Skepthical review: *Unveiling Predictive Neural Signatures of Cognitive Adaptability in Aging Bats: A Multi-Region DTI and Machine Learning Approach*

Summary

The manuscript investigates how brain microstructure and age jointly shape cognitive "strategy" in Egyptian fruit bats. The authors use a three-phase spatial re-learning task to derive four behavioral metrics (short-/long-term relearning speed and perseverative errors), reduce them via PCA into a Cognitive Adaptability Index (CAI), and then model CAI using diffusion tensor imaging (DTI) mean diffusivity (MD) across atlas regions, epigenetic age (DNAmAge), sex, and colony (Sec. 2.1–2.7; Sec. 3.1–3.3). Static prediction models (main effects only) reportedly yield negative cross-validated R^2, while adding DNAmAge×MD interaction terms in ElasticNet highlights negative interactions in Regions 9, 22, and 23, interpreted as age-modulated brain–behavior relationships consistent with an aging-related shift toward short-term flexibility at the expense of long-term memory (Sec. 3.2–3.3; Fig. 12).

The multi-modal integration and the conceptual framing of aging as a potential strategic reconfiguration (rather than a unidimensional decline) are appealing and potentially novel. However, the current manuscript has substantial internal inconsistencies (sample size and feature counts), under-specified imaging and modeling methods, and an inference/validation gap: selected ElasticNet interaction coefficients are described as "statistically significant" despite a $p \gg N$ setting, unclear nesting of preprocessing within cross-validation, limited reporting of interaction-model generalization performance, and no stability/multiple-comparisons control. Additionally, the anatomical identity of the implicated regions is not provided, limiting mechanistic interpretation. Addressing these issues would markedly improve reproducibility, interpretability, and the credibility of the age-modulated conclusions.

Strengths

- Conceptually interesting framing of cognitive aging as a shift in strategy (flexibility vs stability/long-term memory), rather than a simple decline, and an explicit attempt to formalize this via CAI (Sec. 1; Sec. 3.1; Sec. 4).
- Behavioral paradigm is clearly motivated: a three-phase re-learning design that naturally separates short-term updating from longer-term re-learning/perseveration (Sec. 2.1).
- Multi-modal dataset (behavior + DNAmAge + region-wise DTI MD + demographics) in a non-traditional long-lived mammalian model, which could be broadly valuable (Sec. 2.1–2.4).
- Use of regularized regression and Random Forests with cross-validation acknowledges the high-dimensional/small-sample nature of the predictors (Sec. 2.6–2.7).

- Transparency in reporting that "static" models produce negative cross-validated R^2 (i.e., worse than mean-only baseline) is an important and honest diagnostic (Sec. 3.2).
- Initial exploration of age-modulated effects via explicit DNAmAge×MD interactions and visualization via partial dependence is aligned with the core hypothesis (Sec. 2.7; Sec. 3.3).

Major issues

1. **Internal inconsistencies in sample size and feature dimensionality undermine interpretability and reproducibility. Methods state a final analytical cohort of $N = 31$ and MD extracted from 82 regions (Sec. 2.1–2.4.3; Sec. 2.2), whereas Results and figure captions refer to $N = 33$ (Sec. 3; Fig. 1), and Sec. 3.2 reports modeling with MD from 24 regions. The sex distribution is also inconsistent between Methods and Fig. 1. These discrepancies directly affect model dimensionality (p), regularization behavior, and any claims about selected regions/interactions.**

   Recommendation: Provide a single, definitive accounting of: (i) N per modality (behavior, DNAmAge, DTI), (ii) N used for CAI/PCA, (iii) N used for each model (static ElasticNet, static RF, interaction ElasticNet, PDP analyses), and (iv) exclusions with reasons (QC failures, missing files, aborted behavior), ideally as a table or CONSORT-style flow diagram in Sec. 2.4.3. Harmonize N and sex counts/percentages across Sec. 2–3 and figure captions. Explicitly state the exact number of MD regions used in each analysis stage, and justify any reduction from 82 to 24 (QC, missingness, atlas subset, variance filtering), including a list of retained regions (Appendix is fine).

2. **DTI acquisition, preprocessing, atlas registration, and MD extraction are under-specified, preventing evaluation of data quality and anatomical validity (Sec. 2.2). Key details are missing (scanner/field strength; TR/TE; voxel size; b-values; #directions; distortion/motion/eddy correction; tensor fitting method; registration steps; QC; partial volume mitigation; hemisphere handling). Without these, it is difficult to interpret MD as "microstructural integrity" or to assess potential artifacts driving region effects.**

   Recommendation: Expand Sec. 2.2 substantially (or add an Appendix) to include full acquisition parameters and preprocessing steps (skull stripping, motion/eddy/susceptibility correction, tensor fitting algorithm/software, registration pipeline to atlas/template, QC criteria and exclusion counts). Describe how atlas labels were transformed to individual space, whether ROI erosion/thresholding was used to reduce boundary contamination/CSF partial volume, and whether hemispheres were averaged. Report MD units/scaling and any normalization across subjects.

3. **CAI definition/interpretation is internally inconsistent and construct validity is not sufficiently justified. Methods state PC1 loadings were intended to be negative for all four metrics so higher CAI indicates "better adaptability" (Sec. 2.5.2), but Results show mixed-sign loadings and interpret CAI as a trade-off axis (high CAI = better long-term but worse short-term; Sec. 3.1; Table 1). Moreover, it is not established that a single scalar PC1 is the most appropriate outcome vs a 2D (PC1/PC2) characterization or direct modeling of the four metrics.**

   Recommendation: Rewrite Sec. 2.5.2 and Sec. 3.1 to match the actual CAI used: define CAI as the signed PC1 score (explicitly state whether PC1 was multiplied by -1) and whether it is a "trade-off/strategy axis" vs a monotone "adaptability" score. Report PCA details (N used; eigenvalues; variance explained by PC1/PC2; full loading matrix at least for PC1–PC2; scree plot/biplot in supplement). Provide a brief construct-validity argument for why PC1 is the target outcome, and consider adding a robustness analysis: (i) predict PC1 and PC2 separately, or (ii) predict the four behavioral metrics via multivariate/multi-task regression, or (iii) define *a priori* composites for "short-term flexibility" (Phase 2) and "long-term memory/stability" (Phase 3) and show conclusions are consistent.

4. **Cross-validation pipeline and potential information leakage are not sufficiently specified. It is unclear whether CAI PCA, feature standardization, and interaction-term construction were performed globally once or nested within the outer LOOCV folds (Sec. 2.5–2.7). If PCA/scaling were computed on the full dataset prior to CV, this can bias generalization estimates and feature selection, especially with $N \approx 31$ and high-dimensional predictors.**

   Recommendation: In Sec. 2.5–2.7, explicitly state the exact order of operations inside the outer LOOCV loop. Best practice is: within each outer training fold, fit PCA for CAI (if CAI is treated as derived from data), compute scaling parameters, construct DNAmAge×MD interactions, tune hyperparameters via inner CV, fit the final model, then evaluate on the held-out subject. If CAI PCA was computed once on the full cohort, acknowledge this as a limitation and add a sensitivity analysis recomputing CAI within folds (or show empirically that results are unchanged). Confirm all preprocessing is recomputed within training folds to avoid leakage.

5. **Inference and evidential framing for DNAmAge×MD interactions are currently not supportable as written. The interaction model is a $p \gg N$ search (e.g., demographics + 82 MD main effects + 82 DNAmAge×MD interactions; Sec. 2.7.1), yet selected non-zero ElasticNet coefficients are described as "statistically significant" (Sec. 2.7.1; Sec. 3.3; Table 2) without a defined hypothesis-testing framework, multiple-comparisons control, or coefficient stability assessment. Additionally, the manuscript does not clearly**

report the cross-validated performance of the interaction-augmented model; if CV R^2 remains low/negative, strong mechanistic claims are not warranted.

Recommendation: Replace "statistically significant" with "selected/non-zero under ElasticNet" unless a formal inferential procedure is added. Report full out-of-sample performance for the interaction model (LOOCV R^2 plus MAE/RMSE) and provide uncertainty (e.g., bootstrap over subjects or repeated CV where feasible). Add at least one robustness/inference layer appropriate for high-dimensional selection: (i) permutation testing of the entire modeling pipeline to assess whether performance exceeds chance; (ii) stability selection / bootstrap inclusion frequencies for interactions (how often each DNAmAge×MD term is selected across resamples); and/or (iii) an *a* priori ROI interaction test set with FDR correction. Temper Sec. 3.3 and Sec. 4 to present interaction findings as exploratory/hypothesis-generating pending replication.

6. **Random Forest methodology and interpretation (especially age-stratified PDPs) are under-specified and may overstate evidence for moderation. RF hyperparameters beyond n_estimators/max_features are not reported (Sec. 2.6.2; Sec. 3.2–3.3), and the stability of permutation importance is unclear. Moreover, splitting bats into "younger/older" by median DNAmAge for PDP comparisons yields small groups (≈ 15 each) and can create unstable apparent differences; PDPs also assume feature independence and can be misleading in correlated predictors (Sec. 2.7.2; Sec. 3.3; Fig. 12).**

   Recommendation: Report the complete RF setup (tuned ranges and final values for max_depth, min_samples_leaf, min_samples_split, bootstrap, random_state, etc.) and provide stability checks for feature importance across seeds/resamples (Appendix acceptable). For moderation claims, clarify whether PDPs were computed from a single global model or separate models per age group; report group *N*s and data density across MD ranges; and soften language to "suggestive." Consider adding ICE curves or uncertainty bands, or using ALE plots. If the key claim is interaction/moderation, prioritize interaction-capable models (with stability/inference) over visual PDP group contrasts.

7. **Anatomical interpretability is currently blocked because regions are referenced only by atlas indices (Regions 9, 22, 23; also 3, 4, 19) with no mapping to anatomical structures or tissue types (Sec. 2.2; Sec. 3.2–3.3). This prevents readers from assessing biological plausibility (e.g., hippocampal/striatal/frontal involvement) and connecting to cognitive aging literature.**

   Recommendation: Provide an atlas index→anatomical label table (main text or Appendix) and identify whether each ROI is gray/white matter. In Sec. 3.2–3.3 and figure captions, refer to both index and anatomical name (e.g., "Region 22 (hippocampal

formation, if applicable)"). Add a figure visualizing the implicated ROIs on the atlas/template. In Sec. 4, discuss these structures in relation to re-learning, perseveration, and known age-related microstructural changes.

8. **Confounding, representativeness, and limitations are not comprehensively addressed given the small N and multi-modal integration. Potential confounds include scan/batch effects, time lag between behavior and imaging, health/body condition, housing/handling, and colony-specific differences; MD is non-specific and DNAmAge measurement error/clock calibration is not described in enough detail (Sec. 2.3; Sec. 4).**

   Recommendation: Add a dedicated limitations paragraph/subsection in Sec. 4 addressing: small sample size and selection instability; $p \gg N$ interaction search; MD's biological non-specificity and partial volume sensitivity; DNAmAge uncertainties and (if available) correlation with chronological/minimum age; possible batch/time-between-measures confounds; and generalizability across colonies/sex. Where possible, report key correlations (DNAmAge with any chronological estimates; DNAmAge with colony/sex; time between scan and behavior) and consider adding batch/time covariates or sensitivity analyses.

Minor issues

1. Behavioral protocol details and metric edge cases are not fully specified, affecting reproducibility and metric well-posedness (Sec. 2.1; Sec. 2.5–2.5.1). RSI is undefined if a bat never makes a correct entry; PEI is undefined if total entries are 0, and both metrics may be noisy for very low entry counts. Termination rules and opportunities for errors across phases are not fully described.

   Recommendation: Expand Sec. 2.1 and Sec. 2.5–2.5.1 with: trials/time limits per phase, stopping criteria, inter-phase intervals, handling of aborted sessions, and coding of events. Explicitly define conventions for "no correct entry" and "zero entries" cases (e.g., censoring, assigning maximum RSI, exclusion) and report how many subjects/phases were affected. Consider a robustness check excluding low-entry phases or adding exposure normalization (e.g., controlling for total entries/time).

2. Epigenetic age (DNAmAge) methods are described too briefly to interpret the measure as an aging proxy (Sec. 2.3). Platform, preprocessing, clock training/validation, and expected error are not reported; the relationship to chronological age (if known) is not shown.

   Recommendation: In Sec. 2.3, add assay platform and preprocessing steps; cite the epigenetic clock model and its validation performance in this or related bat species; and, if any chronological/minimum age information exists, report DNAmAge–chronological age correlation and bias. If chronological age is unknown, state explicitly that DNAmAge is treated as a relative aging index.

3. Missing-data handling and data integration are described procedurally but not quantitatively (Sec. 2.4.1–2.4.3). Without counts of dropped records per merge/QC step, it is hard to assess selection bias (e.g., older animals disproportionately missing DTI).

   Recommendation: Add a quantitative merge/QC summary in Sec. 2.4.3 (counts per dataset, counts after joins, reasons for exclusions). Briefly compare the final multimodal subset to the full behavioral cohort on DNAmAge/sex/colony to assess representativeness.

4. Interaction-term construction and scaling are not fully detailed (Sec. 2.7.1). It is unclear whether interaction features were re-standardized after multiplication and whether all corresponding main effects were retained, which affects interpretability and regularization.

   Recommendation: Clarify in Sec. 2.7.1: (i) standardization of DNAmAge and MD prior to multiplication, (ii) whether interactions were standardized afterward, (iii) that DNAmAge and MD main effects remain in the model alongside interactions, and (iv) the total predictor count in each interaction model.

5. Figures and captions sometimes omit essential quantitative context (Ns per panel, performance metrics for each model, uncertainty intervals) and have legibility/mislabeling issues (notably around Figs. 9–12 and reported figure references in Sec. 3.2–3.3).

   Recommendation: Increase font/axis label sizes; add N annotations per analysis; report model performance (CV R^2, MAE/RMSE) directly in relevant figure panels/captions; and include uncertainty/stability summaries for coefficients/importances. Fix mis-references (e.g., "Figure 3.2" vs Fig. 9) and ensure Fig. 12 caption/panels match the regions discussed in Sec. 3.3.

6. Terminology such as "static features" is used without a concise definition (Sec. 1; Sec. 2.6; Sec. 3.2), and causal/mechanistic phrasing sometimes outpaces the evidence given negative R^2 and exploratory selection (Sec. 3.3; Sec. 4).

   Recommendation: Define "static" explicitly (e.g., main effects only: DNAmAge, MD, sex, colony; no interactions) near first use. Review Sec. 3.3 and Sec. 4 for wording that implies confirmatory evidence; align phrasing with exploratory modeling and reported generalization.

7. Ethics and animal welfare reporting is minimal (Sec. 2). Standard animal research reporting typically includes named oversight body, protocol identifiers, anesthesia/monitoring details for imaging, and stress-minimization steps for behavior.

   Recommendation: Add an Ethics subsection or expand Sec. 2 with committee name(s), protocol numbers, guideline compliance, anesthesia/analgesia and monitoring for scanning, and behavioral welfare measures (habituation, criteria for stopping).

8. Broader theoretical context could be strengthened: the flexibility–stability trade-off, exploration–exploitation, set-shifting, and compensatory recruitment in aging have substantial cross-species literature that would help position the CAI framing (Sec. 1; Sec. 4).

   Recommendation: Expand Sec. 1 and Sec. 4 with a few targeted references and explicit connections to related constructs and findings in cognitive aging and white-matter microstructure literature, including caution about MD interpretation across tissue types.

Very minor issues

1. Formatting/style inconsistencies (section numbering around Sec. 2.7, region naming conventions such as "Region_3" vs "Region 3", bracketed region lists in running text, inconsistent capitalization of variables) slightly reduce readability (Sec. 1–4).

   Recommendation: Standardize section heading hierarchy, region naming, and variable formatting throughout; use consistent typography for the interaction symbol (×) and consistent decimal precision in text/figures.

2. Typographical artifacts and awkward line breaks appear in places (e.g., words split across lines), and some figure cross-references appear incorrect (e.g., Sec. 3.2 referring to the wrong figure).

   Recommendation: Proofread carefully to remove broken words/line-break artifacts and correct all figure/table references.

3. Median split rule for DNAmAge groups is not fully specified for values exactly at the median (Sec. 2.7.2).

   Recommendation: State deterministic handling of DNAmAge equal to the median (assign to Younger/Older or exclude) and report final group sizes.

4. Affiliation text in the unstructured description appears placeholder-like (if present in the manuscript metadata).

   Recommendation: Ensure author affiliations and institutional details are real and journal-appropriate before final submission.

5. Keywords and minor presentation choices could be made more content-specific and less method-generic.

   Recommendation: Revise keywords to emphasize scientific content (e.g., cognitive flexibility/stability, epigenetic aging, DTI/MD, Egyptian fruit bats) rather than generic statistical terms.

Key statements and references

- • **Precise biological age for each bat was estimated using the DNAmAge-Bat.Rousettus.aegyptiacus_Skin epigenetic clock, which leverages DNA methylation patterns to provide an accurate approximation of chronological age in bats and thereby overcomes limitations of traditional age estimation methods in wild-derived animals.**
- *Reference(s):* DNAmAgeBat.Rousettus.aegyptiacus_Skin
- • **Static features comprising epigenetic age (DNAmAge), sex, origin colony, and Mean Diffusivity (MD) values from 24 brain regions yielded poor predictive performance for the Cognitive Adaptability Index (CAI), with Leave-One-Out Cross-Validated R^2 values of -0.05 for ElasticNet and -0.21 for Random Forest, indicating performance worse than predicting the mean CAI for all subjects.**
- *Reference(s):* $R^2 = -0.05$, $R^2 = -0.21$
- • **In ElasticNet models that included interaction terms between standardized DNAmAge and standardized regional MD values for all 82 brain regions, three statistically significant negative interaction coefficients were identified—DNAmAge × MD Region_22 (-0.017), DNAmAge × MD Region_23 (-0.007), and DNAmAge × MD Region_9 (-0.005)—demonstrating that the association between higher MD (lower tissue integrity) in these regions and CAI becomes increasingly negative with advancing epigenetic age.**
- *Reference(s):* DNAmAge × MD Region_22, DNAmAge × MD Region_23, DNAmAge × MD Region_9
- • **Partial Dependence Plot analyses stratified by age group (Younger vs. Older, defined by a median split on DNAmAge) showed clearly divergent slopes and shapes for the relationship between regional MD and predicted CAI in the key regions identified by the interaction analysis, providing visual evidence that the influence of these regions' microstructure on cognitive adaptability is dynamically modulated by age rather than remaining static across the lifespan.**
- *Reference(s):* DNAmAge

Mathematical consistency audit

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

Maths relevance: light

The paper contains relatively little explicit mathematics (few/no displayed equations). The main analytic content is in the definitions of behavioral indices, PCA-based construction of CAI, descriptions of regression models with standardized predictors, interaction terms, and interpretation of coefficients and cross-validated R^2. The key internal-consistency problems are definitional/notation mismatches (CAI meaning vs PCA loadings; number of brain-region features; cohort size) and unsupported use of the term 'statistically significant' for ElasticNet-selected coefficients.

Checked items

1. ⚠ **RSI_STM definition (count before first correct)** (Sec. 2.5.1, p.4)
    - **Claim:** RSI_STM equals the number of incorrect entries in Phase 2 before the first correct entry in Phase 2.
    - **Checks:** definition consistency, well-posedness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Phase 2 entries are temporally ordered, A 'correct entry' exists (or a convention is defined if not)
    - **Notes:** The metric is mathematically clear if at least one correct entry occurs. The manuscript does not specify what happens if the bat never enters the correct box in Phase 2 (undefined/infinite unless a convention is set).
2. ⚠ **PEI_STM definition (proportion to previous correct box)** (Sec. 2.5.1, p.4)
    - **Claim:** $PEI_{STM} = \frac{\text{entries to P1-correct box during Phase 2}}{\text{total entries in Phase 2}}$.
    - **Checks:** definition consistency, well-posedness, sanity/limits
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Total entries in Phase 2 is positive
    - **Notes:** As a ratio, PEI_{STM} should lie in $[0, 1]$ if the denominator is nonzero. The manuscript does not specify handling when Phase 2 contains zero entries (division by zero).
3. ⚠ **RSI_LTM / PEI_LTM definitions** (Sec. 2.5.1, p.4)
    - **Claim:** RSI_LTM is the number of incorrect entries before the first correct entry in Phase 3; PEI_{LTM} is the fraction of Phase 3 entries to the Phase-2-correct box.
    - **Checks:** definition consistency, well-posedness
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Temporal ordering in Phase 3, At least one Phase 3 entry; convention if none, Convention if no correct entry occurs
    - **Notes:** Same edge-case omissions as short-term metrics: division by zero for PEI_{LTM} and undefined RSI_{LTM} if no correct entry occurs are not addressed.
4. ✔ **PCA standardization step** (Sec. 2.5.2, p.4)

- **Claim:** Each of the four behavioral metrics is standardized (mean 0, SD 1) before PCA to ensure equal weighting.
- **Checks:** definition consistency, sanity/limits
- **Verdict:** PASS; confidence: high; impact: minor
- **Assumptions/inputs:** Each metric has nonzero variance across subjects included in PCA
- **Notes:** Standardizing variables before PCA is internally consistent with the stated goal of equalizing scales. The only caveat (not addressed) is if any metric has zero variance, standardization would be undefined.

5. ✖ **CAI interpretation rule vs reported PCA loadings** (Sec. 2.5.2 (p.4) vs Table 1/Sec. 3.1 (p.6))

    - **Claim:** Methods: higher CAI represents 'better adaptability' with negative loadings for all four metrics; Results: PC1 loadings are mixed-sign and CAI represents a trade-off (high CAI = worse short-term, better long-term).
    - **Checks:** definition consistency, notation consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** CAI is the (possibly sign-flipped) PC1 score
    - **Notes:** Methods asserts a monotone direction (all four loadings negative after sign alignment), but Table 1 shows PEI_{STM} and RSI_{STM} positive while RSI_{LTM} and PEI_{LTM} are negative, and the narrative interprets a trade-off. This contradicts the stated CAI alignment criterion.

6. ✔ **Consistency of trade-off interpretation with Table 1 signs** (Table 1 and paragraph below it, Sec. 3.1, p.6)

    - **Claim:** High CAI corresponds to higher PEI_{STM} and RSI_{STM} (poorer short-term) and lower RSI_{LTM} and PEI_{LTM} (better long-term).
    - **Checks:** algebra/logic, sign consistency
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** Metrics are standardized with higher RSI/PEI meaning worse performance
    - **Notes:** Given the reported loadings (PEI_{STM} +, RSI_{STM} +, RSI_{LTM} −, PEI_{LTM} −), the stated qualitative interpretation matches the signs: increasing PC1 increases short-term error indices and decreases long-term error indices.

7. ✖ **Cohort size used for analyses** (Sec. 2.4.3 (p.3) and Sec. 3 (p.5) / Fig. 1 caption)

    - **Claim:** The final cohort is $N = 31$, but Results/Fig. 1 state $N = 33$.
    - **Checks:** definition consistency, notation consistency
    - **Verdict:** FAIL; confidence: high; impact: critical

- **Assumptions/inputs:** N refers to the set of subjects used in the corresponding analysis
- **Notes:** The manuscript provides two different cohort sizes without specifying modality-specific Ns. This undermines the mathematical definition of the datasets underlying PCA and regression models.

8. ✖ **Number of MD regional features in static models** (Secs. 2.2/2.6.1 (pp.3–4) vs Sec. 3.2 (p.7))

    - **Claim:** MD is extracted for 82 regions and used as predictors, but Results state static modeling used MD from 24 regions.
    - **Checks:** notation consistency, definition consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** A single intended feature set for the static models
    - **Notes:** The predictor dimension is inconsistent across sections. If 24 is a subset, the selection/filtering rule is missing; if not, the Results statement conflicts with Methods.

9. ✔ **Negative cross-validated R^2 interpretation** (Sec. 3.2, p.7)

    - **Claim:** Negative R^2 indicates the model performs worse than predicting the mean CAI for all subjects.
    - **Checks:** algebra/logic
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** R^2 computed as $1 - SSE/SST$ on held-out predictions vs true values
    - **Notes:** Under the standard definition, cross-validated R^2 can be negative when prediction error exceeds the baseline mean predictor's error.

10. ✔ **Interaction term construction with standardized variables** (Sec. 2.7.1, p.5)

    - **Claim:** For each region, create an interaction feature DNAmAge * MD_region using standardized DNAmAge and standardized MD.
    - **Checks:** definition consistency, notation consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Both DNAmAge and MD_region are standardized using the same training data within each CV fold (or otherwise consistently)
    - **Notes:** The interaction feature definition is algebraically coherent. The manuscript does not specify fold-wise vs global standardization (a methodological detail), but symbolically the construction is consistent.

11. ✔ **Interpretation of negative age×MD coefficient** (Sec. 3.3, p.9 (text below Fig. 11/Table 2))

- **Claim:** A negative coefficient for DNAmAge × MD_Region implies the MD–CAI relationship becomes more negative at higher ages, so in older bats higher MD is more strongly associated with lower CAI.
- **Checks:** algebra/logic, sign consistency
- **Verdict:** PASS; confidence: high; impact: moderate
- **Assumptions/inputs:** Linear model includes main effects for age and MD and an interaction term, DNAmAge increases with age (after standardization, older corresponds to larger values)
- **Notes:** In a linear model $y = ... + b_2\,MD + b_3\,(Age \times MD)$, the slope $\partial y/\partial MD = b_2 + b_3\,Age$ decreases with Age if $b_3 < 0$. Thus the stated direction of moderation is mathematically correct.

12. ⚠ **Use of the term 'statistically significant' for ElasticNet-selected interactions** (Sec. 2.7.1, p.5 and Table 2/Sec. 3.3, p.9)
    - **Claim:** Non-zero ElasticNet interaction coefficients are described as 'statistically significant'.
    - **Checks:** definition consistency, missing derivation/criteria
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** Some inferential procedure exists beyond regularization-based selection
    - **Notes:** No hypothesis-testing framework, p-values, confidence intervals, or selection-stability criterion is defined. Without this, the phrase 'statistically significant' is not mathematically grounded in the document.
13. ⚠ **PDP evidence vs identified interaction regions** (Sec. 3.3, p.9 and Fig. 12 caption)
    - **Claim:** PDPs corroborate interaction findings for Regions 9/22/23, but the figure/caption indicates four regions and appears to include Regions 3 and 4.
    - **Checks:** notation consistency, definition consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** The figure panels correspond to the regions named in the caption/text
    - **Notes:** There is a mismatch between the set of regions claimed to be supported (9/22/23) and the set indicated by the figure caption/panels (four regions, seemingly including 3 and 4). Clarification is needed to make the logical support chain consistent.

Limitations

- The provided PDF content contains almost no explicit equations; many model details (objective functions, exact R^2 definition, PCA sign convention, inferential procedures) are described narratively, limiting the ability to verify derivations step-by-step.

- Several key claims depend on unspecified conventions (handling of missing/undefined behavioral metrics, feature filtering from 82 to 24 regions, and the meaning of 'statistical significance' under ElasticNet). These omissions prevent full analytic verification.
- Figure text/panel labels are only partially legible in the embedded images, so some figure-based consistency checks (e.g., exact region IDs in Fig. 12) remain uncertain.

Numerical results audit

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

Of 13 numeric checks, 9 passed and 4 failed. Failures are concentrated in cross-section consistency for cohort N (31 vs 33), sex distribution (Methods vs Figure 1), and brain-region feature counts (82 vs 24). Internal arithmetic for Methods cohort composition and several logical/format checks (percent sums, sign/order checks, PCA loading normalization) passed within stated tolerances.

Checked items

1. ✔ **CAND-01** (Page 3 (Section 2.4.3 Cohort Definition and Descriptive Statistics))
    - **Claim:** Sex Distribution: 18 Male (58.1%), 13 Female (41.9%) with Total Subjects (N): 31.
    - **Checks:** parts_vs_total_and_percentages
    - **Verdict:** PASS
    - **Notes:** Counts sum to total; stated percentages and their sum match within tolerance.
2. ✔ **CAND-02** (Page 3 (Section 2.4.3 Cohort Definition and Descriptive Statistics))
    - **Claim:** Origin Colony: 17 Aseret (54.8%), 14 Herzeliya (45.2%) with Total Subjects (N): 31.
    - **Checks:** parts_vs_total_and_percentages
    - **Verdict:** PASS
    - **Notes:** Counts sum to total; stated percentages and their sum match within tolerance.
3. ✔ **CAND-03** (Page 3 (Section 2.4.3 Cohort Definition and Descriptive Statistics))
    - **Claim:** DNAmAge (Years): Mean: 9.78, SD: 1.83, Range: 6.62 - 15.07.
    - **Checks:** range_mean_feasibility
    - **Verdict:** PASS
    - **Notes:** Logical consistency holds (min $\leq$ mean $\leq$ max, sd ≥ 0).
4. ✔ **CAND-04** (Page 5 (Section 3 RESULTS opening paragraph + Figure 1 caption))

- **Claim:** Results section states cohort comprised $N = 33$ Egyptian fruit bats; Figure 1 caption also says $N = 33$.
- **Checks:** repeated_constant_consistency
- **Verdict:** PASS
- **Notes:** Values match.

5. ✖ **CAND-05** (Page 3 (Section 2.4.3) vs Page 5 (Section 3 RESULTS opening))
   - **Claim:** Methods report final analytical cohort consisted of 31 bats, but Results state cohort comprised $N = 33$ bats.
   - **Checks:** cross_section_sample_size_consistency
   - **Verdict:** FAIL
   - **Notes:** Sample sizes differ across sections.
6. ✔ **CAND-06** (Page 5 (Figure 1 caption))
   - **Claim:** Figure 1 reports sex distribution: 63.6% male, 36.4% female ($N = 33$).
   - **Checks:** percentages_sum_to_100
   - **Verdict:** PASS
   - **Notes:** Percentages sum to ~ 100 within tolerance.
7. ✖ **CAND-07** (Page 5 (Figure 1 caption) vs Page 3 (Section 2.4.3))
   - **Claim:** Sex distribution differs between Figure 1 (63.6% male, 36.4% female; $N = 33$) and Methods (18 male 58.1%, 13 female 41.9%; $N = 31$).
   - **Checks:** cross_section_percentage_consistency
   - **Verdict:** FAIL
   - **Notes:** Figure 1 implied counts and/or N differ from Methods counts (cross-section inconsistency).
8. ✔ **CAND-08** (Page 5 (Section 2.7.2: median split statement))
   - **Claim:** Median DNAmAge used for split is 9.77 years; Methods descriptive mean lists 9.78 years (nearby).
   - **Checks:** numeric_proximity_sanity_check
   - **Verdict:** PASS
   - **Notes:** Mean and median are close within sanity tolerance; (and within range when available).
9. ✔ **CAND-09** (Page 7 (Section 3.2 Predictive modeling with static features))
   - **Claim:** Cross-validated R^2 values: ElasticNet ($R^2 = -0.05$) and Random Forest ($R^2 = -0.21$) are negative.
   - **Checks:** sign_check_and_ordering
   - **Verdict:** PASS

- **Notes:** Sign and ordering conditions satisfied.

10. ✔ **CAND-10** (Page 6 (Table 1 PCA Loadings for CAI))
    - **Claim:** PCA loadings listed: PEI_{STM} +0.507, RSI_{STM} +0.275, RSI_{LTM} −0.499, PEI_{LTM} **−0.647**.
    - **Checks:** pca_loading_norm_check
    - **Verdict:** PASS
    - **Notes:** Sum of squared loadings consistent with normalization within tolerance.
11. ✖ **CAND-11** (Page 3 (Section 2.2 + Section 2.6.1) vs Page 7 (Section 3.2))
    - **Claim:** Number of brain regions: Methods repeatedly state 82 regions, but Results Section 3.2 states MD values from 24 distinct brain regions.
    - **Checks:** cross_section_constant_consistency
    - **Verdict:** FAIL
    - **Notes:** Counts inconsistent across sections.
12. ✖ **CAND-12** (Page 5 (Abstract) vs Page 7 (Section 3.2) vs Page 3-4 (Methods))
    - **Claim:** Abstract/Methods indicate MD extracted from 82 brain regions for 31 bats, but Results static model describes MD from 24 regions.
    - **Checks:** cross_section_sample_and_feature_consistency
    - **Verdict:** FAIL
    - **Notes:** Region counts inconsistent across sections (e.g., 82 vs 24).
13. ✔ **CAND-13** (Page 9 (Table 2 Significant Interaction Coefficients from the Elastic-Net Model))
    - **Claim:** Interaction coefficients listed: DNAmAge×MD_Region_22 **−0.017**; DNAmAge×MD_Region_23 **−0.007**; DNAmAge×MD_Region_9 **−0.005**.
    - **Checks:** ordering_and_sign_check
    - **Verdict:** PASS
    - **Notes:** All negative and absolute magnitudes strictly descending.

Limitations

- Audit is restricted to the provided PDF parsed text; no underlying datasets (DNAmAge values, MD matrices, CAI scores) are available to recompute reported statistics or model metrics.
- Figures are not used for numeric extraction unless the numbers are explicitly written in the parsed text; plot-based values (coefficients, importances, PDP slopes) are treated as unavailable.
- Several checks are limited to logical/consistency validation (e.g., mean within range) rather than full recomputation due to missing raw inputs.