nature portfolio

Peer Review File

Higher-order epistasis drives evolutionary unpredictability toward new antibiotic resistance

Corresponding Author: Professor Erdal Toprak

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this manuscript, the authors investigate a complete set of selected mutations in the beta-lactamase TEM-1. Using a pooled barcode-sequencing-based assay, they determine the fitness of all genotypes under multiple concentrations of ampicillin (the original substrate) and aztreonam (a poor substrate). Their goal is to identify and predict evolutionary pathways toward extended-spectrum beta-lactamases (ESBLs) using regression of epistatic interactions at multiple orders and machine learning. Most mutations reduce fitness in ampicillin compared to the wild type, yielding a relatively smooth landscape. Fitness in aztreonam increases most when multiple mutations with idiosyncratic epistatic interactions are combined, creating a rugged landscape that requires higher-order epistasis and larger training sets.

This work addresses a highly relevant and timely problem. The combinatorial dataset described here is an impressive contribution to the large-scale evaluation of fitness landscapes in single proteins. It is especially valuable since the authors made the dataset publicly available. Understanding evolutionary trajectories and trade-offs in beta-lactam resistance is clinically relevant for combating antibiotic resistance. The experimental and theoretical work is substantial, and the conclusions are mostly sound. However, the manuscript needs further improvement. Descriptions should be clarified, and the discussion of experimental and analytical choices should be extended. It should also be compared to existing, similar datasets.

Major comments:

• Choice of mutations
The 13 loci and 18 variants chosen are only described as being clinically identified. The authors should describe the selection process in more detail and explain whether the variants are linked to aztreonam resistance or extended-spectrum beta-lactamases (ESBLs). What is known about the clinical isolates in which these mutations were observed? Were the relevant antibiotics used in the hospitals where the isolates were collected? Additionally, does laboratory evolution under aztreonam select for some of the same mutations?

• Fitness measurements
The authors use a non-standard measure of fitness with the AUC, which is justified by the non-monotonic growth curves in beta-lactams. However, clearer validation and description of measurement sensitivity and precision are necessary.
Does the AUC measurement in competition align with the growth phenotype data of single strains? Drug degradation and other interactions in the pooled assay may make fitness measurements less quantitative. It would be important to validate the results with direct fitness measurements in monoculture, at least for a few mutants. Selecting a few dozen colonies and comparing their fitness in the pooled assay with their individual dose-response curves would further validate this dataset.
The authors claim a typical standard deviation of 10% between replicates (line 163), yet they never demonstrate this, for example, by plotting correlations between replicates.
Figure S3 shows that some genotypes have very noisy fitness values, jumping from the minimum to the mean trend from one concentration to another. What is the reason for this? Does this occur infrequently enough that it does not affect results such as landscape graphs and epistasis measurements?
Could dose-response curves be extracted from these data for each genotype, perhaps by normalizing to the wild type or the mean, to generate a composite resistance measure? This could be a more robust way to infer a quantitative fitness proxy from the data (e.g., the IC_{50}). One problem with the current approach is that, due to the steep dose-response curve of beta-

lactams, most mutants likely exhibit an all-or-nothing growth response to the drug, which is somewhat obscured by using the AUC and provides limited quantitative information necessary for epistasis analysis. The regression analysis may work better if IC_{50} is used instead of AUC-fitness. The authors should have all the necessary data to test this, and analyzing it would strengthen the manuscript.

• Choice of concentrations analyzed
Understandably, not all measured concentrations were analyzed and presented. Nevertheless, it would be useful to evaluate whether epistatic interactions are consistent across concentrations and if the general patterns from Figures 4 and 5 hold as further validation of the approach.
Another place where this issue arises is in ll. 222, which describes Figure 2C and notes the absence of a clear global peak observed at low and high concentrations. However, contrary to the text, this highly fit global peak is only absent at a single concentration: 36 µg/ml. Since the result hinges on a single concentration, it is unclear whether it is true or potentially due to measurement noise, especially given the threshold-based node definition.

• Predicting genotype fitness
The authors describe two methods for predicting the fitness levels of genotypes: regression with increasing epistatic order and machine learning. However, they never discuss or compare the quality of the predictions, nor do they address the advantages or disadvantages of these methods for future use. Further discussion of the generality of this dataset and approach is necessary. The machine learning section of the main text is somewhat obscure. Aztreonam predictability seems much lower than ampicillin, suggesting difficulty in precisely predicting adaptation to novel phenotypes, which would be very interesting. How do results from aztreonam selection compare to those from other beta-lactams? How does the choice of mutations affect the results, especially when considering pipeline beta-lactams without clinically known mutations? The claim that this general approach can be used to make predictions for beta-lactams in the pipeline (line 428) seems slightly exaggerated. It seems that we first need to identify which mutations are relevant, which may differ for new beta-lactams, before applying the approach.
The last section (lines 364) requires substantial further explanation. It is difficult for readers unfamiliar with DCA, the resulting Hamiltonians, and LGL to follow. It is important to explain these concepts more clearly in the main text, not just in the Methods section.

• Discussion
Currently, this is mainly a summary of the results. The authors should discuss their work in the context of the relevant literature and clarify the broader implications. In particular, how does the landscape structure compare to that in other similar studies, such as references 7 and 8?

Minor comments:

Bold text is used inconsistently throughout the text. For instance l. 171 Figures 1E-F is bold while in l. 187 Figure 2 is not.

l. 253 and Figure 3B for AZT disagree whether 164H is among the enriched mutations.

ll. 538 describing panels G and especially H seems to be wrong as both contain AZT and AMP.

ll. 546 and 593 and Figure 5 and extended data Figures 1 and 2: The captions for the heat maps in Figures 5B and 5C, as well as the extended figures, seem incorrect. The rows show the 18 genotypes, not the columns. The left heatmap in the extended figures is not described. The bars on the right side of the heatmaps are also not explained. Which heatmaps and beeswarms are ordered by row and column? The order of the beeswarm rows doesn't match that of the heatmaps, which makes comparison difficult. It would be easier if they were in the same order.
Is the heatmap color scale correct? The beeswarm SHAP values for R244S in AMP show negative values for the presence of the mutation, but the red color seems to suggest that its presence has a positive impact on fitness, as described in lines 347 and 350.

l. 568: What is the meaning of the letter colors in Figure 6A?

l. 636: The caption for Figure S1 describes wild-type letters as grey. This should be black.

ll.753: Fitness measurements also in 3000 µg/ml AZT and 50000, 12500, 3125 µg/ml AMP are mentioned. However, these fitness measurements are never shown. Why is that?

Methods: The analysis of sequencing data to determine AUC is not described further. How are barcodes matched to their respective genotypes when there are errors in the sequencing reads? Are all barcodes from one genotype pooled together?

Figure 1: Caption for chemical structures is missing.

Figure 1E+F: Concentration units differ between y-label and plot annotation.

Figure 2B,C and S4: It would be helpful to highlight the wild-type containing node to judge accessible peaks from this node.

Figure 4C is missing a color bar because the color bar in Figure 4D seems to be different.

Figure S6: Why is it only shown for AMP? Add the correlation coefficients and p-values to the plot to justify the claim of significance.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
In this study, Gaszek and colleagues create a library of mutants of E. coli with the TET-1 gene to measure how its sequence changes resistance to beta-lactam antibiotics. They comprehensively mutate 13 selected residues (selected based on prior studies finding association with resistance in clinical isolates), and measure the fitness of these libraries at different concentrations of two beta-lactam drugs. Sequence barcoding is used as a readout at five time points. They analyse the fitness of these mutants to determine the predictability of mutation in the gene, and found substantial contributions from epistasis -- both of which differed by the two drugs tested.

This seems to be a well executed study, both in the experimental design (where I have less expertise) and the analysis. The mutation saturation approach, although not covering every residue in the protein, is nevertheless a comprehensive one. The study answers some important evolutionary questions on evolvability and predictability (order of resistance mutations). There are also some interesting potential applications for AMR prediction and forecasting. Generally, the study does a good job of clearly acknowledging its limitations.

I do not have any significant concerns with the study, but do have comments on the presentation of the work, and some analyses which would be needed to achieve impact outside of these experimental conditions.

Major:
1) The clinical relevance of this study could be brought out better.
- How prevalent are the chosen mutations in the population / clinical isolates? Do they show homoplasy?
- The AUC-fitness metric is well motivated. If it is possible to relate its value to current clinical resistant/sensitive thresholds for the tested drugs somewhere, this would help readers put these numbers in context of actual treatment.
- The definition of the WT sequence seems completely arbitrary. Is it the most common allele in the population? Or just in the reference strain? Is there a specific biological meaning/interpretation of the WT allele?
2) A bit more is needed to achieve the aim of forecasting.
2a) How do trajectories/predictability change from one MIC step to the next i.e. can you compare your graphs between concentration steps? What is the path for a sequence to each higher resistance step, not optimising within a given concentration? This would be a very interesting problem to be able to address.
2b) Some of the machine learning aspects could be simplified to support the claims. I would suggest also adding a lasso regression (e.g. with the glmnet package). Then compare (in e.g. a table) the predictive accuracy of linear models, with L1 regularisation, and decision trees. If you could then also show the predictive accuracy accuracy at each training sample size?
3) This is less in my area of expertise, but I did note that in the experimental design, but presence of local cheaters not being a problem is simply asserted, evidence/data is not given to support this important claim. Are there any data, or other studies which support this claim?
4) There are a few issues with the presentation directly:
- The conclusion only briefly restates parts of the results, and in my view needs rewriting.
- Given my view of the WT being arbitrary, the long section on the optimality of the WT is not warranted.
- The description of the epistatic interactions a bit muddled. Probably starting with the overall results from the linear model(s), rather than talking about specific sequences/mutations would help improve this.

Minor comments:
- Mention the species being studied in the abstract
- l72-74. A citation is needed to support this statement.
- l79. A 'novel' beta-lactam. The word novel is ambiguous here, please rephrase.
- l90. 'Using these models, it is now possible to predict fitness of any given TEM-1 variant under β-lactam selection.'. This statement is too strong, as fitness cannot be predicted that accurately, not all TEM-1 variants were constructed (just from selected sites), and this is in a given genetic background and set of experimental conditions. Prediction of these variants is a useful outcome to note, but it would also help readers to note these caveats.
- l108. It's unclear whether L21F was also included, or just L21P -- please note if both were included.
- Relatedly, on l335, "10% of the data yielded a promising RMSD of ~0.3 which is sufficient for classifying mutants as resistant for clinical purposes". Why is this RMSD sufficient? What sensitivity/specificity does this give?
- l249. Why were the top 1% of fitness analysed further? Is there any intuition for this cutoff, or is it simply what was feasible?

- l369. It would be helpful to explain the family couplings and local fields in more biological terms, especially in terms of the experimental setup and the phenotype which was measured.
- Is there a structure available for the mutated protein? Plotting the mutations on this structure, and if possible interpreting strongly epistatic mutations in 3D space would be a good addition. Adding the PDB or AlphaFoldDB ID of this structure to figure 1 would be helpful.
- l382. "A mild but significant correlation". Please state the strength and significance quantitatively.
- The latent space in figure 6 and related analyses is not well explained or interpreted.
- Fig1c. Why not plot these coordinates accurately, rather than evenly spaced? Plotting the mutated residues along with any annotated domains would be more useful.
- Fig3a. Why are there no high fitness AZT / low fitness AMP mutants?
- Fig3b. Add a key for red/black/gray in the plot.
- Fig4. 0.5 R^2 threshold. What is this a threshold on?
- Fig4. The multipanel layout here is not helpful, as G and H appear out of order.
- Fig5. "Columns represent the 18 studied mutatations". There appear to be 553 columns?
- Fig6D. From where are the 'native' beta-lactamases found? In which database/species? (also the 'DCA Hamiltonian' label on this figure is in an odd location). I can see some of this in the methods (l900-902), but it's not clear from the figure or results.

Reviewer #4

(Remarks to the Author)
This manuscript presents a careful, data-rich study of evolutionary trajectories in TEM-1 β-lactamase. By constructing a combinatorially complete library spanning 55,296 genotypes and quantifying fitness under both ampicillin and aztreonam across multiple concentrations, the authors assemble one of the largest empirical antibiotic-resistance fitness landscapes to date. The graph-theoretic coarse-graining used to visualize ruggedness and evolutionary “fate diversity” is particularly innovative, and the core conclusion that selection for a novel function can uncover extensive higher-order epistasis that is largely muted under a native-substrate regime is of broad interest.

However, several points require clarification and/or additional analysis to fully support the manuscript’s claims about “unpredictability” and to improve mechanistic interpretability of the reported epistasis.

i) The manuscript uses AUC-Fitness under ampicillin as a quantitative proxy for loss-of-function, defining AUC-Fitness as the log10-transformed area under the normalized pooled growth curve over 12 hours and motivating this choice by the non-monotonic growth dynamics under β-lactam selection, while also noting pooled-culture effects such as “cheaters.” This is a reasonable operational in vitro metric, but the manuscript also explicitly references clinical context, including framing 36 μg/mL aztreonam as slightly above a clinical cutoff and discussing resistance prediction in clinically oriented terms. Given this, the relationship between AUC-Fitness and clinically used resistance metrics (MIC/IC50 or S/I/R categorization), and more broadly to in vivo/clinical fitness, needs to be clarified. The authors should either provide validation linking AUC-Fitness to MIC/IC50 for a representative panel of variants measured in monoculture, or they should clearly limit claims to in vitro pooled-competition fitness landscapes and add a dedicated Discussion paragraph explaining why, and to what extent, AUC-Fitness is expected (or not expected) to translate to monoculture MIC and clinical outcomes.

ii) The authors argue that the aztreonam landscape is less predictable due to higher-order epistasis and support this in part with LightGBM performance evaluated on held-out data, including the statement that 10% training yields RMSD ~0.3 and would be sufficient to classify mutants as “resistant” for clinical purposes. Random train/test splits in combinatorial genotype spaces often primarily measure interpolation because many test genotypes lie at small Hamming distance from training genotypes, which can inflate apparent predictability. To support a claim of intrinsic unpredictability (or to demonstrate genuine generalization by ML), the authors should perform at least one mutation-stratified or block-holdout validation. For example, excluding all variants containing a focal mutation (e.g., E104K) from training and then predicting them tests whether the model learns transferable epistatic rules rather than memorizing local neighborhoods. A distance-based split enforcing a minimum Hamming distance between training and test sets would serve a similar purpose. Reporting results under such schemes for both ampicillin and aztreonam would substantially strengthen the predictability narrative.

iii) The graph analysis reports a striking non-monotonic dependence of the aztreonam landscape topology on drug concentration and states that at intermediate concentrations a global peak “disappears as it becomes absorbed into a large connection node,” thereby losing influence on evolutionary trajectories. This is counter-intuitive and currently under-explained. The authors should clarify the biological meaning of this “absorption.” In particular, it is important to disambiguate whether the effect reflects fitness differences between the former peak and surrounding genotypes falling below the neutrality threshold used in coarse-graining, whether it reflects genuine biological flattening due to concentration-dependent assay dynamics, or whether it is sensitive to the choice of neutrality definition (which depends on a threshold set from the no-drug condition). A useful addition would be a quantitative diagnostic showing how the peak’s fitness advantage over its neighborhood changes with concentration and a robustness check demonstrating whether the peak’s disappearance persists under reasonable perturbations of the neutrality cutoff.

iv) The manuscript identifies specific epistatic drivers under aztreonam selection and emphasizes strong context dependence among particular residues, while noting that these effects are much weaker under ampicillin. Although the statistical association is compelling, the paper currently lacks a structural or biophysical explanation for why these residues display positive epistasis for aztreonam but not for ampicillin. The authors should include at least a minimal structural

analysis mapping the highlighted residues onto a TEM-1 crystal structure and offering a plausible mechanistic hypothesis (e.g., active-site geometry, loop dynamics, substrate positioning, stability trade-offs), ideally with an aztreonam-docked pose if feasible, or with reference to established ESBL structural literature consistent with the proposed mechanism. This would substantially improve the mechanistic depth and interpretability of the epistasis claims.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors addressed most of my previous points. The additional detailed analyses and supplementary figures improved the manuscript by providing further support for their conclusions, which are now discussed and contextualized in the Discussion section. However, a few points remain to be addressed.

Major comments:

- Re. R1.2
“Distribution histograms per concentration and %CV tables are shown in new Supplementary Figure S16.”: I only see the distribution histograms but not the described %CV tables.

Figure S17 clearly makes the point of the two detection limits, but some elements need to be confirmed. In my version, open circles indicating low read variants were not visible, possibly due to overlay with other variants. If this is the case, highlighting the open circles is confusing. In the aztreonam dose-response curve, there is also a gray band that is not explained in the caption.

In the cheater paragraph, only one citation was added, which is inconsistent with the rebuttal and the language in the paragraph.

Please ensure that the caption of Figure S7 is correct. The figure axis describes the mean AUC across concentrations, but the caption (incorrectly?) describes the AUC at the reference concentration.

- Re. R1.3
Figure S18 does not align with the corresponding caption and rebuttal descriptions. Therefore, I was unable to assess this additional analysis.

Minor comments

- Bold typesetting is still used inconsistently (l. 171 vs l. 187).

- l. 636 Fig S1 WT letters gray → black: I actually meant to change the caption, since in the original figure, the WT residues were shown in black, but described as gray in the caption. Therefore, the original figure should be restored and given the correct caption.

- Methods barcode matching: The authors describe expanded methods in their response, but the Methods section does not seem to have changed.

- The determination of the IC50 for the AUC comparison is not described and should either be added to the methods section or be easily accessible and fully described in the github material with the other wet lab protocols.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have responded thoroughly to all of my comments. I enjoyed reading the new conclusion.

(Remarks on code availability)
I appreciate sharing the reads via github is not feasible. Did the authors explore OSF or zenodo for this data? SRA or ENA may also accept the read data for archiving. If possible, this would be preferable to data available upon request.

Reviewer #4

(Remarks to the Author)
The authors have adequately addressed all of my previous comments and concerns. The revisions have improved the clarity and quality of the manuscript, and I do not have any further major issues to raise.

In my view, the manuscript is now suitable for publication.

(Remarks on code availability)

Point-by-point Response to Reviewers

Manuscript: Gaszek, Yildiz, Meng et al. “Higher-order epistasis drives evolutionary unpredictability toward novel antibiotic resistance” (Nature Communications, round 1 revision).

We thank all four reviewers for their careful evaluation and suggestions. Below we address each comment in turn, with our response and the location of any new text or figure. Revisions span seven broad categories: (a) monoculture IC50/MIC validation for a representative variant panel (n = 13 for AMP, n = 18 for AZT); (b) mutation-stratified ML validation addressing interpolation-vs-generalization concerns; (c) a full model-class comparison across linear, Lasso, decision-tree and LightGBM variants; (d) classification-metric grounding of the RMSD ≈ 0.3 claim; (e) extended-figure reporting of pairwise epistasis, prediction densities and R^2 curves across all measured concentrations; (f) a rewritten Discussion section comparing our landscape structure to refs 7, 8 and the broader combinatorial-β-lactamase literature; (g) clarification of every cited line number, figure caption and typographical inconsistency. The new computational analyses are reproducible from the data and code in the public repository at https://github.com/msadikyildiz/Gaszek_Yildiz_Meng_2025; raw data underlying the remaining display items are available from the corresponding authors on reasonable request.

Reviewer 1

We thank the reviewer for the careful reading and for recognising the scale and public-health relevance of the dataset. The reviewer’s concerns prompted several substantive additions: a monoculture IC50/MIC validation panel for a representative panel of variants (Supplementary Figures S7 and S8); a quantitative replicate-reproducibility analysis (Supplementary Figure S16); full-concentration grids of pairwise epistasis and predicted-vs-measured densities (Supplementary Figures S11 and S12), with the R^2-by-order decomposition reported in Supplementary Figure S9; sensitivity/specificity grounding for the classification claim (Supplementary Figure S15); plain-language expansions of the DCA and LGL analyses; and every figure-caption and typographical fix the reviewer listed.

Major comments

R1.1 — Choice of mutations and clinical context.

Comment. The 13 loci and 18 variants chosen are only described as being clinically identified. The authors should describe the selection process in more detail and explain whether the variants are linked to aztreonam resistance or extended-spectrum beta-lactamases (ESBLs). What is known about the clinical isolates in which these mutations were observed? Were the relevant antibiotics used in the hospitals where the isolates were collected? Additionally, does laboratory evolution under aztreonam select for some of the same mutations?

Response. Seventeen of the 18 substitutions in the library occur in clinically isolated TEM-family variants, including several ESBLs such as TEM-3, TEM-10, TEM-24, and

TEM-52. These substitutions map to 12 positions in the mature enzyme. The remaining position, L21, lies in the TEM-1 signal peptide (residues 1-23), which is cleaved during periplasmic export of the mature enzyme. We originally intended to include the clinically observed L21F substitution, but L21P was generated during library construction. We retained L21P as a construction-control site to preserve combinatorial completeness across all 13 target loci, and this is now stated explicitly in the Methods.

We also added a homoplasy paragraph to the Results. This paragraph notes that the mature-enzyme substitutions have recurred independently across TEM evolutionary lineages (Salverda et al. 2010; Bradford 2001), providing independent support that these positions lie on mutational paths repeatedly used during TEM/ESBL evolution rather than representing arbitrary laboratory choices.

Detailed phenotypic histories for TEM-family clinical isolates remain limited. The Lahey β-lactamase repository has been migrated to the NCBI β-lactamase data resources (https://www.ncbi.nlm.nih.gov/pathogens/beta-lactamase-data-resources/), which catalogues sequence and nomenclature but does not provide detailed phenotypic data, treatment history, or antibiotic-use context for the contributing isolates. We have added a caveat noting this limitation, and the revised Discussion now scopes clinical translation explicitly.

To our knowledge, systematic combinatorial laboratory-evolution experiments under aztreonam selection have not been published for TEM-family β-lactamases. Weinreich et al. 2006 selected TEM-1 under cefotaxime, another oxyimino β-lactam, and recovered a subset of the positions studied here, notably R164 and E240, consistent with our empirical enrichment pattern. We cite this connection in the revised Discussion.

R1.2 — Fitness measurements, drug degradation, 10 % reproducibility.

Comment. The authors use a non-standard measure of fitness with the AUC, which is justified by the non-monotonic growth curves in beta-lactams. However, clearer validation and description of measurement sensitivity and precision are necessary. Does the AUC measurement in competition align with the growth phenotype data of single strains? Drug degradation and other interactions in the pooled assay may make fitness measurements less quantitative. It would be important to validate the results with direct fitness measurements in monoculture, at least for a few mutants. The authors claim a typical standard deviation of 10% between replicates (line 163), yet they never demonstrate this, for example, by plotting correlations between replicates. Figure S3 shows that some genotypes have very noisy fitness values, jumping from the minimum to the mean trend from one concentration to another. What is the reason for this? Could dose-response curves be extracted from these data for each genotype, perhaps by normalizing to the wild type or the mean, to generate a composite resistance measure (e.g., the IC_{50})? The regression analysis may work better if IC_{50} is used instead of AUC-fitness.

Response. We have added monoculture IC50/MIC measurements for a representative variant panel against both drugs (n = 13 for AMP, n = 18 for AZT; Supplementary Figures S7 and S8). Each variant's AUC-fitness correlates monotonically with its monoculture IC50 (Pearson r = 0.88 for AMP, r = 0.80 for AZT), and the rank ordering is

preserved. These results support the use of pooled AUC-fitness as a proxy for the monoculture resistance phenotype within the validated envelope.

At the highest bactericidal ampicillin concentrations, drug degradation by highly active variants could in principle suppress the pooled signal. However, all mutants are at approximately 10^{-5} frequency and typically die before significantly lowering the bulk drug concentration. The monoculture validation confirms that the pooled measurement at those concentrations recovers the single-variant ranking, and we discuss this in the revised Methods.

We have updated the claim at line 163 with quantitative per-genotype numbers. The revised text reports median coefficient of variation of 23-26% on the viable subset at informative concentrations (AMP 12.2-195 µg/mL, AZT 0.44-1.33 µg/mL), with pooled-replicate Pearson r between 0.32 and 0.75 across conditions. Distribution histograms per concentration and %CV tables are shown in new Supplementary Figure S16. Reproducibility decreases at sub-MIC and supra-MIC extremes, as expected from plate-to-plate stochasticity and extinction-driven undersampling, and we discuss this in the revised text.

The concentration-dependent narrowing of quantifiable signal follows from two detection limits operating at opposite ends of the dose axis. At very low β-lactam concentrations (below or at the edge of the active range) the fitness differences among variants collapse toward the wild-type value, so biologically real small effects fall below the measurement floor set by replicate noise. At very high bactericidal concentrations almost every variant dies, leaving a small number of survivors to dominate the dynamic range of the pooled assay; those survivors accumulate disproportionate read counts and their relative fitness is artificially inflated, while the much larger pool of extinct or near-extinct variants sits at the sequencing-detection floor and its true fitness differences are not resolvable. The %CV curves in Supplementary Figure S16 and the dose-dependent fluctuations visible in the per-genotype dose-response traces in Supplementary Figure S17 are both direct manifestations of these two floors. The informative regime lies between them; our reference concentrations for the main-text analyses (AMP 781 µg/mL, AZT 36 µg/mL) were chosen inside this window, and the full-grid supplementary analyses (Supplementary Figures S9, S11 and S12) span both floors to document the transition rather than hide it.

The Figure S3 fluctuations are concentrated at high β-lactam concentrations where mutants go extinct from drug-induced killing, producing low sequencing counts and noisy AUC-fitness estimates. We have added a revised diagnostic version of Figure S3 as new Supplementary Figure S17, which renders low-read points (open markers) per-variant, shows a concentration-dependent limit-of-detection band below the dead-mutant curve, and overlays ±SD per variant (the original Figure S3 is retained). After this revision only 1.1 % of plotted points are flagged; the flagged rows are concentrated among top-20-AMP-composite variants running on the AZT panel (an AMP-selected-then-AZT-tested orthogonality effect), not in the WT, TEM-1^{dead}, single-mutant, clinical-allele or top-20-AZT categories. A library-invariance analysis at AMP 781 µg/mL shows that the pairwise biochemical-epistasis matrix is essentially identical on the clean subset (Pearson r = 1.00), the order-resolved R^2 differs by at most 0.05 at any epistatic order,

and the SHAP driver ranking is largely preserved (Spearman ρ = 0.87); these results (Supplementary Figure S17) indicate that the low-read tail does not materially alter the conclusions and support retaining the `majority_low` filter.

On extracting dose-response curves: in our pooled assay, per-genotype dose-response curves do not follow a standard Hill function (Figure S3, colored traces). As beta-lactam concentration rises, less-fit variants become extinct, opening nutrient space for a small group of highly-fit variants whose frequencies rise at higher drug concentrations, producing non-monotonic curves. The per-genotype values are therefore relative fitness, and extracting IC50s from these curves would conflate competitive expansion with intrinsic sensitivity. The monoculture panel (Supplementary Figures S7 and S8) provides the intrinsic measurements for the representative variant panel.

R1.3 — Consistency across concentrations.

Comment. Understandably, not all measured concentrations were analyzed and presented. Nevertheless, it would be useful to evaluate whether epistatic interactions are consistent across concentrations and if the general patterns from Figures 4 and 5 hold as further validation of the approach. Another place where this issue arises is in ll. 222, which describes Figure 2C and notes the absence of a clear global peak at low and high concentrations. However, contrary to the text, this highly fit global peak is only absent at a single concentration: 36 µg/ml. Since the result hinges on a single concentration, it is unclear whether it is true or potentially due to measurement noise, especially given the threshold-based node definition.

Response. We have added three Supplementary Figures reporting the full-concentration grid: pairwise-epistasis heatmaps at every non-zero concentration (Supplementary Figure S11, AMP and AZT panels), R^2-by-order curves across all 12 drug × concentration conditions (Supplementary Figure S9), and measured-vs-predicted density panels across epistatic orders 1 to 6 at every concentration (Supplementary Figure S12). AZT additive R^2 declines from 0.42 at 0.44 µg/mL to a minimum of 0.04 at 108 µg/mL (0.06 at 324 µg/mL), a clear ESBL-epistasis signature; higher-order R^2 rises monotonically with order at every condition. AMP 781 µg/mL is the second-hardest AMP condition, AMP 3.1 µg/mL is the hardest (compressed dynamic range at sub-MIC). The pairwise-epistasis pattern drifts continuously with concentration rather than holding to a single fixed template: cross-panel Pearson r spans 0.20 to 0.76 for AMP and 0.01 to 0.80 for AZT, highest between adjacent concentrations and decaying as concentrations separate.

We addressed the l.222 comment on the absent global peak at 36 µg/mL with a quantitative diagnostic (Supplementary Figure S18). The fitness advantage of peak-node genotypes over their one- and two-mutation neighbours was tracked across the full AZT concentration range, and the landscape graph was rerun under 31 neutrality cutoffs (0.15 to 0.45, step 0.01). The disappearance of the global peak at intermediate concentrations is robust (it persists in 30/31 and 31/31 threshold settings at 36 µg/mL), and the fitness-advantage analysis shows genuine concentration-dependent flattening: peak genotypes lose their median 1-mutation-neighbour advantage from ~1.5 log-units at 12 µg/mL to ~0 at higher concentrations.

R1.4 — Predicting genotype fitness; DCA/LGL clarity; pipeline β-lactams.

Comment. The authors describe two methods for predicting the fitness levels of genotypes: regression with increasing epistatic order and machine learning. However, they never discuss or compare the quality of the predictions, nor do they address the advantages or disadvantages of these methods for future use. The machine learning section of the main text is somewhat obscure. How do results from aztreonam selection compare to those from other beta-lactams? How does the choice of mutations affect the results, especially when considering pipeline beta-lactams without clinically known mutations? The claim that this general approach can be used to make predictions for beta-lactams in the pipeline (line 428) seems slightly exaggerated. The last section (lines 364) requires substantial further explanation; it is difficult for readers unfamiliar with DCA, the resulting Hamiltonians, and LGL to follow. It is important to explain these concepts more clearly in the main text, not just in the Methods section.

Response. We have toned down the main-text claim at l.428, explicitly noting that extrapolation to drugs without training data, or to mutations at positions outside our 13-residue panel, requires fresh measurements rather than prediction from the present dataset. We have also added a direct model-class comparison (Supplementary Figure S13): linear, Lasso, decision tree and LightGBM, each with and without pairwise-interaction features. LightGBM is the best-performing model at every training size; pairwise-Lasso closes ~79 % of the additive-to-LightGBM gap on AMP but only ~35 % on AZT, the quantitative signature of non-additive structure in the AZT landscape beyond what pairs alone can express.

The DCA/LGL section has been revised with plain-language expansions of the local fields h_i (site conservation), the family couplings e_ij (pairwise couplings, for instance distant residues that fold closer together), the β-lactamase family MSA/HMM construction, and the VAE-based LGL pipeline. The expansions are inserted inline at the relevant main-text paragraphs.

R1.5 — Discussion and comparison to references 7 and 8.

Comment. Currently, this is mainly a summary of the results. The authors should discuss their work in the context of the relevant literature and clarify the broader implications. In particular, how does the landscape structure compare to that in other similar studies, such as references 7 and 8?

Response. We have rewritten the Discussion as an actual discussion rather than a summary of results, as suggested by the referee. The revised text explicitly compares the landscape structure to ref. 7 (Weinreich et al. 2006, five-substitution combinatorial trajectories in TEM-1 under cefotaxime; we extend to higher-order combinations and quantify drug conditions that produce path-level unpredictability) and ref. 8 (Mira et al. 2015, adaptive landscape of resistance genes reorganises with antibiotic concentration; we extend this observation to 13-site combinatorial space and quantify the effect in AZT). We also contextualise against Firnberg 2014, Jacquier 2013, Stiffler 2015 (TEM-1 pairwise work), Chen 2022 (VIM-2 cross-concentration analog) and Chevereau 2015 (theoretical prediction of drug-dependent epistatic reorganisation). Clinical translation is bounded to the validated envelope (monoculture panel, drugs

tested, positions sampled). The full revised Discussion is in the track-changed manuscript.

Minor comments

Comment	Response
Bold inconsistent (l.171 vs l.187)	Normalised; all main-text `Figure N` references now bold.
l.253 vs Fig 3B on 164H	Caption updated: “variants carrying E104K, R164H/S/N and E240K were enriched over the wild-type allele; among R164 substitutions, R164N and R164S were more strongly enriched than R164H.”
ll.538 panels G/H	Caption corrected to match the panels: (G) coefficient of determination R^2 and (H) Spearman ρ, each versus the highest-order epistatic terms included, with ampicillin (grey) and aztreonam (pink) overlaid in both. (The prior caption mislabelled both G and H as R^2 split by drug.) Added gloss that the $R^2 = 0.5$ reference line is a qualitative benchmark, not a statistical cutoff; panels ordered to read G then H.
Fig 5A/5B rows vs columns + R244S colour	Captions for Fig 5A/5B corrected to “Rows represent the 18 studied mutations and columns correspond to the 553 individual genotypes” (previously inverted). Added a note explaining the R244S red-cell-vs-negative-beeswarm discrepancy as context-dependent SHAP attribution.
l.568 Fig 6A letter colours	Caption now reads: “TEM-1 wildtype amino acids listed on x-axis in black; enriched substitutions within the protein family are shown in red.”
l.636 Fig S1 WT letters grey → black	Wild-type residue letters in Figure S1B recoloured from grey to black and the figure re-embedded.
ll.753 missing MIC concentrations	Methods note added: these were MIC-quantification conditions where insufficient barcode recovery prevented fitness estimation; they appear in Figure S2 but not in the fitness-landscape analyses.
Methods barcode matching	Methods expanded: barcodes matched by exact agreement plus Hamming-1 when

Comment	Response
	unambiguous to a single variant; discarded otherwise. All barcodes for a variant pooled before frequency computation. Implementation at `src/01_preprocess.py`.
Fig 1A chemical-structure caption	Expanded to describe the β-lactam hydrolysis mechanism (four-membered ring, nucleophilic attack, loss of acylation capacity).
Fig 1E/F unit labels	Concentration-axis units normalised to µg/mL across the ampicillin and aztreonam fitness-distribution panels (consolidated into Figure 1E in the revision; the former Figure 1F panel now carries the structural atlas).
Fig 2B/C and S4 WT highlight	Asterisks added to the wild-type-containing nodes.
Fig 4C missing colorbar	Fig 4C and Fig 4D now share a unified colorbar centred at each drug's WT fitness, with tick labels ±3 units relative to WT.
Fig S6 AZT panels + r/p values	Figure S6 now includes aztreonam panels alongside ampicillin, with the Pearson r overlaid on every panel (peak Spearman $r_s = -0.51$).

Reviewer 3

We thank the reviewer for the positive assessment and for the constructive suggestions. The revision adds: a homoplasy paragraph and NCBI-data-availability caveat to bring out clinical relevance; a full model-class comparison (Supplementary Figure S13); sensitivity/specificity grounding for the RMSD ≈ 0.3 classification claim (Supplementary Figure S15); a rewritten Discussion with explicit comparisons to refs 7 and 8; and every presentation fix in the minor comments. The cheaters-not-a-problem assertion is now supported by citations from the Gore and Krug labs, and the WT-is-arbitrary concern is addressed in a dedicated paragraph.

Major comments

R3.1 — Clinical relevance.

Comment. The clinical relevance of this study could be brought out better. How prevalent are the chosen mutations in the population / clinical isolates? Do they show homoplasy? The AUC-fitness metric is well motivated; if it is possible to relate its value to current clinical resistant/sensitive thresholds for the tested drugs, this would help readers put these numbers in context of actual treatment. The definition of the WT

sequence seems completely arbitrary. Is it the most common allele in the population, or just in the reference strain? Is there a specific biological meaning/interpretation of the WT allele?

Response. Seventeen of the 18 substitutions, spanning 12 of the 13 targeted positions, occur in published TEM-family clinical variants and act on the mature enzyme. The remaining substitution, L21P, lies within the TEM-1 signal peptide (residues 1 to 23, cleaved during periplasmic export) and does not alter the mature β-lactamase sequence. The intended substitution at this position was L21F, but L21P was generated during construction. We kept L21P in the combinatorial library as a construction-control position so that the library remained combinatorially complete across all 13 target loci, and we have added a sentence to the Methods noting that L21P is expected to behave neutrally with respect to the mature enzyme's activity, since the signal peptide is cleaved during periplasmic export and is absent from the folded β-lactamase. We have also added a homoplasy paragraph citing Salverda et al. 2010, which documents point-mutation recurrence across independent TEM clinical isolates with convergent accumulation at E104, R164, E240, G238, and R244, and Bradford 2001, which describes parallel emergence of ESBL-associated substitutions across geographically and phylogenetically distinct lineages. These lineages are homoplastic by definition: the same substitutions have arisen independently multiple times, providing orthogonal evidence that the positions and substitutions we chose lie on the physical mutational pathways ESBL evolution has repeatedly traversed.

On relating AUC-fitness to clinical resistance thresholds, we have added monoculture IC50/MIC measurements for a representative variant panel (n = 13 for AMP, n = 18 for AZT) against both drugs (Supplementary Figures S7 and S8). AUC-fitness correlates monotonically with IC50 and preserves rank ordering across both drugs. For clinical translation specifically, we anchor the AZT-resistance analysis at 36 µg/mL, which lies in the selective window where the aztreonam landscape exhibits its strongest higher-order epistatic structure. Explicit calibration of AUC-fitness to clinical breakpoints is out of scope for this revision, but the monoculture data provide a validated mapping for the tested variant panel.

On the WT-is-arbitrary concern, we have added a dedicated paragraph clarifying the basis for our wild-type designation. TEM-1 is the founding member of the TEM family, was first clinically isolated in 1963, and is near-optimal against ampicillin, the substrate for which it was selected in the clinic. Subsequent TEM-family variants, from TEM-2 and TEM-3 through the hundreds of known ESBLs, are defined relative to this reference by sequence comparison. The WT designation therefore reflects both a molecular-phylogenetic root and a functional optimum for the native substrate, rather than a convenience of library construction.

R3.2 — Forecasting.

Comment. A bit more is needed to achieve the aim of forecasting. (2a) How do trajectories/predictability change from one MIC step to the next, i.e. can you compare your graphs between concentration steps? What is the path for a sequence to each higher resistance step, not optimising within a given concentration? (2b) Some of the

machine learning aspects could be simplified to support the claims. I would suggest also adding a lasso regression (e.g. with the glmnet package), then comparing (e.g. in a table) the predictive accuracy of linear models, with L1 regularisation, and decision trees, and showing the predictive accuracy at each training sample size.

Response. On trajectory comparisons across concentrations (point 2a), we tracked the fitness advantage of peak-node genotypes across the AZT concentration range (Supplementary Figure S18). For the 12 µg/mL global peak (1,917 genotypes), the median fitness advantage over 1-mutation neighbours rises from zero (no drug) to ~1.5 log-units at 12 µg/mL, declines to ~0.8 at 36 µg/mL, and approaches zero at higher concentrations. The 108 µg/mL peak (185 genotypes) is largely a subset of the 12 µg/mL peak group; as concentration increases, most members lose their fitness advantage while only a small resistant subset persists. This non-monotonic shift reflects a filtering process within the peak group, consistent with clinical observation that ESBL variants arise stepwise rather than in one large jump.

On ML simplification (point 2b), we have added Supplementary Figure S13 with learning curves from 0.03 % to 99 % training fraction for six model classes: linear additive, L1-Lasso additive, linear with all pairwise-interaction features (171-dim), L1-Lasso with pairwise features, a plain decision tree, and LightGBM (manuscript default). All models are evaluated on the same fixed 5 % held-out test set across 11 training-fraction points with 3 subsample seeds each, for both AMP 781 µg/mL and AZT 36 µg/mL. LightGBM is the best-performing model at every training size; pairwise-Lasso closes ~79 % of the additive-to-LightGBM gap on AMP but only ~35 % on AZT. The drug asymmetry is the quantitative signature of non-additive structure in the AZT landscape beyond what pairs alone can express. A single decision tree is the worst model everywhere (default scikit-learn trees over-fit; negative R^2 on AZT at every training size).

The original choice of LightGBM as the manuscript's predictive model was motivated by four properties specific to our data. First, the fitness landscape exhibits higher-order epistasis in AZT that a linear or pairwise model cannot capture (Supplementary Figure S9 shows additive R^2 0.10 at AZT 36 µg/mL and only 0.21 at pairwise, rising to 0.99 only at K = 9 or higher). Gradient-boosted trees natively represent non-additive combinations without a priori feature engineering. Second, the features are categorical with between 2 and 4 states per position; LightGBM's native categorical handling uses these directly, avoiding the 171-column one-hot expansion that a pairwise Lasso requires. Third, SHAP decomposition (Lundberg and Lee 2017) provides per-mutation and per-genotype attributions that recover the biological interpretability on which Figure 5 depends. Fourth, LightGBM is the empirically best-performing model at every training fraction on both drugs in our now-reported head-to-head comparison; the new Supplementary Figure S13 makes this rationale explicit rather than assumed.

R3.3 — Cheaters evidence and citations.

Comment. This is less in my area of expertise, but I did note that in the experimental design, the presence of local cheaters not being a problem is simply asserted; evidence/data is not given to support this important claim. Are there any data, or other studies, which support this claim?

Response. We have added citations at the cheater passage: Yurtsev et al. 2013, 2016 (Gore laboratory, direct characterization of cheater-resistance cooperative dynamics in β-lactamase-producing communities). In our pooled assay, the cheater phenotype registers as low AUC-fitness (long lag phase, lower final density) rather than the high-density profile of a truly drug-degrading resistant variant, because resistant mutants rapidly consume available nutrients before cheaters can rise to non-trivial frequency.

R3.4 — Discussion and epistasis section.

Comment. There are a few issues with the presentation directly: the conclusion only briefly restates parts of the results, and in my view needs rewriting. Given my view of the WT being arbitrary, the long section on the optimality of the WT is not warranted. The description of the epistatic interactions is a bit muddled; probably starting with the overall results from the linear model(s), rather than talking about specific sequences/mutations, would help improve this.

Response. The Discussion has been rewritten around three findings: (i) drug-asymmetric landscape topology with quantitative R^2 numbers, (ii) relationship to prior combinatorial β-lactamase studies including Firnberg 2014, Jacquier 2013, Stiffler 2015, Chen 2022 (VIM-2) and the theoretical prediction from Chevereau 2015, and (iii) explicit comparison to refs 7 (Weinreich et al. 2006) and 8 (Mira et al. 2015), extending their path-predictability framework to 13 sites and quantifying which drug conditions produce path-level unpredictability. Clinical translation is scoped to the validated envelope. The full revised Discussion is in the track-changed manuscript.

Following the suggestion to lead with the overall linear-model results, we now report the global epistatic-order decomposition the reviewer asked to see first as new Supplementary Figures: (1) global ΔR^2-by-order curves for all 12 drug × concentration conditions, including their concentration-dependent shape (Supplementary Figure S9); (2) the quantitative drug asymmetry at matched stringency (AMP 781 vs AZT 36 primary, AMP 195 vs AZT 12 secondary; 3.5× primary, 3.3× secondary; Supplementary Figure S10); and (3) the bridge to the ML-simplification results (Supplementary Figure S13). The Discussion now leads with this model-level decomposition before the per-mutation interpretation, and the main-text epistasis section references the new global-decomposition figures alongside its existing per-mutation narration. Headline: additive $R^2 = 0.35$ (AMP 781) vs 0.10 (AZT 36), a 3.5× asymmetry; pairwise closes the gap to 2.1×; orders 3 to 6 close it to 1.1×; K = 13 asymptotes between 0.999 and 0.99998.

Minor comments

Comment	Response
Abstract, mention *E. coli*	Added: “evolution of *Escherichia coli* TEM-1 β-lactamase into ESBLs”.
l.72-74 citation	Sarkisyan et al. 2016 added for deep-mutagenesis scaling.
l.79 “novel” β-lactam rephrase	“A β-lactam of a different class to which TEM-1 has negligible native activity (aztreonam, a monobactam).”

Comment	Response
l.90 predict-fitness claim too strong	Softened: “Within the TEM-1^{CML} library and for the two β-lactams tested here, these models enable quantitative prediction of variant fitness, with accuracy that varies with drug identity, training fraction, and required operating point.”
l.108 L21P/L21F clarification	Added: “Only L21P, and not L21F, is included in the TEM-1^{CML}. L21 lies within the signal peptide (residues 1 to 23) cleaved during periplasmic export, so L21P is retained as a library-construction control expected to be neutral at the mature-enzyme level.”
l.335 RMSD ~0.3 sens/spec	Added: sensitivity/specificity at three stringency thresholds (above-WT, top-5%, top-1%) for both drugs. AZT at top-1%: sens 0.15, spec 1.00, PPV 0.99, AUROC 0.99. See new Supplementary Figure S15.
l.249 top-1% cutoff justification	Clarified: “The 1 % cutoff was chosen operationally as a stringent tail (approximately 553 genotypes per drug) sufficient to resolve mutational enrichment without requiring absolute fitness thresholds; qualitatively similar patterns obtain at the top 5 %.”
l.369 DCA biological terms	Expanded: h_i = site-wise conservation, e_ij = pairwise couplings quantifying amino-acid co-occurrence strength between position pairs (for example, distant residues that fold closer together).
PDB structure on figure	The structural panel (now Figure 1F) presents a structural atlas of the library positions on the TEM-1 crystal structure, with the PDB accession (1BTL) cited explicitly in the caption.
l.382 correlation quantification	Quantified: peak Spearman r_s = −0.51; overlaid on all Fig S6 panels.
Latent space in Fig 6 not well explained	Expanded (see R1.4 response).
Fig 1C accurate residue coordinates	The structural atlas (Figure 1F) shows all 13 mutated positions at their crystallographic coordinates on PDB 1BTL, coloured by structural class (active-site pocket rim, omega loop, SDN edge,

Comment	Response
	hinge, distal scaffold), with the omega loop and active-site landmarks annotated.
Fig 3A no high-AZT / low-AMP mutants	Noted in Discussion: most clinical TEM variants retain basal AMP activity; full-concentration grid (Supplementary Figures S9, S11 and S12) shows the few genotypes with this profile.
Fig 3B red/black/grey key	Caption now explains: wild-type residues in black, enriched substitutions in red, non-enriched substitutions in grey.
Fig 4 0.5 R^2 threshold meaning	Caption gloss added: qualitative benchmark, not a statistical cutoff.
Fig 4 G/H out of order	Figure layout reordered.
Fig 5 553 columns, not mutations	Captions for Fig 5A/5B corrected: rows = 18 mutations, columns = 553 individual genotypes.
Fig 6D native β-lactamase source	Added inline description of PF13354 HMM-based MSA construction and label-repositioning.

Reviewer 4

We thank the reviewer for the methodologically focused critique, which prompted four new analyses. For point (i), AUC-Fitness to MIC/clinical mapping: we added monoculture IC50/MIC validation for a representative variant panel (n = 13 for AMP, n = 18 for AZT; Supplementary Figures S7 and S8) and a clinical-translation scoping paragraph in the Discussion. For point (ii), block-holdout ML: we added Leave-One-Mutation-Out validation across all 18 substitutions and Hamming-distance-stratified holdout analysis ($D \in \{3..10\}$) for both drugs (Supplementary Figure S14), identified a biologically interpretable failure mode at R244{C,S}, and added classification metrics at three stringency thresholds in Supplementary Figure S15. For point (iii), peak-absorption robustness: we added a fitness-advantage diagnostic across concentrations and a neutrality-cutoff sensitivity analysis across 31 values (Supplementary Figure S18). For point (iv), structural mechanism: we added a minimal structural analysis mapping the key epistatic residues onto the TEM-1 crystal structure, with reference to established ESBL structural literature.

Major comments

R4(i) — AUC-Fitness to MIC/clinical mapping.

Comment. The manuscript uses AUC-Fitness as a quantitative proxy for loss-of-function and motivates this by the non-monotonic growth dynamics under β-lactam selection, while also noting pooled-culture effects such as “cheaters.” This is a reasonable operational in vitro metric, but the manuscript also explicitly references clinical context,

including framing 36 μg/mL aztreonam as slightly above a clinical cutoff and discussing resistance prediction in clinically oriented terms. Given this, the relationship between AUC-Fitness and clinically used resistance metrics (MIC/IC50 or S/I/R categorization), and more broadly to in vivo/clinical fitness, needs to be clarified. The authors should either provide validation linking AUC-Fitness to MIC/IC50 for a representative panel of variants measured in monoculture, or clearly limit claims to in vitro pooled-competition fitness landscapes and add a dedicated Discussion paragraph explaining why, and to what extent, AUC-Fitness is expected (or not expected) to translate to monoculture MIC and clinical outcomes.

Response. We have added monoculture IC50/MIC measurements for a representative panel of TEM-1 variants (n = 13 for AMP, n = 18 for AZT) against both drugs (Supplementary Figures S7 and S8). Key findings: (a) each variant's AUC-fitness correlates monotonically with its monoculture IC50 across both drugs (Pearson r = 0.88 for AMP, r = 0.80 for AZT); (b) the rank order of variants by AUC-fitness is preserved in IC50; (c) per-concentration AUC-fitness tracks per-concentration growth in monoculture. This supports using the pooled AUC-fitness as a proxy for the monoculture resistance phenotype within the validated envelope of the tested variant panel.

The Discussion has been rewritten to scope in vivo/clinical translation explicitly. Our fitness measurements are pooled-competition AUC-fitness in *E. coli* NEB10-β, a controlled uniform genetic background that is systematically different from clinical polymicrobial contexts. Translation to in vivo β-lactam pharmacology, clinical resistance breakpoints, or antibiotics outside our training panel requires additional measurement. Within the validated envelope, the model identifies the high-fitness tail with high specificity (1.00 in both drugs) and high precision at the relevant stringency (positive predictive value 0.99 for aztreonam at the top 1 %, 0.97 for ampicillin at the top 5 %), the operating regime most relevant to resistance surveillance.

R4(ii) — Block-holdout validation.

Comment. The authors argue that the aztreonam landscape is less predictable due to higher-order epistasis and support this in part with LightGBM performance on held-out data, including the statement that 10% training yields RMSD ~0.3 and would be sufficient to classify mutants as "resistant" for clinical purposes. Random train/test splits in combinatorial genotype spaces often primarily measure interpolation because many test genotypes lie at small Hamming distance from training genotypes, which can inflate apparent predictability. To support a claim of intrinsic unpredictability (or to demonstrate genuine generalization), the authors should perform at least one mutation-stratified or block-holdout validation; for example, excluding all variants containing a focal mutation (e.g., E104K) from training and then predicting them, or a distance-based split enforcing a minimum Hamming distance between training and test sets. Reporting results under such schemes for both ampicillin and aztreonam would substantially strengthen the predictability narrative.

Response. The manuscript's choice of LightGBM as the predictive model reflects four properties of the data: native handling of higher-order non-additive interactions (not capturable by linear or purely pairwise models), native categorical feature support

across 13 positions with 2 to 4 states each, compatibility with SHAP attribution for per-mutation biological interpretation (Figure 5), and empirical superiority at every training fraction in our head-to-head comparison (Supplementary Figure S13). We have performed three stratified-holdout analyses on both drugs (Supplementary Figure S14 and accompanying Supplementary Table):

1. *Leave-One-Mutation-Out (LOMO).* For each of the 18 observed substitutions we removed every variant carrying it from the training set and predicted the held-out variants. On average the generalization gap (LOMO RMSD / random-split RMSD) is modest, 1.25× for AMP and 1.30× for AZT, but LOMO is a deliberately severe test: it removes every training example of a substitution and asks the model to predict that substitution's effect with no in-training support, so a context-dependent substitution is precisely the one that cannot be extrapolated. Consistent with the higher-order epistasis the manuscript documents, the gap is accordingly larger and more widely distributed under aztreonam — negative held-out R^2 at seven substitutions (A237T, M69V, G238S, R244C, R244S, R275L, R275Q) — than under ampicillin (four: A237T, M69V, R244C, R244S). The most severe failures by a wide margin are R244C and R244S, whose exclusion yields RMSD 2.2× (AMP) and 1.7× (AZT) the random baseline with strongly negative held-out R^2 (−2.4 to −2.9); R244 is a triple-state position whose R→C/S effect has no support once its variants are withheld. The remaining substitutions stay within ~30 % of the random-split baseline (for example within 5 % for E104K, R164N and N276D under ampicillin, and Q39K, T265M and N276D under aztreonam).

2. *Hamming-distance stratified holdout ($D \in \{3..10\}$).* We trained on genotypes within Hamming distance D of a reference and tested on those at distance ≥ D. With WT as reference, block-holdout accuracy is within 10 to 25 % of random-split accuracy for D ≥ 6 (training n ≥ 8,850). A sensitivity check using a randomly chosen deep genotype (Hamming-7 from WT) confirms that training-set representativeness, not distance per se, drives generalization.

3. *Classification under LOMO.* AUROC at biologically motivated thresholds (P25 for AMP, P90 for AZT) is retained for most LOMO folds: AMP 0.71 median, 0.58 worst (R244C); AZT 0.63 median, 0.51 worst (R244C).

R244 as a biologically interpretable failure mode. Position R244 admits three amino-acid states (R, S, C) and sits at a stability hotspot interacting with the catalytic Ω-loop. Excluding all R244C variants removes every training instance where position 244 carries a cysteine; the model has no support for the R→C substitution effect. A Lasso regressor (purely additive) LOMO-outperforms LightGBM on R244 specifically, because R244-specific epistasis cannot be extrapolated from training data that never contains R244{C,S}. This identifies R244 as the most extreme locus where the tree model learns local, non-transferable structure; the milder LOMO degradation at the other aztreonam epistatic hubs reflects the same higher-order structure, and the model's interpolation within the measured substitution set (random split) is unaffected.

We have revised the clinical-classification claim at l.335 to specify that RMSD ≈ 0.3 is sufficient to screen the high-fitness tail at stringent cutoffs (sens 0.15, spec 1.00, PPV 0.99 at top-1 % for AZT; AUROC 0.99), except for R244-bearing variants where in-training observation is required. Full classification-metric panels are in Supplementary Figure S15.

R4(iii) — Peak absorption: biological meaning and threshold sensitivity.

Comment. The graph analysis reports a striking non-monotonic dependence of the aztreonam landscape topology on drug concentration and states that at intermediate concentrations a global peak "disappears as it becomes absorbed into a large connection node," thereby losing influence on evolutionary trajectories. This is counter-intuitive and currently under-explained. The authors should clarify the biological meaning of this "absorption" — disambiguating whether the effect reflects fitness differences between the former peak and surrounding genotypes falling below the neutrality threshold used in coarse-graining, genuine biological flattening due to concentration-dependent assay dynamics, or sensitivity to the choice of neutrality definition. A useful addition would be a quantitative diagnostic showing how the peak's fitness advantage over its neighborhood changes with concentration and a robustness check demonstrating whether the peak's disappearance persists under reasonable perturbations of the neutrality cutoff.

Response. We addressed this with two complementary analyses (Supplementary Figure S18).

Fitness advantage over neighbourhood across concentrations. For the 12 µg/mL global peak (1,917 genotypes), the median fitness advantage over 1-mutation neighbours peaks at ~1.5 log-units at 12 µg/mL, declines to ~0.8 at 36 µg/mL, and approaches zero at higher concentrations. The 108 µg/mL peak (185 genotypes) shows an analogous pattern; its peak genotypes are largely a subset of the 12 µg/mL peak group (shared sequence logos). The fitness differences between peak genotypes and their neighbours shrink in absolute terms, not merely relative to a threshold, establishing peak absorption as genuine concentration-dependent flattening. The 12 µg/mL peak group is compositionally heterogeneous: as concentration increases, most members lose their fitness advantage while only a small resistant subset persists.

Threshold sensitivity. We rebuilt the fitness-landscape graphs across 31 neutrality cutoff values (0.15 to 0.45, step 0.01) for all AZT concentrations. The disappearance of the global peak at 36 µg/mL is robust: the peak is absent in 30/31 thresholds. At 324 µg/mL the global-peak node exists in most graphs but contains fewer than 32 genotypes (> 16), consistent with progressive winnowing. Our chosen threshold of 0.4 corresponds to the 99th percentile of fitness differences between 1-mutation neighbours at zero drug, providing a principled calibration point.

R4(iv) — Structural and biophysical mechanism.

Comment. The manuscript identifies specific epistatic drivers under aztreonam selection and emphasizes strong context dependence among particular residues, while noting that these effects are much weaker under ampicillin. Although the statistical association

is compelling, the paper currently lacks a structural or biophysical explanation for why these residues display positive epistasis for aztreonam but not for ampicillin. The authors should include at least a minimal structural analysis mapping the highlighted residues onto a TEM-1 crystal structure and offering a plausible mechanistic hypothesis (e.g., active-site geometry, loop dynamics, substrate positioning, stability trade-offs), ideally with an aztreonam-docked pose if feasible, or with reference to established ESBL structural literature consistent with the proposed mechanism.

Response. We map the key epistatic residues (E104K, R164H/S/N, E240K, A237T, G238S, R244C/S) onto the TEM-1 crystal structure (PDB 1BTL) in a structural atlas that now forms panel F of the revised Figure 1, and interpret them via established ESBL-structural literature (Philippon reviews on Ω-loop dynamics; Palzkill on active-site geometry; Meini review on β-lactamase substrate specificity). E240K and E104K sit at the active-site pocket rim and reorganise the R164-containing Ω-loop; A237 and G238 lie on the SDN/B3 strand that constrains the Ω-loop's conformational freedom; R244 is a stability-critical residue in the context distinguishing pre- and post-ESBL variants, at which R244C/S are strongly deleterious. This is consistent with the SHAP analysis: positive epistasis for aztreonam among E104K/R164H/S/N/E240K, combined with preservation of the wild-type residue at A237/G238/R244, reflects accommodation of a larger, less-hydrolysable substrate without compromising enzyme stability. As the Discussion notes, a static structure locates these residues but cannot quantify their energetic coupling; molecular-dynamics simulation and an experimental aztreonam-bound structure are the definitive next tests and are beyond the scope of this revision.

Summary of new analyses added for Reviewer 4

Analysis	Figure
Monoculture IC50/MIC validation (AMP n = 13, AZT n = 18)	Supplementary Figure S7 (combined) + S8 (per-concentration)
LOMO + Hamming-stratified ML validation	Supplementary Figure S14 (AMP + AZT panels)
Classification metrics at 3 thresholds × 2 drugs × 10 seeds	Supplementary Figure S15 (AMP + AZT panels)
Peak fitness-advantage across concentrations + neutrality-cutoff sensitivity	Supplementary Figure S18
TEM-1 structure + residue mapping + literature synthesis	Figure 1F (structural atlas on PDB 1BTL) + Discussion

All analyses use the manuscript's LightGBM configuration (n_estimators=100, learning_rate=0.1) and reference concentrations (AMP 781, AZT 36 µg/mL). Hyperparameters are fixed across splits so reported differences reflect generalization, not tuning. Full per-fold tables, reproducible code and figures are in the cited sub-folders of the public repository.

Reviewer 2

This manuscript was co-reviewed by one of the reviewers who provided the listed reports. No additional comments were received from Reviewer 2.

Summary of new Supplementary Figures

All 12 new figures are embedded in the revised manuscript docx as tracked insertions, with captions co-located in the Supplementary Figure Captions section and images appended after the original Figure S6 in the final figure-dump section.

#	Content	Reviewer ask
S7	Monoculture IC50 vs mean AUC-fitness (AMP n = 13, AZT n = 18)	R1, R4(i)
S8	Per-concentration IC50 vs AUC-fitness panels (AMP + AZT)	R1, R4(i)
S9	Epistatic-order decomposition of variance (cumulative R^2 + ΔR^2 by order) across all 12 drug × concentration conditions	R1, R3 (epistasis rewrite)
S10	Drug-asymmetry at matched selection stringency	R3 (same)
S11	Pairwise biochemical-epistasis heatmaps across all concentrations (AMP + AZT panels)	R1 (concentration consistency)
S12	Measured vs predicted density panels across all concentrations (AMP + AZT)	R1
S13	Model-class comparison (linear / Lasso / tree / LightGBM) for both drugs	R3(2b) explicit
S14	Block-holdout (LOMO + Hamming-stratified) ML validation for both drugs	R4(ii) explicit
S15	Classification metrics at three thresholds for both drugs	R3, R4 sens/spec
S16	Replicate-pair correlation scatter (AMP + AZT)	R1 (replicate evidence)
S17	Revised dose-response profiles with low-read	R1 (Figure S3 noise)

#	Content	Reviewer ask
	flagging (replaces original Figure S3; AMP + AZT)	
S18	Peak-absorption diagnostic + neutrality-cutoff robustness	R1 l.222, R4(iii)

Code and data availability. All new code is in the public repository at github.com/msadikyildiz/Gaszek_Yildiz_Meng_2025, organized under the sprint-identifier folders (s1_model_comparison, s2_block_holdout, s3_rmsd_justification and so on). The full landscape parquet (Epistasis_Combined.parquet, ~53 MB) is provided in the same repository; monoculture IC50/MIC validation data and wet-lab protocols are included alongside it. The repository will be archived at Zenodo on acceptance. The fitness-landscape, epistasis and machine-learning display items are reproducible from the deposited data and public code; data underlying the remaining display items are available from the corresponding authors on reasonable request.

Point-by-point response: editorial revision

We thank the reviewers and editor. Reviewers 2 and 4 raised no further points; Reviewer 3 is satisfied with one code-availability suggestion, addressed below. We have resolved every remaining Reviewer 1 point and completed the editorial checklist. Line numbers refer to the revised manuscript.

Reviewer 1

Major

Comment (R1.2a): "'Distribution histograms per concentration and %CV tables are shown in new Supplementary Figure S16.': I only see the distribution histograms but not the described %CV tables."

Response: The reviewer is correct that the %CV values were shown only as histograms. We have added a per-concentration coefficient-of-variation table (new Supplementary Table S2) reporting, for each drug and concentration, the number of viable genotypes and the median and interquartile %CV (with the pooled-library values for reference). On the viable subset the median %CV is ~23–27% across the ampicillin series and at low aztreonam concentrations (≤4 µg/mL), and rises with aztreonam concentration (≈38% at the 36 µg/mL reference, higher at 108–324 µg/mL) as variants approach extinction, matching the concentration-dependent reproducibility we describe in the text. The Supplementary Figure S16 caption now names every panel and cites the table.

Comment (R1.2b): "Figure S17 … open circles indicating low read variants were not visible, possibly due to overlay with other variants … there is also a gray band that is not explained in the caption."

Response: We regenerated Supplementary Figure S17 so that low-count measurements are drawn as open markers with a distinct black edge, enlarged and placed above the filled markers, so they are clearly distinguishable. We clarified the caption to define both shaded elements: the light-gray band along the bottom of each panel is the limit-of-detection band (fitness at or below the catalytically inactive TEM-1 dead control), and each variant's narrower shaded envelope is its mean ± s.d. across the non-flagged replicates. Open circles denote low-count measurements (per-replicate barcode read-count sum across the four timepoints below 10 in at least two of three replicates).

Comment (R1.2c): "In the cheater paragraph, only one citation was added, which is inconsistent with the rebuttal and the language in the paragraph."

Response: We have added the second citation (Yurtsev, Conwill & Gore, 2016) alongside Yurtsev et al. 2013, so the plural reference to prior work now matches the citations.

Comment (R1.2d): “Please ensure that the caption of Figure S7 is correct. The figure axis describes the mean AUC across concentrations, but the caption (incorrectly?) describes the AUC at the reference concentration.”

Response: The axis is correct: the x-axis is the mean AUC-fitness averaged across the assayed concentrations. The caption was in error and has been corrected to “mean AUC-fitness averaged across the assayed concentrations for each drug.”

Comment (R1.3): “Figure S18 does not align with the corresponding caption and rebuttal descriptions. Therefore, I was unable to assess this additional analysis.”

Response: We apologize; an incorrect image (a duplicate of Supplementary Figure S4) had been placed as Supplementary Figure S18. We have replaced it with the correct peak-absorption diagnostic and rewritten the caption to match its four panels: (a,b) sequence logos and fitness-advantage box plots for the 12 and 108 µg/mL global-peak groups over their one- and two-mutation neighbours across the aztreonam range; (c) presence/absence of a global peak across 31 neutrality-cutoff values (0.15–0.45); (d) a schematic key. The analysis shows the 12 µg/mL peak advantage declines to ~0 at higher concentrations and that the absence of a global peak at 36 µg/mL is robust across 30/31 cutoffs.

Minor

Comment: “Bold typesetting is still used inconsistently (l. 171 vs l. 187).”

Response: Per the editorial formatting guidance we removed bold typesetting from all in-text figure references throughout, resolving the inconsistency.

Comment: “l. 636 Fig S1 WT letters gray → black: I actually meant to change the caption … the original figure should be restored and given the correct caption.”

Response: Understood, and apologies for the misinterpretation. The original figure is retained (panel B letters are black) and the caption is corrected: it no longer states that wild-type residues are grey. The caption now states that all residue letters are black and that the wild-type residue is the tallest letter at each position (letter height ∝ count).

Comment: “Methods barcode matching: … the Methods section does not seem to have changed.”

Response: We have added the barcode-matching description to the Methods: barcode→genotype assignment via the long-read matchbook that links ~300 distinct barcodes to each genotype; short-read barcode counts are assigned to a genotype by exact match to the matchbook, with reads carrying more than two undetermined barcode positions discarded; all barcodes for a genotype are pooled before computing frequencies. A pointer to src/01_data_preprocessing.ipynb is included.

Comment: “The determination of the IC50 for the AUC comparison is not described …”

Response: We have added the IC50 determination to the Methods: OD600 dose-response on a plate reader; background subtraction; AUC integration; and a four-

parameter Hill (sigmoidal) fit to obtain the IC50. The full protocol is available with the other wet-lab protocols in the GitHub repository.

Reviewer 3 (code availability)

Comment: “I appreciate sharing the reads via github is not feasible. Did the authors explore OSF or zenodo for this data? SRA or ENA may also accept the read data for archiving. If possible, this would be preferable to data available upon request.”

Response: We agree; sequencing reads require deposition in an approved community repository, so the raw barcode sequencing reads (long-read matchbook and short-read barcode-count libraries) have been deposited in the NCBI Sequence Read Archive (an INSDC member) under BioProject PRJNA1370879, and the analysis code has been archived at Zenodo (10.5281/zenodo.21442350). The Data Availability and Code Availability statements now cite these accessions instead of “available upon request.”