

## Supplementary Material 4

Diagnostic Figures, Sensitivity Analyses, and Computational Cost Table

Chen Xu

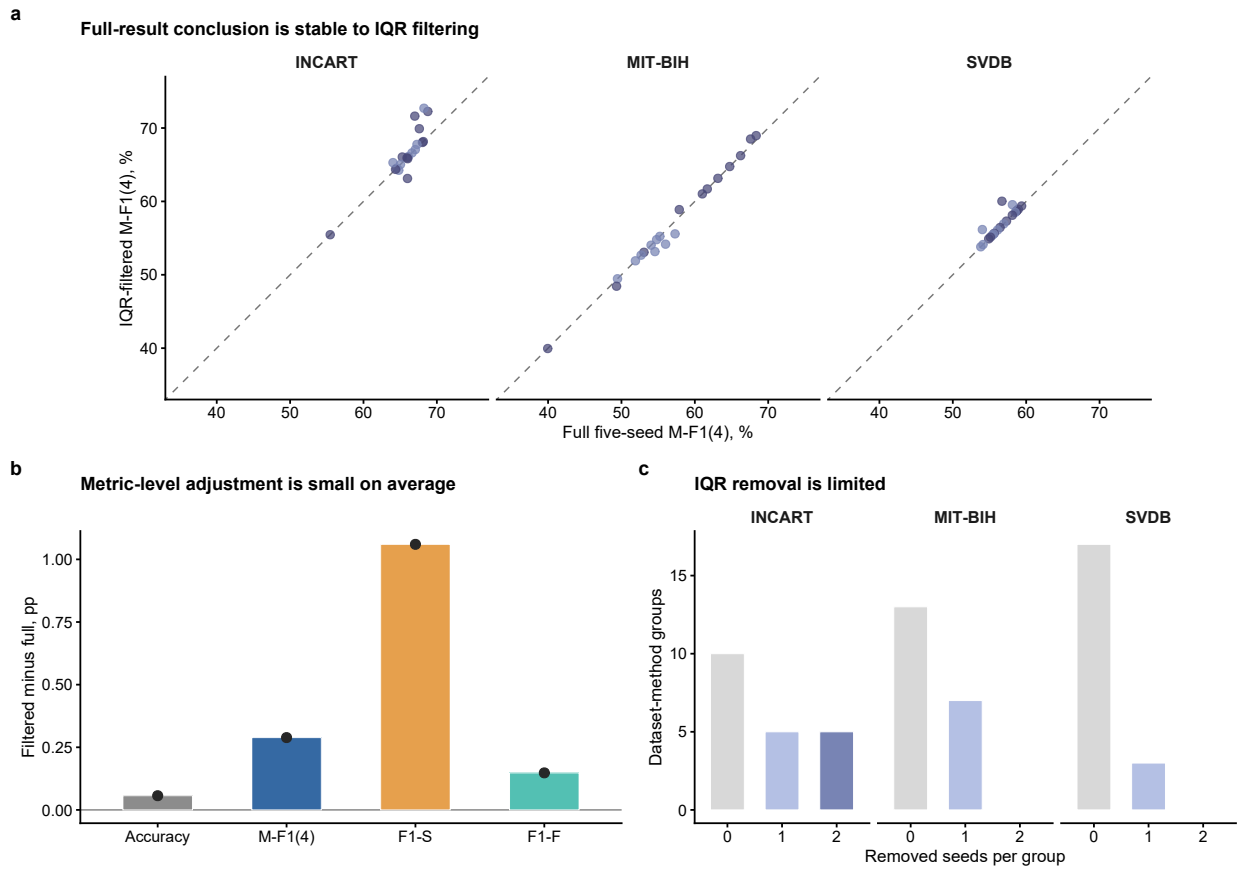
School of Information Science and Engineering, Lanzhou University

Corresponding author: [xchen2024@lzu.edu.cn](mailto:xchen2024@lzu.edu.cn)

ORCID: [0009-0009-8592-4336](https://orcid.org/0009-0009-8592-4336)

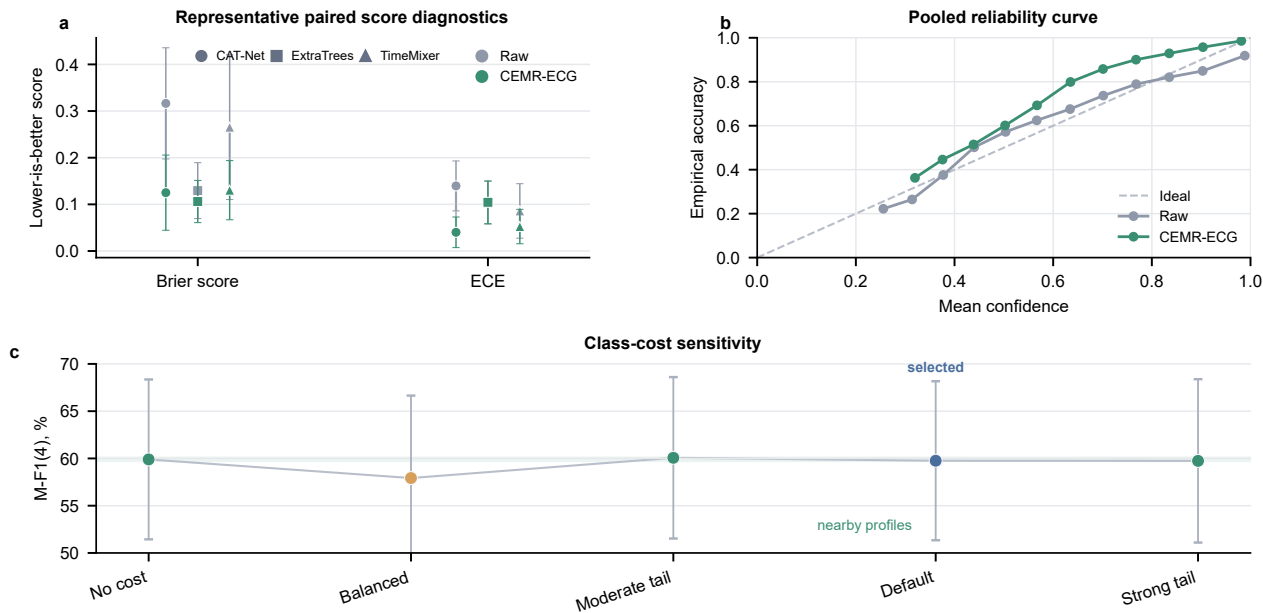
*This Online Resource accompanies the manuscript “Clinical Evidence Morphology–Rhythm for ECG (CEMR-ECG): a traceable evidence framework for imbalanced heartbeat classification” submitted to Physical and Engineering Sciences in Medicine. It collects diagnostic and sensitivity figures that support, but are not required for, the main narrative. The main text retains the framework, primary cross-backbone gains, class-wise behavior, ablation evidence, stability, and score-quality summaries.*

## Supplementary Figure S1. IQR outlier robustness



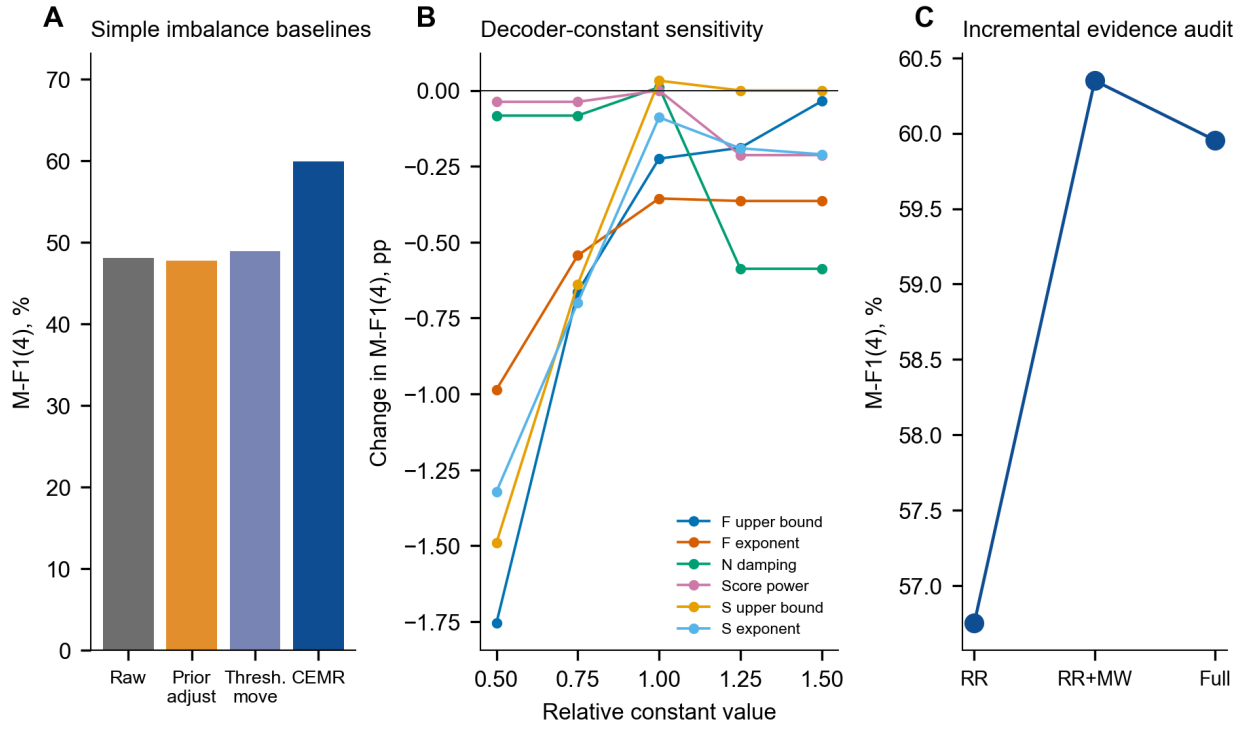
IQR-filtered robustness check. Seed-level outliers were removed within each dataset–backbone group using a  $1.5 \times \text{IQR}$  rule on M-F1(4). The complete five-seed result remains the primary analysis; this figure shows that filtering did not reverse the direction of the framework effect.

# Supplementary Figure S2. Representative calibration sensitivity



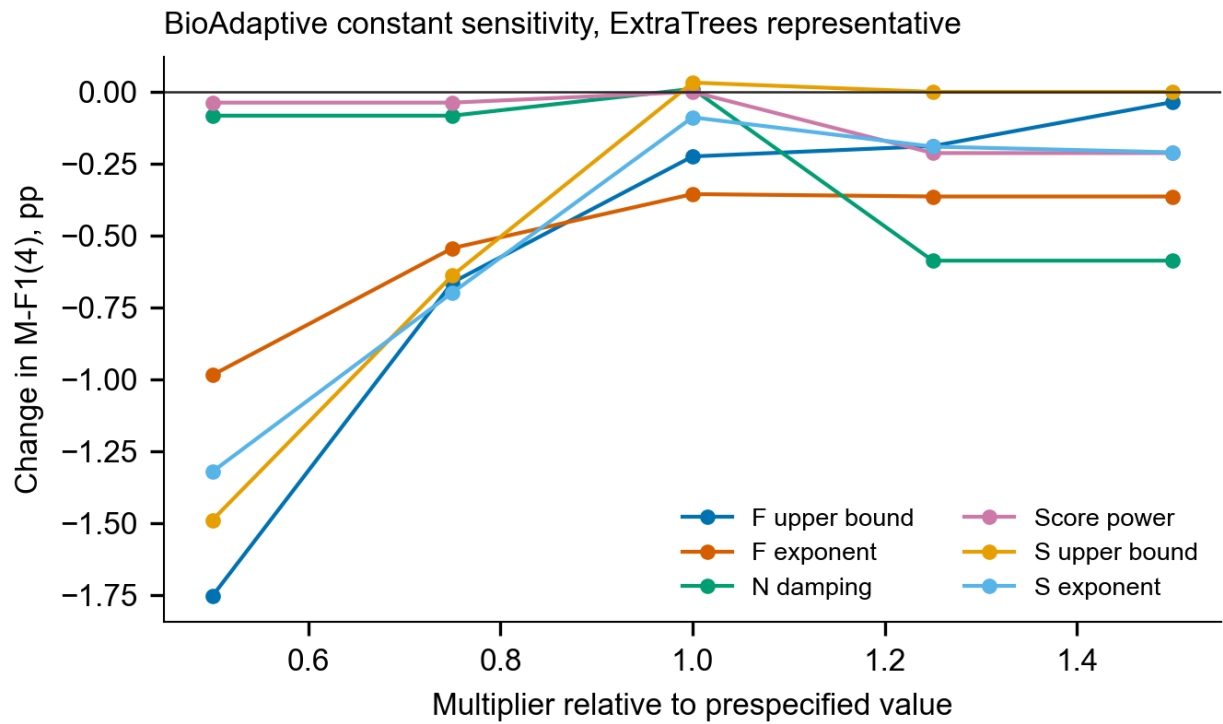
Representative score-quality and calibration sensitivity. This figure provides auxiliary reliability and score-quality diagnostics for representative decision-score interfaces.

## Supplementary Figure S3. Representative baseline and sensitivity checks



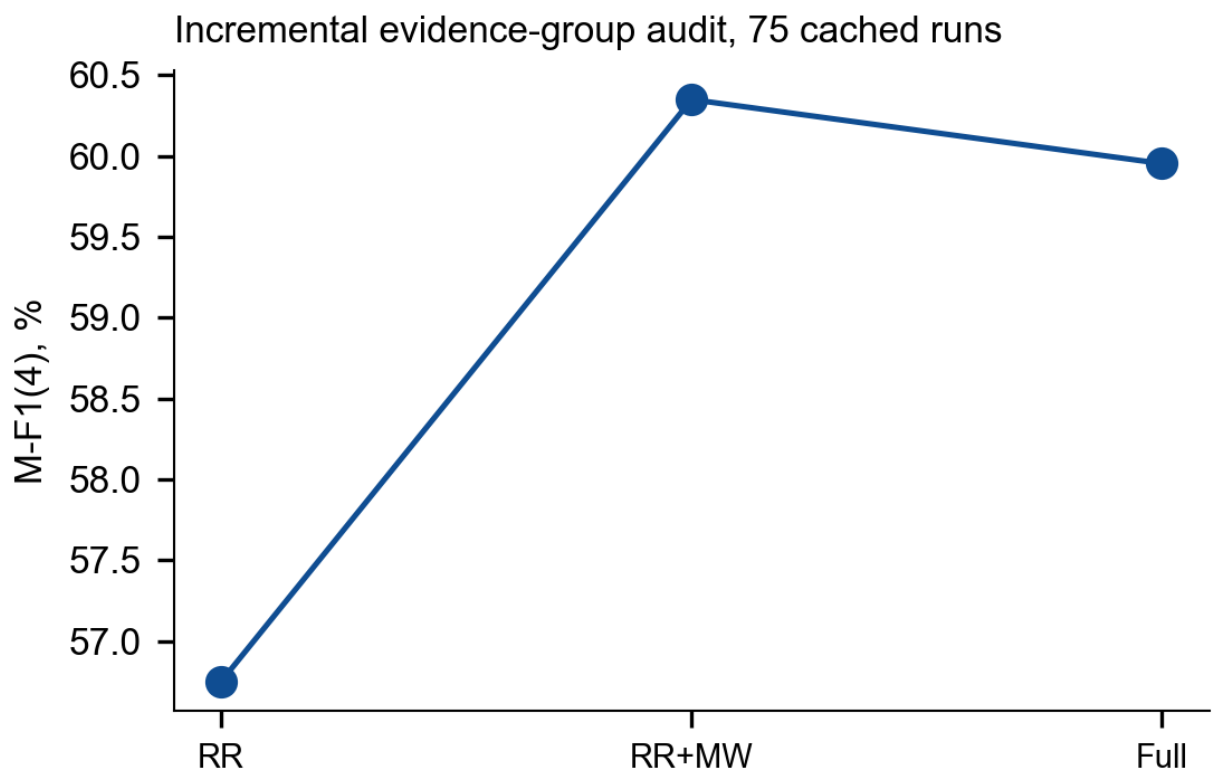
Representative sensitivity and baseline checks. The figure compares raw backbones, validation-selected prior/logit adjustment, validation-selected threshold moving, CEMR-ECG, BioAdaptive constant sensitivity, and incremental evidence-group behavior.

## Supplementary Figure S4. Decoder constant sensitivity



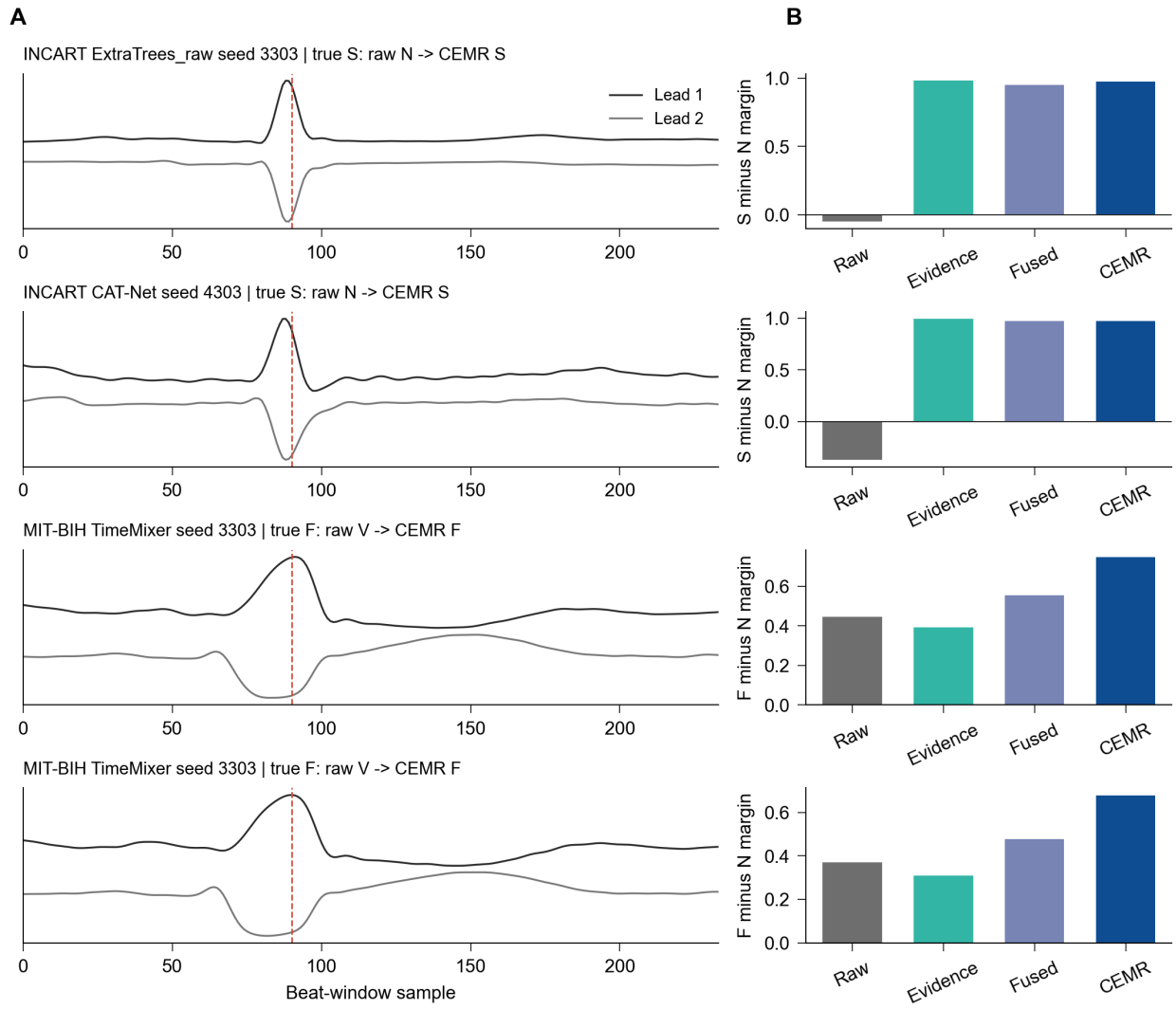
BioAdaptive constant sensitivity for the representative ExtraTrees interface. S/F exponents, S/F upper bounds, score power, and N damping were varied around the fixed defaults.

Supplementary Figure S5. Leave-one-group-in evidence check



Leave-one-group-in evidence check. RR-only evidence supplied the largest aggregate-performance contribution, while morphology/waveform and prototype channels provided named traceability records and partly redundant secondary evidence.

## Supplementary Figure S6. Traceable recovered-beat examples



Traceable recovered-beat examples. Recovered S and F examples show raw, evidence, fused, and final CEMR-ECG class-versus-N margins. The examples are illustrative decision traces from public-database runs and are not patient-level clinical validation.

## Supplementary Table S1. Computational cost summary

*Prediction and training times are reported in seconds and were measured on the experimental compute environment. Medians are computed over all evaluated seed runs pooled across datasets and backbones ( $n = 300$  for the overall summary;  $n = 99$  per family across three datasets and 33 backbone-dataset pairs). Memory usage was not systematically instrumented and is not reported; characterization of peak memory demand and deployment-relevant resource consumption is left for future work.*

**Table S1.** Prediction and training time summary across 300 paired runs.

| Scope                           | Prediction time (s)            |               | Training time (s) |               |
|---------------------------------|--------------------------------|---------------|-------------------|---------------|
|                                 | Raw backbone                   | Full CEMR-ECG | Raw backbone      | Full CEMR-ECG |
| Overall (median; $n = 300$ )    | 1.72                           | 2.73          | 83.4              | 219.7         |
| Q1–Q3                           | 0.43–9.66                      | 1.46–10.56    | 23.3–376.8        | 158.4–504.7   |
| Machine-learning family         | 0.57                           | 1.50          | 28.8              | 166.6         |
| Deep/time-series family         | 3.84                           | 4.76          | 481.1             | 609.8         |
| Evidence layer only             | –                              | 0.93          | –                 | 135.8         |
| Q1–Q3                           | –                              | 0.90–1.08     | –                 | 97.6–155.4    |
| Combined/raw ratio (prediction) | median 1.57 (Q1–Q3: 1.11–3.57) |               |                   |               |
| Combined/raw ratio (training)   | median 2.29 (Q1–Q3: 1.30–6.39) |               |                   |               |

*Notes:* Times are in seconds per run (one dataset–backbone–seed combination). The evidence layer adds a fixed-overhead encoding and fitting step (median 136 s training; median 0.93 s prediction) that is largely independent of the backbone architecture, which explains the larger relative overhead for fast machine-learning backbones. Per-10,000-beat prediction times: raw 0.307 s, evidence 0.175 s, combined 0.475 s (medians over all runs). No memory profiling was performed; the evidence vector dimensionality is 362 features per beat, and the evidence encoder is a single ExtraTrees classifier with 700 trees.