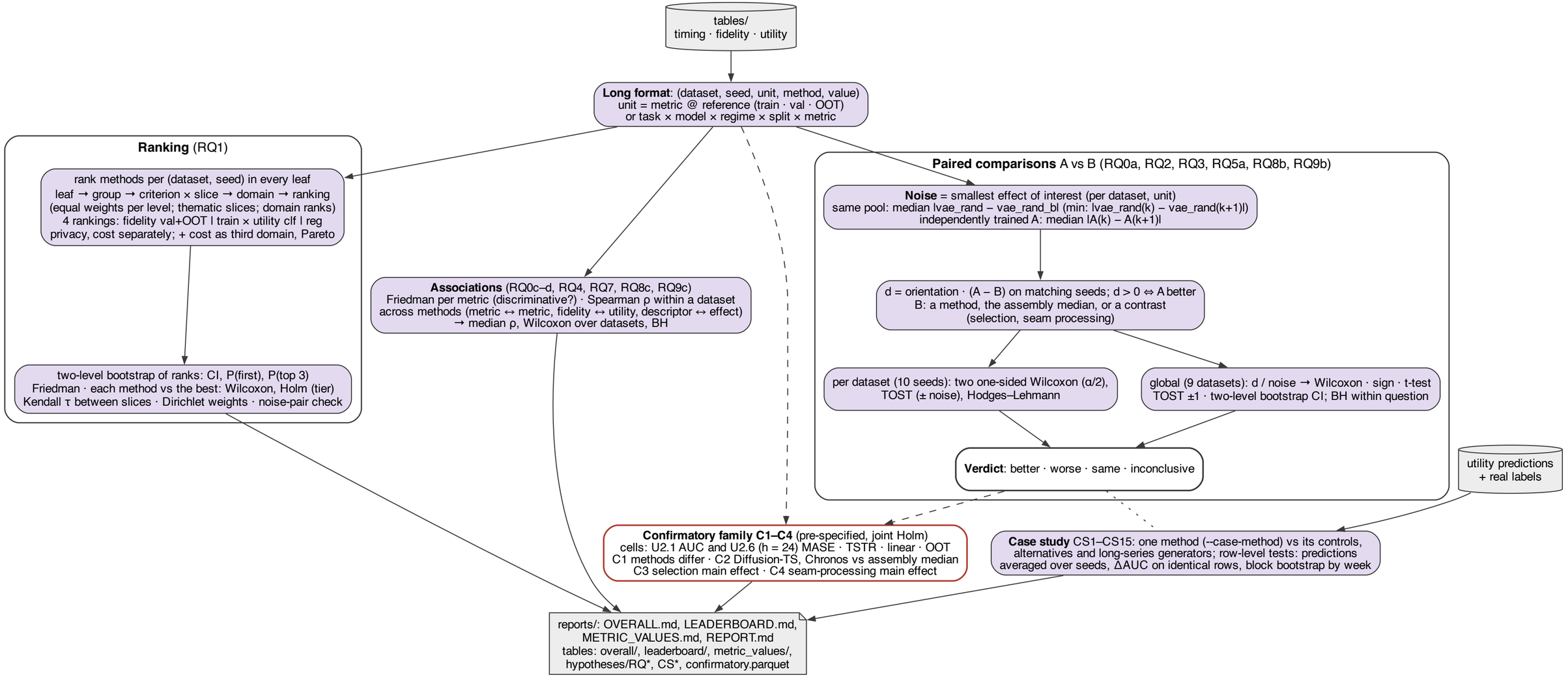

tables/
timing · fidelity · utility
Long format: (dataset, seed, unit, method, value)
unit = metric @ reference (train · val · OOT)
or task × model × regime × split × metric
Ranking (RQ1)
rank methods per (dataset, seed) in every leaf
leaf → group → criterion × slice → domain → ranking
(equal weights per level; thematic slices; domain ranks)
4 rankings: fidelity val+OOT | train × utility clf | reg
privacy, cost separately; + cost as third domain, Pareto
two-level bootstrap of ranks: CI, P(first), P(top 3)
Friedman · each method vs the best: Wilcoxon, Holm (tier)
Kendall τ between slices · Dirichlet weights · noise-pair check
Associations (RQ0c–d, RQ4, RQ7, RQ8c, RQ9c)
Friedman per metric (discriminative?) · Spearman ρ within a dataset
across methods (metric ↔ metric, fidelity ↔ utility, descriptor ↔ effect)
→ median ρ, Wilcoxon over datasets, BH
Paired comparisons A vs B (RQ0a, RQ2, RQ3, RQ5a, RQ8b, RQ9b)
Noise = smallest effect of interest (per dataset, unit)
same pool: median |vae_rand – vae_rand_bl (min: |vae_rand(k) – vae_rand(k+1)|)
independently trained A: median |A(k) – A(k+1)|
d = orientation · (A – B) on matching seeds; d > 0 ⇔ A better
B: a method, the assembly median, or a contrast
(selection, seam processing)
per dataset (10 seeds): two one-sided Wilcoxon (α/2),
TOST (± noise), Hodges–Lehmann
global (9 datasets): d / noise → Wilcoxon · sign · t-test
TOST ±1 · two-level bootstrap CI; BH within question
Verdict: better · worse · same · inconclusive
utility predictions
+ real labels
Confirmatory family C1–C4 (pre-specified, joint Holm)
cells: U2.1 AUC and U2.6 (h = 24) MASE · TSTR · linear · OOT
C1 methods differ · C2 Diffusion-TS, Chronos vs assembly median
C3 selection main effect · C4 seam-processing main effect
Case study CS1–CS15: one method (--case-method) vs its controls,
alternatives and long-series generators; row-level tests: predictions
averaged over seeds, ΔAUC on identical rows, block bootstrap by week
reports/: OVERALL.md, LEADERBOARD.md,
METRIC_VALUES.md, REPORT.md
tables: overall/, leaderboard/, metric_values/,
hypotheses/RQ*, CS*, confirmatory.parquet