أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

In Silico Stress-Testing of Transcriptomic Biomarkers for Deployable Tuberculosis Triage

Yasir Alnakhli, Hugo Tavares, Andries van Tonder

Alnakhli.Yasir@gmail.com, hm533@cam.ac.uk, ajv37@cam.ac.uk

Abstract

Blood host-response signatures can distinguish active tuberculosis from latent infection, but they are discovered in well-resourced laboratories and intended for deployment on low-cost RT-qPCR cartridges. We developed a selection rule that chooses gene panels for worst-case performance under simulated measurement failure — assay noise, probe dropout, coarse quantisation and narrow dynamic range — rather than for peak accuracy under ideal conditions, and then applied the same stress-testing to the selection rule itself. The rule did not survive. Bootstrapping over subjects rather than over cross-validation folds, no pair of panel sizes is distinguishable at $n = 75$; resolving a difference of 0.02 in sensitivity would require between 5,599 and 38,766 subjects. Under eight random seeds the same script returned eight different panel sizes. We therefore take panel size from the device requirement instead: a binomial redundancy argument gives $k = 5$ at a 15% well-failure rate.

A second panel, pre-registered and selected for maximal reproducibility across 400 training sets, transferred significantly *worse* than the original five-gene panel (0.903 versus 0.917, paired DeLong $p = 0.0059$, $n = 180$), with two more genes. The mechanism is measurable: reproducible selection concentrates on mutually correlated genes, and the five-gene panel spans more independent directions (1.43) than the seven-gene one (1.25). An exhaustive deterministic search over all 324,632 five-gene panels returns the same answer on every run, but 293 of the top 300 candidates are statistically indistinguishable. The stress-testing also found four defects in our own pipeline, none of them visible in its output, one of which the original result depended upon. All are reported.

Date: August 23, 2026

Code: https://doi.org/10.5281/zenodo.22020024

Contents

1 Introduction **2**

1.1 The deployment gap . . . 2

1.2 The hypothesis, and what happened to it . . . 2

1.3 Contributions . . . 2

2 Related Work **3**

2.1 Host-response signatures for tuberculosis . . . 3

2.2 Selection stability and cross-validation . . . 3

2.3 Measurement degradation and deployability . . . 3

3 Methodology **3**

3.1 Data . . . 4

3.2 Nested cross-validation . . . 4

3.3 Degradation model . . . 4

3.4 The original selection rule . . . 4

3.5 Stress-testing the rule . . . 5

3.6 Panel size from redundancy . . . 5

3.7 Effective dimension . . . 5

3.8 Pre-registration . . . 5

4 Results **5**

4.1 The original result . . . 6

4.2 The selection rule is not reproducible . . . 6

4.3 The uncertainty was the wrong uncertainty . . . 6

4.4 Panel size is not identifiable at $n = 75$ . . . 6

4.5 Panel size from the device . . . 6

4.6 Reproducible selection does not transfer better . . . 7

4.7 Mechanism: spread, not recurrence . . . 7

4.8 Deterministic design, and the equivalence set . . . 7

4.9 An irreducible ceiling . . . 7

4.10 Defects found in our own pipeline . . . 8

5 Discussion **8**

5.1 Why the rule failed . . . 8

5.2 Reproducibility is not transferability . . . 8

5.3 Determinism is not identifiability . . . 9

5.4 On reporting the defects . . . 9

6 Limitations and Future Work **9**

6.1 Limitations . . . 9

6.2 Future work . . . 10

7 Conclusion **10**

References **11**

1 Introduction

Tuberculosis remains the leading infectious cause of death worldwide. In 2024, 10.7 million people fell ill, 1.23 million died, and 2.4 million were never diagnosed or reported [1]. The diagnostic bottleneck is not detecting exposure but distinguishing *active disease*, which is transmissible and lethal untreated, from *latent infection*, which affects roughly a quarter of the world's population and in most cases never progresses.

Sputum-based tests are the standard of care but fail precisely where the burden is heaviest: in children who cannot produce sputum, in HIV co-infection, and in extrapulmonary disease. Blood host-response signatures sidestep the sampling problem by measuring the patient's immune response rather than hunting the pathogen, and several such signatures have been published [2–5].

1.1 The deployment gap

These signatures are discovered on high-quality RNA sequencing platforms, in well-resourced laboratories, and are usually evaluated on the data in which they were found. Neither condition holds where tuberculosis is actually diagnosed. The intended deployment platform is a low-cost RT-qPCR cartridge — the format already used by Xpert MTB/RIF — which introduces assay noise, probe and well failure, coarse readout quantisation and a narrow dynamic range, in populations the signature has never seen. The World Health Organization's target product profile for a triage test asks for $\geq 90\%$ sensitivity at $\geq 70\%$ specificity [6]; the asymmetry is deliberate, since a false positive costs one confirmatory cartridge while a false negative costs a life and continued transmission.

None of these degradation modes appears in a conventional discovery pipeline.

1.2 The hypothesis, and what happened to it

The project set out to test a single idea: that a gene panel should be selected for its *worst-case* performance under realistic measurement failure, not for its peak accuracy under ideal conditions. Concretely, the rule was to choose the smallest panel whose worst-case degraded sensitivity clears the WHO target by at least one standard error.

Applying that rule produced a five-gene panel, designated Core-5, which validated at 0.879 AUROC on an independent Indian cohort and 0.917 on a Malawian and South African cohort. Both cohorts had been sealed before the panel and its decision threshold were frozen.

We then applied the same stress-testing logic to the selection rule itself. This report documents what that revealed, in three parts. First, that panel size is not identifiable from a discovery-scale cohort, so the rule was reporting noise rather than signal. Second, that a panel selected for maximal reproducibility transferred worse than the original, contrary to a prediction recorded in writing before the sealed cohorts were opened. Third, that a fully deterministic redesign returns the same panel on every run but cannot distinguish it from hundreds of alternatives.

1.3 Contributions

- An identifiability result for panel-size selection, with a quantified sample size requirement: separating adjacent panel sizes to a resolution of 0.02 in sensitivity would require between 5,599 and 38,766 subjects.
- A demonstration, with a pre-registered and failed prediction, that selection stability and cross-cohort transfer are decoupled on a signature dominated by a single expression module.

- A device-driven alternative for panel size based on a closed-form redundancy argument, containing no random component.
- A deterministic panel-design procedure with closed-form degradation, together with the finding that the resulting choice is one of many statistically equivalent options.
- A record of four defects that the stress-testing found in our own pipeline, none of which were visible in its output.

2 Related Work

2.1 Host-response signatures for tuberculosis

Transcriptomic signatures separating active tuberculosis from latent infection have been reported for over a decade. Berry et al. described a 393-gene interferon-inducible whole-blood signature in UK and South African cohorts [2], and the same cohorts remain a standard discovery resource. Kaforou et al. derived smaller signatures in Malawian and South African patients, including HIV-positive individuals [3]. Zak et al. identified a 16-gene signature predictive of progression to active disease in adolescents [5].

Of particular relevance here is the three-gene signature of Sweeney et al. (*GBP5*, *DUSP3*, *KLF2*), which was explicitly designed for portability and has been ported to cartridge formats [4]. It is notable for spanning two directions — two genes up-regulated and one down — where most published signatures, including both panels reported here, move in a single direction.

A recurring observation across this literature is that signatures replicate poorly against one another, and that reported performance drops when a signature is applied outside its discovery data. The identifiability result in this report offers one mechanical explanation: at discovery-cohort sample sizes, the differences between candidate panels are frequently smaller than the uncertainty in estimating them.

2.2 Selection stability and cross-validation

Stability selection [7] selects features by how reproducibly they are chosen across resampled subsets, and provides an upper bound on the expected number of falsely selected variables. Complementary pairs stability selection [8] refines the bound. Both are widely recommended in high-dimensional biomarker settings, and the second panel in this report was constructed accordingly. Our finding that this panel transferred worse than a decorrelated alternative is not a criticism of the method in general, but an observation about its behaviour when the candidate features lie on a single dominant correlated module.

The importance of keeping feature selection inside the cross-validation loop is long established [9, 10]. The pipeline described here follows that discipline for differential expression, normalisation, scaling and gene ranking. It does not, however, place the choice of *panel size* inside the loop, and Section 6 discusses the consequences.

2.3 Measurement degradation and deployability

Work on making genomic classifiers robust to platform change generally addresses cross-platform normalisation or rank-based scoring. Explicit simulation of instrument-level failure modes — probe dropout, quantisation, dynamic-range compression — as a *selection* criterion rather than a post-hoc robustness check is less common, and is the approach this project set out to evaluate.

3 Methodology

3.1 Data

Four publicly available cohorts were used (Table 1). The two Berry cohorts formed the discovery set ($n = 75$). The two external cohorts were downloaded and gene-matched, then left untouched until each panel and its decision threshold had been frozen; neither informed any filter, normalisation constant, differential expression test, gene ranking, panel size or threshold. The GSE37250 analysis uses the HIV-negative subset ($n = 180$).

Cohort	Accession	Platform	ATB	LTBI	Role
Berry — London	GSE107991	RNA-seq	17	16	Discovery
Berry — South Africa	GSE107992	RNA-seq	16	26	Discovery
India	GSE101705	RNA-seq	28	16	Sealed
Kaforou — Malawi / SA	GSE37250	Microarray	97	83	Sealed

Table 1 Cohorts. ATB, active tuberculosis; LTBI, latent infection.

3.2 Nested cross-validation

Five stratified folds, repeated ten times. Every label-aware step occurs on the 60 training samples of each fold: gene filtering (`filterByExpr`), TMM normalisation, edgeR quasi-likelihood differential expression with design ~ cohort + group [11], per-cohort scaler fitting, elastic-net stability selection (50 subsamples of 36 samples, producing a ranking), and ridge fitting for $k = 1, \ldots, 15$ via `cv.glmnet` with $\alpha = 0$ [12]. The 15 held-out samples are projected onto the training genes, scaled with the training scaler, and scored both clean and degraded. Nothing flows backwards.

The nested cross-validation deliberately produces no model that survives: it estimates how the *procedure* behaves and counts which genes recur. The deployed classifier is fitted once afterwards on all 75 discovery samples using the frozen panel.

3.3 Degradation model

Four failure modes are simulated, each applied separately across a severity grid (Table 2). Noise and dropout are stochastic, drawn ten times per fold and averaged; quantisation and compression are deterministic given the fold. Dropout imputes failed wells to the training gene mean.

Stress	Severity grid	Operating point	Stochastic
Assay noise (CV)	$0 \rightarrow 0.50$	0.10	yes
Probe / well dropout	$0 \rightarrow 30\%$	15%	yes
Quantisation (levels)	$256 \rightarrow 4$	16	no
Compression (usable logs)	$8 \rightarrow 2$	4	no

Table 2 Simulated instrument failure modes.

3.4 The original selection rule

Panel size k was chosen as the smallest value whose worst-case degraded sensitivity — the minimum across the four stresses at their operating points, computed per repeat and then averaged — clears the WHO target of 0.90 by at least one standard error. Panel composition was determined by walking down the cross-fold selection frequency ranking, skipping any gene

correlated above $|r| = 0.95$ with one already chosen. This produced Core-5: *GBP1*, *EPSTI1*, *STAT2*, *GBP6*, *FCGR1A*.

3.5 Stress-testing the rule

Three analyses were applied to the rule itself.

Subject-level bootstrap.. The pipeline's reported standard error is the spread across cross-validation repeats divided by $\sqrt{10}$. Repeats re-partition the same 75 subjects and therefore carry no information about sampling uncertainty. We instead resampled subjects with replacement (2,000 draws, stratified by class), recomputing the worst case from stored out-of-fold predictions, and formed paired differences between panel sizes on identical resampled subjects.

Seed sweep.. The full nested section was re-run under eight random seeds, with all other settings fixed, recording the selected panel size and the resulting panel.

Deterministic redesign.. For a linear score $s = \mathbf{w}'\mathbf{x}$, each stress reduces to a diagonal inflation of the within-class covariance, so separation has a closed form:

$$D = \sqrt{\boldsymbol{\delta}'(\Sigma + \Lambda)^{-1}\boldsymbol{\delta}}, \qquad \mathbf{w} \propto (\Sigma + \Lambda)^{-1}\boldsymbol{\delta}, \tag{1}$$

where $\boldsymbol{\delta}$ is the between-class mean difference, Σ the pooled within-class covariance, and Λ the stress-specific diagonal term: σ^2/s_j^2 for noise, $\frac{p}{1-p}(\Sigma_{jj} + \mu_j^2)$ for dropout, and $R_j^2/(12L^2)$ for quantisation. No folds, no stress draws and no cross-validated fitting are required, so the procedure contains no randomness.

3.6 Panel size from redundancy

Where the empirical curve cannot resolve k, panel size is instead taken from the device requirement. If each well fails independently with probability p and the classifier requires at least m functioning genes, the survival probability is a binomial tail, $P(X \geq m)$ with $X \sim \mathrm{Bin}(k, 1-p)$. This contains no seed, no fold and no fitted quantity.

3.7 Effective dimension

Panel redundancy is quantified by the exponentiated Shannon entropy of the correlation eigenvalue spectrum,

$$D_{\mathrm{eff}} = \exp\left(-\sum_i p_i \ln p_i\right), \qquad p_i = \lambda_i \Big/ \textstyle\sum_j \lambda_j, \tag{2}$$

which returns the number of genes when they are mutually independent and 1 when they measure the same thing. A weighted variant substitutes the covariance of the weighted contributions $w_j x_j$, so that a gene with negligible weight contributes a negligible direction.

3.8 Pre-registration

The second panel, Stable-7, and four numerical predictions were written down and countersigned before the sealed cohorts were opened for the second time. The panel was derived from discovery data only, but — as disclosed in Section 6 — after the first validation result was known.

4 Results

4.1 The original result

Core-5 (*GBP1*, *EPSTI1*, *STAT2*, *GBP6*, *FCGR1A*) achieved 0.879 AUROC on the Indian cohort and 0.917 on the Malawian and South African cohort. All five genes are interferon-stimulated: *GBP1* and *GBP6* are guanylate-binding proteins induced by IFN-γ, *STAT2* transduces type-I interferon signalling, and *FCGR1A* is the high-affinity IgG receptor on activated monocytes. *GBP5* and *STAT1* were excluded despite higher selection frequency, as near-duplicates of *GBP1* ($r = 0.99$ and 0.98).

4.2 The selection rule is not reproducible

Run under eight random seeds with all other settings fixed, the same script returned $k \in \{6, 9, 11, 11, 11, 12, 12, 14\}$. The rule returns the position of the first crossing of a threshold by a noisy curve; because that position is an integer index, a small perturbation in any single point relocates the answer by several genes.

4.3 The uncertainty was the wrong uncertainty

Sensitivity in the discovery set rests on approximately 33 active cases, giving a binomial standard error near $\sqrt{0.9 \times 0.1/33} = 0.052$. An independent bootstrap estimate agrees: the 95% interval at $k = 5$ spans 0.727 to 0.952, implying a standard error of 0.057. The pipeline's reported standard error, being the spread across repeats of the same 75 subjects, was roughly an order of magnitude smaller. The rule was therefore thresholding at a resolution the data cannot support.

Restoring subject-level uncertainty, $P(\text{worst-case sensitivity} \geq 0.90)$ peaks at 0.673 raw and 0.420 after isotonic smoothing. The admissible set — the panel sizes whose lower confidence bound clears the target — is empty. We state this claim carefully: it is not that no panel can meet the WHO target, but that these data, under this degradation model, cannot demonstrate that any size does.

4.4 Panel size is not identifiable at $n = 75$

Table 3 gives paired bootstrap comparisons between panel sizes. Every interval crosses zero.

Comparison	Difference	95% CI		n required
$k = 3$ vs $k = 5$	−0.152	−0.424	+0.003	34,231
$k = 5$ vs $k = 7$	+0.011	−0.091	+0.121	8,437
$k = 5$ vs $k = 10$	−0.010	−0.112	+0.061	5,599
$k = 7$ vs $k = 15$	+0.078	−0.109	+0.346	38,766

Table 3 Paired bootstrap over subjects. The final column is the sample size required to resolve a difference of 0.02 in worst-case sensitivity.

The only substantial point estimate is $k = 3$ versus $k = 5$, consistent with the redundancy model below: three genes survive a 15% well-failure rate intact only 61% of the time.

4.5 Panel size from the device

With $p = 0.15$ and $m = 3$ functioning genes required, the survival probability is 0.614 at $k = 3$, 0.890 at $k = 4$, **0.973** at $k = 5$, 0.994 at $k = 6$ and 0.999 at $k = 7$. Five genes is therefore the smallest panel surviving intact 95% of the time, in agreement with the frozen panel size and obtained without reference to any performance curve.

4.6 Reproducible selection does not transfer better

Stable-7 (*GBP1*, *GBP5*, *STAT1*, *EPSTI1*, *RTP4*, *SERPING1*, *BATF2*) comprises the genes selected in at least 80% of 400 training sets (8 seeds × 10 repeats × 5 folds). With $q = 10.75$ average selected features and $p = 730$ candidates, the Meinshausen–Bühlmann bound gives an expected 0.26 falsely selected genes at that threshold.

Cohort	n	Core-5	Stable-7	paired DeLong
India	44	0.879	0.902	$p = 0.096$
Malawi / South Africa	180	**0.917**	0.903	**p = 0.0059**

Table 4 Head-to-head validation. Both panels frozen before unsealing; scored on identical subjects with a single probe mapping.

On the better-powered cohort the less reproducible panel wins, with two fewer genes, contradicting the recorded prediction. The result is robust to the handling of below-detection array values: $p = 0.0059$ with flooring ($n = 180$) and $p = 0.0077$ with dropping ($n = 166$). The two cohorts do not conflict — India at $n = 44$ cannot separate the panels, which is the identifiability finding recurring.

4.7 Mechanism: spread, not recurrence

Panel	Genes	Unweighted D_{eff}	Weighted	Least active gene's share
Core-5	5	1.43	1.42	0.145
Stable-7	7	1.25	1.26	0.089

Table 5 Effective dimension. The weighted variant confirms neither panel relies on a near-inert gene.

Core-5 spans more independent directions with two fewer genes, and the advantage survives weighting by each gene's contribution to the score. The mechanism is straightforward: selection frequency rewards genes that rank together, and on a correlated module the genes that rank together are correlated, so stability selection concentrates within a single direction. Applying the correlation cap at fixed panel size raises reproducibility but lowers spread — mean pairwise Jaccard between seeds falls from 0.94 (uncapped) to 0.65 (capped) at $k = 7$.

4.8 Deterministic design, and the equivalence set

Replacing Monte Carlo degradation with the closed forms of Equation 1 and searching exhaustively over all $\binom{35}{5} = 324{,}632$ five-gene panels drawn from the top candidates by univariate d' returns the same panel on every run, with no seed set anywhere. Verified by re-execution.

However, under a subject bootstrap, 293 of the top 300 candidates are statistically indistinguishable from the best. Only *EPSTI1* and *STAT2* appear in more than half of the equivalent panels (0.91 each); no other gene exceeds 0.49.

4.9 An irreducible ceiling

Five of 28 confirmed active cases in the Indian cohort mount no detectable response, with mean panel z of −0.84 against +0.70 for the remainder at identical sequencing depth. Host-response signatures detect the response, not the pathogen; this is a limit of the approach rather than of any particular panel.

4.10 Defects found in our own pipeline

Four defects were identified by tracing reported values back to the objects that produced them. None was visible in the pipeline's output; each produced plausible, correctly formatted numbers.

1. **A failure mode absent from its own grid.** The dropout operating point (0.15) was not a member of the dropout severity grid, so the lookup assembling the worst case returned no dropout rows and the minimum was taken over three stresses rather than four. Dropout is the binding stress. With it restored, worst-case sensitivity at $k = 5$ is 0.808, not ≥ 0.90.
2. **A checkpoint overwritten without provenance.** A later run at $k = 15$ replaced the stored results, so comparisons ran against a fifteen-gene panel labelled as the five-gene one.
3. **A processing rule excluding only controls.** Negative background-subtracted array intensities became `NaN` under $\log_2(x + 1)$, removing 14 latent controls and zero active cases.
4. **A figure disagreeing with its own caption.** A scatter plot displayed all 44 subjects beneath a caption reading 28.

5 Discussion

5.1 Why the rule failed

The original rule failed for a reason that is structural rather than numerical, and which generalises beyond this dataset.

First, it compared a curve against the wrong uncertainty. Cross-validation repeats re-partition the same subjects, so their spread measures Monte Carlo variability, not sampling variability. Averaging over repeats shrinks the former and leaves the latter untouched. Because the reported standard error was roughly an order of magnitude smaller than the true sampling error, the rule was reading structure that the data does not contain.

Second, a first-crossing threshold is a discontinuous functional. The output is an integer index, so it cannot move by a small amount: a perturbation of 0.02 in one point relocates the selected size by several genes. Such estimators admit no useful variance bound, and cannot be repaired by estimating the curve more precisely. They must be replaced.

Third, the flatness of the curve is partly biological. The candidate space is dominated by a single interferon module, so panels of different sizes carry nearly identical information. Additional subjects would measure that flatness more precisely rather than remove it.

5.2 Reproducibility is not transferability

The clearest finding of the project is that a panel selected for maximal reproducibility transferred significantly worse than one selected with an explicit decorrelation constraint, and did so with two more genes.

The mechanism is measurable rather than speculative. Selection frequency is a popularity statistic: it rewards features that are chosen together across resamples. On a correlated module, features chosen together are features correlated with each other. Maximising reproducibility therefore maximises redundancy, and redundancy reduces the number of independent directions a panel spans. The correlation cap that produced Core-5 buys that spread at a measured cost in reproducibility.

This does not invalidate stability selection, which remains well founded and carries a false-discovery bound we relied upon. It identifies a regime — a single dominant correlated module — in which its objective and the deployment objective diverge. Given that most published

tuberculosis signatures are interferon-driven, that regime is likely to be common rather than exotic.

5.3 Determinism is not identifiability

The deterministic redesign achieves what it set out to: the same panel on every run, with no seed anywhere in the selection path. It does not, however, make the choice meaningful. Under a subject bootstrap, 293 of 300 candidate panels are statistically indistinguishable, and only two genes appear in a majority of them.

For a device team this is arguably useful rather than disappointing. If the statistics cannot separate these panels, choosing among them on manufacturability — probe design, melting-temperature matching, transcript abundance, freedom to operate — costs nothing in accuracy. The appropriate deliverable is a characterised menu with an explicit tie-breaking rule, not a claimed optimum.

5.4 On reporting the defects

Four defects in our own pipeline are reported in Section 4. Two observations follow.

Each produced numbers that were plausible, correctly formatted and within the expected range. None triggered a warning. They were found by tracing values back to the objects that generated them rather than trusting the fields that labelled them. The most consequential — a stress silently absent from its own severity grid — was the foundation of the original headline result.

That is the project's argument applied to itself. A pipeline that looks solid until stressed is exactly what the method was built to detect, and this pipeline was not an exception. Publishing the defects is therefore consistent with the claim rather than an embarrassment alongside it.

6 Limitations and Future Work

6.1 Limitations

Two panels from one discovery set.. Stable-7 was derived from discovery data only, so its validation estimate is unbiased for that panel. The decision to derive a second panel, however, was made after Core-5's validation result was known. Two attempts on the sealed cohorts are disclosed here; a third was deliberately not made.

The mechanism rests on a single comparison.. The link between effective dimension and transfer is supported by one pair of panels on one cohort. It is offered as a falsifiable prediction — that at equal panel size, panels spanning more independent directions transfer better — not as an established result.

Panel size selection is not nested.. Differential expression, scaling and gene ranking occur inside the training fold, but k is chosen by inspecting out-of-fold performance across all candidate sizes. Reported performance at the selected k is therefore optimistic. Moving the choice into an inner loop is described below.

Eight seeds is few. for the stability estimates, and the subsets used for the invariance check comprised as few as two runs each.

Bootstrap intervals rest on one repeat.. The stored out-of-fold predictions are re-initialised each repeat, so the subject bootstrap uses a complete prediction set from a single repeat rather than an

average over ten. This is conservative: the intervals shown are somewhat wider than a ten-repeat average would give, and the conclusion does not depend on it.

Below-detection handling was changed.. Negative array intensities are floored rather than dropped, departing from the transform that produced the originally reported figure. Both are reported and they agree.

The redundancy model assumes. that three functioning genes are required. That is a claim about the classifier, not a measurement.

Discovery-equivalent is not deployment-equivalent.. The equivalence set is measured on discovery data. Core-5 and Stable-7 were both plausible on 75 subjects and differed significantly on 180.

6.2 Future work

Score dropout with the failure mask available.. The degradation model imputes failed wells to the training mean and never informs the classifier which wells failed. A real cartridge knows: absence of amplification is an observable event. The redundancy model predicts that five genes survive a 15% failure rate 97% of the time, yet measured sensitivity falls substantially — a discrepancy that reconciles only if the damage arises from scoring with corrupted inputs rather than from losing genes. Rescoring with renormalisation over surviving weights is inexpensive and would test this directly. It is the single highest-value open experiment.

Move panel-size selection inside the fold.. Choosing k by inner cross-validation on training data only would make the reported performance unbiased for the whole procedure and would turn the stability of k into a measured output, obtainable from one run rather than a seed sweep.

Reconsider candidate-pool construction.. Ranking candidates by univariate d' selects for membership in the dominant module, which is the opposite of the independence the objective rewards. Achievable effective dimension rose from 1.49 to 1.76 as the pool widened from 25 to 50 genes while best separation stayed fixed, indicating that the truncation rather than the objective was the binding constraint. Ranking on residual d' after projecting out the first principal component reached more independent genes, but the resulting panel was unstable across plausible values of an undeclared threshold and is not reported as a result.

More cohorts rather than more subjects.. Panel size requires thousands of subjects to resolve, but panel *composition* was resolvable at $n = 180$. The variance that matters for deployment is between populations and platforms, which a single larger cohort cannot expose.

Test the dimension hypothesis prospectively. , at equal panel size, on a cohort not yet used.

7 Conclusion

We set out to build a selection rule that chooses transcriptomic biomarker panels for their behaviour on low-cost diagnostic hardware rather than for peak accuracy in a laboratory, and then stress-tested that rule with the same tools it applies to panels.

The rule does not work as specified. Panel size is not identifiable from a discovery-scale cohort: no two sizes are distinguishable at $n = 75$, and separating adjacent sizes would require between 5,599 and 38,766 subjects. Size should therefore be taken from the device requirement, where a binomial redundancy argument gives five genes at a 15% well-failure rate — with no seed, no fold and no fitted quantity.

A second panel, pre-registered and selected for maximal reproducibility, transferred significantly worse than the first, contrary to a prediction recorded in writing. The mechanism is measurable: reproducible selection concentrates on mutually correlated genes, and the five-gene panel spans more independent directions than the seven-gene one. Reproducibility of feature selection and generalisation are decoupled on a signature dominated by a single expression module, and in this case they trade off.

A deterministic redesign returns the same panel on every run but cannot distinguish it from hundreds of alternatives, which reframes the deliverable: a characterised menu of equivalent panels with an explicit tie-breaking rule is more honest, and more useful to a device team, than a claimed optimum.

Finally, the method found four defects in the pipeline that implemented it, none of them visible in its own output, one of which the original headline result depended upon. That is the project's argument demonstrated on itself, and it is reported in full rather than repaired quietly.

The contribution is therefore not a new tuberculosis signature. It is an identifiability result with a quantified sample-size requirement, a measured counterexample to standard advice on stability selection, and a device-driven alternative for panel size — together with an audit trail that allows every claim to be checked against the code that produced it.

References

[1] World Health Organization. *Global Tuberculosis Report 2025*. Tech. rep. Reporting 2024 data. Geneva: World Health Organization, 2025.

[2] Matthew P. R. Berry, Christine M. Graham, Finlay W. McNab, et al. "An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis". In: *Nature* 466 (2010), pp. 973–977.

[3] Myrsini Kaforou, Victoria J. Wright, Tolu Oni, et al. "Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study". In: *PLoS Medicine* 10.10 (2013), e1001538.

[4] Timothy E. Sweeney et al. "Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis". In: *The Lancet Respiratory Medicine* 4.3 (2016), pp. 213–224.

[5] Daniel E. Zak, Adam Penn-Nicholson, Thomas J. Scriba, et al. "A blood RNA signature for tuberculosis disease risk: a prospective cohort study". In: *The Lancet* 387 (2016), pp. 2312–2322.

[6] World Health Organization. *High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting*. Tech. rep. Geneva: World Health Organization, 2014.

[7] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B* 72.4 (2010), pp. 417–473.

[8] Rajen D. Shah and Richard J. Samworth. "Variable selection with error control: another look at stability selection". In: *Journal of the Royal Statistical Society: Series B* 75.1 (2013), pp. 55–80.

[9] Christophe Ambroise and Geoffrey J. McLachlan. "Selection bias in gene extraction on the basis of microarray gene-expression data". In: *Proceedings of the National Academy of Sciences* 99.10 (2002), pp. 6562–6566.

[10] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7 (2006), p. 91.

[11] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22.