Skepthical review: *Critical Assessment of a Multimodal Pipeline for Studying Cognitive Resilience in Aging Bats: Insights from Data Integration Failures*

Summary

The manuscript is framed as a multimodal study of cognitive resilience in aging Egyptian fruit bats (Rousettus aegyptiacus), intending to integrate DNA methylation age (DNAmAge), MRI-derived total brain volume (TBV) from $b = 0$ DTI images, and spatial-memory metrics from a multi-phase foraging paradigm via a staged pipeline (data harmonization → behavioral feature engineering → TBV quantification → regression with an age×TBV interaction; Sec. 1, Sec. 2.1–2.4). In practice, the pipeline fails primarily at behavioral feature extraction: the Excel parsing assumptions (e.g., presence/structure of an 'Absolute_Time' column and consistent sheet layouts) do not match the actual files, leading to empty event tables and degenerate behavioral outcomes (e.g., all latencies fixed at $10,800\,\mathrm{s}$; error counts all zero; Sec. 2.2, Sec. 3.2.1). TBV extraction appears to be technically salvageable after correcting an initial 4D-vs-3D NIfTI assumption, but the imaging workflow is under-specified and the resulting TBV values are not cleanly integrated with valid behavioral measures (Sec. 2.3, Sec. 3.2.2). Consequently, the planned regression models are fit on constant/invalid outcomes, producing NaN coefficients or pathological fit indices (e.g., $R^2 = -\infty$), so no biological inference is supported (Sec. 2.4, Sec. 3.3). As a methodological "failure case study," the paper has potential value, but it currently reads partly like an internal postmortem: key diagnostics and raw-format evidence are not shown, Methods/Results are sometimes conflated, inconsistencies remain in cohort statistics and model description, and the lessons learned are not yet distilled into reusable guardrails/checklists and reproducible artifacts.

Strengths

- Unusually candid end-to-end documentation of failure propagation in a multimodal workflow, including concrete failure signatures (KeyError, empty parsed tables, zero-variance metrics, NaN coefficients, $R^2 = -\infty$) and their downstream consequences (Sec. 2.2, Sec. 3.2–3.3).
- Clear articulation of the intended staged pipeline (harmonization → feature engineering → TBV quantification → regression with interaction) and explicit listing of planned covariates (Sex, Origin) and moderation hypothesis (Sec. 1, Sec. 2.1–2.4).
- Cohort harmonization effort is described with a final integrated set of **33** bats (from **41**) and an attempt to track metadata, behavioral references, methylation age, and imaging references (Sec. 2.1, Sec. 3.1).
- The manuscript's premise—multi-omics (DNAmAge) + neuroimaging + cognition in a long-lived bat model—is compelling and could be impactful once the pipeline is made robust (Sec. 1, Sec. 4.1–4.2).

- The paper already contains the seeds of useful best practices (format verification, early plausibility checks, staged QC) and can be developed into a practical guide with relatively targeted additions (Sec. 3.4, Sec. 4.4).

Major issues

1. **Manuscript genre and audience are not consistently defined: the text alternates between an empirical aging/brain/cognition study and a methodological failure-case report, which blurs what readers should take away and what standards of evidence apply (Sec. 1, Sec. 4.1–4.4).**

   Recommendation: Decide and signal the genre explicitly in Sec. 1 and Sec. 4: if this is primarily a methods/failure case study, state up front that no biological inference is attempted due to invalid downstream data, and restructure the Introduction/Conclusions to emphasize (i) the intended scientific question only insofar as it motivates pipeline requirements and validation criteria, and (ii) the deliverable as documented failure modes + guardrails/checklists. Alternatively, if aiming for an empirical paper, the behavioral extraction must be fixed and results re-run; the current version should not present inferential modeling outputs as results.

2. **Behavioral feature-extraction failure—the central failure mode—is described too generically to be reproducible or broadly instructive. Readers cannot see what the Excel files actually look like, what \texttt{read_excel} returned, or which concrete discrepancies caused the "0 entries parsed" outcome (Sec. 2.2, Sec. 3.2.1).**

   Recommendation: Augment Sec. 2.2 and Sec. 3.2.1 with concrete diagnostics: (1) a table comparing assumed vs. actual sheet names and column headers (including whitespace/case/merged-cell artifacts); (2) an anonymized snippet (header row + 5–10 rows) from 1–2 representative files; (3) the exact \texttt{pandas.read_excel()} parameters used (engine, header, skiprows, dtype) plus the resulting column-name list printed by the script; (4) a mapping from each discrepancy to the observed failure signature (KeyError, empty dataframe, downstream constants). Also include at least one manual verification (e.g., hand-count events in one file) demonstrating that "0 entries" is a parsing artifact rather than absence of behavioral events.

3. **Operational definitions of behavioral metrics and edge-case handling are under-specified, which weakens the methodological value even if parsing were fixed. In particular, the meaning of "Absolute_Time" (wall-clock vs elapsed), phase start references, and the hard-coding of $10,800\,\mathrm{s}$ latency ($3\,\mathrm{h}$) imply censoring/right-truncation that is not acknowledged as a modeling issue (Sec. 2.2; also ambiguity noted in Sec. 2.2 latency definition).**

Recommendation: In Sec. 2.2, define each metric with unambiguous computation rules, units, and time origin: specify whether latency is (time of first correct entry − phase start time) and where phase start time is obtained. Explicitly describe handling of: no-correct-entry trials (right-censoring vs imputation), repeated rapid entries, missing box IDs, invalid action codes, and non-monotonic timestamps. If $10,800\,\mathrm{s}$ is a censoring threshold, label it as such and state how censoring would be modeled (or, if not modeled, justify and discuss bias). Provide a single summary table of all metrics (names, phases, definitions, units, edge-case rules) and ensure names are consistent throughout (Sec. 2.2, Sec. 3.2.1, Sec. 3.3).

4. **Neuroimaging/TBV quantification is inconsistently described and scientifically fragile as written. Methods assume 4D DTI with multiple $b = 0$ volumes and prior skull-stripping, while Results indicate the inputs were 3D preprocessed images; TBV is computed via intensity > 0 voxel counting, which is sensitive to nonzero background and imperfect skull stripping. There is also an internal inconsistency about whether TBV extraction failed (empty plots) or succeeded but was unusable due to behavior (Sec. 2.3, Sec. 3.2.2; Fig. 4 caption vs text).**

   Recommendation: Rewrite Sec. 2.3 to clearly separate initial assumptions from the actual data encountered: report actual NIfTI dimensionality, voxel sizes, and what preprocessing had already been applied (motion/eddy, skull stripping, registration), including tools if known. Then justify and QC the TBV approach: show background intensity distributions/histograms, confirm that nonbrain voxels are truly zero (or adopt a more defensible brain mask–based volume), and include visual overlays for a subset. In Sec. 3.2.2, report how many bats yield valid TBV values after QC, plus mean/SD/range, and explicitly reconcile why any TBV-related plots are empty (e.g., merge/filtering bug, NaNs, plotting code) if TBV exists.

5. **Methods vs Results are partially conflated, and Sec. 2.4 appears truncated/misplaced (including a stray fragment). The regression formula is also repeated in an odd location (Sec. 3.1), making it unclear what was planned vs what was executed, and on which dependent variables (Sec. 2.4, Sec. 3.1, Sec. 3.3).**

   Recommendation: Reorganize to cleanly separate (i) intended modeling plan (Sec. 2.4) from (ii) what was actually run and why it is invalid (Sec. 3.3). Fix Sec. 2.4 truncation and present the full model specification once (display equation), clearly listing intended outcomes and assumptions. In Sec. 3.3 add a compact table per dependent variable: N, missingness, variance/unique values, censoring fraction, whether the model was fit, and whether outputs were invalid (NaN, singular, $-\infty$ R^2). Remove the formula from Sec. 3.1 unless explicitly needed for Results narrative.

6. **Statistical modeling is demonstrably invalid given constant/degenerate outcomes, but the manuscript does not turn this into explicit, automated "fail-fast" guardrails; additionally, there is at least one internal inconsistency suggesting a model where $\mathrm{DNAmAge}{\mathrm{scaled}}predictsitself("predict\mathrm{DNAmAge}") (Sec. 3.3 vs Sec. 2.4).}} reported a significant effect of $\mathrm{DNAmAge}_{\mathrm{scaled}**

   Recommendation: First, correct the dependent-variable naming/error in Sec. 3.3 to match the model in Sec. 2.4 (or explicitly provide the alternative model if DNAmAge was ever an outcome). Second, add explicit pre-model validation criteria in Sec. 3.3/Sec. 4.4 (and ideally in code): checks for variance > 0, minimum unique values, plausible ranges, missingness thresholds, and censoring proportion; if violated, abort model fitting and emit a structured error. Explicitly state that any apparent significance under constant outcomes is a red-flag artifact and should not be reported as a result.

7. **Reproducibility is not yet adequate for a paper whose core contribution is pipeline robustness/failure analysis: readers cannot inspect code, environment, logs, or a minimal reproducer of the Excel-format problem (Sec. 2.1–2.4, Sec. 3.2.1).**

   Recommendation: Provide (preferably in a public repository and/or Supplement): (1) scripts/notebooks with a commit hash, (2) computational environment (Python + package versions; OS), (3) representative log outputs showing the parsing failures and validation summaries, and (4) a small synthetic dataset that mimics the problematic Excel structure (e.g., multi-row headers, merged cells, shifted columns) so others can reproduce the failure and the fix without access to sensitive raw data. If raw data cannot be shared, state constraints explicitly and provide synthetic stand-ins + schema documentation.

8. **The manuscript's lessons are currently somewhat ad hoc and lightly connected to established best practices in multimodal data organization/QC (e.g., BIDS-like conventions, schema validation, unit testing, standardized neuroimaging QC). This limits "bigger picture" impact (Sec. 1, Sec. 3.4, Sec. 4.4).**

   Recommendation: Expand Sec. 1 and Sec. 4.4 to situate the case study in existing methodological frameworks: cite and briefly map observed failure modes (format assumptions, silent empty parses, dimension mismatches, degenerate outcomes) onto best practices such as schema validation, unit/integration tests, staged QC reports, and standardized data layouts (e.g., BIDS principles where applicable). Then translate that into a concrete checklist/table: Assumption → Quick validation → Failure signature → Automated guardrail → Remediation.

9. **Internal inconsistencies in cohort summaries and reporting reduce credibility and make it harder to track what dataset underlies each stage: sex counts differ (18/15 vs 19/14) and DNAmAge summary differs (9.84 ± 1.91 vs 9.60 ± 1.74 with same range). These are small but foundational bookkeeping errors for a pipeline-validation paper (Sec. 2.1 vs Sec. 3.1; Fig. 1/caption).**

   Recommendation: Recompute and reconcile all cohort descriptors from the final analytic dataframe used for harmonization (the **33** bats). Report the definitive sex/origin counts and DNAmAge mean/SD/range once (and reuse consistently across Sec. 2.1, Sec. 3.1, and Fig. 1). If differences arise from rounding or different inclusion filters (e.g., before/after exclusions), label them explicitly and show a short included-vs-excluded comparison table.

10. **Figures are currently not optimally serving the methodological narrative: several are low-resolution, some panels are redundant due to constants/emptiness, and some plots can be misleading (e.g., regression lines on invalid outcomes; empty TBV panels without clear cause) (Figs. 1–6; Sec. 3.2–3.3).**

    Recommendation: Export figures as vector (PDF/SVG) or $\geq$ **300** dpi, enlarge fonts, add panel labels, axis units, and sample-size annotations. Replace redundant constant/empty multi-panels with compact diagnostic tables (min=max, SD=0, \%missing) and/or more informative pipeline diagnostics (e.g., printed column-name mismatches, post-merge missingness heatmaps). Remove/gray out fitted regression lines where modeling is invalid and annotate panels explicitly as "artifact of parsing failure."

Minor issues

1. Key biological/methodological context is underspecified for DNAmAge and its interpretation: the epigenetic clock source is not clearly cited/defined, and it is unclear whether DNAmAge derives from skin (variable names suggest this) and what that implies for brain-aging inference (Sec. 1, Sec. 2.1).

   Recommendation: Add citations and a short description of the DNAmAge estimator (tissue, platform, training context, expected error), and clarify how tissue choice affects interpretation (e.g., systemic vs brain-specific aging). Define DNAmAge units and any preprocessing (e.g., scaling, residualization) in Sec. 2.1/2.4.

2. Subject inclusion/exclusion and missing-data handling beyond missing DTI are not fully described; selection bias cannot be assessed (Sec. 2.1, Sec. 3.1).

   Recommendation: Explicitly list all planned and actual exclusion criteria for methylation, behavior, and imaging. State whether any behavioral sessions were incomplete/corrupted (even if parsing failed). Provide a small included vs excluded summa-

ry (age/sex/origin) to assess potential bias.

3. Metric naming and terminology are inconsistent (e.g., perseveration vs perseverance; binary vs count variables sharing names), which complicates linking definitions to plots and models (Sec. 2.2, Sec. 3.2.1, Sec. 3.3).

   Recommendation: Standardize variable names (single snake_case convention) and ensure each metric has a unique name consistent across text, figures, and any equations. Add a one-page metric dictionary table (name, phase, definition, units).

4. Ethics/animal welfare and approvals are not mentioned despite animal behavioral testing and imaging (Sec. 2).

   Recommendation: Add an Ethics statement (end of Sec. 2 or separate subsection) with institutional approvals/protocol identifiers and guidelines followed. If the study is secondary analysis and approvals are not retrievable, state this clearly and describe governance constraints.

5. The paper would be strengthened by explicitly listing stop/go criteria after each pipeline stage (harmonization, behavior parsing, TBV extraction, model fitting), rather than only narrating failures after the fact (Sec. 3.1–3.3, Sec. 4.4).

   Recommendation: Add a short staged QC protocol: for each stage, specify required outputs, minimal summary stats/plots, and explicit thresholds that must be satisfied to proceed (e.g., $\geq \boldsymbol{X}$ non-missing events per phase; outcome variance $> \mathbf{0}$; TBV within plausible range; successful merges with $\geq \boldsymbol{Y}$ subjects).

6. Title/abstract emphasis on "multimodal data integration" may overpromise relative to the actual integration demonstrated (primarily a single regression specification) (Sec. 1, Sec. 2.4).

   Recommendation: Either adjust the title/abstract to emphasize "pipeline validation/failure analysis in a multimodal context," or briefly expand Sec. 2.4/Sec. 4.4 to discuss alternative integration strategies that were planned/possible (without presenting results), clearly separating them from executed analyses.

7. Abbreviations and some task details are sparse for readers unfamiliar with the foraging paradigm (TBV, DTI, STM/LTM; phase structure, box layout, reversal definition) (Sec. 1, Sec. 2.2).

   Recommendation: Define all abbreviations at first use and add a concise task schematic description (boxes, phases, durations/delays, what constitutes an entry, how reversals are encoded) so behavioral variables are interpretable independent of the code.

8. Software environment details are missing (versions of Python/pandas/nibabel/statsmodels; OS), which matters for a reproducibility-focused paper (Sec. 2.1–2.4).

Recommendation: Add a brief "Computational environment" subsection in Sec. 2 listing software and package versions, and indicate where code/logs are archived.

9. Tone occasionally uses strong language (e.g., repeated "catastrophic"), which may distract from the technical message (Sec. 3.4, Sec. 4.4).

   Recommendation: Edit for a neutral technical tone (e.g., "critical failure mode," "pipeline-breaking assumption," "invalid downstream inference") and reserve stronger wording for clearly defined failure criteria.

Very minor issues

1. Typographical/formatting inconsistencies (e.g., "skull-scimming" vs skull-stripping; broken words due to line breaks; inconsistent heading formatting such as a stray leading "#"; inconsistent LaTeX unit formatting) (Sec. 2.3, Sec. 3.2.2, Sec. 4.4).

   Recommendation: Proofread and standardize: correct typos, remove broken-word line breaks, harmonize heading styles, and use consistent unit formatting (e.g., $\mathbf{0.25\,mm^3}$).

2. Scaling ("z-scoring") is mentioned without an explicit formula or clarity on the sample used to compute mean/SD (Sec. 2.4).

   Recommendation: State the scaling explicitly as $\boldsymbol{(x - \text{mean})/\text{SD}}$ and specify whether mean/SD are computed on the final **33**-bat analytic cohort per variable (and after exclusions/missingness handling).

3. Minor figure accessibility issues (inconsistent panel labeling, dense gridlines, unclear category ordering) (Figs. 1–6).

   Recommendation: Add panel letters to all multi-panel figures, simplify grids, order categories logically (e.g., by phase), and ensure readability in grayscale/colorblind-safe palettes.

Mathematical consistency audit

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

Maths relevance: light

The paper contains limited formal mathematics: definitions of engineered behavioral metrics, a voxel-volume-based TBV computation, standardization (z-scoring), and a multiple-linear-regression model with an interaction term. No multi-step derivations are presented; the primary audit focus is internal consistency of definitions, variable roles, and dimensional/unit logic.

Checked items

1. ✔ **Phase duration conversion** (Sec. 2.2, p.3)

- **Claim:** A **3**-hour phase duration equals $10,800$ seconds and is used as a censoring value when no correct entry occurs.
- **Checks:** algebra/arithmetic consistency, definition consistency
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** **1** hour = **3600** seconds, Phase duration is exactly **3** hours
- **Notes:** $3 \times 3600 = 10800$ is correct and consistently used later when describing uniform latencies.

2. ⚠ **Latency-to-first-correct definition uses Absolute_Time** (Sec. 2.2, p.2–3)
    - **Claim:** Latency is defined as the 'Absolute_Time' of the first correct entry (and set to $10,800$ s if not found).
    - **Checks:** definition consistency, units/dimensional sanity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Column 'Absolute_Time' could represent elapsed time or wall-clock time, Latency should be an elapsed duration from a phase start
    - **Notes:** If 'Absolute_Time' is a wall-clock timestamp, it must be referenced to a phase start time to become a latency; that subtraction is not stated. If it is already elapsed seconds since start, the definition is fine but should be clarified.

3. ✔ **Error count definition (pre-first-correct)** (Sec. 2.2, p.3 (P1_Errors, P2_Errors, P3_Errors))
    - **Claim:** Errors are the number of incorrect entries occurring before the first correct entry in the phase.
    - **Checks:** definition consistency, logical coherence
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** A 'correct' box is defined per phase, Entries are discrete events with identified box number
    - **Notes:** The definition is coherent and consistently described across phases.

4. ✔ **Binary perseveration indicators** (Sec. 2.2, p.3 (P2_Perseveration_Binary; P3_Perseveration_LTM))
    - **Claim:** Binary variables indicate whether the first entry in a phase went to the previously correct box (test1 for test2; test2 for test3).
    - **Checks:** definition consistency, logical coherence
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** A unique 'first entry' exists per phase (or is defined when none exist)

- **Notes:** Definitions are consistent; however, the manuscript does not specify what happens if no entries exist in a phase (edge case), though later Results imply this occurred (0 entries) and values defaulted.

5. ✔ **TBV voxel volume computation** (Sec. 2.3, p.3)

   - **Claim:** Voxel dimensions $0.5\,\mathbf{mm} \times 0.5\,\mathbf{mm} \times 1.0\,\mathbf{mm}$ imply voxel volume $0.25\,\mathbf{mm}^3$; TBV equals non-zero voxel count times voxel volume.
   - **Checks:** units/dimensional consistency, algebra/arithmetic consistency
   - **Verdict:** PASS; confidence: high; impact: moderate
   - **Assumptions/inputs:** Voxel dimensions are in mm, Binary mask counts voxels
   - **Notes:** $0.5 \times 0.5 \times 1.0 = 0.25\,\mathbf{mm}^3$. Multiplying by a voxel count yields $\mathbf{mm}^3$, consistent with a volume measure.

6. ✔ **Averaging $b = 0$ volumes** (Sec. 2.3, p.3)

   - **Claim:** If the first three volumes are $b = 0$ images in a 4D NIfTI, averaging across the 4^{th} dimension yields a single 3D reference image.
   - **Checks:** linear algebra/array-shape logic, definition consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** Input data are 4D with at least three volumes, The first three volumes correspond to $b = 0$
   - **Notes:** The array operation described is logically correct given the assumptions. Results later state the assumption failed for their data, but the math/logic of the operation itself is consistent.

7. ⚠ **Z-scoring standardization** (Sec. 2.4, p.4)

   - **Claim:** DNAmAge and TBV are standardized (z-scored) to reduce multicollinearity and ease interaction interpretation.
   - **Checks:** definition completeness, logical/statistical coherence
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** Standard z-score is $(x - \mathbf{mean})/\mathbf{SD}$ with $\mathbf{SD} > 0$
   - **Notes:** No explicit formula is provided and it is not stated which sample (final cohort) is used to compute mean/SD, though the intent is standard. This does not create a direct contradiction but leaves verification incomplete.

8. ✔ **Regression model specification with interaction** (Sec. 2.4, p.4)

   - **Claim:** For each cognitive outcome, fit: $\text{Cognitive\_Metric} \sim \mathrm{DNAmAge}\{\mathrm{scaled}\} + \mathrm{TBV}\}\} + \mathrm{DNAmAge}\{\mathrm{scaled}\}\times \mathrm{TBV}\$.\}\} + \text{Sex} + \text{Origin}$
   - **Checks:** notation consistency, model structure consistency
   - **Verdict:** PASS; confidence: high; impact: critical

- **Assumptions/inputs:** \text{Cognitive_Metric} is a scalar response per bat, Sex and Origin are encoded as categorical covariates
- **Notes:** The formula matches the narrative description and the stated moderation hypothesis.

9. ✖ **Results describe predicting $\mathrm{DNAmAge}{\mathrm{scaled}} from \mathrm{DNAmAge}$** (Sec. 3.3, p.6)

   - **Claim:** An example model is described as 'attempting to predict $\mathrm{DNAmAge}{\mathrm{scaled}}' and reporting a significant effect of \mathrm{DNAmAge}$.}
   - **Checks:** symbol/variable-role consistency, logical coherence
   - **Verdict:** FAIL; confidence: high; impact: critical
   - **Assumptions/inputs:** The Methods model uses a cognitive metric as dependent variable
   - **Notes:** This statement conflicts with the Methods model (dependent variable should be \text{Cognitive_Metric}). It also implies $\mathbf{DNAmAge}_{\mathrm{scaled}}$ is both response and predictor, which is internally inconsistent as written. Likely a variable-name slip (e.g., should refer to \text{P1_Latency} as dependent), but the text as-is is contradictory.

10. ✔ **Claim about nan outputs for zero-variance dependent variables** (Sec. 3.3, p.6)

    - **Claim:** When the dependent variable has zero variance (constant), regression outputs (coefficients/SE/t/p) become nan.
    - **Checks:** logical/statistical sanity check
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Regression procedure requires an estimable residual variance for standard errors and test statistics
    - **Notes:** A constant response can make variance-based quantities undefined; the narrative is plausible and consistent with the stated data-extraction failure. Exact software behavior is not verifiable from the paper alone, but the direction (ill-posed inference) is consistent.

11. ✖ **Internal consistency of cohort summary descriptions** (Sec. 2.1, p.2 vs Sec. 3.1, p.4)

    - **Claim:** Methods and Results report cohort mean/SD DNAmAge and sex counts for the same final cohort ($n = 33$).
    - **Checks:** definition/statement consistency (non-derivation)
    - **Verdict:** FAIL; confidence: high; impact: minor
    - **Assumptions/inputs:** Both sections refer to the same final analysis cohort

- **Notes:** Methods report mean **9.84** (SD **1.91**), **18** males/**15** females; Results report mean **9.60** (SD **1.74**), **19** males/**14** females. Without recalculating, the issue is that the manuscript provides conflicting descriptive summaries for what appears to be the same cohort.

Limitations

- The manuscript contains very few explicit equations and no step-by-step derivations; most checks are on definitional and dimensional consistency rather than algebraic derivations.
- Only the provided PDF text/pages were available; no supplementary materials, code, or raw tables were included to disambiguate variable definitions such as whether 'Absolute_Time' is elapsed time or wall-clock time.
- Some reported issues (e.g., nan outputs, $-\infty$ R^2) depend on software conventions; this audit only assesses whether the statements are internally coherent given the described data pathologies, not whether a specific library must output exactly those values.

Numerical results audit

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

11 numeric consistency checks were executed; all passed within stated exact/near-exact tolerances. However, two cross-section inconsistencies were identified via check notes: sex counts differ between Methods vs Results/Fig.1, and DNAm age mean/SD differ between Methods vs Results/Fig.1 (range matches in both). Several additional model/data-dependent claims could not be verified from the available numeric detail.

Checked items

1. ✔ **C1** (Methods p.2 §2.1 (Data preparation and harmonization))
    - **Claim:** Final cohort of **33** bats selected from an initial pool of **41** subjects.
    - **Checks:** difference_check
    - **Verdict:** PASS
    - **Notes:** Computed excluded_n = **8**.
2. ✔ **C2** (Results p.4 §3.1 (Cohort description); also Methods p.2 §2.1)
    - **Claim:** From initial pool of **41**, 8 subjects were excluded, leaving final cohort **33**.
    - **Checks:** parts_to_total
    - **Verdict:** PASS
    - **Notes:** Computed final_cohort_n = **33** from initial_pool_n−excluded_n.
3. ✔ **C3** (Methods p.2 §2.1 (EDA cohort characterization))

- **Claim:** Cohort comprised 18 males and 15 females (total 33).
- **Checks:** parts_to_total
- **Verdict:** PASS
- **Notes:** Parts sum to 33. Counts differ vs C4 (males=19, females=14).

4. ✔ **C4** (Results p.4 §3.1 and Fig.1 caption p.5)
   - **Claim:** Cohort consisted of 19 males and 14 females (total 33).
   - **Checks:** parts_to_total
   - **Verdict:** PASS
   - **Notes:** Parts sum to 33. Counts differ vs C3 (males=18, females=15).
5. ✔ **C5** (Methods p.2 §2.1; Results p.4 §3.1; Fig.1 caption p.5)
   - **Claim:** Origins: Aseret $n = 19$ and Herzeliya $n = 14$ (total 33).
   - **Checks:** parts_to_total
   - **Verdict:** PASS
   - **Notes:** Origins sum to 33.
6. ✔ **C6** (Methods p.2 §2.1 (EDA cohort characterization))
   - **Claim:** Mean DNAm age 9.84 years (SD 1.91), range 6.62–15.07 years.
   - **Checks:** range_contains_mean
   - **Verdict:** PASS
   - **Notes:** Checked min≤mean≤max: $6.62 \leq 9.84 \leq 15.07$.
7. ✔ **C7** (Results p.4 §3.1; Fig.1 caption p.5)
   - **Claim:** Mean DNAm age 9.60 years (SD 1.74), range 6.62–15.07 years.
   - **Checks:** range_contains_mean
   - **Verdict:** PASS
   - **Notes:** Checked min≤mean≤max: $6.62 \leq 9.6 \leq 15.07$. Mean/SD differ vs C6 (mean=9.84, sd=1.91).
8. ✔ **C8** (Methods p.3 §2.2 (Phase 1 metric definition); also Results p.5 and Fig.2 caption)
   - **Claim:** Phase duration 3 hours = $10,800$ seconds (used as max latency).
   - **Checks:** unit_conversion
   - **Verdict:** PASS
   - **Notes:** Computed $3 \times 3600 = 10800$ seconds.
9. ✔ **C9** (Methods p.3 §2.3 (Brain volume quantification))
   - **Claim:** Voxel dimensions $0.5\,\mathrm{mm} \times 0.5\,\mathrm{mm} \times 1.0\,\mathrm{mm}$ imply voxel volume $0.25\,\mathrm{mm}^3$.
   - **Checks:** multiplicative_consistency

- **Verdict:** PASS
- **Notes:** Computed $dx \times dy \times dz = 0.25$.

10. ✔ **C10** (Methods p.3 §2.2 (Behavioral feature engineering))
    - **Claim:** Nine quantitative spatial memory metrics are listed across three phases (2 in Phase 1, 4 in Phase 2, 3 in Phase 3).
    - **Checks:** count_consistency
    - **Verdict:** PASS
    - **Notes:** Summed phase counts: $2 + 4 + 3 = 9$.
11. ✔ **C11** (Results p.5 §3.2.1; Fig.2 caption p.9; Fig.5 caption p.7)
    - **Claim:** Uniform maximum latencies reported as $10,800$ seconds for all subjects across phases (P1, P2, P3).
    - **Checks:** repeated_constant_consistency
    - **Verdict:** PASS
    - **Notes:** Compared constants: 10800 vs 10800.

Limitations

- Only parsed text from the PDF was available; no tables of raw values or model summaries were included to enable recomputation of reported statistics (e.g., regression coefficients, p-values, R^2).
- Figures are referenced but their numeric content is not extractable without reading plot pixels; such checks were excluded per instructions.
- Many claims (e.g., missingness checks, file dimensionality, number of non-zero voxels, TBV distributions) depend on underlying datasets (CSV/Excel/NIfTI) not included in the PDF, so they cannot be verified here.