AcneCascade-SEM · acne pathogenesis

Faija · faijazamil1@gmail.com

1 · PROBLEM

what and why

Acne pathogenesis is described qualitatively as a cascade — C. acnes strain/sebum shift → dysbiosis → host immune activation — but no model has actually inferred the causal ordering and cross-modal lag structure from real multi-omic data. A 2024 Cell Host & Microbe study profiled 1,234 C. acnes isolate genomes with integrated genome, transcriptome, and metabolome data, showing that skin disease state and body site shape strain-level genomic and functional differences — but the analysis is cross-sectional and correlational. Existing causal-discovery methods (DYNOTEARS, Rhino) assume a single modality and synchronized integer-lag sampling, which does not fit three asynchronous omic layers propagating through skin at different rates. This project tests whether a cross-modal, lag-aware causal model can recover the true cascade ordering from existing public data.

2 · APPROACH

model / method, and why this one

AcneCascade-SEM has three components: (1) per-modality encoders — lightweight autoencoders for strain-genomic profile, metabolomic output, and host transcriptional signature into a shared latent space; (2) a modality-pair lag operator using disease-severity stage as a validated pseudotime axis; (3) a sparse directed graph layer across modalities, with acyclicity enforced only between modalities — not within — since feedback loops are biologically expected. This design was chosen over naive correlation/mediation analysis because its output is directly falsifiable against known biology: recovering the assumed clinical ordering validates the method; recovering a different ordering is a novel, reportable finding.

3 · DATASET

source, size, preprocessing

Public data only — no new sequencing required. The Cell Host & Microbe multi-omic C. acnes dataset (1,234 isolates: genome, transcriptome, metabolome; healthy/atopic dermatitis/acne-stratified), plus supplementary host-transcriptome acne datasets on NCBI GEO.

4 · METRICS

how you’ll know it worked

how you’ll know it worked Whether the recovered causal ordering matches the clinically assumed cascade (validation) or diverges from it (novel finding). Statistical comparison of the learned graph against a null/random graph baseline. Robustness checks on the pseudotime substitution (e.g. varying severity-stage binning).

5 · COMPUTE

training time, memory, dependencies

Small-scale tabular/omic ML. Modest cloud GPU access for encoder training and hyperparameter sweeps (GPU-days to GPU-week scale). Open-source stack: PyTorch, scikit-bio, causal-discovery libraries.

6 · DELIVERABLES

model, paper, demo

- · Cross-modal causal graph with reported ordering and confidence.
- · Code for encoders, lag operator, and graph layer.
- · Write-up comparing recovered cascade vs clinical assumption.
- · Robustness analysis of the pseudotime-as-lag substitution.

7 · RISK

what breaks first, and plan B

what breaks first, and plan B Pseudotime substitution is a methodological compromise, not true longitudinal sampling. Plan B: report the learned graph descriptively (difference from null) rather than claiming temporal causality outright. Structural risk: the researcher does not code solo. Primary ask is a technical collaborator to implement encoders, lag operator, and graph layer while the researcher leads scientific direction and write-up.