Correction Notes: Tahoe-100M Conservation–Divergence Manuscript

This document records the full remediation history for "Two largely independent axes of drug transcriptional response in Tahoe-100M" — what was originally claimed, what a reviewer identified as wrong, the mechanism of each error, what was recomputed, and the final corrected numbers. It is the appropriate home for the revision-history narrative that was previously threaded through the manuscript itself; the manuscript now reports only final, corrected results, with brief pointers back to this document.

1. Coherence–depth claim: an under-corrected multiple-testing procedure

Original claim. At equal core size, proximal-target drugs were reported to have significantly more coherent conserved cores (pathway-enrichment coherence) than distal-target drugs: Mann–Whitney $p = 0.002$, joint-model $\beta = +0.076$, $p = 0.004$, $\Delta R^2 = 1.3\%$ over core size, $n = 168$ depth-labelled enrichable drugs. This was originally presented as a headline finding — evidence that target transcriptional depth organizes context-robustness via pathway coherence.

What a reviewer found wrong. The per-drug Benjamini–Hochberg (BH) procedure used to call a pathway "enriched" first selected pathways with an observed overlap ≥ 2 genes, and only then BH-corrected across that pre-filtered subset. This understates the effective multiple-testing universe: the correct correction denominator is the full tested pathway library (6,026 size-filtered GO/Reactome/KEGG gene sets per drug), not the small subset that happened to clear an overlap filter first.

Mechanism verification. An independent re-run of the archived procedure exactly reproduced the original per-drug `n_sig` and `coherence` values (100% agreement), confirming the bug was real and not a misreading of the original code.

Recomputation. BH correction was re-run honestly across the full size-filtered 6,026-pathway library, per drug, at FDR < 0.1. This nearly halved both the enrichable population (374 → 132 drugs with ≥ 1 BH-significant pathway, down from an originally reported 324) and the depth-labelled enrichable cohort available for the coherence–depth test (168 → 77 drugs: 35 proximal, 35 distal, 7 nuclear-receptor).

Final corrected result. Under the corrected procedure, the proximal-vs-distal coherence effect is **no longer significant**: R^2 rises from 0.457 (core size alone) to 0.468 (+ depth), $\Delta R^2 = 1.1\%$, coefficient-test $p = 0.234$, Mann–Whitney-on-residuals $p = 0.165$ ($n = 77$). The coherence–depth relationship does not survive honest correction and is not reported as a manuscript finding (Supplementary Figure S2 shows the original-vs-corrected comparison directly).

Replacement finding. Bypassing the coherence intermediate entirely and regressing each empirical axis directly on target depth + log(core size) + n_cell_lines (OLS, $n = 157$ depth-labelled drugs with both axis values): depth has no significant direct effect on context-robustness (proximal coefficient $p = 0.601$, $\Delta R^2 = 0.8\%$), but a real, moderate direct effect on dose-emergence (proximal coefficient = −0.093, $p = 0.0013$, $\Delta R^2 = 6.3\%$), robust to cluster-robust standard errors and leave-one-mechanism-class-out refitting (proximal coefficient range [−0.107, −0.081] across 26 refits, always $p < 0.01$). This direct-regression result is now the manuscript's mechanistic-depth result (§3.6).

2. Stress-axis variance partition: order-dependent (Type I) sums of squares

Original claim. A sequential (Type I) decomposition of per-sample stress-axis magnitude across the full 50-cell-line data gave SS(cell line) ≈ 1,228, SS(drug) ≈ 731, SS(dose) ≈ 12, and this was used to derive a scalar "drug/dose" ratio of approximately 98× (drug effect roughly 98 times the dose effect).

What was wrong. Type I (sequential) sums of squares are order-dependent: the variance attributed to each factor depends on the order the factors are entered into the model, so a single precise fold-ratio computed from them is not a valid or reproducible summary of relative factor importance. The originally reported sequential sums of squares are internally consistent (they were reproduced exactly on re-run) but are the wrong basis for a "drug vs. dose" comparison.

Recomputation. The partition was re-fit as a marginal (Type II, no interactions) model and then as a fully crossed (Type III) model with drug×dose and cell-line×dose interaction terms, sum-to-zero contrasts, and 1,283 parameters.

Final corrected result. Under the Type III model, cell line and drug remain the dominant factors (partial η^2 = 0.660 and 0.525 respectively, versus 0.018 for dose). However, the drug×dose interaction sum of squares (≈343) is ~29-fold the dose main-effect sum of squares (≈12), and the cell-line×dose interaction (≈19) is also non-negligible relative to the dose main effect. This means dose's effect on the stress axis is highly drug-specific, and the originally reported "drug/dose ≈ 98×" ratio is not statistically defensible once this interaction is accounted for. The manuscript now reports only the partial-η^2 comparison from the Type III model (§3.4) and does not restate a single "drug/dose" ratio.

3. Permutation-null calibration of core-gene calls: a scoped check versus the actual reporting stringency

Original check. An initial calibration check of the core-gene-calling thresholds (z ≥ 2 "loose" and z ≥ 3 "medium"/"strict") was run on a scoped sample of 6 drugs, calling `drug_core_periphery()` without explicitly specifying `effect_min`/`z_thr`. This silently used the function's **loose** default (effect ≥ 0.5, z ≥ 2.0) — a materially weaker threshold than the manuscript's actual headline reporting stringency (medium: effect ≥ 1.0, z ≥ 3.0).

What a reviewer found insufficient. A reviewer correctly judged that a 6-drug scoped check at the loose (non-headline) threshold could not certify the calibration of the headline context-robustness ranking, which is reported at the medium stringency across all 277 context-robustness drugs.

Recomputation. A full-library calibration check was run across all 277 context-robustness drugs (5 independent null replicates per drug) at the manuscript's actual medium reporting stringency, alongside the original loose-threshold check for comparison.

Findings, both retained. - At the **loose** threshold (the setting of the original 6-drug check), the null distribution is measurably non-Gaussian and right-skewed (skew ≈ 0.9–1.5; Kolmogorov–Smirnov/Shapiro–Wilk reject normality at $p < 0.001$), with z ≥ 2 type-I error running ~1.9× nominal and implied maximum FDR reaching ~14–45% for robust drugs and >100% of the reported core-gene count for labile drugs (e.g., Simotinib, Hydroxyfasudil). This is a genuine

property of the loose threshold and is retained as a caveat on any analysis using it (§3.8). - At the manuscript's actual **medium** stringency, the full 277-drug recalibration is materially better-calibrated: 253 of 277 drugs (91%) have zero empirical false-core genes across 5 null replicates; the worst case (Belzutifan) has an implied maximum FDR of 2.5%; and the null-adjusted, residualized ranking correlates with the reported `core_frac_resid` ranking at Spearman ρ = 0.99 (n = 277). Bootstrap resampling (200 cell-line resamples) confirms the extreme illustrative drugs used in Figs 1–3 keep their positions and non-overlapping confidence intervals.

Conclusion. The manuscript's headline medium-stringency context-robustness ranking (Figs 2, 3, 5, 6) is treated as well-calibrated. The loose-threshold calibration concern applies specifically to analyses run at that looser, non-headline setting, and individual core-gene *lists* at the loose threshold should still be read as descriptive rather than FDR-controlled.

4. Dose-emergence / stress-axis pilot-to-full-scale coverage transition

Original scope. The dose-emergence classifier and stress-axis decomposition were first developed on a 22-cell-line pilot subset of Tahoe-100M, before the full 50-cell-line release was used for the context-robustness axis and target-depth analyses.

What changed. Both analyses were re-run on the full 50-cell-line release once it was adopted as the primary scale for all three analyses (context-robustness, dose-emergence, stress-axis), so that all reported numbers in the manuscript are drawn from a single, consistent 50-line pseudobulk collapse.

Comparison of scales. The per-drug low-dose-emergent fraction agrees closely between the 22-line pilot and the full 50 lines (Spearman ρ = 0.64, n = 379; median 0.43 at 50 lines vs. 0.44 at 22 lines pilot). The two scales agree qualitatively on the dose-emergence classifier but differ on the stress-axis variance partition, where the larger cell-line panel changes the ranking of variance components (see Correction 2 above; §3.4). The manuscript reports the 50-line values as primary throughout and notes the pilot value only where it differs materially.

5. Enrichment background gene-count reconciliation (not an error — documented for clarity)

Two different background gene counts appear in the Methods for different steps of the enrichment pipeline: 12,310 baseMean-filtered expressed genes (§2.1, the genome-wide expression filter) and 11,151 Tahoe-expressed-gene symbols (§2.7, the subset of those 12,310 genes that maps into the core/periphery label table used for enrichment testing). These are both-correct denominators for different pipeline steps, not an inconsistency, but are recorded here because the two counts were flagged during review as needing an explicit reconciliation note (retained in Methods §2.7).

6. Coverage mismatch in the axis-independence test (267/268-drug denominator)

Issue identified. The near-independence test between context-robustness and dose-emergence (§3.5) draws on 268 drugs classifiable on both axes, but per-drug cell-line coverage in this set is uneven (ranging 17–50 lines, median 49); only 114 of the 268 drugs have full 50-line coverage. An

initial version of this analysis did not explicitly check whether sparse-coverage drugs were driving the apparent near-independence result.

Recomputation. Coverage-matched subsets were computed: the full 268-drug set, a ≥48-line subset (n = 197), and the full-coverage 114-drug subset.

Final result. All three coverage strata give the same qualitative reading (frac_low-vs-robustness ρ near zero and non-significant at every threshold: –0.03 at n=268, –0.03 at n=197, –0.15 at n=114). Sparse-coverage drugs are not manufacturing the apparent independence; this coverage-sensitivity check is now reported directly in §3.5 rather than as a separate correction.

Summary of superseded values

Quantity	Originally reported	Final corrected
Coherence–depth effect (Mann–Whitney)	p = 0.002, n = 168	p = 0.165 (residuals), not significant, n = 77
Coherence–depth ΔR^2	1.3%	1.1% (not significant)
Drugs with ≥1 BH-significant pathway	324 (reported)	132
Depth-labelled enrichable cohort	168	77
Stress-axis "drug/dose" ratio	~98× (Type I, order-dependent)	not defensible as a scalar ratio; partial η^2 = 0.53–0.66 (cell line/drug) vs. 0.02 (dose), Type III
Core-gene calibration (loose threshold)	not separately calibrated	type-I error ~1.9× nominal, max FDR ~14–45% (robust), >100% (labile)
Core-gene calibration (medium/ headline threshold)	not separately calibrated	91% of drugs with zero empirical false-core genes, max FDR 2.5%, ρ = 0.99 vs. reported ranking

Data and code

All recomputed tables referenced above (`enrichment_metrics_corrected.csv`, `tahoe_stress_axis_variance_partition_corrected.csv`, `calibration_type1_error_check.csv`, `calibration_qq_zscore_normality_check.csv`, `calibration_bootstrap_core_stability.csv`, `calibration_check_summary.csv`) are archived alongside the manuscript's other derived tables and analysis code on Zenodo: https://doi.org/10.5281/zenodo.21449974 (concept DOI, always resolves to the latest version; current version is v2, https://doi.org/10.5281/zenodo.21531168).