A178 · Agentic In Silico Testing of LLM-Generated Biomedical Hypotheses: A CAR-T Biomarker Case Study

Presenter: Yunmai Wang — Computational Biology and Biomedical Informatics, Yale University

Authors: Yunmai Wang, Brian Ondov, Hua Xu

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Although large language models (LLMs) can rapidly generate biomedical hypotheses, experimental validation remains a bottleneck. Public resources such as GEO enable in silico screening before laboratory experiments. Yet free-text hypotheses may conflate mechanisms and clinical claims or involve multiple genes and outcomes. We present an agentic pipeline that converts free-text hypotheses into atomic subclaims, normalizes entities and qualifiers, and routes relations to executable, provenance-preserving evidence workflows.

As a case study, we examined the LLM-generated hypothesis that "the ratio of CD8+ T-cell abundance to GRB10 expression determines CAR-T efficacy and may serve as an independent outcome biomarker." The framework decomposed this claim into three prespecified components: higher CD8+ abundance predicts favorable response; lower GRB10 expression predicts favorable response; and the ratio outperforms either component alone.

Our framework automatically identified and retrieved seven public single-cell cohorts and evaluated these components using prespecified endpoints. Lower GRB10 showed modest discrimination (median AUC, 0.588; IQR, 0.570–0.629), with favorable-direction AUCs in six cohorts. The ratio modestly exceeded GRB10 alone in all seven cohorts (median ΔAUC, 0.027); no paired bootstrap confidence interval excluded zero. The ratio showed an exploratory signal for early CAR-T response in two cohorts, accompanied by antigen-responsive GRB10 expression, but this signal did not extend to durable response or long-term persistence. Evidence is needed to establish the ratio as a critical or independent predictor. Nevertheless, the results support GRB10 as a plausible biomarker component and demonstrate the feasibility of automated in silico evaluation for screening LLM-generated hypotheses.

Agentic In Silico Testing of LLM-Generated Biomedical Hypotheses: A CAR-T Biomarker Case Study

YUNMAI WANG*[1], Brian Ondov[2], Hua Xu[2]

[1] *Computational Biology and Biomedical Informatics, Yale University, New Haven, CT, USA*

[2]*Department of Biomedical Informatics and Data Science, Yale School of Medicine, Yale University, New Haven, CT, USA*

**Presenting author*

Although large language models (LLMs) can rapidly generate biomedical hypotheses, experimental validation remains a bottleneck. Public resources such as GEO enable in silico screening before laboratory experiments. Yet free-text hypotheses may conflate mechanisms and clinical claims or involve multiple genes and outcomes. We present an agentic pipeline that converts free-text hypotheses into atomic subclaims, normalizes entities and qualifiers, and routes relations to executable, provenance-preserving evidence workflows.

As a case study, we examined the LLM-generated hypothesis that "the ratio of CD8+ T-cell abundance to *GRB10* expression determines CAR-T efficacy and may serve as an independent outcome biomarker." The framework decomposed this claim into three prespecified components: higher CD8+ abundance predicts favorable response; lower *GRB10* expression predicts favorable response; and the ratio outperforms either component alone.

Our framework automatically identified and retrieved seven public single-cell cohorts and evaluated these components using prespecified endpoints. Lower *GRB10* showed modest discrimination (median AUC, 0.588; IQR, 0.570–0.629), with favorable-direction AUCs in six cohorts. The ratio modestly exceeded *GRB10* alone in all seven cohorts (median ΔAUC, 0.027); no paired bootstrap confidence interval excluded zero. The ratio showed an exploratory signal for early CAR-T response in two cohorts, accompanied by antigen-responsive *GRB10* expression, but this signal did not extend to durable response or long-term persistence. Evidence is needed to establish the ratio as a critical or independent predictor. Nevertheless, the results support *GRB10* as a plausible biomarker component and demonstrate the feasibility of automated in silico evaluation for screening LLM-generated hypotheses.

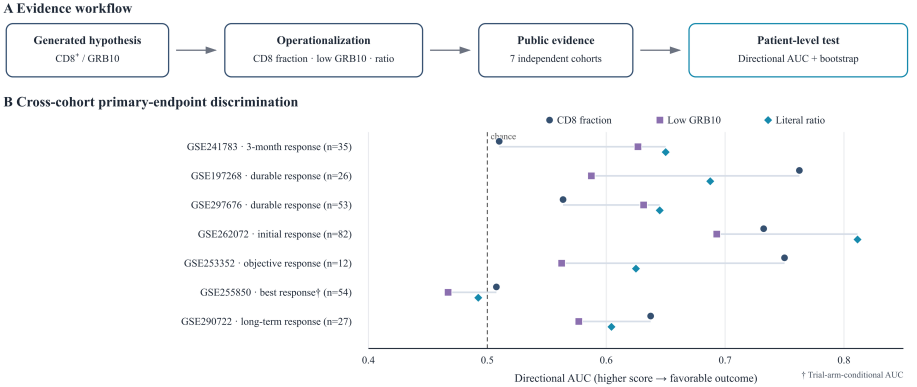

Figure 1. Agent-assisted decomposition and cross-cohort testing of the CD8⁺/*GRB10* hypothesis. Directional AUCs are shown for one primary endpoint per cohort; the dashed line marks chance discrimination. The ratio exceeded low *GRB10* in 7/7 cohorts but exceeded CD8 fraction in only 3/7; no paired bootstrap interval for incremental AUC excluded zero.