continuous latent, and approximately 0.5 after thresholding to binary presence/absence. Roughly half of a defense system's variance therefore sits *between* lineage clusters and is untouched by any within-cluster shuffle.

A.2 Defense and anti-defense content

Host defense presence is obtained by thresholding a phylogenetically autocorrelated latent at a per-system quantile chosen to hit a target prevalence drawn uniformly from $[0.12, 0.55]$. Thresholds must be computed column-wise: supplying a vector of quantiles to a single `numpy.quantile` call returns a cross-product rather than per-system thresholds, an error that produces a silently mis-shaped defense matrix. Element anti-defense content is drawn independently per family with prevalence in $[0.10, 0.45]$.

A.3 Calibration

Intercepts for both f and g are solved by bisection so that the realised marginal rates match their targets regardless of the swept parameter. Without this, sweeping (for example) CRISPR prevalence would also shift the base establishment rate, confounding the sweep with a prevalence change. Realised rates are recorded in every output record and can be audited: at the default operating point a typical replicate gives $P(Z{=}1) = 0.350$ against a target of 0.35 and $P(Y{=}1 \mid Z{=}1) = 0.218$ against a target of 0.22.

A.4 Invariants

Two invariants are checked on every simulated dataset. First, $P(Z{=}1 \mid Y{=}1) = 1$ exactly, since an element cannot establish in a host it never met. Second, $P(Z{=}1 \mid S{=}1) \approx 1$, with the small shortfall attributable entirely to the false-match rate ε; the measured value at the default configuration is 0.9977 with $\varepsilon = 10^{-4}$.

B Metric ceilings under sparsity

The planted interaction block $\mathbf{W}$ is sparse by design, so most of its entries are tied at zero in the ground truth. Rank correlation against a tie-heavy vector cannot reach unity even for a perfect estimator. We compute the ceiling directly by scoring the truth against itself with ties broken by negligible random jitter: for the default configuration (86 zeros of 96 entries) the ceiling is 0.5302 with a standard deviation of 0.0000 across 200 draws.

This is not a technicality. In an early run, the oracle, LEDGER and the naive baseline all scored 0.529–0.530 on this metric — that is, all three were exactly at the ceiling and the metric had no discriminative power whatsoever. Reporting those numbers as "moderate recovery" would have badly misdescribed a perfect score and would have hidden the real differences between the methods, which are large. We therefore report support recovery (AUROC, AUPRC, precision at k) and Spearman restricted to true nonzeros for the interaction block, and report the computed ceiling alongside any tie-prone Spearman.

C Reproduction

The full study is reproduced by:

```
R=ledger.sweeps
for S in e1 e1b e1c e1d e5 \
         misspec ablations; do
  python -m $R --sweep $S \
    --reps 20 \
    --out experiments/$S.jsonl
done
python -m ledger.figures
python -m ledger.report
```

Each fit writes one JSON record containing its complete configuration, all metrics, its convergence status and its realised marginal rates. All tables and figures are generated from those records by `ledger/report.py` and `ledger/figures.py`; no number in this paper was transcribed by hand. Bibliography entries are verified against the Crossref registry by `python -m ledger.verify_refs paper/references.bib`.