Skepthical review: *Exploratory Multi-Modal Investigation of Brain Microstructure and Epigenetic Aging in Egyptian Fruit Bats: Identifying Phenotypes of Resilience and Vulnerability*

Summary

The manuscript reports an exploratory multi-modal study in Egyptian fruit bats aiming to relate neuroimaging and behavior to epigenetic age (DNAm age) and to define resilient vs. vulnerable aging phenotypes (Sec. 1, Sec. 2.2–2.4). Substantial dataset constraints prevented the originally proposed analyses: MRI was provided as 3D (not 4D), so the planned diffusion-weighted signal variability (DW-SV) metric could not be computed (Sec. 2.2.1–2.2.3, Sec. 3.2.1), and the derived behavioral metrics (Exploration Entropy; Navigational Redundancy) were identically zero across animals/phases, making the behavioral modality unusable (Sec. 2.2.1–2.2.3, Sec. 3.2.1). The final integrated dataset included 31 bats with DNAm age, atlas-parcellated regional mean MRI intensities (24 regions + global mean), and demographics (Sex, Origin) (Sec. 2.1.2, Sec. 2.3.1–2.3.2, Sec. 3.1). The authors fit an Elastic Net model with LOOCV to predict DNAm age, but performance was worse than a mean predictor ($R^2 = -0.101$; MAE ≈ 1.405 years; Sec. 2.4.2, Sec. 3.3.1). Despite this, the manuscript inspects non-zero coefficients as putative "brain-aging signatures" and defines "Resilient/Vulnerable" phenotypes via LOOCV residual quartiles (Sec. 3.3.2–3.3.3). The work is most valuable as a transparent feasibility/lessons-learned pipeline paper; however, biological interpretations (regional "signatures" and residual-based phenotypes) are currently not supported given unclear imaging signal definition/normalization, a likely behavioral extraction failure, under-specified (and potentially non-nested) model tuning, small N relative to p, and the non-predictive model outcome (Sec. 3.3–3.4, Conclusions).

Strengths

- Candid reporting of major data limitations (3D MRI instead of 4D; degenerate behavioral features; poor predictive performance) without attempting to hide null/negative results (Sec. 1, Sec. 3.2–3.3, Conclusions).
- Clear intent and conceptual framing for a multi-modal aging pipeline in a long-lived, non-traditional model organism, which is valuable even when results are negative (Introduction, Sec. 2.2–2.4).
- Concrete data curation/harmonization steps (ID matching, filtering to complete cases) that can be reused by future studies (Sec. 2.1.1–2.1.2).
- Use of regularized regression (Elastic Net) and out-of-fold prediction via LOOCV is directionally appropriate for small-n, multi-feature settings, and key performance metrics are reported (Sec. 2.4.2, Sec. 3.3.1).

- Residual-based phenotyping is presented with a clear mathematical definition (Sec. 2.4.3), which could be a useful template once a validated predictor exists.

Major issues

1. **The manuscript's interpretive claims about identifying "brain-aging signatures" and defining "Resilient/Vulnerable" phenotypes are not supported by the evidence because the predictive model is non-informative ($R^2 = -0.101$; Sec. 3.3.1) and the reported imaging–DNAm associations are weak (Sec. 3.2.2).** Interpreting non-zero Elastic Net coefficients from a poor/unstable model as biological "signatures," and describing residual-quartile groups ($N \approx 8$ each) as phenotypes without uncertainty/effect sizes, risks overinterpretation (Sec. 3.3.2–3.3.3, Sec. 3.4, Conclusions).

   Recommendation: Reframe Sec. 3.3.2–3.3.3, Sec. 3.4, and Conclusions so coefficient patterns and residual-based grouping are explicitly presented as workflow demonstrations and hypothesis generation only (not evidence of signatures/phenotypes). If any region-level comparisons remain in the main text, add effect sizes and uncertainty (e.g., bootstrap CIs), clearly label them exploratory, and avoid phenotype language unless the predictor shows at least modest validity/calibration. Consider moving the residual-based phenotype subsection to an appendix as a worked example.

2. **The neuroimaging feature used—atlas-based regional mean "signal intensity" from a single 3D volume—is not sufficiently defined, controlled, or justified.** The manuscript does not clearly specify what the 3D image represents (e.g., b_0 , FA/MD map, T1/T2, magnitude image), nor does it describe essential preprocessing steps (bias-field correction, skull stripping/brain masking, registration type, QC) or intensity normalization/harmonization. Without these, between-bat intensity differences may reflect scanner/session scaling, coil loading, registration errors, or other nuisance variation rather than biology (Sec. 2.3.1–2.3.2, Sec. 3.2.2).

   Recommendation: Add a dedicated subsection in Sec. 2.3 detailing MRI acquisition and preprocessing: scanner/field strength, sequence/contrast, TR/TE, voxel size, whether these are raw images or diffusion-derived maps, and why the term "DTI" is appropriate (or remove it). Describe brain masking, bias-field correction, atlas registration (rigid/affine/nonlinear; software; interpolation), and QC (e.g., visual checks, registration failures). Implement and report an intensity normalization strategy across subjects (e.g., within-brain z-scoring, histogram matching, reference-tissue scaling) and re-run core analyses (Sec. 3.2.2, Sec. 3.3.1) to assess robustness. Provide a table listing the 24 atlas regions with anatomical names, voxel counts, and label IDs used in figures/text.

3. **Behavioral metrics (Exploration Entropy; Navigational Redundancy) being exactly zero for all bats and phases is highly suggestive of a data extraction/parsing bug or an empty-sequence failure mode, yet the manuscript**

does not provide diagnostic evidence to distinguish technical failure from true behavioral invariance (Sec. 2.2.1–2.2.3, Sec. 3.2.1). This undermines the multi-modal premise and leaves a key modality unresolved.

Recommendation: Extend Sec. 2.2 and Sec. 3.2.1 with a systematic behavioral pipeline audit: (1) report per-bat/per-phase counts of events, sequence lengths, and number of unique boxes visited; (2) show 2–3 representative raw table snippets (rows/columns) from the xlsx files and the parsed sequences to verify correct column names, action codes, timestamps, and box IDs; (3) validate entropy/redundancy on toy synthetic sequences that must yield non-zero values; (4) explicitly report how cases with zero post-discovery actions are handled; and (5) if a bug is found, recompute behavioral features and re-run multi-modal analyses. If the task truly yields near-deterministic behavior, add simpler robust summaries (e.g., latency to first discovery, total visits, perseveration/repeat count) as fallback features and discuss experimental causes.

4. **Key aspects of the Elastic Net modeling and evaluation pipeline are under-specified and may be methodologically flawed for small n ($N = 31$) and correlated predictors.** It is unclear whether (i) predictors were standardized within each training fold (to prevent leakage), (ii) alpha/l1_ratio and regularization strength were selected via nested CV (inner tuning within each LOOCV training set), (iii) categorical variables (Sex, Origin) were encoded consistently, and (iv) missing values/outliers were handled (Sec. 2.4.2, Sec. 3.3.1). Coefficient interpretation is especially fragile without stability analysis.

   Recommendation: Expand Sec. 2.4.2 to fully specify: exact design matrix columns; encoding of Sex/Origin; preprocessing/scaling; missing-data handling; software and versions; and hyperparameter search ranges. Implement leakage-free training by fitting scalers/encoders inside each training fold only. Prefer nested CV (inner CV/grid search for hyperparameters; outer LOOCV for evaluation). Add model sanity checks: permutation test (shuffle DNAm age and re-fit) and coefficient stability (bootstrap or selection frequency across folds). Report additional metrics (RMSE, correlation r, calibration slope/intercept) alongside R^2/MAE in Sec. 3.3.1.

5. **Residual-quartile "Resilient/Vulnerable" phenotypes are not meaningful when derived from a model with negative predictive value; residuals mainly reflect model error/noise. Additionally, the paper does not clarify the biologically standard alternative—DNAm age acceleration relative to chronological age—because chronological age availability/usage is unclear (Sec. 2.4.3, Sec. 3.3.3; also target column suggests skin clock).**

   Recommendation: Revise Sec. 2.4.3 and Sec. 3.3.3 to avoid resilience/vulnerability labels unless a validated predictor exists. If chronological age exists, add it explicitly and analyze DNAm age acceleration (DNAmAge residualized on chronological age, with sex/origin covariates as appropriate), and report basic clock validation in this cohort (DNAm vs chronological correlation and error). If chronological age is unavail-

able, state this prominently and tone down any "biological age discrepancy" interpretation. If stratification is still desired for exploration, use model-free approaches (e.g., clustering/PCA on normalized imaging features) and then test association with DNAm age, clearly labeled exploratory.

6. **There is a persistent mismatch between the paper's stated goals (DW-SV + advanced behavior + multi-modal prediction) and what was actually implemented (regional mean intensity + demographics), which risks overselling the contribution and confusing readers about what was tested vs. planned (Abstract, Introduction, Sec. 3.4, Conclusions).**

   Recommendation: Revise the Abstract, Introduction, and Conclusions to separate (i) planned aims (DW-SV from 4D DWI; behavioral metrics) from (ii) achieved analyses (3D intensity + demographics). State explicitly that the dataset as provided cannot test the DW-SV hypothesis. Reposition the manuscript as a feasibility/analysis pipeline report under real-world constraints, and list concrete requirements for future data collection (true 4D diffusion with bvals/bvecs; behavior design to elicit variability; larger N).

7. **Small sample size ($N = 31$) relative to the number of candidate predictors (24 regions + global + Sex + Origin; plus potential one-hot expansion) and multiple exploratory views (correlations, coefficient inspection, residual grouping) creates substantial risks of overfitting, unstable feature selection, and false positives. The manuscript acknowledges some limitations but does not quantify instability or address multiple comparisons systematically (Sec. 3.2.2–3.3.2, Sec. 3.4).**

   Recommendation: Add a focused limitations/statistics paragraph in Sec. 3.4 (or a dedicated Limitations subsection) quantifying the n-to-p challenge and explicitly stating that region-level findings are exploratory. Apply multiple-comparison control where univariate tests are presented (e.g., FDR). Include stability/sensitivity analyses (bootstrap coefficients; repeated CV seeds where applicable; permutation baselines). Consider dimension reduction (e.g., PCA on normalized regional intensities) as a more stable exploratory alternative and report whether any principal components associate with DNAm age.

Minor issues

1. Cohort counts and summary statistics are internally inconsistent across Table 1 and Results Sec. 3.1 (e.g., Sex 24/18 vs 23/18; Origin 23/18 vs 22/19; and $\mathbf{24 + 18 = 42}$ while total is 41). Mean±SD age/DNAm age summaries also differ (Table 1: $\mathbf{9.87 \pm 1.98}$ vs Sec. 3.1: $\mathbf{9.60 \pm 1.74}$) without clear labeling of which variable/cohort each refers to (Sec. 3.1; Table 1).

Recommendation: Audit and reconcile all cohort summaries: ensure subgroup counts sum to totals and match between Table 1/Table 2 and Sec. 3.1. Clearly label whether reported mean±SD refers to chronological age vs DNAm age and whether it is computed on the initial $N = 41$ cohort or the final $N = 31$ complete-case cohort.

2. Feature-count arithmetic is inconsistent (e.g., "26 potential features" while listing 24 regional + 1 global + sex + origin, which totals 27 before categorical encoding details; Sec. 3.3.2 and related statements).

   Recommendation: Explicitly list the final design matrix columns and their count (including whether Sex/Origin are binary or one-hot encoded, and whether global mean is included). Update all feature-count statements to match the actual implementation.

3. Atlas region references are often numeric (e.g., "region 14") and figures use generic/non-anatomical labels, limiting interpretability (Sec. 3.2.2–3.3.2; Figures 8–12).

   Recommendation: Provide a region key (main text table or supplement) mapping label IDs to anatomical names and voxel counts. Use anatomical labels in coefficient and group-difference plots (or include both ID and name).

4. Subject selection from the initial cohort ($N = 41$) to the final integrated cohort ($N = 31$) is described but not assessed for selection bias (Sec. 2.1.2, Sec. 3.1).

   Recommendation: In Sec. 3.1, compare included vs excluded bats on DNAm age (and chronological age if available), Sex, and Origin (simple tests and descriptive stats). Report whether filtering could bias conclusions.

5. Figures are frequently hard to interpret due to missing/unclear axis units, inconsistent N labels, small fonts/low resolution, and missing statistical context (CIs/p-values/performance metrics). Some figures are redundant given degenerate behavioral data (Figures 3–4) and several plots would benefit from colorblind-safe palettes (Figures 1–4, 7–12).

   Recommendation: Make figures publication-ready: add units, explicit N, clearer panel titles, and legible fonts; export as vector/high-resolution. Move or condense behavior-zero-variance figures to the supplement. Where inferential statements are implied, add uncertainty (CIs) and clearly state whether any p-values are reported and how multiple comparisons are handled.

6. Target-variable naming is inconsistent and includes typographical artifacts (e.g., 'DNAmAgeBat.Rousettus.aegyptiacus_Skin' vs variants containing 'Skip'/spacing). This makes it harder to reproduce the pipeline (Sec. 2.1.1, Sec. 2.4.2, Sec. 3).

   Recommendation: Standardize to one canonical target name in the manuscript (e.g., DNAmAge) and provide a single mapping to the original dataset column name in Sec. 2.1/2.4.2. Remove stray 'Skip' text and spacing inconsistencies.

7. The manuscript would benefit from clearer context on (i) epigenetic clocks in bats (including tissue specificity—here apparently skin), (ii) expectations for MRI correlates of DNAm age, and (iii) why the chosen modeling approach is appropriate given constraints (Sec. 1, Sec. 3.4).

   Recommendation: Add a brief Related Work/Context subsection (end of Sec. 1) summarizing prior bat epigenetic clock work, typical age-acceleration analyses, and MRI-aging correlates. Explicitly discuss why skin DNAm age may weakly relate to brain measures and how that affects expectations.

8. Multiple-comparison considerations are not addressed where many regions are examined (correlation heatmaps; region-wise group contrasts), increasing false-positive risk (Sec. 3.2.2, Sec. 3.3.3; Figures 8–12).

   Recommendation: Where region-wise p-values are presented (or implied), apply FDR (or explicitly state that no inferential testing is being claimed). Prefer reporting effect sizes + CIs for a pre-declared small set of regions if any biological interpretation is retained.

9. Front-matter metadata appears to contain placeholders or mismatched keywords (e.g., "Astronomy data analysis/modeling"; placeholder affiliations), which undermines credibility and discoverability (front matter before Sec. 1).

   Recommendation: Replace keywords with relevant terms (epigenetic aging, bats, MRI/diffusion MRI, elastic net, multimodal integration) and remove any placeholder author/affiliation text.

Very minor issues

1. Typos and formatting artifacts appear throughout (broken words across line breaks, stray '#' characters/extra periods in headings, inconsistent quotation style around "Resilient/Vulnerable", inconsistent spacing in math and units) (Sec. 1–3, Conclusions).

   Recommendation: Proofread and clean formatting globally: fix broken hyphenation/line-break artifacts, normalize headings, standardize quotes for labels, and ensure consistent numeric formatting (e.g., '9.60 \pm 1.74$ years').

2. Entropy definition omits the standard convention for zero-probability states ($0 \cdot \log(0) = 0$), which is needed for mathematical completeness (Sec. 2.2.2).

   Recommendation: Add a one-line clarification: terms with $p(i) = 0$ contribute 0 (or restrict the sum to i with $p(i) > 0$).

3. Figure caption and referencing style is inconsistent (some captions/panel headings appear as stray lines; some references use varied phrasing) (Sec. 3.2.2–3.3.3).

Recommendation: Standardize captions to begin with 'Figure X.' and ensure consistent in-text citations ('Figure X'). Fold any stray headings into the appropriate captions.

Key statements and references

- • **The study employed an Elastic Net regression model with Leave-One-Out Cross-Validation to predict DNA methylation (DNAm) age from 24 regional Mean Signal Intensity metrics, Global Mean Signal Intensity, and demographic factors (Sex and Origin) in a final cohort of 31 Egyptian fruit bats, but the model showed poor predictive performance with $R^2 = -0.101$ and Mean Absolute Error = 1.405 years, indicating predictions were worse than using the cohort mean DNAm age.**
- *Reference(s):* (none)
- • **Across the 31 bats in the final analysis cohort, Global Mean Signal Intensity derived from 3D MRI showed no significant linear association with DNAm age, with a Pearson correlation coefficient $r = -0.039$ and $p = 0.837$, suggesting this global neuroimaging metric is not related to epigenetic age in this dataset.**
- *Reference(s):* (none)
- • **A comprehensive correlation analysis between DNAm age and the 24 atlas-defined regional Mean Signal Intensity values, as well as the global mean, revealed generally weak linear relationships, with Pearson correlation coefficients indicating that neither regional nor global mean signal intensity robustly correlates with epigenetic age in these Egyptian fruit bats.**
- *Reference(s):* (none)
- • **In the final Elastic Net model trained on all 31 subjects, 12 of 26 predictors (24 regional Mean Signal Intensities, Global Mean Signal Intensity, Sex, Origin) received non-zero coefficients, with regions 14 (coefficient = 0.157), 6 (0.131), and 13 (0.093) emerging as the strongest positive predictors of higher epigenetic age, and regions 12 (−0.083) and 10 (−0.082) as the strongest negative predictors, while Sex had a small positive coefficient (0.003), indicating only weak and highly exploratory regional brain-aging signatures.**
- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains a small number of explicit formulas/definitions (Shannon entropy for exploration, a ratio-based redundancy metric, LOOCV residual definition) and qualitative descriptions of Elastic Net/LOOCV. There are no extended derivations. The main internal mathematical concern is an inconsistency in the stated number of candidate predictors versus the enumerated feature list.

Checked items

1. ✔ **Exploration Entropy (Shannon) definition** (Sec. 2.2.2, p.3)
    - **Claim:** Defines visitation entropy as $H = -\sum_{i=1}^{6} p(i)\log_2(p(i))$, where $p(i)$ is the proportion of entries into box i.
    - **Checks:** algebra, normalization/constraints, limiting/sanity cases
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** There are 6 boxes (indexed $i = 1..6$). $p(i) \geq 0$ and $\sum p(i) = 1$ based on proportions of total entries.
    - **Notes:** The formula matches Shannon entropy in bits (base-2 logarithm). Limiting cases behave correctly: if all visits are to one box, $H = 0$; if uniform over 6 boxes, $H = \log_2(6)$.
2. ⚠ **Entropy well-definedness when $p(i) = 0$** (Sec. 2.2.2, p.3)
    - **Claim:** The entropy sum is written over all $i = 1..6$ without specifying handling for boxes with $p(i) = 0$.
    - **Checks:** domain/definition consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** Some boxes may be unvisited, implying $p(i) = 0$.
    - **Notes:** As written, $\log_2(0)$ is undefined. In practice one uses the convention $p \log p := 0$ when $p = 0$, or sums only over $p(i) > 0$. The paper does not state this; add clarification to make the definition mathematically complete.
3. ✔ **Navigational Redundancy (post-discovery) ratio definition** (Sec. 2.2.3, p.3)
    - **Claim:** Defines Post-Discovery Redundancy as (# incorrect entries after discovery)/(total # entries after discovery), with value set to 0 if there are no post-discovery entries.
    - **Checks:** algebra, normalization/constraints, edge cases
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Counts are nonnegative integers. Incorrect entries after discovery $\leq$ total entries after discovery when total> 0.
    - **Notes:** The ratio is bounded in $[0, 1]$ when the denominator is positive. The special-case definition when the denominator is 0 avoids division-by-zero and is a standard, mathematically coherent convention.

4. ✔ **Regional Mean Signal Intensity definition** (Sec. 2.3.2, p.3–4)

    - **Claim:** For each atlas region label, compute the mean MRI signal intensity over voxels in that region; compute global mean over all non-zero atlas voxels.
    - **Checks:** definition consistency, edge cases
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Atlas labels define disjoint (or at least well-defined) voxel sets per region. Non-zero atlas voxels correspond to a brain mask.
    - **Notes:** Definitions are internally coherent. No explicit formula is provided, but the described operations uniquely define regional and global averages given a mask.

5. ✔ **LOOCV fold training size** (Sec. 2.4.2, p.4)

    - **Claim:** In each LOOCV fold, one bat is held out and the model is trained on the remaining $N - 1$ bats.
    - **Checks:** algebra, procedure/logic consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Total sample size is N.
    - **Notes:** The described LOOCV procedure is logically and algebraically consistent with standard leave-one-out cross-validation.

6. ✔ **Age residual definition and sign interpretation** (Sec. 2.4.3, p.4)

    - **Claim:** Defines Age Residual = Actual DNAmAge − Predicted Age (from LOOCV), and interprets bottom quartile as 'actual much younger than predicted' (residual very negative) and top quartile as 'actual much older than predicted' (residual very positive).
    - **Checks:** algebra, sign consistency, procedure/logic consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Predicted age is the LOOCV out-of-fold prediction for each subject.
    - **Notes:** The residual definition and the phenotype directionality are consistent: Actual $<$ Predicted implies negative residual (younger-than-predicted), and Actual $>$ Predicted implies positive residual (older-than-predicted).

7. ✔ **Phenotype stratification by quartiles** (Sec. 2.4.3, p.4; Sec. 3.3.3, p.8)

    - **Claim:** Uses 25th and 75th percentile residual thresholds to define Resilient and Vulnerable groups (bottom and top quartiles).
    - **Checks:** normalization/constraints, definition consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Residual distribution is computed across all N subjects with LOOCV predictions.

- **Notes:** Quartile-based grouping is mathematically well-defined. (It is descriptive and does not require additional derivation to be internally consistent.)

8. ✖ **Predictor feature count inconsistency (26 vs enumerated list)** (Sec. 3.3.2, p.8)
    - **Claim:** States the model selected 12 out of 26 potential features, enumerated as (24 regional mean intensities, Global Mean Signal Intensity, Sex, Origin).
    - **Checks:** counting/dimensional consistency, notation/definition consistency
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Each listed item corresponds to at least one column in the design matrix.
    - **Notes:** The enumerated list totals 27 (24 + 1 + 1 + 1) if each is a single feature. The text says 26. This discrepancy indicates unclear feature construction (e.g., whether Global Mean Signal Intensity was included; whether Origin was encoded into multiple dummies; or whether one of these was excluded). This should be reconciled because it affects the stated model specification and interpretability of coefficient plots.

Limitations

- The document provides no explicit Elastic Net objective function, penalty definitions, or notation for encoding categorical predictors (Sex/Origin), preventing verification of the exact mathematical model beyond high-level description.
- Most quantitative inconsistencies in cohort statistics are reported numbers rather than derived expressions; this audit flags internal consistency but does not attempt numerical validation.
- Several referenced figures/tables are not fully legible in the provided parsed text, limiting cross-checks between captions and the main text where exact values might clarify definitions.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Out of 12 automated numeric consistency checks, 8 passed and 4 failed. Failures were concentrated in cross-section inconsistencies for initial cohort descriptors (sex/origin counts; age mean ±SD) and in a feature-count breakdown that does not sum to the stated total.

Checked items

1. ✔ **C01_subject_count_sex_origin_table1_sumcheck** (Page 3, Table 1 (Initial Cohort Characteristics))

- **Claim:** Table 1 reports Number of Subjects = 41; Sex (Male/Female) = 23/18; Origin (Aseret/Herzliya) = 22/19.
- **Checks:** parts_vs_total
- **Verdict:** PASS
- **Notes:** Sex sum matches total; Origin sum matches total.

2. ✔ **C02_initial_age_range_consistency_table1** (Page 3, Table 1 (Initial Cohort Characteristics))
    - **Claim:** Age (Mean ± SD) = $\mathbf{9.87 \pm 1.98}$ yrs; Age (Min - Max) = $\mathbf{6.62 - 15.07}$ yrs.
    - **Checks:** range_sanity_check
    - **Verdict:** PASS
    - **Notes:** Basic inequality sanity checks.
3. ✖ **C03_initial_cohort_sex_origin_in_results_vs_total** (Page 4, Results §3.1 (Cohort characteristics and data curation))
    - **Claim:** Results text states initial dataset comprised 41 bats and reports 'balanced distribution across sexes (24 males, 18 females) and origin colonies (23 Aseret, 18 Herzliya)'.
    - **Checks:** parts_vs_total
    - **Verdict:** FAIL
    - **Notes:** Sex sum does not match total; Origin sum matches total.
4. ✖ **C04_table1_vs_results_initial_sex_origin_mismatch** (Page 3 Table 1 vs Page 4 Results §3.1)
    - **Claim:** Table 1 lists Sex=23/18 and Origin=22/19, while Results §3.1 lists Sex=24/18 and Origin=23/18 for the same initial cohort size 41.
    - **Checks:** repeated_constant_consistency
    - **Verdict:** FAIL
    - **Notes:** Pairwise equality check across sections.
5. ✖ **C05_initial_age_mean_sd_table1_vs_results** (Page 3 Table 1 vs Page 4 Results §3.1)
    - **Claim:** Table 1 reports Age mean±SD = $\mathbf{9.87 \pm 1.98}$ yrs, while Results §3.1 reports mean DNAm age = $\mathbf{9.60 \pm 1.74}$ years (same stated initial cohort of 41).
    - **Checks:** repeated_statistic_consistency
    - **Verdict:** FAIL
    - **Notes:** Compared reported mean/SD across sections using abs_tol.
6. ✔ **C06_data_availability_counts_from_venn_text** (Page 5, Results §3.1 and Figure 2 caption/text)

- **Claim:** All 41 subjects had metadata and behavioral data files; 33 possessed DTI scans; 31 possessed complete data across all three modalities.
- **Checks:** set_cardinality_consistency
- **Verdict:** PASS
- **Notes:** Checked subset inequalities and derived missing DTI count.

7. ✔ **C07_final_cohort_sex_origin_sumcheck** (Page 5, Results §3.1 (final cohort description; Table 2 referenced but not shown in text))
    - **Claim:** Final cohort $N = 31$ with sex distribution 19 males, 12 females; origin 16 Aseret, 15 Herzliya.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** Sex sum matches total; Origin sum matches total.
8. ✔ **C08_final_cohort_age_range_vs_mean_sd_sanity** (Page 5 (final cohort mean±SD) and Page 3 (overall min-max provided for initial cohort))
    - **Claim:** Final cohort mean age 9.41 ± 1.64 years; initial cohort min-max 6.62–15.07 years. Check final mean is within stated overall min-max (sanity).
    - **Checks:** range_sanity_check
    - **Verdict:** PASS
    - **Notes:** Sanity check: final mean within initial cohort range.
9. ✖ **C09_feature_count_total_26_breakdown** (Page 8, Results §3.3.2)
    - **Claim:** Model had 26 potential features: 24 regional mean signal intensities + Global Mean Signal Intensity + Sex + Origin.
    - **Checks:** parts_vs_total
    - **Verdict:** FAIL
    - **Notes:** Checked sum of feature components vs reported total.
10. ✔ **C10_selected_feature_count_vs_total** (Page 8, Results §3.3.2)
    - **Claim:** Elastic Net selected 12 out of 26 potential features as having non-zero coefficients.
    - **Checks:** ratio_sanity_check
    - **Verdict:** PASS
    - **Notes:** Bound check only.
11. ✔ **C11_phenotype_group_sizes_sum_to_final_n** (Page 8, Results §3.3.3)
    - **Claim:** Phenotype counts: Resilient $N = 8$, Vulnerable $N = 8$, Nominal $N = 15$ (final cohort $N = 31$).
    - **Checks:** parts_vs_total

- **Verdict:** PASS
- **Notes:** Checked phenotype group sizes sum to final cohort N.

12. ✔ **C12_quartile_group_size_expectation_check** (Page 8, Results §3.3.3 (quartile-based classification with $N = 31$))
    - **Claim:** Bottom quartile classified as Resilient ($N = 8$) and top quartile as Vulnerable ($N = 8$) for $N = 31$.
    - **Checks:** quantile_count_sanity
    - **Verdict:** PASS
    - **Notes:** Checked quartile group sizes against $n_{\text{total}} * \textbf{quartile}_{\text{f}}\textbf{raction}$ with abs rounding tolerance.

Limitations

- Checks are limited to arithmetic/logical consistency of numbers explicitly stated in the provided PDF text; no underlying datasets (MRI voxels, behavioral logs, per-subject predictions) are available for recomputation.
- Figure-based numeric verification is constrained because extracting precise plotted values from images is out of scope; only numbers explicitly written in text/captions are used.
- Some statistical quantities (e.g., p-values, R^2 / MAE) cannot be independently verified without raw data or explicit intermediate values.