Prediction Errors in Human Mitochondrial tRNAs Are Structure-Dependent: MitoTIP Over-Calls in Base-Paired Stems and Under-Calls in Unpaired Loops

Thikra M. Almuqbil
KAUST Academy, Bioinformatics program, University of Cambridge
Mentor: Hugo Tavares, Department of Genetics, University of Cambridge

Table of Contents

Abstract 2
1 Introduction 2
1.1 Mitochondrial tRNAs as a source of human disease 3
1.2 An untested assumption in current practice 3
1.3 Aims and approach 4
1.4 Scope and boundary conditions 4
2 Methodology 4
2.1 Data sources 4
2.2 Extraction 5
2.3 Region definitions 7
2.4 The audited variant set and a defect in the benchmark 8
2.5 Independent validation sets 9
2.6 Statistics 9
2.7 Derivation of the region-specific thresholds 9
2.8 Reproducibility 10
3 Results 10
3.1 Validation of the structural atlas 10
3.2 Distribution of disease variants across structural regions 10
3.3 Structural dependence of misclassification 11
3.4 Robustness of the regional finding 13
3.5 Recomputation of the published performance figures 13
3.6 Confirmation on independent variant sets 14
3.7 Candidate threshold rules and held-out performance 14
3.8 Derivation of the base-paired threshold from the benign distribution 16

3.9 External validation and the specificity cost 17
3.10 The false-negative to false-positive trade 18
3.11 Syndrome-level distribution of recovery 18
3.12 Label quality among base-paired false positives 20
3.13 Candidate discriminative features 20
3.14 Comparison with an independently implemented predictor 22
4 Discussion 22
4.1 Mechanism of the structural asymmetry 22
4.2 Anticipated and unanticipated findings 23
4.3 Scope of the contribution 23
4.4 Implications for the evaluation of variant-effect predictors 24
5 Limitations 24
6 Future Work 25
7 Conclusion 26
References 26

Abstract

Pathogenic mtDNA variants are enriched in mitochondrial tRNA genes, and MitoTIP is the most frequently used in silico predictor when a novel variant occurs in one of these genes. For every possible single-base substitution in all 22 genes, MITOMAP gives a MitoTIP score, using a single pathogenicity threshold for the entire molecule. However, a tRNA is not homogeneous: base-paired stems and unpaired loops are subject to different structural constraints. I constructed a position-specific structural atlas of all 22 human mt-tRNAs by means of the printed Mamit-tRNA numbering tables and applied it to study the context-dependence of MitoTIP misclassification. It does. In the tool's own labeled set, base-paired variants are mis-scored in 23% of cases, unpaired variants in 1.6% of cases, a 14-fold difference (Fisher exact p=9×10^(-18)), and the two error types are not mixed: every stem error is a false positive and every loop error a false negative. This is about calibration, not discrimination. Fitting region-specific thresholds to the training table alone increases detection of 29 held-out, ClinGen-confirmed pathogenic variants from 75.9% to 89.7 %, at a measured cost of unpaired specificity falling from 100% to 94 % against population-frequent variants. The correction is a recalibration, not a new model; it shifts the blind spot of the diagnosis; it does not eliminate it.

1 Introduction

1.1 Mitochondrial tRNAs as a source of human disease

The human mitochondrial genome encodes 22 transfer RNAs, and they are an outsized source of disease. They make up only 5-10% of the mtDNA genome but account for about 70-75% of disease-causing mtDNA variants (Petropoulou et al., 2025). Consequently, the clinical picture is wide: one and the same gene family is responsible for MELAS, MERRF, Leigh syndrome, Kearns–Sayre syndrome, non-syndromic sensorineural hearing loss, cardiomyopathy and tubulointerstitial kidney disease. Thus, a variant of uncertain significance in one of these genes is a common and important finding, and the question of what to do with it is answered in practice by an in silico predictor.

That predictor is MitoTIP (Sonney et al., 2017). MITOMAP gives a MitoTIP score for each single-base substitution possible at each position in all 22 genes (Lott et al., 2013), and the clinician does not so much run the tool as look up the answer. Published scores are accompanied by quartile assignments and the top two quartiles are read as pathogenic – i.e. one raw-score cut-off, applied identically from the acceptor stem to the T-loop.

1.2 An untested assumption in current practice

This project questions that uniform treatment. A tRNA is not a string without features. It folds into a cloverleaf and then into an L-shape and the positions within it do structurally different jobs: base-paired stem positions hold the fold together through Watson–Crick pairing, while unpaired loop positions carry the recognition and modification chemistry that lets the molecule be charged, read and processed (Salinas-Giegé et al., 2015). A substitution that disrupts a helix and a substitution that changes a loop nucleotide are not the same kind of event, and there is no a priori reason to expect a predictor to be equally reliable in both.

MitoTIP is not itself blind to structure; its score has a secondary-structure component, and this is exactly why the question is worth asking carefully rather than treating any structural signal as a discovery. What hadn't been tested, as far as I could see, was if the errors of the tool are evenly spread across the molecule. That is a different question than whether the score uses structure, and it is the question a clinician's confidence rests upon.

There is an independent predictor, PON-mt-tRNA (Niroula & Vihinen, 2016) which gives pathogenicity rates for each structural region. It is used here as a second opinion, but with a caveat stated up front and repeated in Section 3.14: it shares the same structural prior, and so cannot independently corroborate a structural claim about MitoTIP.

1.3 Aims and approach

The work is two-level. First, a structural atlas of all 22 human mt-tRNAs at the position level, derived by parsing the printed Mamit-tRNA numbering tables (Pütz et al., 2007) such that each mtDNA coordinate has a structural region label taken from a published source rather than inferred by arithmetic. The second uses that atlas as an analytical lens on MitoTIP: where its errors sit, whether that distribution is an artifact of how the analysis set was assembled, and whether a region-aware threshold does better on variants the tool was never fitted on.

What is claimed and what is not claimed is worth mentioning exactly. The point is not that MitoTIP is a bad predictor. It is that a single threshold applied across a structurally heterogeneous molecule is the wrong instrument for it, that the errors that result are systematic and directional rather than random, and that a region-specific threshold measurably reduces missed pathogenic variants on independent data at a cost that is measured and reported rather than assumed away.

1.4 Scope and boundary conditions

Two boundaries are placed before rather than after analysis, and both are found in each of the tables that follow.

First, the analysis is restricted to the non-anticodon body of the molecule. We exclude the three codon-reading nucleotides at canonical positions 34–36, because a substitution there changes which codon is read—a different failure mode from destabilizing a fold, and one that would confound a structural comparison. The four flanking positions of the anticodon loop (32, 33, 37, 38) are retained, as they do not read the codon and are similar to other unpaired positions that bear modifications.

Second, and a consequence of the design of MitoTIP's own benchmark, each performance figure computed on that benchmark is a training-set number. MitoTIP's Table S1, 'Variants used in optimization (Supplementary)' Section 3.5 discusses why this weakens a naive performance claim, but actually strengthens, not weakens, the regional finding.

2 Methodology

2.1 Data sources

Table 1 lists everything the analysis reads. Three of the four primary sources publish their tables as print layout, prose or pixels rather than as machine-readable files, so a substantial part of the work was extraction (Section 2.2).

Table 1. *Data sources and what each contributes. Accession date for all MITOMAP material was 11 August 2026; the GenBank frequency set shipped with the MitoTIP score file comprises 66 801 sequences collected 15 July 2026.*

Source	Contribution	Form
Mamit-tRNA numbering tables (Pütz et al., 2007)	Structural region label for every position of all 22 genes	Two PDFs, no text table
MitoTIP supplementary S1–S4 (Sonney et al., 2017)	689 labelled variants used to optimise the tool; its published thresholds; its own documented error cases	Word tables, print layout
MITOMAP MitoTIP score table (Lott et al., 2013)	6 016 scored variants: three substitutions and one single-base deletion per position, plus GenBank frequency	Tab-separated
MITOMAP confirmed / reported RNA mutations	Held-out ClinGen-confirmed set; a weaker "reported" tier; allele context for repair	CSV export
Petropoulou *et al.* 2025	331 disease variants across 26 named syndromes	Running prose; no supplementary data
PON-mt-tRNA S6, S7 (Niroula & Vihinen, 2016)	Region-level pathogenicity rates from an independent predictor	Supplementary tables

The Mamit-tRNA PDFs were retrieved through the Internet Archive (snapshots 2019-11-13 and 2021-06-22) because the original host was not reachable while this work was done. That is a genuine reproducibility caveat and the snapshot dates are part of the provenance, not a convenience.

2.2 Extraction

2.2.1 The structural atlas

The two Mamit-tRNA numbering PDFs defeat `camelot`, `tabula` and `pdfplumber`'s own table extractor, for two independent reasons. There are no ruling lines around cells, what looks like a table is digits positioned over coloured background rectangles, and the structural-region labels are drawn as 45°-rotated text, one glyph at a time, whose bounding boxes interleave vertically with the data rows and destroy line grouping.

The parser therefore works at the character level, using `pdfplumber` (Singer-Vine & The pdfplumber contributors, 2024) to read each glyph's page coordinates and text matrix:

1. A glyph is rotated if and only if its text matrix carries a non-zero shear term ($|m_1| > 0.01$; the tolerance avoids floating-point equality). Upright glyphs are data, rotated glyphs are labels.

2. Rotated labels are reassembled along their own diagonal. Text at $-45°$ advances equally in $+x$ and $-y$, so $x + y$ is nearly invariant within one label and differs sharply between labels. Chaining on x alone fails, because all twelve labels begin at the same x.

3. Regions are assigned by two-dimensional geometry: wide coloured rectangles are stem bands, the unshaded gaps between them are loops, and each label is assigned to the block containing its vertical centre. Narrow rectangles spanning a single gene column are gene-specific overrides and are checked *first*; merging them into the wide bands mislabels a row across every gene.

4. Genes are assigned by coordinate-range membership rather than column index. Column position only bootstraps a per-gene [min, max], after which every number is reassigned by asking which range contains it. This is direction-agnostic, so the eight light-strand genes whose coordinates run antiparallel to tRNA numbering need no special handling, and it self-checks: a value outside every range is flagged.

The output is 1508 gene-position records across 22 genes, covering 1504 distinct mtDNA positions, exactly the position set MitoTIP scores, with no difference in either direction.

2.2.2 The phenotype axis

Petropoulou *et al.* publish no supplementary data files; variants are named in running prose inside disease sections. The phenotype label is therefore *the section a variant is named in*, and the paper's own numbered heading hierarchy supplies the taxonomy.

Prose mining is normally fragile. It is defensible here because a deliberately permissive pattern (`[ACGT]\d{3,5}[ACGT]`, with no word boundaries, since the PDF text layer has lost its spaces) is paired with two independent, data-driven checks. The legacy notation carries the reference allele, which must equal the rCRS base at that coordinate (Andrews et al., 1999) or its complement; and the prose names a gene, which the atlas assigns independently from the coordinate. A token failing either check is dropped and named in the log rather than guessed at. Of the reference-base checks, 458 matched directly, and 14 matched as complements. All 14 in light-strand genes, which the atlas predicts independently, with none unexplained.

This recovered 331 distinct variants from 472 variant–phenotype mentions, of which 464 came from a disease section; 317 of the 331 (95.8%) fall inside the analysis set.

2.2.3 Optical character recognition of the review figures

The review also displays variants inside cloverleaf figures, which are 300–375 dpi raster images with no text layer. Those were rendered and read with Tesseract (Smith, 2007), but the purpose was methodological rather than acquisitive: I wanted to know whether MITOMAP's tables already contained what the figures show, because if they did, the atlas should be built from the machine-readable source and the figures need never be read again.

They do: 283 of 307 recovered figure variants (92.2%) are already present in MITOMAP once strand orientation is normalised. Measured against a hand-verified benchmark of all 24 labels on two panels, OCR recall was 17/24 (71%), with one invented token that both safety checks flagged. **No analysis script reads the OCR output**, so OCR error cannot propagate into any reported number; the file is retained as the evidence for the decision.

One error in this comparison is worth recording, because accepting it would have reversed the conclusion. An earlier version matched the figure notation against MITOMAP literally and reported 62.5% coverage, which made every light-strand variant appear absent: the figures draw each tRNA on its own strand, so light-strand labels are complements of the reference sequence. Normalising orientation raises coverage to 92.2%. The uncorrected figure would have supported the opposite reading — that the figures constitute an independent source worth reading systematically — and it is reported here rather than silently repaired.

2.3 Region definitions

All region assignment is by table lookup, never by arithmetic on mtDNA coordinates. The justification is empirical: domain lengths vary between genes (the acceptor stem is the only constant at 14 nt, while the D-stem ranges from 0 to 8), mt-tRNA-Ser(AGY) lacks a D-arm entirely, eight genes are transcribed from the light strand, and canonical numbering is non-consecutive with lettered insertions and unnumbered loop positions. Hard-coded position ranges would mislabel.

Definitions live in a single JSON configuration read by both the Python and R code, so the two languages cannot drift. The primary set, `d_body_excl_anticodon_triplet`, contains **1420** of the 1508 positions. Three alternative definitions are computed in the same pass so that the sensitivity analysis cannot be forgotten or run under only one definition.

The one sanctioned use of position numbers is the anticodon triplet at conventional positions 34–36. It is defensible because those positions carry a *printed* conventional number in the source table, and because the pipeline independently verifies that positions 34–36 spell the expected anticodon in all 22 genes and halts if any disagrees. That check is what makes the exception safe.

2.4 The audited variant set and a defect in the benchmark

MitoTIP's S1 table contains 689 labelled variants: 38 pathogenic and 651 benign, counts the extraction asserts before writing anything. Reading it required column-wise parsing — the table places one pathogenic column beside nine benign columns, so a row-wise read pairs each pathogenic variant with nine unrelated benign ones.

Sixty-five of the benign entries are malformed: the allele given is identical to the rCRS reference base, so the entry describes no substitution and cannot be keyed to a score. A notation audit found that the strings `del`, `ins`, `:` and `-` occur zero times in the table, so the notation cannot express an indel at all; there are correspondingly no deletions in either class, which is a consequence of the notation rather than a scientific exclusion.

Two mechanisms for the defect were tested against their own base rates: a typesetting twin, where the reference letter was printed instead of the allele (55.4% of the malformed entries against a 28.7% base rate, odds ratio 3.09, $p = 3 \times 10^{-5}$), and a flattened deletion the notation could not carry (29.2% against 2.2%, odds ratio 18.3, $p = 2.4 \times 10^{-12}$). Both are supported, and they do not partition the set — eight entries carry both signals — so for those the variant *class*, not merely the allele, is unrecoverable.

Recovery therefore used one symmetric rule with no preference between variant classes: take MITOMAP's recorded changes at the position, drop what MitoTIP cannot score, drop any allele already present as a well-formed entry, and see what survives. This resolved 18 of the 65 — ten as deletions, eight as substitutions. Nine remain contested, where both readings survive, and are left unassigned. **Forty-seven of the 65 stay excluded even in the most permissive arm**: the defect is mitigated, not solved.

Because the recovery is a judgement, four analysis arms are carried through every downstream calculation (Table 2). `SD_both` is the primary arm; `A_conservative` assumes nothing and is reported alongside.

Table 2. *The four analysis arms. Sensitivity is identical across all four because no recovered entry is pathogenic; the arms differ only in how many benign variants they admit.*

Arm	TP	FN	FP	TN	Sens. (%)	Spec. (%)
`A_conservative` (624 variants)	33	5	67	520	86.8	88.6
`S_substitution` (632)	33	5	69	526	86.8	88.4
`D_deletion` (634)	33	5	71	526	86.8	88.1
`SD_both` (642) — **primary**	33	5	73	532	86.8	87.9

Applying the region filter to the primary arm leaves **622** variants inside the analysis set (317 base-paired, 305 unpaired), which is the denominator for every regional rate reported below. Variants are always counted as distinct (position,allele) pairs

rather than as join rows, because four mtDNA positions belong to two overlapping genes and correctly appear twice.

2.5 Independent validation sets

Two sets were assembled that MitoTIP was not fitted on.

Held-out pathogenic. MITOMAP's confirmed RNA mutations carry ClinGen classifications assigned after MitoTIP was published. Those absent from S1 form a held-out set of **29** testable pathogenic variants. Being pathogenic-only, this set can measure sensitivity and nothing else: with no negatives in it, no false positive is even definable. A second, larger tier of **251** MITOMAP-"reported" variants is used as a weaker-labelled check.

Population-tolerated. The negative class comes from the GenBank counts shipped with the score file: **101** variants carried by at least 50 apparently healthy complete mitochondrial genomes. This is a proxy for benign, not a curated benign call, so the frequency cut was varied (≥20, ≥50, ≥100) to check that nothing depends on where the line falls. Pooling the two gives one external set of 130 variants in which both error types are measurable.

2.6 Statistics

Regional error rates are compared with Fisher's exact test, two-sided. Proportions carry exact Clopper–Pearson 95% intervals (Clopper & Pearson, 1934), chosen because several regional counts are small and normal-approximation intervals would be wrong there. Discrimination is summarised by the area under the ROC curve, computed as the Mann–Whitney U statistic with tie-corrected variance, which is threshold-free and therefore sidesteps the quartile-versus-cutoff question entirely. Robustness of the regional result was checked by leave-one-gene-out (22 refits), across the four region definitions, and across the four analysis arms.

Predicted folding free-energy changes were computed with ViennaRNA (Lorenz et al., 2011): $\Delta\Delta G = \Delta G_{\text{mutant}} - \Delta G_{\text{wild-type}}$, so a positive value means the substitution weakens the fold.

2.7 Derivation of the region-specific thresholds

Two thresholds were derived by two *different* rules, because the two score distributions have different shapes. This asymmetry is deliberate and is the methodological core of the proposal, so Section 3.8 reports the evidence for it rather than asserting it here.

- **Unpaired regions.** Within the training set the two classes are completely separated: the highest-scoring benign loop variant is at 9.9526 and the lowest-scoring pathogenic one at 11.2541, with nothing in between. Every cutoff inside that empty band gives identical training performance, so no optimisation criterion can choose among them. The threshold was set at the

top edge of the band, min(pathogenic) $= \mathbf{11.25}$, which is the most conservative point in it.

- **Base-paired regions.** Here the classes overlap — the highest benign score, 16.4194, sits above the lowest pathogenic score, 16.0901 — so no separating cutoff exists. The threshold was therefore taken from the *benign distribution alone*, as quantile(benign, 0.90) $= \mathbf{13.989}$ over 289 benign stem variants. A rule that never looks at the pathogenic variants cannot be overfitted to the 28 of them that exist.

Both thresholds were fitted on the training table only and then applied, unchanged, to the validation sets of Section 2.5.

2.8 Reproducibility

The pipeline is 17 numbered R scripts plus two Python helpers and one figure script, run by a single command (`Rscript run_all.R`) under R 4.3.3 (R Core Team, 2024) with `dplyr` (Wickham et al., 2023), `tidyr`, `jsonlite` and `ggplot2` (Wickham, 2016). A complete run from the raw files takes about one minute on a laptop, single-threaded.

Figures never compute statistics: the figure script reads only the result tables, so no number can exist in a figure without existing in a CSV first. Every script writes a plain-text log naming each variant it excluded and why. Code, data, the extraction pipeline and the exploratory analyses are public at https://github.com/Thikra-Almuqbil/mt-tRNA-mitotip.

3 Results

3.1 Validation of the structural atlas

The parser yielded 1508 gene-position records in 22 genes. It was subjected to three checks before anything was built.

The strongest is extrinsic. The gene coordinate ranges from the PDF geometry alone were compared with the annotated loci of the revised Cambridge reference sequence (Andrews et al., 1999): all 22 are exact matches. This isn't an internal consistency check -- this is two independent sources agreeing. The parser never saw those loci. Named anchors behave as expected - m.3243 lands in tRNA-Leu(UUR) at the D-loop, m.8344 in tRNA-Lys at the T-loop. A cross-check against the canonical Sprinzl numbering disagrees at only 2 of 1,302 positions, both in Ser(AGY) and both explained by gene-specific shading in the source PDF, which I confirmed by cropping the page and inspecting it.

3.2 Distribution of disease variants across structural regions

It is useful to establish where the disease is before asking where the predictor fails. Figure 1 shows the 331 review-derived variants against structural region and

syndrome.

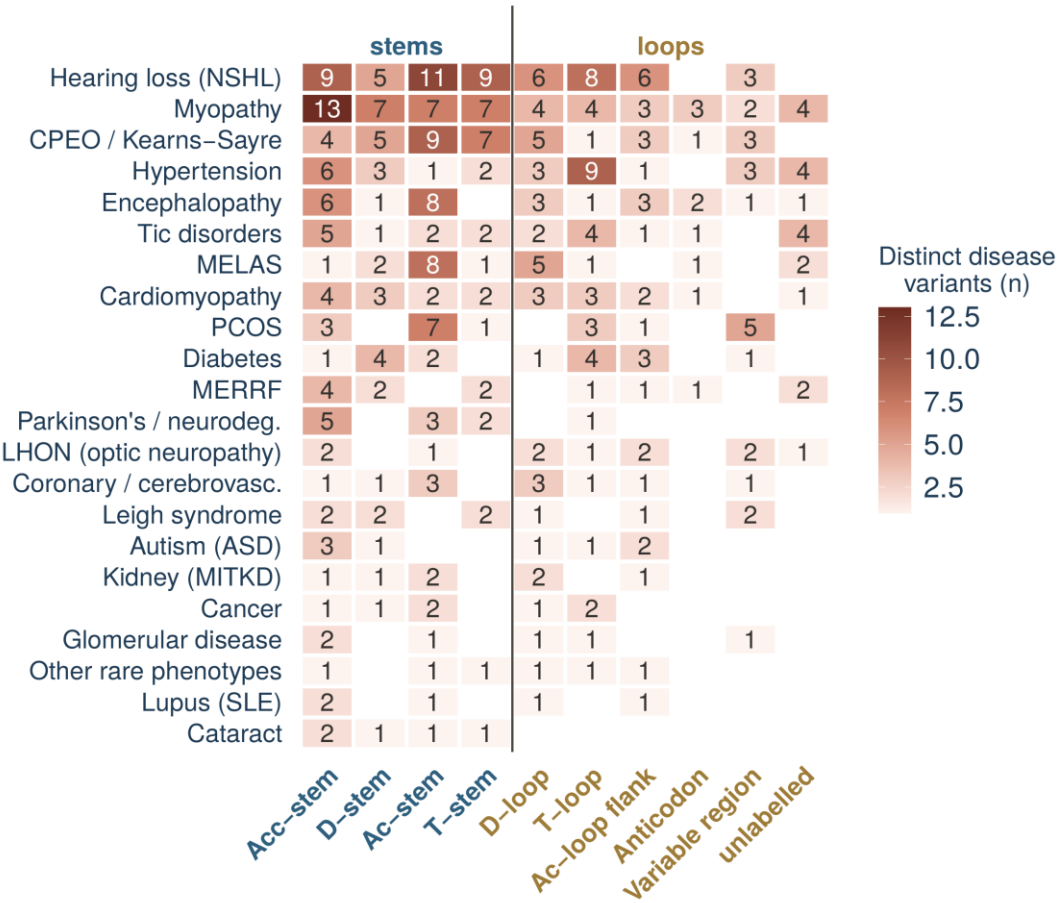


Figure 1: *Disease variants of different mt-tRNAs per structural region and syndrome (n=331; Petropoulou et al. 2025). The horizontal axis is also the stem/loop key and the region labels are colored by structural family. Samples variants were seen in both stems and loops in all syndromes. There are variants of both families for every syndrome. So the structural position tells you how a variant breaks the tRNA, not which disease follows -- and this matters for interpreting everything after this, because it means structural asymmetry reported below is a property of the predictor and not of the disease biology.*

3.3 Structural dependence of misclassification

On the 622 auditable variants inside the analysis set, misclassification depends strongly on pairing status (Figure 2, Table 3).

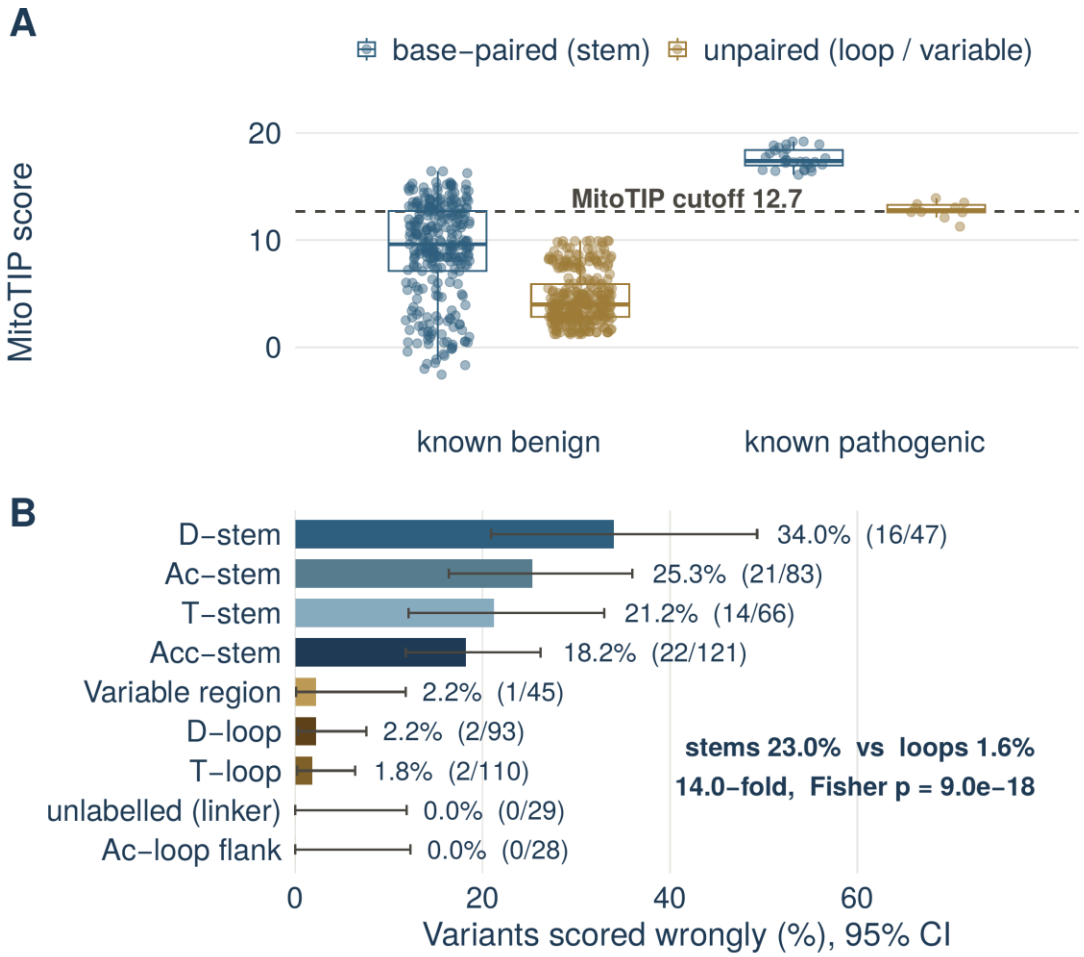


Figure 2: ***(A)*** *MitoTIP score distribution by class label and pairing status, with the tool's own cutoff (12.7) as a dashed line. Unpaired variants sit systematically lower than base-paired ones in both classes.* ***(B)*** *Misclassification rate per structural region with exact Clopper–Pearson 95% confidence intervals; counts are errors over variants tested. The stem intervals and the loop intervals do not overlap.*

Table 3. *Misclassification by structural region, primary arm. FP, false positive; FN, false negative. Note the final column: the direction of error is absolute.*

Region	Pairing	Tested	Errors	Rate (%)	FP/FN
D-stem	paired	47	16	34.0	16 / 0
Ac-stem	paired	83	21	25.3	21 / 0
T-stem	paired	66	14	21.2	14 / 0
Acc-stem	paired	121	22	18.2	22 / 0
Variable region	unpaired	45	1	2.2	0 / 1
D-loop	unpaired	93	2	2.2	0 / 2
T-loop	unpaired	110	2	1.8	0 / 2
Ac-loop flank	unpaired	28	0	0.0	0 / 0

Region	Pairing	Tested	Errors	Rate (%)	FP/FN
unlabelled	unpaired	29	0	0.0	0 / 0
All stems	paired	317	73	**23.0**	73 / 0
All loops	unpaired	305	5	**1.6**	0 / 5

23% vs 1.6% misclassification of base-paired positions vs unpaired positions: 14-fold difference, Fisher exact two-sided p=9x10^(-18).

This is the more informative half of the result, the other being the absolute direction of error. MitoTIP over-calls where the molecule is base-paired and under-calls where the molecule is not. There are no exceptions to the 73 false positives and five false negatives, respectively. A predictor with random errors would not produce a clean double dissociation of this sort. This is the basis for the calibration interpretation developed in Section 4.
Here normalization is important. A denominator is variants tested per region; raw error counts would tell us how many variants a region happens to contribute, not anything about the predictor.

3.4 Robustness of the regional finding

Three ways it could be, each tested.

Leave-one-gene-out. Refitting with each of the 22 genes removed in turn gives rate ratios from 12.5 to 17.8, with a worst-case $p = 1.4 \times 10^{-15}$. No single gene carries the result.

Region definition. Across the four definitions the ratio runs 12.3 to 15. Redrawing the category boundaries moves the number without threatening the conclusion.

The recovery judgement. Across the four arms of Table 2 the ratio runs 12.9, 13.4, 13.5 and 14. The most important row is `A_conservative`, which admits *no* recovered entry at all and still gives 12.9-fold. The finding does not depend on the repair step being right.

3.5 Recomputation of the published performance figures

The only quantitative performance claim made in the MitoTIP paper is a sensitivity and specificity of 74% at a single score cutoff. Recalculating at the same cut-offs on the present scoring file gives 86.8% and 88.9% sensitivity and specificity at 12.8 versus published 74.4% and 74%. The training-set AUC is 0.971. The quartile rule used elsewhere in this project corresponds with a raw cutoff at 12.668, so the comparison is like-for-like on the threshold.

These figures should not be interpreted as a validation of MitoTIP and the difference between them and the published values is not evidence of an improved tool. The S1 table is labeled “Variants used in optimization”, so any figure based on it — the publication’s and mine — is a training-set figure. Furthermore, the published values are obtained by a leave-one-out procedure while the scores distributed today are

final fitted scores, a difference enough to justify a discrepancy of this magnitude. The recomputation is reported for a narrow purpose: to establish that the pipeline correctly processes the tool's own data. It does, and the strongest evidence of this is that it reproduces all five of MitoTIP's published false negatives from raw downloads, independently.

It is worth stating plainly that the same caveat cuts the other way for the regional finding. The incorrectly scored stem variants are the variants the model was trained on. If there is a systematic error on points in the training data, it is a limitation of the feature set rather than poor generalization. Thus the fact that the benchmark is based on training data actually strengthens Section 3.3 rather than weakens it.

3.6 Confirmation on independent variant sets

Two independent tiers were tested (Section 2.5).

On the 29 held-out, ClinGen-confirmed pathogenic variants, MitoTIP misses **6 of 7** unpaired variants (85.7%) against **1 of 22** base-paired ones (4.5%) — a 19-fold difference, $p = 9.9 \times 10^{-5}$. On the larger and weaker "reported" tier of 251 variants the same asymmetry appears: 97 of 102 unpaired variants missed (95.1%) against 32 of 149 base-paired (21.5%), 4.4-fold, $p = 2.3 \times 10^{-34}$.

Overall held-out detection under the current single threshold is 75.9% (22/29). Missed diagnoses are where under-calling occurs in loops.

3.7 Candidate threshold rules and held-out performance

Figure 3 tests five threshold rules against the held-out set.

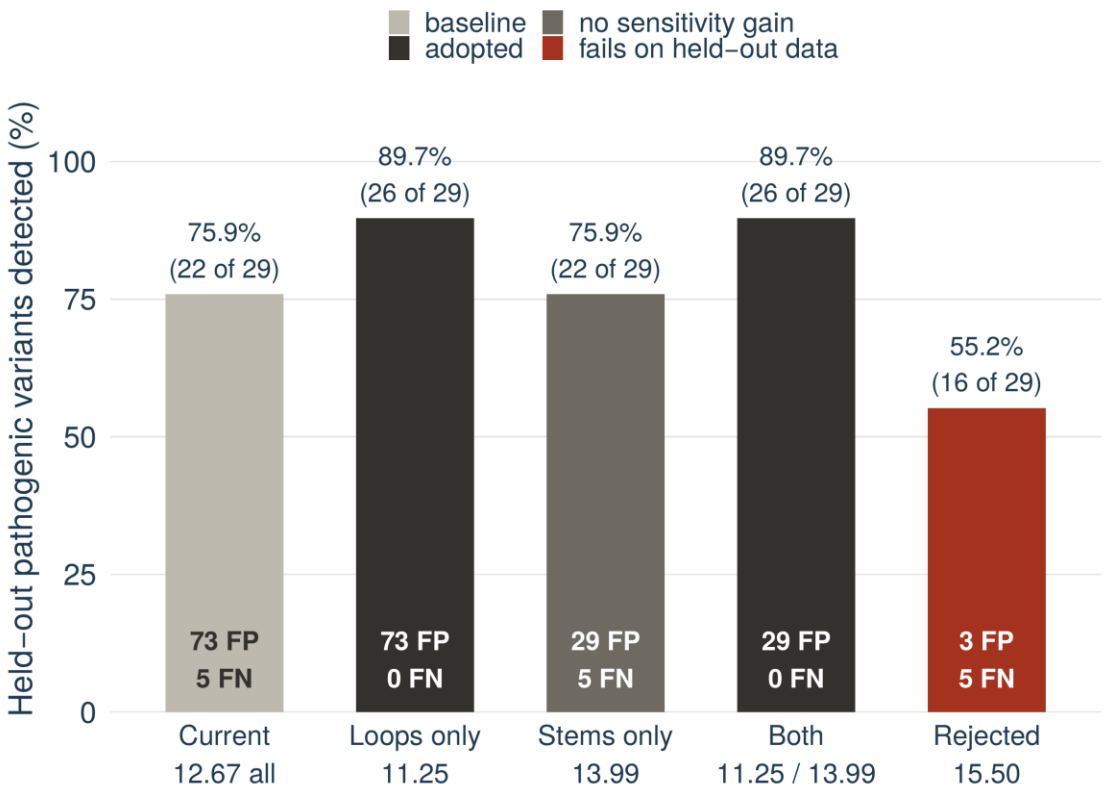


Figure 3: *Detection of 29 held-out ClinGen-confirmed pathogenic variants under five threshold rules. Counts inside each bar are errors on the 622-variant training set (FP, false positive; FN, false negative). "Stems only" is a counterfactual included to show what the base-paired change does and does not buy; "Rejected" is the training-set optimum for stems.*

- **Lowering the unpaired threshold to 11.25** raises held-out detection from 75.9% to **89.7%** (26 of 29) and costs nothing on the training set: false negatives fall from 5 to 0 and no false positive is added.
- **Raising the base-paired threshold to 13.989 adds no sensitivity at all.** Held-out detection stays at 75.9%. What it buys is specificity: training false positives fall from 73 to 29, and training specificity rises from 87.5% to 95%.
- **The two changes are independent**, acting on disjoint sets of variants, so applying both gives 89.7% with 29 false positives. Reported as a single step this would invite the reader to assume both knobs contributed to the 89.7%. Only one did.
- **The training-set optimum for stems, 15.5, was tested and rejected.** It looks excellent in-sample and collapses to **55.2%** (16 of 29) held out — worse than doing nothing. Pushed further to 16.09 it reaches 48.3%.

The rejected rule is plotted rather than discarded because it is the clearest available evidence that in-sample optimisation fails on the base-paired side.

3.8 Derivation of the base-paired threshold from the benign distribution

Of these, the one most likely to be challenged is this design choice, so the reasoning is set out in full.

The lowest-scoring pathogenic stem variant in the training data is at 16.0901. So, stem sensitivity is 100% for all cutoffs up to 16.09. Specificity increases monotonically over the same range. Any criterion that rewards a good sensitivity-specificity trade (Youden's J, maximum accuracy, F_1, minimum false positives) therefore reduces to "maximise specificity" and is at the top of the sweep, where held-out sensitivity is 48.3%.

The consequence is general rather than specific to any particular criterion: no in-sample criterion can choose a defensible base-paired threshold on these data. This is a property of the score distribution. From this the printed threshold sweep follows and requires no argument. This is why the base-paired threshold is calculated only from the benign distribution, as a rule that never sees the 28 pathogenic base-paired variants can never be overfit to them.

The opposite problem is with the unpaired side. At that point the classes are completely separated by a band 1.30 wide, so that every criterion ties across the whole band and none can separate the points inside it. The choice of the upper edge of the band is a conservatism argument, not an optimization result – and the sweep confirms it is not knife-edge: thresholds 11.25, 11.00, 10.50 and 10.00 all give the same 89.7%. Detection only reaches 93.1% at 9.50 and there it starts to add false positives for training.

3.9 External validation and the specificity cost

Since the held-out set is pathogenic-only, there is no way to measure a false positive. The pooled external set of 130 variants (29 ClinGen confirmed pathogenic, 101 population tolerated) can.

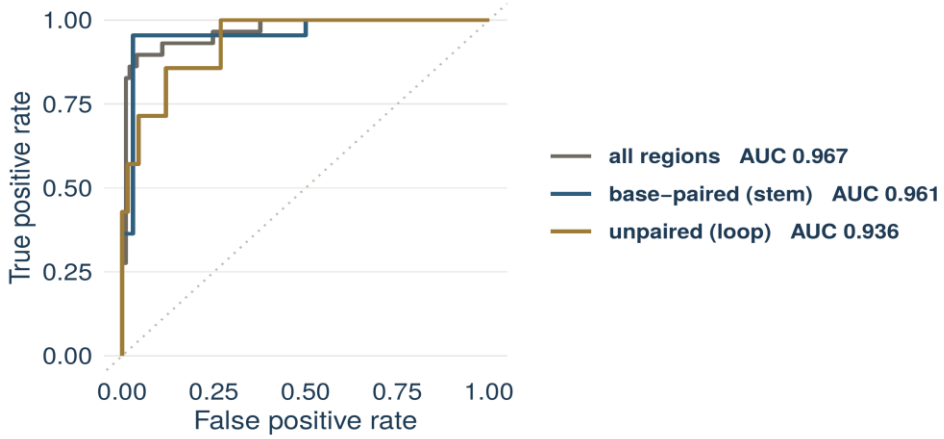


Figure 4: *Receiver operating characteristic on 130 variants MitoTIP was not fitted on: 29 ClinGen-confirmed pathogenic and 101 carried by at least 50 healthy GenBank genomes. AUC values are given for the pooled set and within each structural context.*

Discrimination on unseen data is good in *both* structural contexts: AUC 0.967 overall ($p = 2.0 \times 10^{-14}$), 0.961 base-paired, 0.936 unpaired. Varying the frequency cut that defines the tolerated class gives AUC 0.960 (≥20, $n = 157$), 0.967 (≥50, $n = 101$) and 0.971 (≥100, $n = 62$); restricting to MitoTIP-unlabelled variants only gives 0.938. Nothing turns on where the line is drawn.

Together with Section 3.3 this pinpoints the failure exactly. The score does a good job ranking unseen variants within each structural context, but a single cutoff does not do a good job across both. It is an issue of calibration and not discrimination, a single threshold across two offset distributions, which is clear directly from Figure 2A.

The price is calculated for the same set (Table 4). The unpaired specificity drops from 100% to 94% (4 false positives out of 67 tolerated unpaired variants). Base-paired specificity is not going anywhere.

Table 4. *Specificity against population-tolerated variants, before and after the correction. The unpaired column is the entire cost of the proposal.*

Rule	Pairing	False positives	Specificity (%)
Current single threshold (12.674)	unpaired	0 / 67	100.0
Current single threshold (12.674)	paired	1 / 34	97.1
Region-specific (11.25 / 13.989)	unpaired	4 / 67	94.0
Region-specific (11.25 / 13.989)	paired	1 / 34	97.1

3.10 The false-negative to false-positive trade

Having both classes in one external set exposes an uncomfortable arithmetic, and should be reported rather than buried: the correction produces eight errors before and eight errors after. One for one, four false negatives become four false positives.

There are three reasons why the trade is favorable and none of them is that the error count went down:

1. **The denominators differ.** Four missed diagnoses out of 29 pathogenic variants is a 13.8% failure rate on the class that matters clinically; four false alarms out of 101 tolerated variants is 4%.

2. **Youden's *J* rises from 74.9 to 84.7**, because sensitivity gains 13.8 points (75.9% to 89.7%) while specificity loses 4 (99% to 95%).

3. **The costs are not symmetric.** For a screening step feeding clinical review, a false positive is a variant that gets looked at; a false negative is a diagnosis that does not happen.

3.11 Syndrome-level distribution of recovery

Aggregate sensitivity does not say who benefits. Figure 5 resolves the change by syndrome across both evidence tiers.

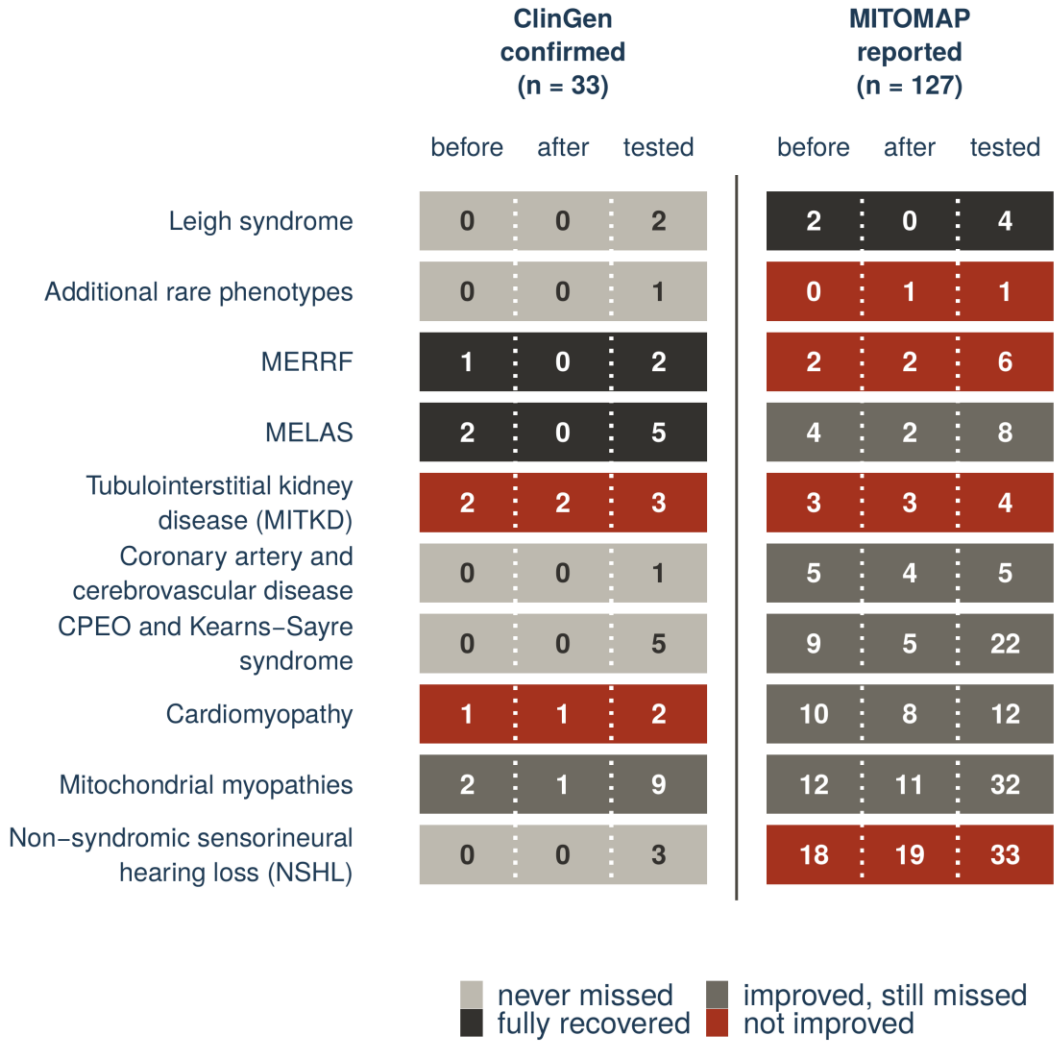


Figure 5: *Pathogenic variants missed per syndrome, before and after the correction, in each of the two independent levels of evidence. Within a tier, before is the number missed under MitoTIP's existing single threshold, after is the number missed under the region-specific thresholds, and tested is the number of variant-syndrome entries for that syndrome. Each is caused by a variant reported in multiple syndromes, so we count entries, not variants. n below each tier heading is the total for that tier. The verdict for the syndrome as a whole is determined by cell color.*

Pooling the tiers, the correction clears **Leigh syndrome** entirely (2 missed to 0, of 6 entries) and recovers most missed variants in **MELAS** (6 to 2, of 13) and **CPEO / Kearns–Sayre** (9 to 5, of 27), with smaller gains in MERRF (3 to 2), mitochondrial myopathies (14 to 12) and cardiomyopathy (11 to 9). It does **not** help tubulointerstitial kidney disease (5 to 5, of 7), and non-syndromic sensorineural hearing loss ends marginally *worse* (18 to 19, of 36).

The variants found cluster in the classical encephalomyopathies where the causal changes are located in loops. It is dominated by syndromes working via tertiary folding. What is missed. The blind spot moves, it does not close.

3.12 Label quality among base-paired false positives

An obvious alternative explanation is that the 73 stem false positives are actually pathogenic variants incorrectly recorded as benign in 2017. We tested this by asking whether they are enriched for variants that are now named in MITOMAP's disease lists: 8.2% of stem false positives are named compared to 10.6% of true negatives, p=0.66. There is no enrichment. But whatever the false positives are, they are not focused on mislabeling.

3.13 Candidate discriminative features

Both remaining error classes were probed for usable features (Figure 6).

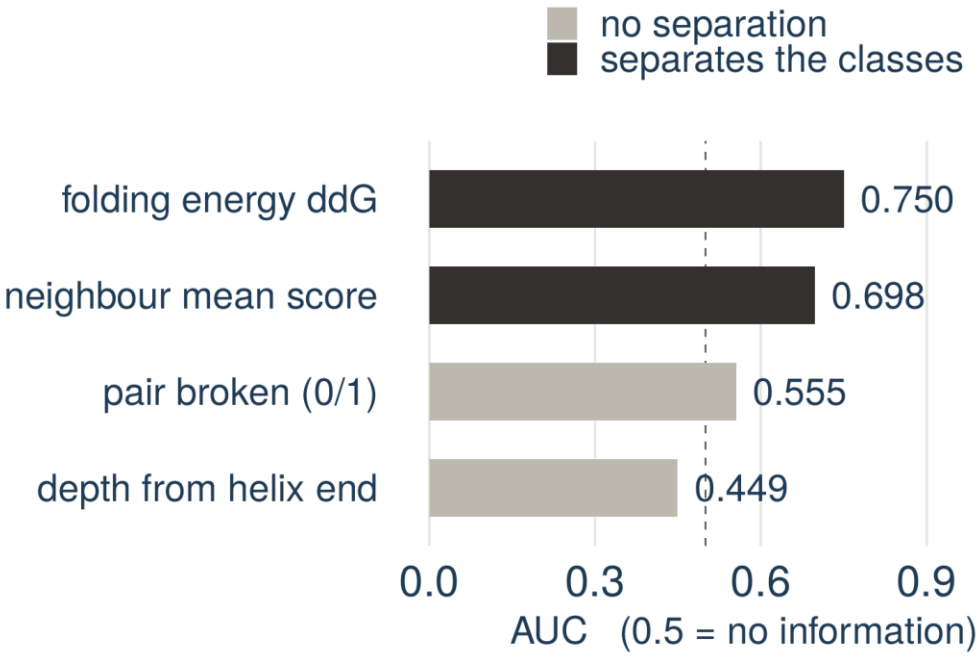


Figure 6: *Ranking of candidate features for separating pathogenic from benign base-paired variants. The dashed line indicates no information. AUC values are in 28 pathogenic vs 282 benign stem variants.*

Predicted folding destabilisation separates the classes among base-paired variants (AUC 0.750, $p = 1.2 \times 10^{-5}$), as does the mean MitoTIP score of the flanking positions (0.698, $p = 5.7 \times 10^{-4}$). Two simpler structural descriptors do not: whether the base pair is broken at all (0.555, $p = 0.26$) and depth from the end of the helix (0.449, $p = 0.35$).

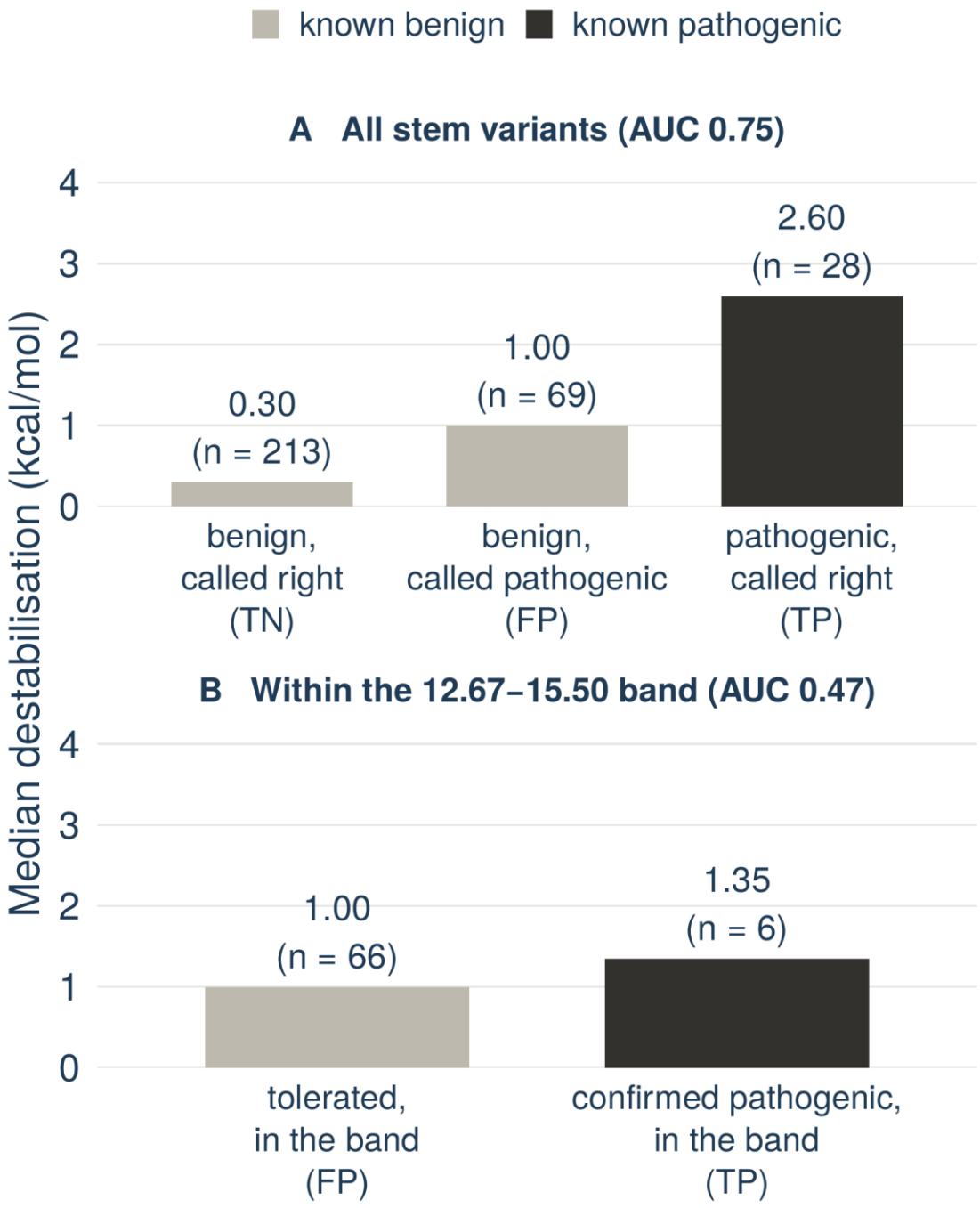


Figure 7: *Median predicted folding destabilisation by classification outcome.* ***(A)*** *All base-paired variants: destabilisation increases from correctly called benign, through false positives, to correctly called pathogenic.* ***(B)*** *Within the 12.67–15.50 score band, where the contested variants sit, the same feature carries no information.*

Figure 7 adds the part that matters for what *cannot* be done. Across all base-paired variants, median $\Delta\Delta G$ rises from 0.30 kcal/mol for correctly called benign variants ($n = 213$) through 1.00 for false positives ($n = 69$) to 2.60 for correctly called

pathogenic ones ($n = 28$). Inside the 12.67–15.50 band, however, the feature does not separate the classes at all (AUC 0.465, $p = 0.78$; 6 pathogenic against 66 tolerated). Thus destabilization is a real pathogenic mechanism in stems, but it fails to rescue borderline cases – which is the mechanistic reason why the base-paired threshold cannot simply be raised and is consistent with the held-out failure of the 15.50 rule.

The residual misses were also grouped according to the molecular function annotated to the region of each variant to disrupt; the association with whether a variant remains missed is significant (p=0.007, n=211, 20 000 permutations). Those classes are structural annotations from the literature, not experimental assignments. Section 5 says so again.

3.14 Comparison with an independently implemented predictor

PON-mt-tRNA is also an independently implemented predictor that reports pathogenicity rates by structural region (Niroula & Vihinen, 2016) and its regional pattern is in the same direction as MitoTIP's. That is not corroboration for three things:

1. It has the same prior structure. Of its twelve features, two are RNA secondary structure, so the structural claim made by one tool cannot be independently confirmed by the other.

2. Its published axis is a rate of pathogenicity, not a rate of error. These are different quantities and the comparison cannot tell us "both tools share a bias" from "PON is right".

3. It's not its headline accuracy that matters. The reported 1.00 accuracy requires inputs of segregation, biochemical and histochemical evidence. As a purely computational predictor it is 0.69 accuracy, MCC 0.39. It also has a retired web server, so the only per-variant scores available are for the 46 variants in its supplementary tables, 4 of which overlap the audited set.

Two tools independently implemented, and leaning on structure in the same direction, is the defensible sentence. This report is based on nothing.

4 Discussion

4.1 Mechanism of the structural asymmetry

The pattern in Section 3.3 is not arbitrary, and the mechanism is directly illustrated in Figure 2A: unpaired variants are systematically scored lower than base-paired variants in both classes. Thus a single cutoff drawn across two offset distributions is at the same time too high for one of them and too low for the other, and its errors are in opposite directions on either side. That is what a calibration failure looks like, and why the two types of error do not mix.

Once it is stated, the biology behind the offset is not surprising. Most stem substitutions will be scored as consequential by a predictor that considers the disruption of Watson-Crick pairing, but a helix can tolerate a single mismatch and still fold, so a large fraction of those variants are in fact benign, and are thus false positives. Conversely, pathogenic loop variants are often chemically subtle: they disrupt a modification, a recognition contact or a tertiary interaction without destabilizing the secondary structure to a measurable degree, so that a structure-weighted score reads them as acceptable. The architecture of the molecule is the bias of the tool.

This also explains why the two thresholds could not be derived in the same way (see Section 3.8). The discriminating signal is there, but the cutoff is just misplaced, so the unpaired class distributions are separated. The base-paired distributions overlap because a real subset of pathogenic stem variants is actually indistinguishable from benign ones on the features available.

4.2 Anticipated and unanticipated findings

The direction was anticipated prior to analysis of the full error distribution, from the regional rates of PON-mt-tRNA, from MITOMAP's own comments on m.8344, and from the class imbalance in the training table. The prediction was that un-paired regions would be under called. That order matters: it's what separates a prediction from a story told after the fact.

Three things had not been anticipated.

The absolute. A directional bias was plausible**;** zero exceptions in either direction across 622 variants was not. That is the most striking one number in the project**,** and it is why the calibration reading is hard to avoid.

That bought no sensitivity to the base-paired change. I thought that a **region-specific** rule would help both sides. It doesn't: all of the **held-out** gain is from the unpaired threshold and the base-paired threshold is just specificity. It would have been easy to have reported the two as one step, but that would have been misleading, which is why the stem-only counterfactual is plotted in Figure 3.

That the raw error count does not go down. 8 errors before. 8 errors after. I didn't see that coming and it was uncomfortable to find out because the headline sensitivity gain is real and the total error count is unmoved. The defense has to be made on unequal denominators and unequal costs rather than on the count and that argument is set out in the Results rather than left implicit.

4.3 Scope of the contribution

An overclaim would be to present 11.25 and 13.989 as the correct thresholds for mt-tRNA variant interpretation. The MitoTIP score is fixed, published, and deterministic, across a closed space of 6 016 possible variants. This is a better case for transportability than most calibration exercises can make. Transportability was

tested rather than assumed, with 89.7% on unseen ClinGen variants, and AUC 0.967 where both classes are new. But the labels aren't fixed. ClinGen calls more of them. The unpaired threshold is at the edge of a band defined by 10 pathogenic variants. The training optimum on the base-paired side has already been shown to collapse out of sample.

The defensible claim is therefore more modest, and more useful for that. A single threshold applied everywhere is demonstrably wrong, and region-specific thresholds are demonstrably better on data the tool was not fitted on; but the two values are a demonstration rather than a standard. What is the transport method: split the molecule by pairing status, put the unpaired cutoff at the edge of the empty band separating the classes, and get the base-paired cutoff from a benign-only quantile so it cannot be overfit to the small number of pathogenic variants available.

That's why the rules and not only the values are reported, and why the whole pipeline regenerates from raw sources with one command. Anyone adopting this should re-derive both numbers against the current label set using those rules.

4.4 Implications for the evaluation of variant-effect predictors

Two general points follow. Neither is unique to MitoTIP.

A single global threshold on a structurally heterogeneous molecule is a modeling assumption, not a neutral choice, and seldom stated as such. Where a score distribution varies systematically across structural contexts, aggregate sensitivity and specificity will hide compensating errors - as in the present case, where an overall AUC of 0.971 coexists with a 14-fold difference in error rate between two halves of the same molecule.

Second, evaluating a tool on its own optimization set is more misleading than it first appears, and in a more subtle way than simple optimism. Not only does it boost performance, it also changes the meaning of a residual error. An error on training data can't be explained by poor generalization, so it points at the feature set instead. That is why Section 3.3 is presented as a statement of the strengths and limitations of MitoTIP, rather than a performance figure.

5 Limitations

The limitations below are ordered by how much they constrain the conclusion, not by how easy they are to admit.

The benchmark is the tool's own optimization suite. MitoTIP was fit on the MitoTIP S1 table, so no performance figure computed on it represents performance on unseen variants and none is presented as such in this report (Section 3.5). The same overlap creates a partial circularity. MitoTIP already includes a secondary-structure component, so the observation that its errors follow structure partially validates its claim. Neither objection accounts for the asymmetry — the direction of

error is absolute, with zero errors of the wrong type in either class — but both are why the independent tiers of Section 2.5 carry the burden of argument rather than the training-set figures, and both constrain how surprising it should be.

The unpaired side is dependent on few pathogenic variants. The unpaired error rate is 1.6%, which corresponds to 5 errors in 305 variants, as opposed to 317 variants supporting the base-paired side. The unpaired threshold lies at the edge of a band defined by ten pathogenic loop variants; the addition of one new low-scoring pathogenic loop variant would shift it. The labeled set is imperfect at source, with 65 of its benign entries malformed and 47 unresolved after repair (Section 2.4). That is why we operate with four arms of analysis in all calculations and why we report A_conservative (which forbids any repaired entry whatsoever) in parallel with the main arm throughout.

Population frequency is a proxy of tolerance not a benign invitation. The variant carried by 50 healthy genomes may still be pathogenic with low penetrance. The MitoTIP GenBank set is not a neutral population baseline: MITOMAP mentions uneven continental sampling and inclusion of sequences from patients with mitochondrial disease. Changing the frequency cut changes the AUC between 0.960 and 0.971 so the conclusion is independent of where the line is drawn but the negative class and hence the measured specificity cost is still an approximation. The benchmarked space is narrower than the output of the tool: S1 contains no deletion, so the 1 504 single-base deletions MitoTIP scores have never been validated and insertions are by construction out of scope.

6 Future Work

The results indicate a specific next step, not a general need for more data.

Another threshold, a pattern. Section 3.8 demonstrates that no in-sample threshold criterion can work for the base-paired side and Figure 7B shows that the single feature that does have real discriminative power loses it precisely where the contested variants are located. Both point in the same direction: the remaining errors need features used all the time, not some third cutoff. Two previously measured here candidates are predicted folding destabilization (AUC 0.750) and neighborhood score context (0.698).

Features that are not modeled by MitoTIP at all. Two are worth naming, the first the more interesting as it maps onto the finding. Under-calling is observed at unmodelled post-transcriptional modification sites. Modifications are enriched in loops. For example, the MERRF variant at position 34 of tRNA-Lys acts by blocking a taurine modification, which could not be represented by any secondary-structure feature. Similarly, aminoacylation identity elements, the discriminator base and the acceptor-stem positions read by a synthetase are invisible to a fold-stability model.

Weighted conservation of base pairs, not of positions. This is a refinement, not a new feature, as MitoTIP already implements interspecies conservation. A G:C to A:U

substitution at both partners keeps the pair, but both bases are changed. Thus, conservation of a position counts as a variable site for the case of full structural conservation. Scoring conservation at the pair level would allow one to distinguish compensated changes from those that are truly disruptive.

An impending test. The cleanest possible validation is to fix the two thresholds now and test them on mt-tRNA variants that ClinGen classifies in the future. This turns a retrospective calibration exercise into a prediction and all that is needed is patience.

7 Conclusion

The errors in prediction of human mitochondrial tRNAs by MitoTIP are structural errors, not random errors. Against the tool's own labeled set, 23.0% of base-paired variants are misclassified against only 1.6% of unpaired ones, a 14.0-fold difference ($p=9.0\times10-18$) that survives leave-one-gene-out refitting, four region definitions and four analysis arms. The two error types do not mix at all: every stem error is a false positive and every loop error a false negative. This is due to calibration rather than discrimination because unpaired variants score systematically lower in both classes and one cutoff cannot fit both distributions. Region-specific thresholds fitted on the training table alone raise detection of 29 held-out, ClinGen-confirmed pathogenic variants from 75.9% to 89.7%, with the recovery concentrated in the classical encephalomyopathies. The entire gain comes from the unpaired threshold, while the base-paired threshold buys specificity and no sensitivity. The cost is not taken for granted but measured on an independent set: unpaired specificity drops from 100% to 94.0%, four false positives among 67 population-tolerated variants, and the raw error count does not drop at all. This is still a recalibration, not a new model and it moves the diagnostic blind spot rather than closing it – tubulointerstitial kidney disease is still missed and hearing loss ends up marginally worse. The method behind the pair of numbers is transferable, not the numbers themselves: split the molecule by pairing status, place the unpaired cutoff at the edge of the empty band between the classes, and derive the base-paired cutoff from a benign-only quantile so that it cannot be overfitted to the handful of pathogenic variants available. If one sentence is to be remembered, it is that a single global threshold on a structurally heterogeneous molecule is a modeling assumption, not a neutral default, and that an aggregate AUC of 0.971 was hiding a 14-fold difference in error rate between two halves of the same molecule.

References

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (1999). Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, *23*(2), 147. https://doi.org/10.1038/13779

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413. https://doi.org/10.1093/biomet/26.4.404

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, *6*, 26. https://doi.org/10.1186/1748-7188-6-26

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., Procaccio, V., & Wallace, D. C. (2013). mtDNA variation and analysis using MITOMAP and MITOMASTER. *Current Protocols in Bioinformatics*, *44*(1), 1.23.1–1.23.26. https://doi.org/10.1002/0471250953.bi0123s44

Niroula, A., & Vihinen, M. (2016). PON-mt-tRNA: A multifactorial probability-based method for classification of mitochondrial tRNA variations. *Nucleic Acids Research*, *44*(5), 2020–2027. https://doi.org/10.1093/nar/gkw046

Petropoulou, A., Kypraios, N., Rizopoulou, D., Kouvela, A., Maniatis, A., Anastasopoulou, K., Anastogianni, A., Korfiatis, T., Grafanaki, K., Stamatopoulou, V., & Stathopoulos, C. (2025). Mitochondrial tRNA-derived diseases. *International Journal of Molecular Sciences*, *26*(24), 12023. https://doi.org/10.3390/ijms262412023

Pütz, J., Dupuis, B., Sissler, M., & Florentz, C. (2007). Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures. *RNA*, *13*(8), 1184–1190. https://doi.org/10.1261/rna.588407

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Salinas-Giegé, T., Giegé, R., & Giegé, P. (2015). tRNA biology in mitochondria. *International Journal of Molecular Sciences*, *16*(3), 4518–4559. https://doi.org/10.3390/ijms16034518

Singer-Vine, J., & The pdfplumber contributors. (2024). *Pdfplumber: Plumb a PDF for detailed information about each text character, rectangle and line*. https://github.com/jsvine/pdfplumber.

Smith, R. (2007). An overview of the Tesseract OCR engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, *2*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Sonney, S., Leipzig, J., Lott, M. T., Zhang, S., Procaccio, V., Wallace, D. C., & Sondheimer, N. (2017). Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. *PLOS Computational Biology*, *13*(12), e1005867. https://doi.org/10.1371/journal.pcbi.1005867

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr