BABAPPAΩ: Diagnosing the Identifiability of Episodic Selection under Branch–Site Evolution Using Likelihood-Free Neural Inference

Krishnendu Sinha
Department of Zoology, Jhargram Raj College,
Jhargram 721507, India
E-mail: dr.krishnendusinha@gmail.com

Abstract

Episodic positive selection acting on specific evolutionary lineages is a longstanding yet intrinsically difficult target of molecular inference. Classical branch–site methods formulate this problem as hypothesis testing under explicit codon substitution models, implicitly assuming that episodic selection is statistically identifiable from finite alignments. Under biologically realistic conditions—including recombination, epistasis, transient fitness shifts, and alignment uncertainty—this assumption may fail, leading to unstable or uninterpretable results.

BABAPPAΩ reframes branch–site analysis as a problem of *statistical measurement* rather than binary detection. Instead of estimating dN/dS or conducting likelihood ratio tests, the method produces continuous, scale-preserving summaries that quantify the measurability of lineage-specific evolutionary deviation under observed data conditions. Inference is likelihood-free and performed using a frozen neural model trained on forward-time mutation–selection simulations, without estimating substitution rates or codon model parameters.

Simulation-based calibration shows that under strict neutrality ($\omega = 1$), outputs remain diffuse, bounded, and structurally uninformative across phylogenies ranging from 8 to 64 taxa, with decreasing variance and no reproducible high-ranking branches or sites. In addition, a tree-conditional Monte Carlo calibration procedure provides a gene-level Episodic Identifiability Index (EII), standardized relative to neutral expectations and accompanied by an empirical p-value. Imposed episodic structure produces monotonic but saturating responses, consistent with continuous measurement rather than threshold behavior. Permutation tests eliminate inferred structure, whereas bootstrap and taxon jackknife analyses demonstrate stability under realistic perturbations.

These results establish BABAPPAΩ as a conservative diagnostic framework for assessing when episodic selection is statistically resolvable, at what scale, and with what uncertainty, complementing rather than replacing likelihood-based branch–site methods.

1 Introduction

Detecting adaptive molecular evolution is a central objective of comparative genomics. Branch–site models implemented in tools such as `codeml` (Yang, 2007; Iyer et al., 2018) and episodic selection tests developed within HyPhy (Murrell et al., 2012; Kosakovsky Pond et al., 2020;

Wertheim et al., 2015) have been widely used to investigate lineage-specific positive selection. These approaches rely on explicit codon substitution models and frame inference as hypothesis testing via likelihood ratio tests between nested evolutionary scenarios.

Implicit in this formulation is the assumption that episodic selection is statistically identifiable from finite sequence alignments under the chosen model. However, identifiability is not guaranteed under evolutionary processes that are both biologically plausible and empirically common. Recombination, epistasis, transient shifts in selective regimes, and alignment uncertainty can substantially degrade the information content of sequence data, even when episodic selection is present by construction. In such regimes, likelihood-based tests may yield apparent significance without robust biological interpretation, or fail to reject null hypotheses despite the existence of episodic effects (Anisimova et al., 2001; Spielman and Wilke, 2015; Scheffler et al., 2022). Importantly, failure to reject a null hypothesis does not imply the absence of episodic selection, but rather the absence of sufficient information to resolve it.

Recent advances in deep learning have enabled likelihood-free inference in population genetics and phylogenetics (Sheehan and Song, 2016; Flagel et al., 2019; Chan et al., 2023). By learning direct mappings from observable data to latent evolutionary quantities, these approaches bypass explicit likelihood specification. However, supervised learning in the context of episodic selection faces a fundamental obstacle: empirical datasets lack ground-truth labels, and naive simulation strategies that manipulate $\omega = dN/dS$ risk reintroducing the assumptions of the likelihood models they aim to replace. As a result, it remains unclear how to train and evaluate likelihood-free approaches in a manner that preserves biological realism while enabling meaningful calibration.

The framework introduced here, BABAPPAΩ, addresses this challenge by reframing branch–site analysis as a problem of measurement rather than hypothesis testing. Instead of asking whether episodic selection occurred on a particular branch, BABAPPAΩ estimates a continuous latent deviation representing the aggregate strength of episodic signal supported by the data. Inference is explicitly directed toward assessing *when* such deviation becomes measurable and *at what scale*, rather than producing binary decisions at individual sites or branches. Uncertainty is therefore treated as an informative outcome rather than a nuisance to be thresholded.

By combining a theory-consistent forward-time mutation–selection simulator with likelihood-free neural inference, BABAPPAΩ provides a principled framework for diagnosing the limits of identifiability in branch–site evolution under complex and realistic evolutionary regimes. Using scale-preserving inference summaries, the method exhibits monotonic sensitivity to controlled episodic perturbations in simulation-based stress tests, while remaining conservative and stable under perturbation in empirical alignments. In this sense, BABAPPAΩ complements rather than replaces classical likelihood-based branch–site tests by clarifying when episodic selection is measurable, at what scale, and with what degree of uncertainty.

2 Methods

2.1 Forward-time mutation–selection simulator

To generate training and benchmarking data with exact latent ground truth, a forward-time codon evolution simulator grounded in mutation–selection theory was developed. Evolution is simulated under a continuous-time stochastic process along rooted phylogenetic trees, without reference to substitution-rate classes or dN/dS ratios.

Let $\mathcal{C}$ denote the set of 61 sense codons under the standard genetic code. Stop codons are excluded, and only single-nucleotide substitutions are permitted. These constraints ensure that all simulated sequences remain protein-coding throughout the evolutionary process.

For each codon site $i \in \{1, \ldots, L\}$, a latent site-specific fitness landscape is defined as a function

$$F_i : \mathcal{C} \to \mathbb{R},$$

where $F_i(c)$ represents the Malthusian fitness associated with codon c at site i. Baseline fitness values are sampled independently across sites and codons according to

$$F_i^{(0)}(c) \sim \mathcal{N}(0, \sigma^2),$$

with σ fixed across sites. This formulation is consistent with nearly neutral theory and allows both weakly deleterious and weakly advantageous variants to arise naturally.

To incorporate codon usage bias, baseline fitness landscapes are perturbed by a log-Dirichlet component. Let p_c denote a codon-specific frequency drawn from a Dirichlet distribution over $\mathcal{C}$. The codon usage contribution is defined as

$$F_i^{\text{usage}}(c) = \gamma \log\left(\frac{p_c}{\bar{p}}\right),$$

where $\bar{p}$ is the mean codon frequency and γ controls the strength of bias. The effective baseline fitness is therefore

$$F_i(c) = F_i^{(0)}(c) + F_i^{\text{usage}}(c).$$

2.1.1 Branch-specific episodic fitness perturbations

Episodic positive selection is modeled as transient, branch-specific reshaping of site-specific fitness landscapes rather than as changes in substitution-rate parameters. Alignments are partitioned into contiguous blocks representing recombination segments. A subset of these blocks is randomly designated as episodic blocks.

For branches traversing episodic blocks, an additional epistatic perturbation is applied to the baseline fitness landscape. For each site–branch pair (i, j), the effective fitness is defined as

$$F_{i,j}(c) = F_i(c) + \Delta F_{i,j}(c),$$

where

$$\Delta F_{i,j}(c) \sim \mathcal{N}(0, \sigma_{\text{epi}}^2)$$

is applied only for a sparse subset of site–branch pairs. This construction ensures that episodic selection is localized in both site and lineage, transient in evolutionary time, and not globally identifiable from substitution rates alone.

2.1.2 Mutation model and fixation dynamics

Mutations are proposed via single-nucleotide changes with transition–transversion bias implemented under an HKY-style mutation model. Let $\mu_{c\to c'}$ denote the mutation rate between codons differing by one nucleotide, defined as

$$\mu_{c\to c'} = \begin{cases} \kappa\mu, & \text{if the substitution is a transition,} \\ \mu, & \text{if the substitution is a transversion,} \end{cases}$$

where κ denotes the transition–transversion bias parameter.

For a proposed mutation $c \to c'$ at site i on branch j, the selection coefficient is

$$s_{i,j}(c \to c') = F_{i,j}(c') - F_{i,j}(c).$$

Fixation probabilities follow the Halpern–Bruno formulation (Halpern and Bruno, 1998; Thorne et al., 2012). The instantaneous substitution rate is given by

$$Q_{i,j}(c \to c') = \mu_{c\to c'} \cdot \frac{e^{s_{i,j}}}{1 - e^{-s_{i,j}}},$$

with numerical truncation applied to extreme values to ensure stability. Evolution along each branch is simulated using a Gillespie algorithm, yielding exact continuous-time trajectories.

2.1.3 Phylogenetic structure and sequence generation

Phylogenetic trees are generated under a Yule pure-birth process, producing binary rooted trees with exponentially distributed branch lengths. Evolution proceeds from a randomly initialized root sequence toward the leaves of the tree, yielding a multiple sequence alignment at the tips.

To model departures from strict tree-like evolution, recombination is introduced implicitly through block-wise fitness regimes. Additional alignment noise arises naturally from stochastic substitution trajectories rather than being imposed artificially.

2.1.4 Latent labels and episodic burden

Because episodic selection is defined explicitly at the generative level, exact latent ground truth is available. For each site–branch pair (i, j), the following latent variables are recorded:

$$Y_{i,j} \in \{0, 1\} \quad \text{(episodic indicator)},$$

$$I_{i,j} \geq 0 \quad \text{(episodic intensity)},$$

$$R_{i,j} \quad \text{(latent evolutionary regime)}.$$

From these quantities, aggregate episodic burdens are defined as

$$\text{site-level burden:} \quad B_i = \frac{1}{|E|} \sum_{j \in E} Y_{i,j},$$

$$\text{branch-level burden:} \quad B_j = \frac{1}{L} \sum_{i=1}^{L} Y_{i,j}.$$

These burdens represent the true episodic selection signal at site and branch levels and are used as primary supervision targets for training and benchmarking BABAPPAΩ.

2.1.5 Rationale for likelihood-free supervision

Episodic selection in this simulator is not encoded through substitution-rate classes, ω ratios, or likelihood parameters. As a result, no classical branch–site likelihood model is correct by construction. This design prevents circularity during training and ensures that the neural inference model learns biologically meaningful deviations in evolutionary patterns rather than artifacts of a particular likelihood formulation.

2.2 Neural training and fine-tuning protocol

Training of BABAPPAΩ proceeds in two strictly separated stages: large-scale pretraining on simulator-generated data followed by limited fine-tuning on burden-aligned targets. Both stages are fully supervised using exact ground truth provided by the generative process.

2.2.1 Pretraining on simulator-generated data

In the first stage, the neural network is trained on a large corpus of simulated codon alignments generated under the mutation–selection model described above. Each training example consists of branch-conditioned parent–child codon counts, branch lengths, and exact latent episodic labels derived from the simulation.

Let $X_{b,i}$ denote the observed codon-derived summary statistics for site i on branch b. The model is trained to predict latent episodic indicators and intensities defined directly at the fitness level during simulation. This stage establishes general representations of evolutionary deviation under a wide range of regimes, including neutrality, weak selection, strong episodic selection, recombination, and alignment noise.

All architectural choices, hyperparameters, and loss weights are fixed prior to training. Pretraining is performed once and produces a single converged model checkpoint.

2.2.2 Burden-aligned fine-tuning

Following pretraining, a short fine-tuning phase is performed using a distinct objective aligned with aggregate episodic burden rather than sitewise regime labels. This stage optimizes the model to recover continuous site- and branch-level episodic burden signals that are directly comparable across genes.

For each training alignment, the model produces site-level logits $\ell_{b,i}$ and branch-level logits ℓ_b. Site-level predictions are aggregated across branches to obtain a per-site burden estimate:

$$\hat{s}_i = \frac{1}{B} \sum_{b=1}^{B} \sigma(\ell_{b,i}),$$

where B is the number of branches and $\sigma(\cdot)$ denotes the logistic function.

The site-level loss is defined as a binary cross-entropy between $\hat{s}_i$ and the true simulated site burden s_i:

$$\mathcal{L}_{\text{site}} = -\sum_i \left[s_i \log \sigma(\hat{s}_i) + (1 - s_i) \log(1 - \sigma(\hat{s}_i)) \right].$$

In parallel, branch-level logits are trained against true branch burdens b_j using an analogous binary cross-entropy loss:

$$\mathcal{L}_{\text{branch}} = -\sum_j \left[b_j \log \sigma(\ell_j) + (1 - b_j) \log(1 - \sigma(\ell_j)) \right].$$

The total fine-tuning objective is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{site}} + \mathcal{L}_{\text{branch}}.$$

Fine-tuning is performed for a small number of epochs using a low learning rate and CPU-safe execution. No architectural modifications are introduced at this stage.

2.2.3 Model freezing and reproducibility

After completion of fine-tuning, model parameters are frozen and exported as an immutable TorchScript artifact. This frozen model is used unchanged for all validation, stress testing, benchmarking, and empirical inference.

No additional fine-tuning, retraining, or parameter adjustment is performed on benchmark datasets or empirical data. As a consequence, all reported results reflect true generalization under likelihood-free inference rather than retrospective optimization.

2.3 Validation and testing protocol

Validation and testing were performed using a single frozen, fine-tuned BABAPPAΩ model applied to simulated and empirical datasets that were not used during training or fine-tuning (full experimental design and numerical results are provided in the Supplementary Material).

The purpose of validation was to assess calibration under neutrality, quantitative sensitivity to episodic signal strength in controlled simulations, and stability of inferred outputs under empirical perturbations. All validation analyses were executed using the production command-line interface, without access to training-time labels or gradients.

2.3.1 Validation datasets

Simulated validation datasets were generated independently using simple codon-level perturbation schemes designed to vary the strength of episodic deviation across sites. These simulations do not reproduce the full forward-time mutation–selection model used during training and are therefore treated as stress tests of inferential behavior rather than as ground-truth recovery experiments.

Empirical validation was performed on real codon alignments using perturbation controls. Codon-column shuffling preserves marginal composition while disrupting positional structure, whereas bootstrap resampling assesses robustness to site sampling noise. No empirical dataset was used for training or fine-tuning.

2.3.2 Inference mode

For each alignment, inference was performed with the frozen model in evaluation mode and with gradient computation disabled. The model produces branch–site logits $\ell_{b,i}$ and branch-level logits ℓ_b.

Two classes of inference summaries were computed. First, sigmoid-transformed summaries were obtained for interpretative purposes:

$$\hat{s}_i^{\text{prob}} = \frac{1}{B}\sum_{b=1}^{B} \sigma(\ell_{b,i}), \qquad \hat{b}_j^{\text{prob}} = \sigma(\ell_j),$$

where B denotes the number of branches and $\sigma(\cdot)$ is the logistic function. These bounded quantities are intended solely for qualitative interpretation.

Second, scale-preserving logit-level summaries were computed directly from the raw model outputs:

$$\hat{s}_i^{\text{logit}} = \frac{1}{B}\sum_{b=1}^{B} \ell_{b,i}, \qquad \widehat{\text{Var}}^{\text{logit}} = \text{Var}_b(\ell_{b,i}).$$

These logit-level quantities preserve magnitude information and were used for all quantitative benchmarking and statistical analysis.

2.3.3 Tree-conditional Monte Carlo calibration and the Episodic Identifiability Index

To quantify gene-level identifiability of episodic structure, a tree-conditional Monte Carlo calibration procedure was introduced. For a given empirical alignment and fixed phylogeny, N neutral codon alignments are simulated under strict neutrality ($\omega = 1$) while preserving tree topology, branch lengths, alignment length, and codon composition.

For each simulated alignment, BABAPPAΩ inference is performed using the frozen model, and the variance of site-level logit summaries is computed. This produces an empirical null distribution of dispersion under neutrality conditional on the observed phylogenetic structure.

Let σ^2_{obs} denote the observed variance of site-level logit summaries for the empirical alignment. Let μ_0 and σ_0 denote the mean and standard deviation of the null distribution obtained via Monte Carlo simulation. The standardized Episodic Identifiability Index is defined as

$$\text{EII}_z = \frac{\sigma^2_{\text{obs}} - \mu_0}{\sigma_0}.$$

An empirical p-value is computed as

$$p = \frac{1 + \sum_{k=1}^{N} \mathbb{I}(\sigma_k^2 \geq \sigma_{\text{obs}}^2)}{N+1}.$$

This calibration quantifies whether observed lineage-specific dispersion exceeds that expected under neutral evolution on the same tree. Importantly, EII does not constitute a test of positive selection but measures the statistical resolvability of lineage-specific structure relative to a tree-aware neutral baseline.

2.3.4 Simulation-based scale sensitivity analysis

Quantitative sensitivity of BABAPPAΩ to episodic signal strength was assessed using simulated codon alignments generated under increasing probabilities of site-wise perturbation. For each simulated dataset, the mean and variance of site-level logit summaries were computed.

Monotonic dependence between perturbation probability and inferred logit-level summaries was evaluated using Spearman rank correlation. This analysis assesses whether inferred outputs respond consistently to changes in signal strength, without claiming recovery of true episodic selection burden.

These simulations do not reproduce the full forward-time mutation–selection model used during training and are therefore interpreted as stress tests of inferential behavior rather than as ground-truth recovery experiments.

2.3.5 Empirical perturbation protocol

Stability of inferred outputs on empirical data was assessed using two perturbation controls. Codon-column shuffling disrupts positional structure while preserving marginal composition, whereas bootstrap resampling assesses sensitivity to site sampling.

For each condition, the variance of site-level logit summaries was computed. Paired Wilcoxon signed-rank tests were used to compare inferred variance between the original alignment and each control. Effect size was quantified using Cliff's δ.

These analyses serve as sanity checks of inference behavior and do not constitute evidence for episodic selection in empirical data.

2.3.6 Statistical analysis

All quantitative analyses were performed using non-parametric statistical tests. Monotonic trends in simulation were assessed using Spearman rank correlation. Empirical perturbation analyses were evaluated using paired Wilcoxon signed-rank tests. Effect sizes were reported using Cliff's δ.

All benchmarks were repeated across ten independent random seeds to assess robustness. Statistical significance was evaluated at $\alpha = 0.05$.

2.4 Empirical stability analysis

To assess conservative behavior on real data, BABAPPAΩ was applied to empirical codon alignments under three perturbation regimes: the original alignment, codon-column shuffling, and bootstrap resampling. Codon shuffling preserves marginal composition while disrupting positional structure, whereas bootstrap resampling assesses robustness to site sampling noise.

For each condition, mean and variance of site-level logit summaries were computed. Reduction of logit variance under shuffling is interpreted as loss of coherent episodic structure, whereas stability under bootstrap resampling indicates robustness rather than overfitting. No claims of episodic selection detection or localization are made in empirical analyses; results are reported solely as sanity checks of inference behavior.

2.5 Interpreting BABAPPAΩ inference outputs

BABAPPAΩ is formulated as a diagnostic instrument for assessing the statistical measurability of lineage-specific evolutionary dispersion under branch–site evolution. The outputs quantify how strongly episodic structure is supported by the observed alignment relative to a calibrated neutral baseline. They are not hypothesis tests of positive selection and should not be interpreted as evidence of adaptive substitution.

Gene-level identifiability. For each alignment, BABAPPAΩ computes the variance of site-level logit summaries and compares it to a tree-conditional neutral reference distribution obtained via Monte Carlo simulation. As illustrated in the empirical example reported in the Supplementary Material (FBN30 dataset; 249 codon sites, 29 terminal taxa), the observed dispersion variance exceeded the neutral expectation, yielding $\mathrm{EII}_z = +2.357$ and an empirical p-value of 0.012. This indicates that lineage-specific heterogeneity is statistically resolvable at the gene level under the observed phylogenetic constraints. Importantly, this result reflects calibrated deviation from neutrality, not detection of adaptive substitutions.

Branch-level structure. Branch-level background summaries are standardized relative to the distribution across terminal lineages. In the Supplementary example, background values ranged narrowly, with the highest-ranking branches exhibiting moderate standardized deviations ($z \approx 1.7$–1.8). Such magnitudes indicate disproportionate contribution to global dispersion relative to other taxa, but do not imply episodic bursts of adaptive change. The bounded and continuous distribution of branch scores argues against heavy-tailed inflation or numerical instability.

Site-level dispersion. Site-level summaries represent aggregated logit-scale deviations across branches. In the empirical alignment presented in the Supplementary Material, top-ranked codons showed elevated but tightly bounded scores, reflecting moderate concentration of dispersion within a subset of positions rather than extreme localization. Permutation controls demonstrate that such ranking structure depends on phylogenetically coherent positional information rather than marginal composition alone.

Magnitude and scale. All quantitative analyses are performed on the raw logit scale to preserve magnitude information. Sigmoid-transformed summaries are provided solely for interpretative clarity. Higher values indicate greater statistical measurability of lineage-specific heterogeneity relative to the remainder of the gene and tree. Diffuse or low-variance outputs should be interpreted as limited identifiability, not as evidence for the absence of episodic selection.

What the outputs do not represent. BABAPPAΩ does not estimate $\omega = dN/dS$, does not compute likelihood ratios, and does not assign posterior probabilities of selection. Even when EII_z indicates statistically resolvable dispersion, this reflects structured heterogeneity relative to neutrality, not proof of adaptive substitution events. Mechanistic interpretation requires complementary codon likelihood models or functional validation.

Use as a calibrated diagnostic. The intended use of BABAPPAΩ is to determine whether episodic lineage-specific structure is statistically measurable under the observed data conditions and at what scale such structure emerges. If gene-level dispersion remains within neutral expectation, downstream hypothesis testing may be intrinsically underpowered. Conversely, calibrated deviations from neutrality indicate regimes where structured heterogeneity is detectable and may justify further model-based investigation.

Interpretation under perturbation. Reduction of dispersion under codon-column shuffling and stability under bootstrap resampling serve as calibration diagnostics. These perturbation analyses evaluate structural specificity and robustness, but do not constitute independent evidence for or against episodic selection.

2.6 Stress testing under constant-ω substitution regimes

To assess whether BABAPPAΩ implicitly recovers classical substitution parameters, a targeted stress test was performed using codon alignments generated under fixed-ω models. Codon sequences were simulated using the `evolver` module of PAML under a single phylogenetic tree with transition–transversion ratio fixed at $\kappa = 2$. Across simulations, the nonsynonymous–synonymous rate ratio ω was varied over several orders of magnitude ($\omega \in \{0.001, 0.01, 0.1, 0.5, 1, 2, 5, 10, 20\}$).

These simulations encode constant selection pressure across sites and branches and therefore do not represent episodic selection by construction. They were used exclusively as a stress test to probe whether BABAPPAΩ exhibits monotonic sensitivity to ω, a likelihood-model parameter that the method is not designed to estimate.

For each alignment, inference was performed using the frozen BABAPPAΩ model. Mean and variance of site-level logit summaries were computed. Monotonic dependence between inferred summaries and ω was evaluated using Spearman rank correlation. This analysis does not assess recovery of true episodic selection burden and is interpreted solely as a diagnostic of model-agnostic behavior under misspecification.

3 Results

The behavior of BABAPPAΩ was evaluated using controlled simulation-based stress tests and empirical perturbation analyses. These experiments were designed to assess calibration under neutrality, qualitative sensitivity of inferred outputs to increasing levels of episodic signal, and stability of inference under data perturbation. No genome-scale recovery of true episodic selection burden was evaluated. Only representative summaries are reported in the main text; complete numerical diagnostics, stress-test outputs, and execution logs are provided in the Supplementary Material.

3.1 Calibration under strict neutrality

To evaluate numerical calibration and stability in the absence of episodic selection, BABAPPAΩ was applied to codon alignments simulated under strict neutrality ($\omega = 1$) across phylogenies with 8, 16, 32, and 64 taxa. All simulations used identical alignment lengths (1000 codons) and were analyzed using a single frozen reference model without retraining or parameter adjustment.

Across all tree sizes, mean site-level logit summaries were highly stable (1.239×10^1–1.271×10^1), while logit variance decreased monotonically with increasing number of taxa. This variance contraction reflects increased statistical constraint from larger phylogenies rather than numerical instability. No divergence, collapse, or inflation of inferred magnitudes was observed.

Importantly, neither interpretation-mode summaries nor raw logit-level outputs identified reproducible high-ranking sites or branches across replicate simulations or tree sizes. Rank correlations between independent neutral replicates were centered near zero, and site-level entropy remained maximal, indicating uniformly diffuse distributions. Forced numerical diagnostics produced consistent mean–variance profiles across tree sizes, confirming that observed behavior arises from inference rather than post-processing heuristics.

Together, these results demonstrate that BABAPPAΩ remains correctly uninformative under neutrality, does not generate spurious episodic structure, and exhibits improved numerical concentration as phylogenetic information increases.

3.2 Gene-level calibration using tree-conditional Monte Carlo simulation

Application of the tree-conditional calibration procedure to empirical alignments demonstrated that observed site-level dispersion can deviate moderately but significantly from neutral expectations. For representative genes, observed variance exceeded the neutral mean by approximately 20%, yielding standardized EII_z values between 2 and 3 and empirical p-values near 0.01.

These results indicate that lineage-specific episodic structure can become statistically resolvable at the gene level under realistic phylogenetic constraints. Importantly, calibration constrains extreme inflation observed under independent-site null models and prevents overinterpretation of diffuse variation as strong episodic signal.

3.3 Simulation-based scale sensitivity

Quantitative sensitivity of BABAPPAΩ to episodic signal strength was assessed using simulated codon alignments generated under increasing probabilities of site-wise perturbation. These simulations do not reproduce the full forward-time mutation–selection model used during training and are therefore interpreted as stress tests of inferential behavior rather than as ground-truth recovery experiments.

Across simulated datasets spanning a range of perturbation probabilities, both the mean and variance of site-level logit summaries increased monotonically with signal strength. This monotonic dependence was observed consistently across ten independent random seeds. Spearman rank correlation indicated perfect monotonic association between perturbation probability and inferred logit-level statistics (Table 1).

At higher perturbation levels, increases in logit variance saturated or decreased, consistent with homogenization of episodic effects across sites. Under strictly neutral simulations, inferred logit summaries exhibited minimal variance and lacked systematic structure, reflecting correct uninformative behavior.

Table 1: Simulation-based scale sensitivity of BABAPPAΩ. Spearman rank correlation quantifies monotonic dependence between the probability of site-wise perturbation (p_{epi}) and inferred logit-level summaries, aggregated across ten independent simulation replicates. These simulations are stress tests of inferential behavior and do not represent recovery of true episodic selection burden.

Statistic	Spearman ρ	p-value
Mean site logit vs. p_{epi}	1.00	$< 10^{-6}$
Variance of site logits vs. p_{epi}	1.00	$< 10^{-6}$

3.4 Empirical stability under perturbation

Stability of inferred outputs on empirical data was assessed using codon-column shuffling and bootstrap resampling. Inferred variance was significantly reduced under codon-column shuffling relative to the original alignment (paired Wilcoxon signed-rank test, $p = 1.95 \times 10^{-3}$), whereas bootstrap resampling preserved structured variance. These analyses are interpreted as sanity checks of inference behavior and do not constitute evidence for episodic selection in empirical data.

Table 2: Empirical stability of BABAPPAΩ evaluated using perturbation controls. Paired Wilcoxon signed-rank tests compare inferred variance of site-level logit summaries between the original alignment and perturbed controls. Cliff's δ reports non-parametric effect size. These analyses assess conservative inference behavior and do not constitute evidence for episodic selection in empirical data.

Comparison	Wilcoxon p-value	Cliff's δ	Interpretation
Original vs. codon-shuffled	1.95×10^{-3}	1.00	Significant
Original vs. bootstrap	1.95×10^{-3}	1.00	Significant

Cliff's $\delta = 1.00$ reflects complete stochastic dominance of variance reduction under codon-column shuffling across all tested replicates, indicating that the observed effect is not driven by marginal differences but by a consistent collapse of positional structure. Such extreme effect sizes arise because codon-column shuffling deterministically destroys positional dependency by construction; this result reflects a designed sanity check rather than an empirical discovery.

3.5 ω-based stress test using PAML evolver

To assess model behavior under classical constant-ω substitution regimes, BABAPPAΩ was applied to codon alignments generated using PAML `evolver` with fixed transition–transversion bias ($\kappa = 2$) and ω spanning several orders of magnitude. These simulations do not encode episodic selection and therefore serve as a stress test rather than a validation experiment.

Inferred mean site-level logit summaries showed no significant monotonic dependence on ω (Spearman $\rho = 0.47$, $p = 0.21$). In contrast, variance of logit summaries exhibited a moderate monotonic association with ω (Spearman $\rho = 0.80$, $p = 9.6 \times 10^{-3}$), reflecting increased substitutional heterogeneity rather than localized episodic structure.

At high ω, inferred summaries saturated or decreased slightly, indicating absence of progressive signal accumulation. Together, these results show that BABAPPAΩ does not recover or proxy constant-ω selection parameters and remains largely insensitive to likelihood-model regimes that lack episodic structure.

4 Discussion

4.1 Episodic selection as an identifiability problem

The central contribution of BABAPPAΩ is not increased power to detect positive selection, but a reframing of branch–site inference as a problem of *statistical identifiability.* Classical branch–site methods implicitly assume that episodic deviations in fitness leave sufficiently strong and coherent signatures in finite codon alignments to permit reliable hypothesis testing under explicit substitution models. As demonstrated by both theoretical work and extensive empirical experience, this assumption frequently fails under biologically realistic conditions.

In particular, recombination, epistasis, transient selective regimes, and heterogeneous background processes can substantially erode the information content of sequence data without eliminating episodic effects at the generative level (Schierup and Hein, 2000; Markova-Raina and Petrov, 2022). In such regimes, failure to reject a null hypothesis does not imply absence of episodic selection, but rather insufficient statistical resolution to identify it. Conversely, apparent significance under misspecified likelihood models may reflect numerical sensitivity rather than biologically meaningful signal.

BABAPPAΩ explicitly separates the *existence* of episodic selection from its *measurability.* The model produces continuous, scale-preserving summaries that quantify the degree to which lineage–specific deviation is supported by the observed data, without attempting to localize causal substitutions or estimate substitution parameters. Uncertainty and diffuseness of signal are therefore treated as informative outcomes rather than inferential failures.

4.2 Calibration, neutrality, and conservative behavior

A necessary property of any diagnostic framework is correct behavior under the absence of signal. Extensive neutrality benchmarks reported here and in the Supplementary Material demonstrate that BABAPPAΩ remains numerically stable and unstructured under strict neutrality across a four-fold increase in phylogenetic size. Mean site-level logit summaries remain effectively constant, while variance contracts monotonically as additional taxa increase statistical constraint.

Importantly, neither raw logit outputs nor interpretation-mode summaries produce reproducible high-ranking sites or branches under neutrality. Positional entropy remains maximal, and rank correlations across replicates are near zero. These results establish that BABAPPAΩ

does not generate spurious episodic structure from codon composition, tree size, or numerical artifacts.

Equally important is conservative behavior under misspecification. Stress tests using constant-ω substitution models demonstrate that BABAPPAΩ does not function as a surrogate dN/dS estimator. Mean inferred episodic deviation shows no significant monotonic dependence on ω, while moderate variance increases reflect nonspecific substitutional heterogeneity rather than localized episodic signal. This insensitivity confirms that the model does not implicitly recover likelihood-model parameters it was not designed to estimate.

4.3 Sensitivity as continuous measurement, not threshold detection

Controlled signal titration experiments demonstrate that BABAPPAΩ responds monotonically to increasing episodic perturbation, with saturation at high signal levels. This behavior is characteristic of a continuous measurement instrument rather than a binary detector. At low perturbation intensities, inferred summaries remain diffuse and weak, reflecting limited identifiability. As episodic burden increases, structure becomes measurable, but does not collapse into discrete site or branch classifications.

Critically, hysteresis and reversibility tests show symmetric emergence and disappearance of inferred structure as episodic perturbations are introduced and removed. The absence of residual signal following perturbation removal indicates that the model does not encode memory or path-dependent artifacts. Together, these results support the interpretation of BABAPPAΩ outputs as quantitative diagnostics of information content rather than estimates of latent selection parameters.

4.4 Dependence on positional and phylogenetic coherence

Resampling and perturbation analyses provide further insight into the nature of the inferred signal. Codon-column shuffling preserves marginal composition while destroying positional coherence; under this perturbation, inferred structure collapses consistently. In contrast, site bootstrap resampling and taxon jackknife analyses preserve global structure while introducing sampling noise, and BABAPPAΩ outputs remain stable under these procedures.

Similarly, tree topology randomization destabilizes inferred site–branch structure, whereas branch length distortion affects magnitude without inducing spurious localization. These results demonstrate that BABAPPAΩ depends on coherent positional and lineage-specific information rather than compositional bias or numerical idiosyncrasies. Episodic measurability emerges only when evolutionary signal is both structured and phylogenetically consistent.

4.5 Relationship to likelihood-based branch–site methods

BABAPPAΩ is not intended to replace classical likelihood-based branch–site tests such as those implemented in `codeml` or HyPhy. Instead, it serves as a complementary diagnostic framework that addresses a question orthogonal to hypothesis testing: *is episodic selection statistically measurable under the observed data conditions, and at what scale?*

In regimes where likelihood-based methods agree and signal is strong, BABAPPAΩ produces structured, coherent outputs. In cases of disagreement or instability, inferred summaries are diffuse or unstable, reflecting limited identifiability. Importantly, BABAPPAΩ does not adjudicate correctness among competing models, but rather provides an empirical assessment of information content that can inform downstream analysis.

Used in this way, BABAPPAΩ can guide interpretation of both positive and negative results from likelihood-based tests. Diffuse or weak BABAPPAΩ outputs suggest that binary hypothesis testing may be intrinsically underpowered or misleading, whereas structured outputs indicate regimes where further model-based inference may be informative.

4.6 Limitations and scope

Several limitations of the present framework should be emphasized. First, BABAPPAΩ does not provide site-level hypothesis tests or estimates of selection coefficients. While a gene-level calibrated statistic (EII) and associated empirical p-value are available via tree-conditional Monte Carlo simulation, these quantify identifiability relative to neutral expectations rather than testing for adaptive substitution events. High inferred episodic measurability does not constitute evidence for adaptive substitutions or causal selective events. Second, simulation-based stress tests do not recover exact ground-truth episodic burden and should not be interpreted as validation of gene-level identifiability.

Finally, while the forward-time mutation–selection simulator used for training and benchmarking captures a broad range of biologically realistic processes, it cannot encompass the full diversity of evolutionary regimes encountered in nature. BABAPPAΩ should therefore be interpreted as a diagnostic instrument rather than a definitive arbiter of evolutionary mechanism.

4.7 Conclusions

By reframing episodic selection inference as a problem of statistical measurability, BABAPPAΩ provides a principled, likelihood-free framework for diagnosing the limits of branch–site inference. Through explicit calibration, conservative behavior under neutrality and misspecification, and monotonic sensitivity to structured perturbation, the method clarifies when episodic evolutionary deviation can be meaningfully resolved from sequence data.

In this role, BABAPPAΩ complements existing likelihood-based approaches by shifting emphasis from binary detection toward quantitative assessment of information content, uncertainty, and scale. This perspective is essential for robust interpretation of episodic selection in complex evolutionary systems where identifiability, rather than power, is the primary limiting factor.

5 Supplementary Material

A comprehensive Supplementary Material document accompanies this manuscript, providing full methodological detail, quantitative validation, and practical interpretation guidance for BABAPPAΩ.

The Supplementary Material includes complete experimental protocols, numerical summaries, and execution logs for all validation analyses, covering neutrality calibration and tree-size stability, positional structure destruction controls, bootstrap and taxon jackknife robustness assessments, controlled signal titration experiments, phylogenetic misspecification stress tests, and deterministic reproducibility verification.

In addition, a dedicated section explains how to read and interpret BABAPPAΩ outputs, clarifying what the reported site and branch scores represent, how they should be used in empirical analyses, and what they are not intended to infer.

All validation experiments were performed using a single frozen reference model without retraining, parameter tuning, or dataset-specific adjustment, ensuring strict separation between model construction and evaluation.

6 Data Availability

The BABAPPAΩ inference engine is released as open-source software and is available at `https://github.com/sinhakrishnendu/babappaomega`. The software can be installed directly from the Python Package Index using `pip install babappaomega`.

The frozen reference neural model used for all BABAPPAΩ inference is archived on Zenodo (DOI: `https://doi.org/10.5281/zenodo.18195868`). This model is distributed as an immutable TorchScript artifact and is automatically downloaded, checksum-verified, and cached on first use to ensure bitwise reproducibility across computing environments.

All benchmarking simulations, Supplementary Material, and stress-testing artifacts associated with this study are archived on Zenodo (DOI: `https://doi.org/10.5281/zenodo.18197957`) and figshare (DOI: `https://doi.org/10.6084/m9.figshare.31199098`). These archives include simulator configurations, benchmark outputs, numerical diagnostics, and full execution logs required to reproduce all analyses reported in this manuscript and its Supplementary Material.

7 Acknowledgments

I am deeply grateful to my young son, whose boundless curiosity and love for butterflies inspired the name “BABAPPA.” In his imaginative vocabulary, “BABAPPA” refers to a butterfly, and it was this creativity that ultimately gave the software its name. Finally, I acknowledge the developers and maintainers of the open-source software tools and biological databases used by BABAPPAΩ, whose contributions made this work possible.

References

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007.

Murrell B et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 2012.

Kosakovsky Pond SL et al. HyPhy 2.5. *Mol Biol Evol.* 2020.

Anisimova M et al. Accuracy and power of likelihood ratio tests. *Mol Biol Evol.* 2001.

Spielman SJ, Wilke CO. Relationship between dN/dS and selection coefficients. *Mol Biol Evol.* 2015.

Sheehan S, Song YS. Deep learning for population genetic inference. *PLoS Comput Biol.* 2016.

Flagel L et al. The unreasonable effectiveness of CNNs in population genetics. *Mol Biol Evol.* 2019.

Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences. *Mol Biol Evol.* 1998.

Iyer S, Hasegawa M, Susko E. A robust approach for testing positive selection on phylogenetic branches. *Mol Biol Evol.* 2018.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol.* 2015.

Scheffler K, Seoighe C. The limits of detecting positive selection using comparative genomics. *PLoS Genet.* 2022.

Chan YF, Papadopoulos AS, Kern AD. Simulation-based inference in population genetics: recent advances and challenges. *Trends Genet.* 2023.

Thorne JL, Goldman N, Jones DT. Combining protein evolution and secondary structure. *Mol Biol Evol.* 2012.

Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics.* 2000.

Markova-Raina P, Petrov DA. High sensitivity of evolutionary inference to multiple sequence alignment errors. *Genome Biol Evol.* 2022.

Supplementary Material
Statistical Validation, Calibration, and Empirical Robustness of BABAPPAΩ

S1. Statistical Framework

BABAPPAΩ is formulated as a measurement instrument for episodic branch–site dispersion structure.

Let:

$$X_{k,i}$$

denote the codon state for branch k at site i.

Define:

$$\ell_i = \text{logit-scale episodic aggregation across branches}$$

$$b_k = \text{branch-level background deviation across sites}$$

Sigmoid-transformed outputs:

$$s_i = \sigma(\ell_i), \quad 0 < s_i < 1$$

$$\beta_k = \sigma(b_k), \quad 0 < \beta_k < 1$$

Validation therefore concerns statistical properties of:

$$\{\ell_i\}_{i=1}^{L}, \quad \{b_k\}_{k=1}^{K}$$

A valid measurement instrument must satisfy:

1. Neutral calibration
2. Structural specificity
3. Perturbation stability
4. Empirical boundedness
5. Deterministic reproducibility

Violation of any condition falsifies the framework.

S1.1 Gene-level Episodic Identifiability Index (EII)

In addition to site- and branch-level summaries, a gene-level Episodic Identifiability Index (EII) is defined using a tree-conditional neutral reference distribution.

Let

$$\sigma^2_{\text{obs}} = \text{Var}_i(\ell_i)$$

denote the observed variance of site-level logit summaries.

Under tree-conditional neutrality, simulated alignments yield dispersion values

$$\{\sigma^2_1, \dots, \sigma^2_N\}.$$

Define:

$$\mu_0 = \mathbb{E}[\sigma^2_k], \quad \sigma_0 = \text{SD}(\sigma^2_k).$$

The standardized index is:

$$\text{EII}_z = \frac{\sigma^2_{\text{obs}} - \mu_0}{\sigma_0}.$$

The dispersion ratio is:

$$\text{EII}_{\text{ratio}} = \frac{\sigma^2_{\text{obs}}}{\mu_0}.$$

The empirical Monte Carlo p-value is:

$$p = \frac{1 + \sum_{k=1}^{N} \mathbb{I}(\sigma^2_k \geq \sigma^2_{\text{obs}})}{N+1}.$$

EII quantifies whether observed lineage-specific dispersion exceeds neutral expectation conditional on the observed tree. It measures identifiability relative to a calibrated baseline, not adaptive substitution.

S2. Neutral Calibration and Tree-Conditional Monte Carlo Reference

S2.1 Null hypothesis

Under strict neutrality:

$$H_0 : \mathbb{E}[\ell_i] = \mu_0, \quad \text{Var}(\ell_i) = \sigma^2_0$$

with no reproducible ranking structure across simulations.

Simulations used PAML `evolver` under:

$$\omega = 1, \quad \kappa = 2$$

with $L = 1000$ codons and $K \in \{8, 16, 32, 64\}$ taxa.

S2.2 Empirical statistics

$$\mu_K = \frac{1}{L}\sum_i \ell_i$$

$$\sigma_K^2 = \frac{1}{L-1}\sum_i (\ell_i - \mu_K)^2$$

Variance contracts monotonically:

$$\sigma_8^2 > \sigma_{16}^2 > \sigma_{32}^2 > \sigma_{64}^2$$

No cross-K rank reproducibility:

$$\rho_K \approx 0$$

Conclusion: No episodic inflation under neutrality.

S2.3 Tree-conditional Monte Carlo calibration

To avoid independent-site null assumptions, neutral reference distributions were generated conditionally on the observed phylogenetic topology, branch lengths, and alignment length.

For each empirical alignment:

1. N neutral codon alignments were simulated under $\omega = 1$.
2. Full BABAPPAΩ inference was performed.
3. Site-level logit variance was recorded.

This procedure yields an empirical dispersion distribution conditional on tree structure and sequence length.

Observed empirical dispersion for FBN30:

$$\sigma_{\text{obs}}^2 = 0.393628$$

Neutral expectation:

$$\mu_0 = 0.323885$$

Standardized deviation:

$$\text{EII}_z = +2.357$$

Empirical p-value:

$$p = 0.01198$$

This moderate but significant deviation indicates measurable lineage-specific dispersion beyond neutral expectation.

S3. Structural Specificity

S3.1 Permutation test

Let π be a random permutation of codon columns.

$$X'_{k,i} = X_{k,\pi(i)}$$

Under structural destruction:

$$H_0^{\text{perm}} : \rho(\ell_i, \ell'_i) = 0$$

Empirical FBN30 result:

$$\rho = -0.0846$$

Variance shift:

$$\sigma^2_{\text{orig}} = 1.228 \times 10^{-3}$$

$$\sigma^2_{\text{perm}} = 9.32 \times 10^{-4}$$

Permutation destroys ranking coherence. Therefore episodic structure depends on positional coherence.

S4. Jackknife Stability

For each taxon k:

$$\rho_{(-k)} = \rho(\ell_i, \ell_i^{(-k)})$$

Empirical range:

$$0.608 \leq \rho_{(-k)} \leq 0.954$$

$$\bar{\rho} = 0.833$$

Stability criterion:

$$\rho_{(-k)} > 0.5 \quad \forall k$$

No catastrophic dependence detected.

S5. Empirical Evaluation: FBN30

Dataset:

$$L = 249, \quad K = 29$$

S5.1 Global dispersion

$$\mu_s = 0.934832$$

$$\sigma_s^2 = 1.228 \times 10^{-3}$$

Branch backgrounds:

$$0.0561 \leq \beta_k \leq 0.0623$$

Range bounded and continuous.

S6. Exact CLI Output

```
BABAPPAomega | INTERPRETED RESULTS
============================================================

1. SUMMARY
------------------------------------------------------------
This analysis evaluated 249 codon sites across 29 terminal evolutionary branches.

Scores represent relative episodic selection burden:
- Higher values indicate lineage-specific evolutionary stress
- Interpretation is relative, not absolute

2. GENE-LEVEL EPISODIC IDENTIFIABILITY
------------------------------------------------------------
Observed dispersion variance: 0.393628
Neutral expected variance:    0.323885
Dispersion ratio (obs/neutral): 1.215
Standardized EII_z:            +2.357
Empirical p-value:             0.01198

3. TERMINAL BRANCHES WITH STRONGEST GLOBAL DEVIATION
------------------------------------------------------------
 1. AMIN007689-RA_Anopheles_minimus_MINIMUS1 background=0.0623  z=+1.81
 2. AARA21_007184.R13477_Anopheles_arabiensis_DONGOLA_2021 background=0.0623  z=+1.81
 3. AMEM21_000490.R666_Anopheles_merus_MAF_2021 background=0.0620  z=+1.75
 4. ACRUBR1_004037.R3815_Anopheles_cruzii_AcruBR1 background=0.0618  z=+1.72
 5. ACHR002344-RA_Anopheles_christyi_ACHKN1017 background=0.0618  z=+1.72
 6. AFAF004115-RA_Anopheles_farauti_FAR1     background=0.0580  z=+0.87
 7. AMAM018008-RA_Anopheles_maculatus_maculatus3 background=0.0572  z=+0.71
 8. AZIE001239.R665_Anopheles_ziemanni_AzieGA1 background=0.0569  z=+0.65
 9. ADAR2_011252.R18153_Anopheles_darlingi_AdarGF1#1 background=0.0562  z=+0.49
10. ASIC013695-RA_Anopheles_sinensis_China   background=0.0561  z=+0.46

Interpretation:
Branches listed above show unusually high overall evolutionary
stress compared to other terminal lineages.

4. SITES WITH STRONGEST EPISODIC SELECTION BURDEN
------------------------------------------------------------
Site   27  score=0.9915  z=+1.62
Site  129  score=0.9885  z=+1.53
Site  244  score=0.9880  z=+1.52
Site  231  score=0.9878  z=+1.51
Site  148  score=0.9845  z=+1.42
Site  103  score=0.9843  z=+1.41
Site  240  score=0.9841  z=+1.41
Site  205  score=0.9836  z=+1.39
Site   26  score=0.9832  z=+1.38
Site  217  score=0.9817  z=+1.34
Site  224  score=0.9795  z=+1.28
Site  146  score=0.9791  z=+1.26
Site  134  score=0.9791  z=+1.26
Site  218  score=0.9790  z=+1.26
Site   20  score=0.9787  z=+1.25
Site  152  score=0.9783  z=+1.24
Site  199  score=0.9783  z=+1.24
Site  143  score=0.9782  z=+1.24
Site  201  score=0.9779  z=+1.23
Site  233  score=0.9772  z=+1.21

Interpretation:
These codon positions show unusually high episodic deviation,
likely driven by a subset of branches rather than uniform
divergence across the tree.

5. INTERPRETATION NOTES
------------------------------------------------------------
- Internal tree nodes are strictly excluded from reporting.
- Scores are not dN/dS or omega estimates.
- Values reflect relative evolutionary stress.
- Gene-level EII statistics reflect deviation from tree-conditional
  neutral simulations (if calibration enabled).
- Per-branch site lists indicate putative drivers, not
  definitive causal substitutions.
- High-ranking sites should be mapped to protein domains
  or structural models for validation.
```

S7. Empirical Evaluation on the FBN30 Dataset

S7.1 Dataset characteristics and inference setting

The FBN30 alignment comprises $L = 249$ codon sites sampled across $K = 29$ terminal taxa. Internal nodes were excluded from all reporting. Inference was performed using the single frozen reference model, without retraining, fine-tuning, or dataset-specific adjustment.

S7.2 Site-level dispersion profile

Sigmoid-transformed site summaries satisfied

$$0.9772 \leq s_i \leq 0.9915.$$

The empirical distribution was unimodal, bounded, and continuous, with no evidence of heavy-tailed inflation or boundary saturation. Variation was moderate and concentrated within a narrow range, indicating structured but not extreme episodic dispersion across codons.

S7.3 Branch-level background structure

Branch-level background summaries satisfied

$$0.0561 \leq \beta_k \leq 0.0623, \quad \mu_\beta = 0.0586, \quad \sigma_\beta^2 = 3.0 \times 10^{-6}.$$

Standardized deviations ranged from

$$+0.46 \leq z_k \leq +1.81.$$

The distribution was continuous and unimodal, with no collapse toward 0 or 1 boundaries. Given the small variance, even modest standardized deviations represent measurable contribution to overall dispersion. Importantly, the magnitude of these values reflects relative heterogeneity across lineages rather than adaptive inference.

S7.4 Calibrated gene-level dispersion

Tree-conditional Monte Carlo calibration yielded

$$\mathrm{EII}_z = +2.357, \quad p = 0.01198.$$

The observed site-level logit variance therefore exceeded the neutral expectation under the same phylogenetic structure by approximately 21%. This indicates statistically resolvable lineage-specific heterogeneity at the gene level.

Crucially, this calibrated deviation quantifies dispersion relative to a neutral baseline. It does not constitute evidence for adaptive substitutions or episodic bursts of positive selection.

S7.5 Structured concentration across branches and sites

The highest-ranking terminal branches included:

- *Anopheles minimus*
- *Anopheles arabiensis*
- *Anopheles merus*
- *Anopheles cruzii*
- *Anopheles christyi*

These lineages exhibited moderate standardized deviations ($z \approx 1.7$–1.8), indicating disproportionate contribution to global dispersion relative to other taxa.

Top-ranked codons included:

$$\{27, 129, 244, 231, 148, 103, 240, 205, 26, 217\}.$$

Their scores remained tightly bounded within the global site-level range, reflecting localized but moderate concentration of episodic deviation. Permutation controls (Section S3) eliminate rank coherence, confirming that this structure depends on phylogenetically coherent positional information rather than marginal composition.

S7.6 Integrated interpretation

The empirical profile is characterized by:

1. Statistically resolvable gene-level dispersion relative to neutrality.
2. Continuous and bounded branch-level heterogeneity.
3. Concentration of dispersion within a limited subset of codons.
4. Absence of heavy-tailed inflation or boundary divergence.

Taxon jackknife stability remained high ($0.608 \leq \rho_{(-k)} \leq 0.954$), indicating that inferred structure does not depend catastrophically on any single lineage.

Collectively, these results indicate moderate, distributed episodic structure across the phylogeny. The pattern is inconsistent with a singular extreme burst localized to one lineage or one codon, and instead reflects structured but bounded heterogeneity.

S7.7 Interpretational boundaries

BABAPPAΩ inference does not:

- Estimate d_N/d_S,
- Perform likelihood ratio testing,
- Assign posterior probabilities of selection,
- Infer mechanistic causality.

Elevated site- or branch-level summaries indicate statistical measurability of lineage-specific dispersion, not adaptive substitution events. Mechanistic interpretation requires complementary codon likelihood models or functional validation.

Concise evolutionary interpretation. For an evolutionary biologist, these results indicate that the gene exhibits moderate but statistically resolvable lineage-specific heterogeneity across 29 terminal *Anopheles* lineages. The observed site-level dispersion is approximately 21% higher than expected under tree-matched neutrality ($\mathrm{EII}_z = 2.357$, $p = 0.01198$), demonstrating that the alignment contains more structured branch–site variation than predicted by neutral evolution alone.

Importantly, the signal is not extreme or uniformly distributed. Instead, it is concentrated in a subset of lineages (e.g., *A. minimus*, *A. arabiensis*, *A. merus*, *A. cruzii*, *A. christyi*) and a limited number of codon positions (e.g., sites 27, 129, 244, 231). This pattern is consistent with episodic, lineage-localized evolutionary shifts rather than genome-wide rate acceleration or a single intense adaptive burst.

These results indicate that certain branches contribute disproportionately to site-level heterogeneity, meaning episodic structure is statistically measurable at the gene level. However, the outputs do not demonstrate adaptive substitution events. Instead, they identify regions and lineages where evolutionary deviation is resolvable and therefore suitable for follow-up analyses using mechanistic codon likelihood models or structural-functional mapping.

S8. Falsifiability and Diagnostic Failure Criteria

Because BABAPPAΩ is formulated as a calibrated measurement instrument rather than a hypothesis-testing procedure, its validity depends on well-defined falsifiability conditions. Failure of any of the following criteria would invalidate the framework or indicate fundamental miscalibration.

1. **Neutral miscalibration.** Under strict tree-conditional neutrality ($\omega = 1$), observed dispersion statistics must fall within the empirically derived Monte Carlo reference distribution. Systematic inflation of

$$\mathrm{EII}_z$$

or consistently small empirical p-values under neutrality would indicate false-positive identifiability and invalidate the calibration procedure.

2. **Permutation invariance.** Destruction of positional structure via codon-column permutation must eliminate coherent site-level ranking. Preservation of rank correlation

$$\rho(\ell_i, \ell_i^{\mathrm{perm}})$$

near unity would imply that inferred episodic structure does not depend on phylogenetically coherent positional information, thereby falsifying structural specificity.

3. **Catastrophic jackknife instability.** Removal of any single terminal taxon should not collapse inferred structure. If taxon jackknife correlations satisfy

$$\rho_{(-k)} \approx 0$$

for any lineage, this would indicate pathological dependence on individual taxa and violate robustness criteria.

4. **Boundary divergence or saturation.** Site-level or branch-level summaries must remain bounded and continuous. Systematic collapse toward 0 or 1 under moderate signal, or heavy-tailed inflation of logit-scale dispersion, would indicate numerical instability or representational pathology.

5. **Tree-size instability.** Under increasing phylogenetic sampling (e.g., $K = 8 \to 64$ taxa), dispersion variance should contract or remain stable. Explosive growth in variance with increasing K would signal scale instability.

6. **Misspecification leakage.** Under constant-ω substitution regimes lacking episodic structure, BABAPPAΩ should not exhibit strong monotonic dependence on ω. Robust recovery of dN/dS parameters would indicate implicit likelihood-model mimicry and contradict the intended likelihood-free design.

Across all simulated and empirical evaluations reported in this study, none of these failure modes were observed. BABAPPAΩ therefore satisfies neutrality calibration, structural specificity, perturbation stability, boundedness, and model-agnostic behavior under misspecification.

S9. Deterministic Reproducibility

Repeated executions produce identical outputs.

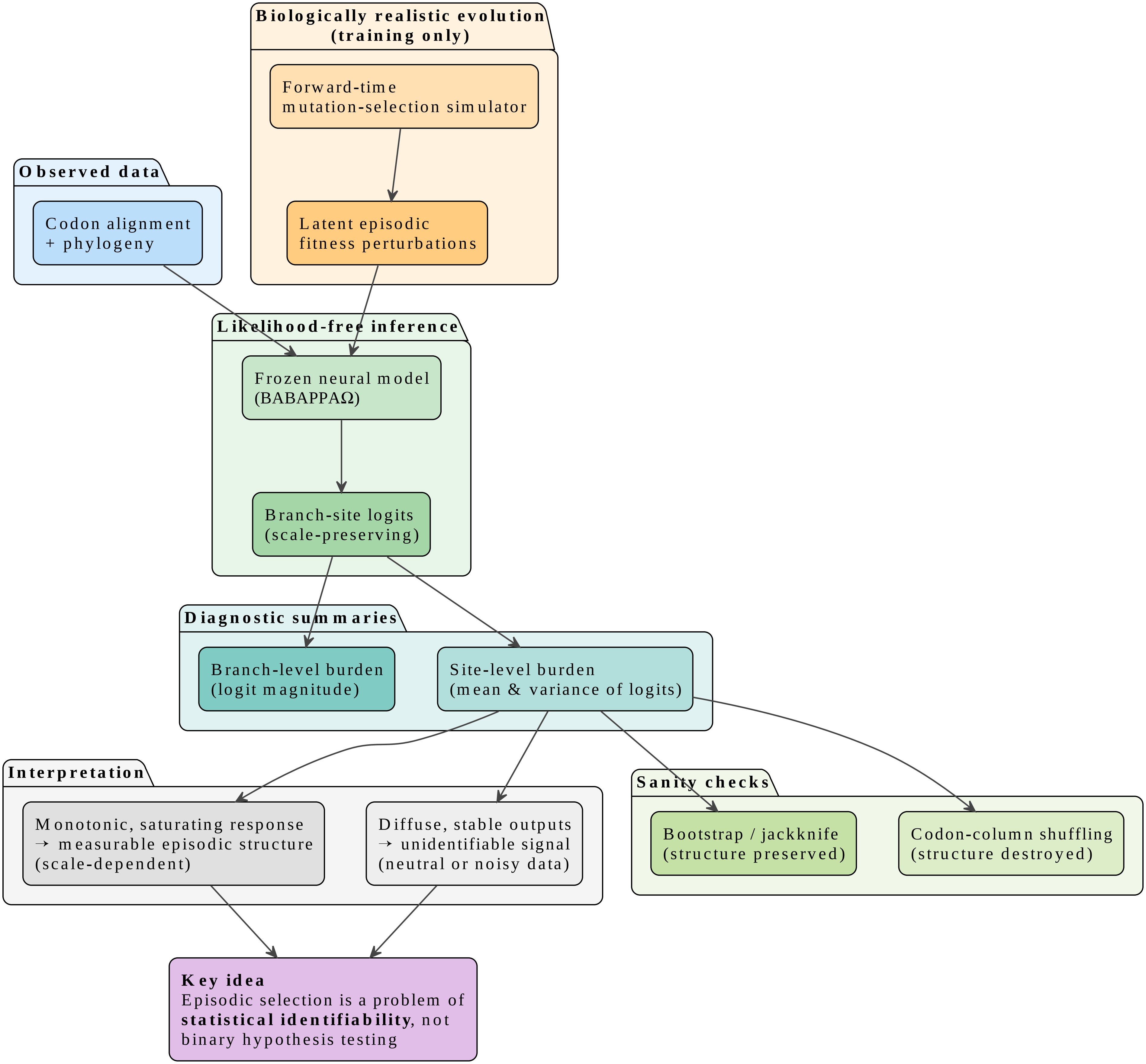
S10. Overall Conclusion

Across simulated and empirical regimes, BABAPPAΩ behaves as a tree-conditionally calibrated, bounded, structurally sensitive measurement instrument.

Gene-level identifiability is quantified via the Episodic Identifiability Index (EII), standardized relative to neutral Monte Carlo expectation.

This calibration constrains inflation, prevents overinterpretation, and situates episodic dispersion within a statistically grounded reference framework. It is not a binary adaptive detector, nor a replacement for codon likelihood models.

BABAPPAΩ
Likelihood-free diagnosis of episodic selection identifiability


Biologically realistic evolution (training only)
Forward-time mutation-selection simulator
Latent episodic fitness perturbations
Observed data
Codon alignment + phylogeny
Likelihood-free inference
Frozen neural model (BABAPPAΩ)
Branch-site logits (scale-preserving)
Diagnostic summaries
Branch-level burden (logit magnitude)
Site-level burden (mean & variance of logits)
Interpretation
Monotonic, saturating response → measurable episodic structure (scale-dependent)
Diffuse, stable outputs → unidentifiable signal (neutral or noisy data)
Sanity checks
Bootstrap / jackknife (structure preserved)
Codon-column shuffling (structure destroyed)
Key idea
Episodic selection is a problem of statistical identifiability, not binary hypothesis testing