TRACE: STUDY-LEVEL TRANSCRIPTOMIC HYPOTHESIS GENERATION WITH BIOLOGICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounded biological reasoning on transcriptomic studies remains a major challenge for genomic and large language models. While existing transcriptomic foundation models learn cell- and gene-level representations, they operate primarily on individual cells and struggle to generate mechanistic explanations for study-level phenomena. General-purpose large language models (LLMs) can generate plausible biological narratives, yet they frequently miss the regulatory pathways and molecular cascades driving transcriptomic signatures. We formulate a task where an LLM is given a disease context alongside descriptions of two biological conditions and must predict the intermediate mechanistic cascades separating them. Using curated condition contrasts from RummaGEO, we construct a benchmark and training dataset of 15,251 human and mouse mechanistic reasoning samples. Each sample maps biological conditions to grounded multi-scale cascades of kinases, transcription factors, and biological processes derived from empirical differential expression enrichment. We then train a TRACE (Transcriptomic Reasoning & Analysis of Condition Expression), an open-source 4B-parameter reasoning model based on Qwen3.5, and conduct rubric-based evaluation of hypothesis generation assessing reasoning trace coherence, final mechanistic summaries, and recovery of top enriched regulatory entities. Across held-out GEO studies and unseen disease contexts, TRACE recovers more empirically supported regulatory terms and mechanistic pathways than open-source and frontier model baselines. Case studies confirm that our model generates supported mechanistic cascades in its reasoning. We release the 15k-sample reasoning dataset, the TRACE model, and the enrichment-to-trace synthesis pipeline to facilitate scalable mechanistic reasoning in transcriptomics.

1 INTRODUCTION

Advancements in omics technologies - spanning transcriptomics, genomics, and perturbation screens - have driven the growth of large-scale publicly-available datasets that complement existing biomedical literature, creating both opportunities and bottlenecks for computational biology (Dai & Shen, 2022). The increasing data scale has driven the development of transcriptomic foundation models that learn rich representations of genes and cells from large-scale RNA sequencing (RNA-seq) data. In single-cell RNA-seq, foundation models have demonstrated strong performance on downstream tasks such as cell-type annotation, perturbation response prediction, and gene-function inference (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Rosen et al., 2023; Adduri et al., 2025). However, as biological datasets become larger and more heterogeneous, there is increasing need for models that not only predict accurately, but also produce outputs that facilitate interpretation of mechanisms and disease processes.

Despite impressive performance, current transcriptomic foundation models have two key limitations. First, most models provide limited interpretability beyond attention analysis of model layers, making it difficult to translate latent representations into mechanistic hypotheses and insights (Kendiukhov, 2026; Baek et al., 2025). Second, most existing approaches focus on modeling individual cells, whereas multi-cellular context is critical for many biologically and clinically relevant questions at the patient and study level, where experimental context, cohort design, and textual metadata all constitute important context for modeling (Bunne et al., 2024). Addressing this gap requires models

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

that formulate natural language hypotheses separating biological conditions, grounding mechanistic explanations in verifiable regulatory cascades and empirical study-level signatures.

Connecting high-level condition descriptions to empirical gene expression requires an appropriate level of biological abstraction. Rather than modeling individual, unannotated gene lists, functional enrichment workflows aggregate differential expression into structured gene sets, pathways, and regulatory processes. Community-curated knowledge bases such as the Gene Ontology (Ashburner et al., 2000) and KEGG (Kanehisa & Goto, 2000) organize genes into canonical biological processes and pathways, while specialized libraries such as ChEA capture transcription factor targets derived from ChIP-seq experiments (Keenan et al., 2019). Enrichment platforms such as Enrichr (Chen et al., 2013) further aggregate hundreds of gene set libraries from diverse sources, including these resources, into a unified framework for enrichment analysis. Complementary approaches combine kinase and transcription factor libraries to infer regulatory pathways linking upstream signaling to downstream biological processes (Chen et al., 2012; Clarke et al., 2018). In parallel, databases like RummaGEO systematically extract differential expression contrasts from tens of thousands of Gene Expression Omnibus (GEO) studies across diverse perturbations and disease phenotypes (Marino et al., 2024). While prior models have used these resources primarily for representation learning or gene function retrieval (Clarke et al., 2026), they have not been leveraged to train reasoning models that explicitly infer the regulatory cascades separating experimental conditions.

Large language models (Brown et al., 2020; Radford et al., 2019; Radford, 2018; Ouyang et al., 2022; Hurst et al., 2024; Team et al., 2023) offer a complementary paradigm through their ability to synthesize scientific context, follow structured tasks, and perform multi-step causal reasoning. When presented with experimental setups or disease states, modern LLMs can readily formulate plausible mechanistic narratives describing potential downstream effects. In functional genomics, however, plausibility alone is insufficient. When tasked with reasoning over actual study-level contrasts, generalist models frequently hallucinate or rely on canonical associations that miss the specific molecular cascades, upstream regulators, and signaling pathways measurably perturbed in transcriptomic experiments. Bridging this divide requires models specifically trained to ground hypothesis generation in empirical transcriptomic signatures.

In this work, we present a framework to construct high-quality mechanistic reasoning samples for biological condition comparison in transcriptomic studies. To ensure that reasoning is grounded in empirical biology, we establish an enrichment pipeline that analyzes differential expression signatures from RummaGEO samples against established gene set libraries. This enrichment pipeline extracts regulatory cascades linking upstream kinases and transcription factors (TFs) to downstream biological pathways, forming enriched term lists which can be used for grounding and evaluating reasoning traces. We use this pipeline to construct a high-quality cross-species mechanistic reasoning dataset with 15,251 samples, spanning diverse biological conditions and focusing on disease and associated pathway contexts. We additionally train and release TRACE, an open-source 4B-parameter model adapted from Qwen3.5, as a hypothesis generation model that reasons about the molecular cascades separating experimental groups. Through rubric-based benchmarking on held-out studies and unseen disease contexts, we demonstrate that TRACE accurately identifies empirically enriched regulatory terms where existing open-source and commercial frontier language models rely on generic associations. Furthermore, we conduct an analysis of the model's generated reasoning, examining term-level statistics across disease categories and providing detailed biological case studies that demonstrate how the predicted cascades align with genuine experimental mechanisms. Our main contributions are:

- **A cross-species mechanistic reasoning dataset:** We construct and release a dataset of 15,251 human and mouse mechanistic reasoning samples derived from RummaGEO condition comparisons, grounding reasoning traces and final mechanistic explanations in empirical transcriptomic enrichment.
- **A scalable enrichment pipeline:** We establish an end-to-end cross-enrichment framework that systematically mines cascades of kinases, transcription factors, and pathways from differential expression signatures, enabling scalable generation of grounded reasoning traces.
- **An open-source reasoning model:** We train and release TRACE (Qwen3.5 4B), demonstrating through rubric-based evaluations that it surpasses strong baselines at recovering enriched regulatory entities and pathways across unseen studies and diseases.

- **Mechanistic reasoning analyses and case studies:** We perform analyses of TRACE-generated hypotheses, quantifying the recovery of regulatory terms across disease contexts and providing biological case studies that support the predicted molecular cascades.

2 Background & Related Work

2.1 Transcriptomic Foundation Models

Recent years have seen rapid progress in transcriptomic foundation models trained on large-scale RNA sequencing data, typically curated from single-cell databases such as Human Cell Atlas (Regev et al., 2017) and CELLxGENE (Program et al., 2024) which span many organisms, tissues, developmental states, and disease contexts. These models learn dense, generalizable representations of genes and cells through self-supervised learning, and are applied to downstream tasks such as cell-type annotation, perturbation prediction, batch correction, and imputation (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Rosen et al., 2023; Adduri et al., 2025; Levine et al., 2023).

Despite these advances, most transcriptomic foundation models are optimized around cell-level prediction and latent embedding quality rather than study-level mechanistic interpretation. In practice, many biologically meaningful questions are framed at the study level, comparing different experimental conditions (e.g., treatment versus control), where interpretation relies on coherent signatures of up- and down-regulated genes. As first established by the original Connectivity Map (Lamb et al., 2006), leveraging these contrast-driven transcriptional signatures remains a fundamental strategy for translating differential expression into mechanistic insights. This creates a mismatch between the dominant modeling paradigm and the structure of large public study repositories such as GEO, motivating approaches that explicitly model study-level signatures and associated metadata.

2.2 Large Language Models

LLMs (Brown et al., 2020; Radford et al., 2019; Radford, 2018; Ouyang et al., 2022; Hurst et al., 2024; Team et al., 2023) have demonstrated strong capabilities in natural language understanding, structured generation, and multi-step reasoning across scientific and biomedical domains (Zhao et al., 2023). Domain-adapted biomedical LLMs further improve terminology handling and factual grounding in life science contexts by leveraging specialized corpora and instruction tuning (Labrak et al., 2024; Bosselut et al., 2024). These developments have made LLMs increasingly attractive for hypothesis synthesis, literature-guided interpretation, and interactive biological analysis workflows.

Within transcriptomics, Cell2Sentence introduced the idea of transforming transcriptomic profiles into natural language sequences, enabling the direct application of LLMs to single-cell tasks without any architectural modifications (Levine et al., 2023; Rizvi et al., 2026). Other works such as GenePT (Chen & Zou, 2024) use frontier LLMs to derive gene and cell embeddings based on NCBI summaries of genes, indirectly leveraging learned LLM representations for genomic tasks. These existing approaches, however, do not provide interpretable reasoning along with their predictions.

Recently, BioReason (Fallahpour et al., 2025) developed a multimodal DNA-LLM reasoning model for gene pathway and disease prediction. There is not yet, however, a foundation model trained to predict study-level effects while generating interpretable reasoning traces grounded in functional enrichment and biological pathways. Our work builds on this direction by combining study metadata, expression-derived signatures, and language-based reasoning.

2.3 Gene Set Signatures

Gene sets and pathway libraries provide a biologically meaningful abstraction for interpreting high-dimensional expression data, with the original GSEA method (Subramanian et al., 2005) serving as a seminal example of gene set and biological pathway analysis on leukemia and lung cancer data. Curated resources such as KEGG, Gene Ontology, and Reactome now serve as core biological references for functional analysis (Kanehisa & Goto, 2000; Ashburner et al., 2000; Fabregat et al., 2018). Aggregated libraries including MSigDB and Enrichr have standardized downstream enrichment workflows and enabled broad comparative analyses across multiple gene set databases (Liberzon et al., 2011; Chen et al., 2013). In parallel, large-scale automated resources such as Rummagene and

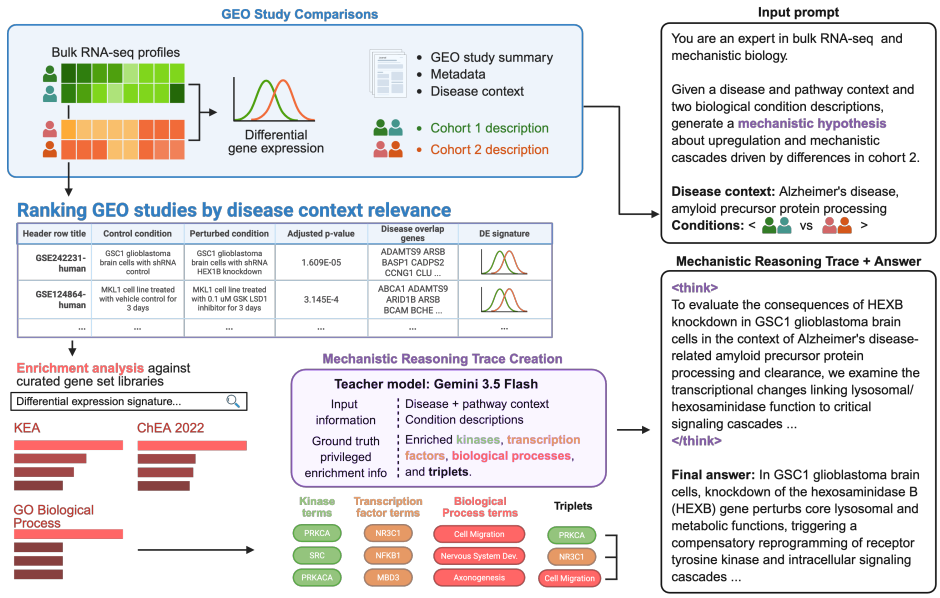


Figure 1: Overview of the TRACE mechanistic hypothesis generation framework. Bulk RNA-seq studies from GEO are ranked by disease-context relevance and enriched against reference libraries (KEA, ChEA, and GO Biological Process) to extract regulatory triplets and top-ranking terms. A teacher model synthesizes step-by-step reasoning traces and final answers conditioned on these empirical cascades, yielding supervised training pairs.

RummaGEO have expanded the available universe of gene signatures mined from publications and GEO studies, respectively (Clarke et al., 2024; Marino et al., 2024). Related high-throughput perturbation resources, including the L1000 connectivity map (Subramanian et al., 2017) and DepMap (Tsherniak et al., 2017), further expand the available landscape of functional signatures.

3 TRACE Framework and Dataset Construction

We present TRACE, a framework that converts disease-relevant differential expression signatures from bulk transcriptomic studies into enriched kinase, TF, and pathway cascades, and uses them to synthesize reasoning traces and mechanistic explanations separating two experimental conditions (Figure 1). For each disease, we (i) paired a Jensen DISEASES gene set with two GO Biological Process context sets, (ii) retrieved RummaGEO signatures enriched for those contexts, (iii) mined kinase–TF–GO cascades from each signature, and (iv) generated teacher reasoning traces conditioned on this hidden enrichment evidence, yielding 15,251 samples used to fine-tune TRACE.

3.1 Disease–GO context pairing

To establish reliable disease contexts for hypothesis sample generation, we paired disease definitions from Jensen DISEASES (1,811 terms) (Pletscher-Frankild et al., 2015) with functional pathway terms from GO Biological Process 2025 (Ashburner et al., 2000) sourced via Enrichr (Chen et al., 2013). We first filtered Jensen DISEASES to disease terms containing between 12 and 400 annotated genes. For each candidate disease gene set, we assessed overlap across all GO Biological Process terms using one-sided Fisher's exact tests against a reference background of 22,000 genes. Candidate disease–GO pairs were required to share at least 5 genes and exhibit a Fisher odds ratio of at least 8.0, from which the two most statistically significant GO terms were retained per disease. We subsequently performed manual review of all disease terms, correcting paired GO terms for 89 of the diseases to ensure biological coherence between each disease and its selected pathway contexts. The final context catalog comprises 263 disease entities, each paired with two GO Biological Process context sets that guide signature retrieval and subsequent cascade construction.

3.2 RANKING RUMMAGEO SAMPLES BY DISEASE CONTEXT RELEVANCE

For each disease context (the Jensen gene set and its two GO terms), we tested every RummaGEO differential expression signature against the three context gene sets using one-sided Fisher's exact tests, with Benjamini–Hochberg FDR correction within each context set. Signatures were retained if they had $\geq$5 overlapping genes, FDR ≤ 0.01 and odds ratio ≥ 2 for at least two of the three sets, ranked by mean rank across sets, and capped at 20 up- and 20 down-regulated signatures per disease and species. For mouse studies, orthologous gene symbols were mapped to human gene symbols to ensure compatibility with Enrichr reference libraries. Condition metadata describing the baseline and perturbed biological states were extracted and standardized from curated RummaGEO study annotations.

3.3 CASCADE MINING AND SAMPLE CONSTRUCTION

For each retained RummaGEO sample for the disease context, we extracted regulatory cascades by performing enrichment analysis of the study's differential expression signature against three complementary gene set libraries: KEA 2015 for upstream kinases (Kuleshov et al., 2021), ChEA 2022 for transcriptional regulators (Keenan et al., 2019), and GO Biological Process 2025 for downstream cellular programs (Ashburner et al., 2000). Enrichments were filtered at FDR ≤ 0.05 with a minimum gene overlap of 2, retaining the top 15 significant terms from each library. To capture coordinated regulatory activity, we mined recurring kinase $\times$ transcription factor $\times$ GO Biological Process triplets across the enriched terms, selecting the top 5 supported triplets alongside the 10 most significant kinases, transcription factors, and pathways from each library as independent lists.

Each remaining RummaGEO sample was then formatted into a mechanistic reasoning sample comprising a student-facing prompt and a grounded teacher completion from Gemini 3.5 Flash-Lite. The student prompt provides only: (1) the disease and associated pathway context, (2) the standardized phrase descriptions of the two experimental conditions, and (3) an instruction requesting a multi-step causal cascade (upstream signaling $\rightarrow$ transcription factor activation $\rightarrow$ downstream biological processes) followed by a cohesive final mechanistic summary. The empirical enrichment terms and extracted triplets were strictly withheld from the model input, serving as hidden ground-truth references to condition teacher reasoning trace synthesis and compute rubric-based scores.

3.4 CURATED DISEASE-CENTRIC REASONING DATASET

From RummaGEO samples, we construct a cross-species dataset for grounded mechanistic hypothesis generation across 242 disease-GO contexts (21 of the 263 diseases had no RummaGEO signatures passing the overlap thresholds). The resulting corpus comprises 15,251 differential expression contrast samples, 7,100 human and 8,151 mouse signatures, from 3,441 unique GEO series (Figure 2A). The catalog covers diverse organ systems and domains, including immunological, metabolic, and neurological disorders (Figure 2B). We report the top enriched kinases, TFs, and biological processes (Figure 2C-E) and their rank across all samples in the dataset (Figure 2F).

3.5 LARGE LANGUAGE MODEL BACKBONE

TRACE is initialized from the open-source Qwen3.5 4B checkpoint, chosen for its instruction-following and reasoning capabilities, and trained with supervised fine-tuning (standard next-token cross-entropy) using its native chat template. Each target completion contains (1) a step-by-step reasoning trace within thought markers describing upstream signaling, transcriptional regulators and downstream pathways, followed by (2) a concise final mechanistic paragraph.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Task definition. We evaluate models on the task of study-level mechanistic hypothesis generation from biological condition descriptions. Given an input prompt containing: (1) the disease biological context, and (2) concise phrase descriptions of the baseline (control) and perturbed biological conditions, each model is tasked with generating an explicit step-by-step reasoning trace followed by a

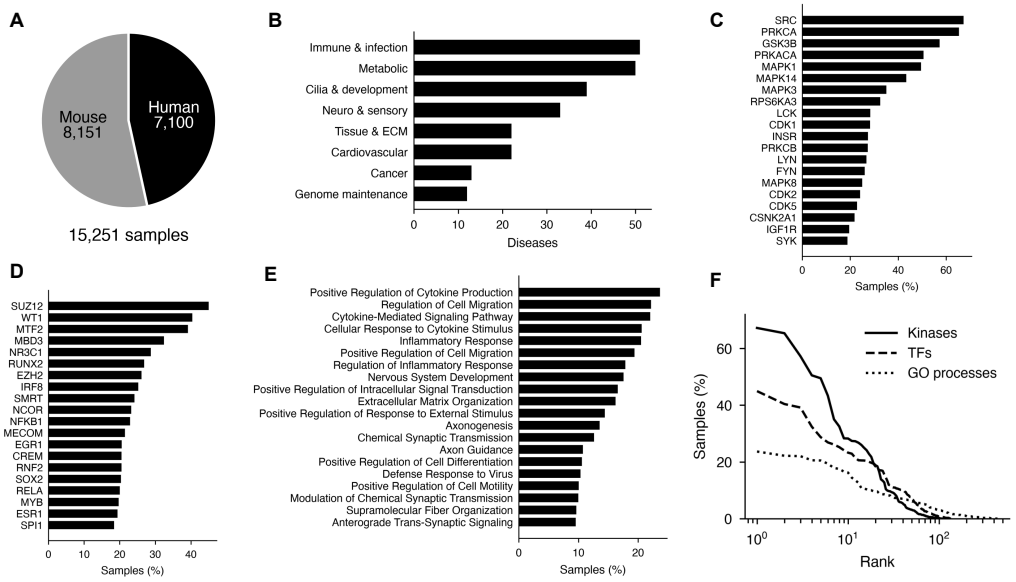


Figure 2: Composition of the disease-centric hypothesis-generation dataset. (A) Sample counts by species for 15,251 samples across 242 diseases and 3,441 GEO series. (B) Diseases assigned to biological systems. (C–E) Most frequent ground-truth enrichment for (C) kinases (KEA 2015), (D) TFs (ChEA 2022), and (E) GO Biological Processes. For each sample, the signature was tested against each library with a one-sided Fisher's exact test (background = 22,000 genes). Terms with $\geq$ 2 overlapping genes and BH-FDR ≤ 0.05 were retained. (F) Rank–frequency curves for all enriched kinases (n = 110), TFs (n = 138), and GO processes (n = 438).

cohesive final mechanistic paragraph. The generation must explain the upstream signaling events, intermediate transcriptional regulators, and downstream cellular processes that separate the two conditions. To test generalizability, evaluations are conducted on two distinct test partitions: a held-out split of 2,451 samples originating from unseen GEO studies within known disease categories, and an out-of-distribution split of 315 samples from five diseases held out entirely from training.

Compared models. We evaluate TRACE against a suite of open-source LLMs and frontier commercial models. We evaluate all models using their standard chat templates under identical system prompts and input contexts. Because modern generalist LLMs are pretrained extensively on biomedical literature and possess broad knowledge of molecular pathways, they are already capable of formulating coherent biological hypotheses separating experimental states. However, unlike TRACE, standard baseline LLMs are not explicitly trained to recover the specific empirical regulatory cascades driving differential expression in bulk transcriptomic studies.

Rubric-based evaluation. We establish a standardized 10-point rubric-based evaluation using Gemini 3.5 Flash-Lite as a strong LLM evaluator that scores each model's generated output against the empirical enrichment ground truth for that RummaGEO sample. The ground truth comprises top-ranking terms from three libraries: KEA 2015 (upstream kinases), ChEA 2022 (transcription factors), and GO Biological Process 2025, alongside co-enriched kinase $\times$ transcription factor $\times$ GO process triplets. The rubric grading prompt and breakdown are provided in Appendix A.1.

4.2 Results

Grounded mechanistic training surpasses generalist LLMs. On held-out GEO studies, TRACE outperforms both open-source models and frontier commercial LLMs on rubric-based mechanistic hypothesis generation (Figure 3, Table 1). TRACE 4B SFT attains an overall score of 5.90 ± 0.08, substantially exceeding the strongest commercial baseline, Gemini 3.5 Flash-Lite (4.28 ± 0.05), and the strongest open-source baseline, Gemma3 4B-it (1.95 ± 0.03). Untuned base models such

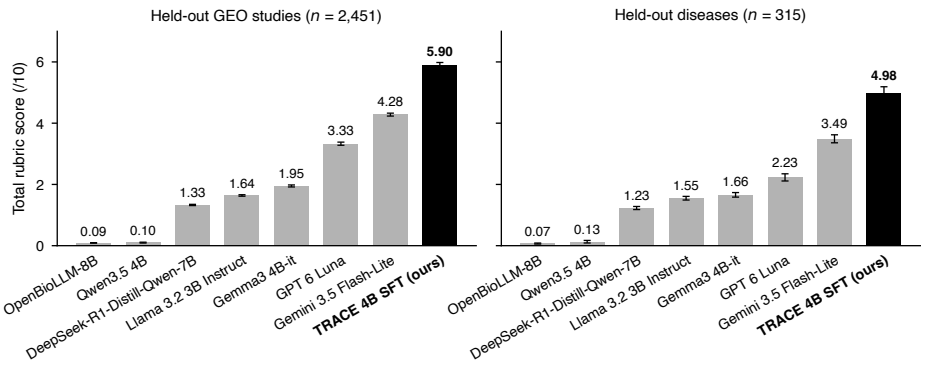


Figure 3: Overall rubric scores on held-out GEO studies ($n = 2{,}451$) and held-out diseases ($n = 315$), with models ordered by performance. Bars show mean total rubric score (out of 10); error bars show standard error. Component scores are reported in Tables 1 and 2.

Table 1: Mechanistic hypothesis generation evaluation on held-out GEO studies ($n = 2{,}451$). Values are mean $\pm$ standard error.

Model	Reasoning	Answer	Enrichment Coverage	Overall
Qwen3.5 4B	0.06 ± 0.01	0.02 ± 0.00	0.02 ± 0.00	0.10 ± 0.01
Gemma3 4B-it	1.27 ± 0.01	0.31 ± 0.01	0.37 ± 0.01	1.95 ± 0.03
Llama 3.2 3B Instruct	1.08 ± 0.01	0.23 ± 0.01	0.33 ± 0.01	1.64 ± 0.02
OpenBioLLM-8B	0.05 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.09 ± 0.01
DeepSeek-R1-Distill-Qwen-7B	0.96 ± 0.01	0.15 ± 0.01	0.23 ± 0.01	1.33 ± 0.02
Gemini 3.5 Flash-Lite	$\mathbf{2.45 \pm 0.02}$	1.00 ± 0.02	0.83 ± 0.02	4.28 ± 0.05
GPT 6 Luna	2.02 ± 0.02	0.71 ± 0.02	0.61 ± 0.01	3.33 ± 0.05
TRACE 4B SFT (ours)	2.41 ± 0.03	$\mathbf{1.72 \pm 0.02}$	$\mathbf{1.77 \pm 0.02}$	$\mathbf{5.90 \pm 0.08}$

as Qwen3.5 4B (0.10 ± 0.01) and domain-adapted biomedical baselines such as OpenBioLLM-8B (0.09 ± 0.01) struggle to recover genuine empirical signatures.

While commercial LLMs achieve comparable reasoning trace scores (2.45 ± 0.02 for Gemini 3.5 Flash-Lite vs. 2.41 ± 0.03 for TRACE) by generating fluent, logically structured reasoning, they fail to recover the ground truth enriched terms. In the final mechanistic answer, TRACE achieves an alignment score of 1.72 ± 0.02 on empirical regulatory triplets compared to 1.00 ± 0.02 for Gemini 3.5 Flash-Lite and 0.31 ± 0.01 for Gemma3 4B-it. Similarly, on independent enrichment coverage, TRACE achieves higher rubric scores for recovery of kinase, TF, and pathway enrichment terms. These results demonstrate that generalist models formulate plausible hypotheses, but consistently miss the empirical signaling cascades and regulatory factors in actual transcriptomic experiments.

TRACE generalizes to unseen disease contexts. To establish that the model learns generalizable mechanistic reasoning rather than memorizing disease-specific pathway associations, we evaluated performance on an out-of-distribution test split comprising held-out diseases never encountered during training (Table 2). TRACE maintains its performance lead, achieving an overall score of 4.98 ± 0.21, compared to 3.49 ± 0.13 for Gemini 3.5 Flash-Lite and 2.23 ± 0.12 for GPT 6 Luna.

Across unseen diseases, TRACE retains strong recovery across both triplet alignment (1.39 ± 0.06) and independent enrichment coverage (1.45 ± 0.06), whereas the commercial baselines drop to 0.75 and 0.65 (Gemini 3.5 Flash-Lite) and 0.38 and 0.33 (GPT 6 Luna). This persistent advantage across unseen disease domains indicates that TRACE learns generalizable rules for inferring upstream kinase signaling, TFs, and downstream functional programs.

Table 2: Mechanistic hypothesis generation evaluation on held-out diseases ($n = 315$). Values are mean $\pm$ standard error.

Model	Reasoning	Answer	Enrichment Coverage	Overall
Qwen3.5 4B	0.08 ± 0.02	0.03 ± 0.01	0.02 ± 0.01	0.13 ± 0.04
Gemma3 4B-it	1.14 ± 0.03	0.20 ± 0.03	0.32 ± 0.03	1.66 ± 0.07
Llama 3.2 3B Instruct	1.03 ± 0.02	0.16 ± 0.02	0.36 ± 0.03	1.55 ± 0.06
OpenBioLLM-8B	0.03 ± 0.01	0.03 ± 0.01	0.01 ± 0.01	0.07 ± 0.02
DeepSeek-R1-Distill-Qwen-7B	0.91 ± 0.02	0.10 ± 0.02	0.21 ± 0.02	1.23 ± 0.05
Gemini 3.5 Flash-Lite	2.10 ± 0.06	0.75 ± 0.04	0.65 ± 0.04	3.49 ± 0.13
GPT 6 Luna	1.52 ± 0.05	0.38 ± 0.04	0.33 ± 0.03	2.23 ± 0.12
TRACE 4B SFT (ours)	$\mathbf{2.14 \pm 0.09}$	$\mathbf{1.39 \pm 0.06}$	$\mathbf{1.45 \pm 0.06}$	$\mathbf{4.98 \pm 0.21}$

We additionally extracted the kinases, TFs, and GO processes named in each final mechanistic answer of our model and compared them with the sample's own enrichment. TRACE does not default to hub regulators, as every kinase and TF enriched in at least 10% of samples is named less often than it is enriched while naming frequency still tracks enrichment frequency across entities (Pearson r = 0.78 for kinases, 0.57 for TFs; Figure 4A). Named pathways are also highly specific. The model names a GO process mainly when the sample's signature is enriched for it: for example, DNA damage response is named in 72% of answers for samples enriched for it, but in fewer than 1% of answers for all other samples (Figure 4B). Across evaluation splits, 67–78% of named kinases, 70–73% of named TFs and 59–75% of named GO processes appear in the sample's own enrichment (Figure 4C). TRACE performs well on held-out GEO studies, likely due to seeing similar enrichment profiles on the same disease within the training set. For held-out diseases, TRACE still performs well, with similar TF grounding in its answers (71% vs 70% in training).

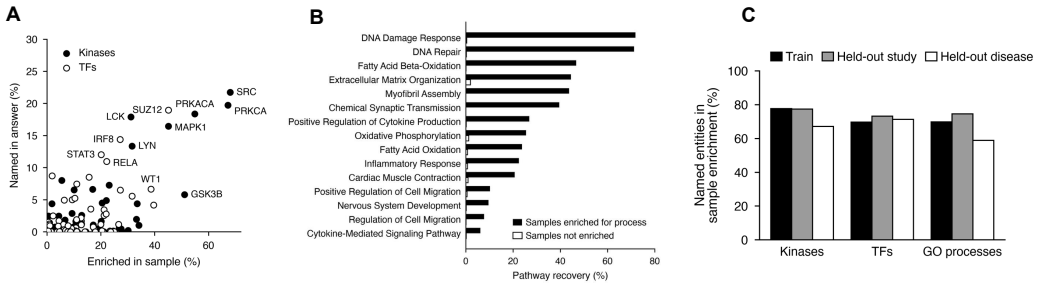


Figure 4: Model predictions are grounded in each sample's enrichment signal. Analyses use the 2,449 samples for which TRACE produced a completed final answer (1,953 of 2,451 held-out study; 250 of 315 held-out disease; 246 train). (A) Fraction of answers that name each kinase (filled) or transcription factor (open) versus fraction of the same samples in which that entity is top-enriched (Pearson r = 0.78 for kinases; 0.57 for TFs). (B) Pathway recovery for the 15 GO Biological Processes named most often by the model. Bars show the fraction of answers that name it among samples whose enrichment includes that process (black) or not (white). (C) Fraction of kinases, TFs, and GO processes named in an answer that also appear in that sample's top-enriched sets, stratified by evaluation split: train (n = 246), held-out study (n = 1,953), and held-out disease (n = 250).

4.3 Reasoning Case Study

Beyond individual samples, the model's proposed cascades are disease-specific. Grouping held-out answers by the biological system of the disease context, the kinases, TFs and GO processes the model names shift in line with each system's biology: for example, ATR, ATM and CHEK1 with E2F1 for genome maintenance disorders, PRKAA2 with PPARG and PPARA for metabolic diseases, and PRKACA with TBX20 and GATA4 for cardiovascular diseases (Figure S1A–C). These system-specific patterns closely track the enrichment observed in the same samples' signatures

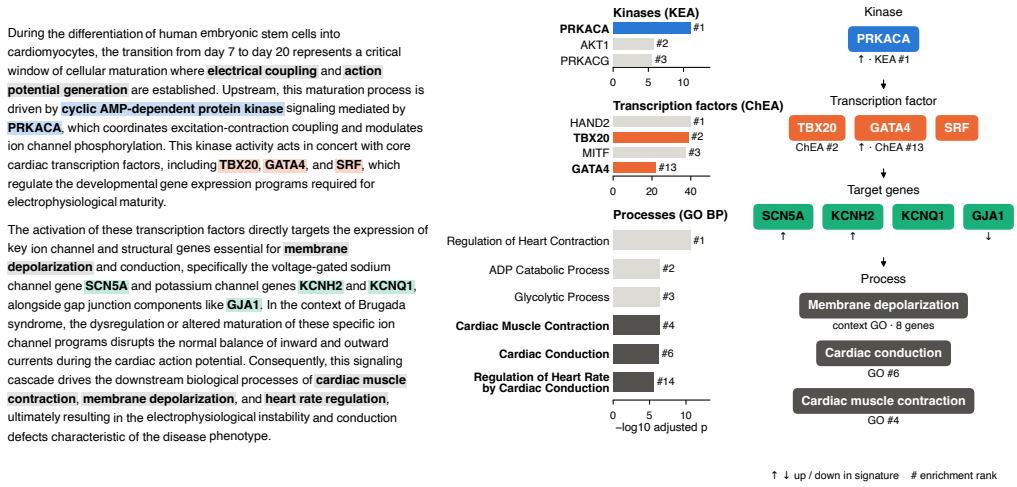


Figure 5: Case study: Brugada syndrome (held-out disease). (left) Model reasoning for human ESC-derived cardiomyocytes at day 20 versus day 7 of differentiation (GSE114373), with kinases (blue), transcription factors (orange), genes (green) and biological processes (gray) highlighted. (middle) Top KEA, ChEA and GO Biological Process enrichment terms for the same sample. (right) Proposed kinase $\rightarrow$ TF $\rightarrow$ target gene $\rightarrow$ process cascade, with each node annotated against the sample: $\uparrow$/ $\downarrow$, up- or down-regulated in the signature; #, rank among the sample's enrichment terms.

(Pearson $r = 0.73$; Figure S1D) and extend to diseases held out entirely from training (Figure S1E). We therefore examined one held-out disease in detail.

To assess reasoning about a disease absent from training, we analyzed a trace for Brugada syndrome (Figure 5), a rare inherited disorder of the heart's electrical activity that can trigger ventricular fibrillation and sudden cardiac arrest, affecting approximately 1 in 2,000 people (Vutthikraivit et al., 2018). For a cardiomyocyte maturation signature (hESC-derived cardiomyocytes, day 20 vs day 7; GSE114373, Pezhouman et al., 2022) the model proposed a PKA (PRKACA) $\rightarrow$ TBX20/GATA4 $\rightarrow$ SCN5A cascade that recovers both the sample's enrichment signal and the known genetics of the disease. PRKACA and TBX20 were the top-ranked kinase and second-ranked transcription factor in the sample's own KEA and ChEA enrichment, and SCN5A, the only gene with definitive evidence for Brugada syndrome (Chen et al., 1998; Hosseini et al., 2018), was up-regulated. The proposed regulators, TBX20 and GATA4, are both Brugada syndrome risk loci in the largest genome-wide association study to date (Barc et al., 2022), GATA4 directly activates the SCN5A promoter in the human heart (Tarradas et al., 2017), and TBX20 maintains ion-channel gene expression in adult cardiomyocytes (Shen et al., 2011). PKA in turn increases sodium current by trafficking Nav1.5 channels to the membrane (Hallaq et al., 2006), and drugs that raise cAMP/PKA activity, such as isoproterenol and cilostazol, are used clinically to suppress electrical storms in Brugada syndrome (Ohgo et al., 2007). The model's ability to recover enriched entities and draw on established literature demonstrates its capacity to propose coherent biological mechanisms for hypothesis generation.

5 Conclusion

We present TRACE, a framework for mechanistic hypothesis generation on transcriptomic studies. Our approach grounds biological reasoning directly in empirical differential expression signatures and enrichment information. We propose a scalable enrichment pipeline that mines enrichment-supported kinases, TFs, and biological processes, and create a curated dataset of 15,251 human and mouse reasoning samples. By training on these structured reasoning traces, TRACE (Qwen3.5 4B) learns to hypothesize about the underlying molecular cascades separating biological conditions. TRACE demonstrates superior recovery of genuine mechanistic terms relative to open-source and commercial frontier LLMs. We release our curated dataset, model checkpoint, and enrichment-to-trace pipeline to support scalable, interpretable hypothesis generation in computational biology.

6 AI USE AND ETHICS

AI USE STATEMENT

In this work, we used generative AI tools to help with editing software code for implementation of our framework and benchmarking. AI was also used to assist with drafting and editing parts of the paper. We have reviewed all AI-assisted work and AI-generated software code and writing, and multiple authors on the work verified the correctness and validity of the AI-assisted code and writing. AI was also used as a strong teacher model for data generation, specifically reasoning trace generation using ground truth enrichment information calculated from public data samples. We take responsibility for the final content of this work, including produced data, code, artifacts, and text.

We have not used generative AI tools for conceptual design of any models or frameworks, and the rest of the required disclosure tasks are not applicable to this work.

REPRODUCIBILITY STATEMENT

Our software code is implemented in Python and organized on GitHub. Upon acceptance of the paper, we intend on doing a full release of dataset samples, code, and trained model checkpoints for the community. We intend on supporting full reproducibility of our main experiments and work, including providing adequate information to the community on implementing and running the experiments.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with state. *BioRxiv*, pp. 2025–06, 2025.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Seungbyn Baek, Kyungwoo Song, and Insuk Lee. Single-cell foundation models: bringing artificial intelligence into cell biology. *Experimental & Molecular Medicine*, 57(10):2169–2181, 2025.

Julien Barc, Rafik Tadros, Charlotte Glinge, David Y Chiang, Mariam Jouni, Floriane Simonet, Sean J Jurgens, Manon Baudic, Michele Nicastro, Franck Potet, et al. Genome-wide association analyses identify new brugada syndrome risk loci and highlight a new mechanism of sodium channel regulation in disease susceptibility. *Nature genetics*, 54(3):232–239, 2022.

Antoine Bosselut, Zeming Chen, Angelika Romanou, Antoine Bonnet, Alejandro Hernández-Cano, Badr Alkhamissi, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, et al. Meditron: Open medical foundation models adapted for clinical practice. 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Edward Y Chen, Huilei Xu, Simon Gordonov, Maribel P Lim, Matthew H Perkins, and Avi Ma'ayan. Expression2kinases: mrna profiling linked to multiple upstream regulatory layers. *Bioinformatics*, 28(1):105–111, 2012.

Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

Mingmei Chen, Hui Liang, Min Wu, Haoyang Ge, Yan Ma, Yan Shen, Shunyuan Lu, Chunling Shen, Hongxin Zhang, Zhugang Wang, et al. Fgf9 regulates bone marrow mesenchymal stem cell fate and bone-fat balance in osteoporosis by pi3k/akt/hippo and mek/erk signaling. *International Journal of Biological Sciences*, 20(9):3461, 2024.

Qiuyun Chen, Glenn E Kirsch, Danmei Zhang, Ramon Brugada, Josep Brugada, Pedro Brugada, Domenico Potenza, Angel Moya, Martin Borggrefe, Günter Breithardt, et al. Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Nature*, 392(6673):293–296, 1998.

Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pp. 2023–10, 2024.

Daniel J B Clarke, Maxim V Kuleshov, Brian M Schilder, Denis Torre, Mary E Duffy, Alexandra B Keenan, Alexander Lachmann, Axel S Feldmann, Gregory W Gundersen, Moshe C Silverstein, et al. expression2kinases (x2k) web: linking expression signatures to upstream cell signaling networks. *Nucleic acids research*, 46(W1):W171–W179, 2018.

Daniel JB Clarke, Giacomo B Marino, Eden Z Deng, Zhuorui Xie, John Erol Evangelista, and Avi Ma'ayan. Rummagene: massive mining of gene sets from supporting materials of biomedical research publications. *Communications Biology*, 7(1):482, 2024.

Daniel JB Clarke, Giacomo B Marino, and Avi Ma'ayan. Gsfm: A gene set foundation model pre-trained on a massive collection of diverse gene sets. *Patterns*, 2026.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

Xiaofeng Dai and Li Shen. Advances and trends in omics technology development. *Frontiers in medicine*, 9:911861, 2022.

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J Maddison, et al. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model. *arXiv preprint arXiv:2505.23579*, 2025.

Susan A Habinowski and Lee A Witters. The effects of aicar on adipocyte differentiation of 3t3-l1 cells. *Biochemical and biophysical research communications*, 286(5):852–856, 2001.

Haifa Hallaq, Zhenjiang Yang, Prakash C Viswanathan, Koji Fukuda, Wangzhen Shen, Dao W Wang, K Sam Wells, Jingsong Zhou, Jianxun Yi, and Katherine T Murray. Quantitation of protein kinase a-mediated trafficking of cardiac sodium channels in living cells. *Cardiovascular research*, 72(2):250–261, 2006.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024.

S Mohsen Hosseini, Raymond Kim, Sharmila Udupa, Gregory Costain, Rebekah Jobling, Eriskay Liston, Seema M Jamal, Marta Szybowska, Chantal F Morel, Sarah Bowdin, et al. Reappraisal of reported genes for sudden arrhythmic death: evidence-based evaluation of gene validity for brugada syndrome. *Circulation*, 138(12):1195–1205, 2018.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Masanori Iwaki, Morihiro Matsuda, Norikazu Maeda, Tohru Funahashi, Yuji Matsuzawa, Makoto Makishima, and Iichiro Shimomura. Induction of adiponectin, a fat-derived antidiabetic and antiatherogenic factor, by nuclear receptors. *Diabetes*, 52(7):1655–1663, 2003.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Alexandra B Keenan, Denis Torre, Alexander Lachmann, Ariel K Leong, Megan L Wojciechowicz, Vivian Utti, Kathleen M Jagodnik, Eryk Kropiwnicki, Zichen Wang, and Avi Ma'ayan. Chea3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic acids research*, 47(W1):W212–W224, 2019.

Ihor Kendiukhov. Systematic evaluation of single-cell foundation model interpretability reveals attention captures co-expression rather than unique regulatory signal. *arXiv preprint arXiv:2602.17532*, 2026.

Maxim V Kuleshov, Zhuorui Xie, Alexandra BK London, Janice Yang, John Erol Evangelista, Alexander Lachmann, Ingrid Shu, Denis Torre, and Avi Ma'ayan. Kea3: improved kinase enrichment analysis via data integration. *Nucleic acids research*, 49(W1):W304–W316, 2021.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

J Lamb, ED Crawford, D Peck, JW Modell, IC Blat, MJ Wrobel, J Lerner, JP Brunet, A Subramanian, KN Ross, M Reich, H Hieronymus, G Wei, SA Armstrong, SJ Haggarty, PA Clemons, R Wei, SA Carr, ES Lander, and TR Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

Jürgen M Lehmann, Linda B Moore, Tracey A Smith-Oliver, William O Wilkison, Timothy M Willson, and Steven A Kliewer. An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ (pparγ). *Journal of Biological Chemistry*, 270(22): 12953–12956, 1995.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: Teaching large language models the language of biology. *bioRxiv*, pp. 2023–09, 2023.

Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

Giacomo B Marino, Daniel JB Clarke, Alexander Lachmann, Eden Z Deng, and Avi Ma'ayan. Rummageo: automatic mining of human and mouse gene sets from geo. *Patterns*, 5(10), 2024.

Takeshi Ohgo, Hideo Okamura, Takashi Noda, Kazuhiro Satomi, Kazuhiro Suyama, Takashi Kurita, Naohiko Aihara, Shiro Kamakura, Tohru Ohe, and Wataru Shimizu. Acute and chronic management in patients with brugada syndrome associated with electrical storm of ventricular fibrillation. *Heart Rhythm*, 4(6):695–700, 2007.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Arash Pezhouman, James L Engel, Ngoc B Nguyen, Rhys JP Skelton, William Blake Gilmore, Rong Qiao, Debashis Sahoo, Peng Zhao, David A Elliott, and Reza Ardehali. Isolation and characterization of human embryonic stem cell-derived heart field-specific cardiomyocytes unravels new insights into their transcriptional and electrophysiological profiles. *Cardiovascular Research*, 118 (3):828–843, 2022.

Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X Binder, and Lars Juhl Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, pp. gkae1142, 2024.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

Jane EB Reusch, Lilliester A Colton, and Dwight J Klemm. Creb activation induces adipogenesis in 3t3-l1 cells. *Molecular and cellular biology*, 20(3):1008–1020, 2000.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Curtis Jamison Perry, Ivan Vrkic, Nicole Mayerli Constante, Zirui Fu, Sizhuang He, et al. Scaling large language models for next-generation single-cell analysis. *BioRxiv*, pp. 2025–04, 2026.

Evan D Rosen, Pasha Sarraf, Amy E Troy, Gary Bradwin, Kathryn Moore, David S Milstone, Bruce M Spiegelman, and Richard M Mortensen. Pparγ is required for the differentiation of adipose tissue in vivo and in vitro. *Molecular cell*, 4(4):611–617, 1999.

Yanay Rosen, Yusuf Roohani, Ayush Agarwal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *BioRxiv*, pp. 2023–11, 2023.

Tao Shen, Ivy Aneas, Noboru Sakabe, Ralf J Dirschinger, Gang Wang, Scott Smemo, John M Westlund, Hongqiang Cheng, Nancy Dalton, Yusu Gu, et al. Tbx20 regulates a genetic program essential to adult mouse cardiomyocyte function. *The Journal of clinical investigation*, 121(12): 4640–4654, 2011.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences*, 102(43):15545–15550, 2005.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Hong Sun, Pouya Saeedi, Suvi Karuranga, Moritz Pinkepank, Katherine Ogurtsova, Bruce B Duncan, Caroline Stein, Abdul Basit, Juliana CN Chan, Jean Claude Mbanya, et al. Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes research and clinical practice*, 183:109119, 2022.

Anna Tarradas, Carlos Mackintosh, Oriol Llorà-Batlle, Alexandra Pérez-Serra, Montserrat Batlle, Félix Pérez-Villa, Thomas Zimmer, Ivan Garcia-Bassets, Ramon Brugada, Pedro Beltran-Alvarez, et al. Transcriptional regulation of the sodium channel gene (scn5a) by gata4 in human heart. *Journal of molecular and cellular cardiology*, 102:74–82, 2017.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Koji Tomiyama, Hiroko Nakata, Hidenori Sasa, Shinya Arimura, Eisuke Nishio, and Yasuhiro Watanabe. Wortmannin, a specific phosphatidylinositol 3-kinase inhibitor, inhibits adipocytic differentiation of 3t3-l1 cells. *Biochemical and biophysical research communications*, 212(1): 263–269, 1995.

Peter Tontonoz, Erding Hu, and Bruce M Spiegelman. Stimulation of adipogenesis in fibroblasts by pparγ2, a lipid-activated transcription factor. *Cell*, 79(7):1147–1156, 1994.

A Tsherniak, F Vazquez, PG Montgomery, BA Weir, G Kryukov, GS Cowley, S Gill, WF Harrington, S Pantel, JM Krill-Burger, RM Meyers, L Ali, A Goodale, Y Lee, G Jiang, J Hsiao, WFJ Gerath, S Howell, E Merkel, M Ghandi, LA Garraway, DE Root, TR Golub, JS Boehm, and WC Hahn. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017.

Wasawat Vutthikraivit, Pattara Rattanawong, Prapaipan Putthapiban, Weera Sukhumthammarat, Prin Vathesatogkit, Tachapong Ngarmukos, and Ammarin Thakkinstian. Worldwide prevalence of brugada syndrome: a systematic review and meta-analysis. *Acta Cardiologica Sinica*, 34(3): 267, 2018.

Leanne Wilson-Fritch, Alison Burkart, Gregory Bell, Karen Mendelson, John Leszyk, Sarah Nicoloro, Michael Czech, and Silvia Corvera. Mitochondrial biogenesis and remodeling during adipogenesis and in response to the insulin sensitizer rosiglitazone. *Molecular and cellular biology*, 23 (3):1085–1094, 2003.

Xianmin Xia and Ginette Serrero. Inhibition of adipose differentiation by phosphatidylinositol 3-kinase inhibitors. *Journal of cellular physiology*, 178(1):9–16, 1999.

Toshimasa Yamauchi, J Kamon, H Waki, Y Terauchi, N Kubota, K Hara, Y Mori, T Ide, K Murakami, N Tsuboyama-Kasaoka, et al. The fat-derived hormone adiponectin reverses insulin resistance associated with both lipoatrophy and obesity. *Nature medicine*, 7(8):941–946, 2001.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

A Appendix

A.1 Rubric-Grading Evaluation

Automated lexical overlap metrics (such as BLEU, ROUGE, or BERTScore) cannot distinguish between genuine molecular mechanisms and ungrounded correlative text. A model that recites common pathway names without causal linkage can achieve favorable token-overlap scores while offering negligible biological insight. To address this, we implement a standardized rubric-based evaluation system administered by an expert LLM evaluator (Gemini 3.5 Flash-Lite; Figure S3-S4).

The grading prompt provides the evaluator with the experimental condition descriptions and four hidden ground-truth enrichment blocks derived from the contrast's empirical differential expression signature: (1) co-enriched kinase $\times$ transcription factor $\times$ GO biological process triplets, (2) independent top-ranking kinases from KEA 2015, (3) independent top-ranking transcription factors from ChEA 2022, and (4) independent top-ranking pathways from GO Biological Process 2025. The evaluator inspects both the student model's intermediate reasoning trace and its final mechanistic summary, enforcing strict penalties against associative list-dumping and ungrounded biological claims.

A.1.1 Scoring Dimensions and Criteria

The total quantitative grade spans 10 points across three primary mechanistic dimensions, supplemented by an auxiliary structural assessment:

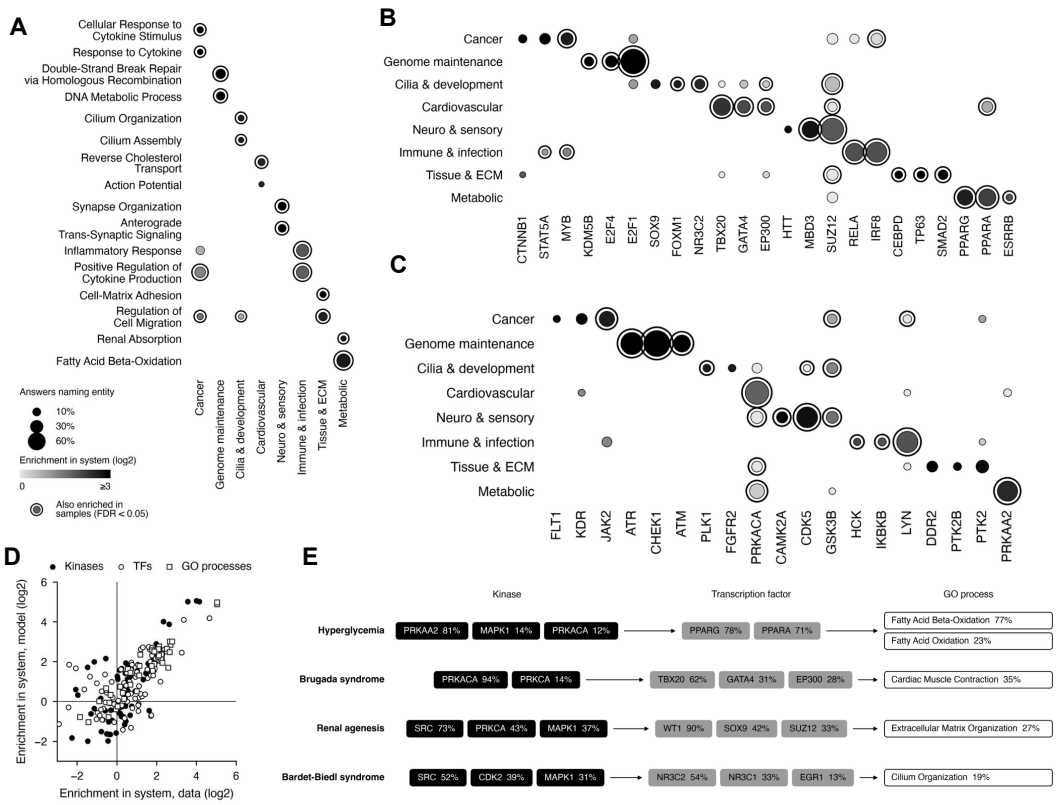


Figure S1: Disease-specific signaling cascades proposed by the model mirror each disease's transcriptional signal. 2,203 held-out answers with a completed final mechanistic answer (1,953 held-out study, 250 held-out disease), grouped by the biological system of the disease context. (A–C) Signature GO processes (A), transcription factors (B), and kinases (C) named in answers for each disease system. Entities are retained if they are at least 2-fold over-represented in that system's answers relative to all answers (one-sided Fisher's exact test, BH-FDR < 0.05). Dot size $=$ share of the system's answers naming the entity; gray level $= \log_2$ fold enrichment in that system; ring $=$ entity also significantly enriched in the same samples' signatures (Fisher's exact test, BH-FDR < 0.05). (D) System-specific enrichment of each entity in model answers versus sample enrichment ($\log_2$ fold enrichment) for all system–entity pairs supported by ≥ 5 samples in both (Pearson $r = 0.73$, $n = 204$). (E) Most frequently named kinases, TFs, and GO processes for held-out diseases with ≥ 20 answers never seen in training (Hyperglycemia $n = 69$, Brugada syndrome $n = 65$, Renal agenesis $n = 60$, Bardet–Biedl syndrome $n = 54$). Percentages are the share of that disease's answers naming each entity.

1. Quality of the Reasoning Trace (4 points). This component evaluates the intermediate step-by-step reasoning trace produced within the thought tags. Full credit (4 points) requires an unbroken, directional molecular cascade tracing the biological transition from experimental perturbation to phenotypic consequence (stimulus $\rightarrow$ kinase activation $\rightarrow$ transcription factor regulation $\rightarrow$ downstream cellular process). Each link must specify the biological mechanism of action (e.g., phosphorylation, nuclear translocation, or promoter binding). Partial credit (2–3 points) is awarded if reference entities or valid pathway steps are mentioned but connected via correlative assertions or unexplained logical leaps. Hallucinated biology, generic disease overviews, or disconnected lists receive 0–1 points.

2. Triplet Alignment in Final Mechanistic Answer (3 points). This dimension measures whether the model's final synthesis centers on at least one empirical kinase $\times$ transcription factor $\times$ GO pathway triplet extracted from the contrast signature. High credit (3 points) requires integrating a supported triplet into a biologically coherent mechanism where the signaling components interact

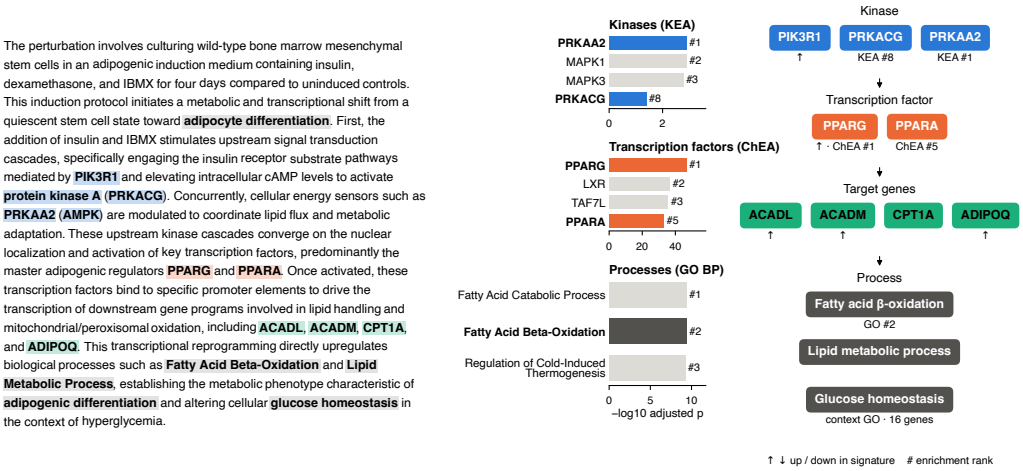


Figure S2: Case study: hyperglycemia (held-out disease). (left) Model reasoning for mouse bone marrow mesenchymal stem cells cultured for four days with versus without adipogenic induction (insulin, dexamethasone and IBMX; GSE252394), with kinases (blue), transcription factors (orange), genes (green) and biological processes (gray) highlighted. (middle) Top KEA, ChEA and GO Biological Process enrichment terms for the same sample's up-regulated signature. (right) Proposed kinase → transcription factor → target gene → process cascade.

in a verified functional manner. Partial credit (1–2 points) is assigned if individual triplet members are named but not joined into a functional unit, or if the transition between the transcription factor and downstream pathway is absent. Low or zero credit is assigned if the model focuses on unsupported pathways or names isolated genes without functional context.

3. Coverage of Independent Top Lists (3 points). This score quantifies the breadth and accuracy with which the model recovers independent top-ranking kinases, transcription factors, and GO biological processes. Points are assigned based on whether top-list entities (or direct gene-family homologs) appear in meaningful biological contexts. If entities are forced into the text as an uncontextualized list, or linked only through generic assertions (e.g., stating that a factor "regulates survival"), credit is withheld.

4. Narrative Coherence and Structure (Auxiliary 1–5 Scale). To differentiate between articulate mechanistic prose and keyword enumeration, the evaluator assigns an auxiliary coherence score ranging from 1 (unstructured enumeration of entities without narrative flow) to 5 (a cohesive, readable biological hypothesis). This metric is analyzed independently to verify that models achieve high mechanistic scores through structured biological narratives rather than exhaustive keyword lists.

A.1.2 Evaluation Execution

The rubric prompt enforces structured JSON output. For each evaluated sample, the judge model returns numerical scores for all categories alongside explicit textual rationales explaining why points were awarded or deducted. This guarantees consistent scoring standards across diverse model architectures, preserves full grading transparency, and enables targeted auditing of model errors across specific biological pathways.

A.2 Additional Reasoning Case Study

Figure S2 contains an additional case study regarding hyperglycemia, a major phenotype of diabetes mellitus, affecting over 500 million adults globally (Sun et al., 2022). It arises in large part from impaired insulin sensitivity in which adipose tissue plays a central role. Hyperglycemia was held out from training, and for bone marrow mesenchymal stem cells induced toward adipogenesis with insulin, dexamethasone and IBMX (GSE252394, Chen et al., 2024) the model proposed an

Standardized Grading Prompt for Rubric-Based Mechanistic Hypothesis Grading

```
You are an expert biologist grading a mechanistic hypothesis written
from bulk RNA-seq condition metadata.

Score the TEXT TO GRADE against the REFERENCE ENRICHMENT LISTS. Do
not reward superficial disease associations or generic buzzwords
unless they are mechanistically connected to the listed entities.  Do
not penalize for omitting Enrichr, KEA, ChEA, or RummaGEO by name.

Rules for partial credit and logical coherence:
- Functional terms:  Unlisted pathology terms, intermediate signaling
molecules, disease phenotypes, close synonyms, gene-family members,
or closely related pathway names may earn partial credit ONLY if they
are mechanistically bridging listed entities to the study conditions.
Mentioning these terms in isolation without a clear biological link
receives zero credit.
- Logical jumps:  Penalize loose associative reasoning or
"name-dropping." If listed kinases, transcription factors, and
pathways are strung together without explaining the causal
biochemical links between them (e.g., asserting A activates B
without establishing how, or jumping from a receptor to an unrelated
phenotype), cap the score at partial credit.  A high score requires
explicit directionality and plausible biological transitions between
steps.

Total points:  10.

(i) Quality of the reasoning trace --- 4 points
Evaluate the step-by-step causal flow from the perturbed condition to
the downstream molecular consequences.
- Full credit:  A continuous, logical cascade (stimulus/perturbation
-> kinase signaling -> transcriptional control -> downstream
cellular/phenotypic process).  Each transition must state the
directional mechanism (e.g., phosphorylation, nuclear translocation,
transcriptional repression).  Unlisted biological context or
pathology terms receive partial credit if they directly bridge these
steps.
- Partial credit:  Lists several reference entities or valid
mechanistic terms, but connects them via hand-waving, correlative
statements, or unexplained jumps in logic.
- Minimal/No credit (0-1 points):  Generic disease summaries,
hallucinated biology, or disconnected keyword lists with no
directional reasoning.

(Continued in next figure...)
```

Figure S3: The structured prompt template employed for automated rubric-based grading. The judge model receives the experimental condition definitions alongside the hidden reference enrichment lists (triplets, top kinases, top TFs, and top GO terms), scoring the student output across causal step-by-step reasoning, triplet recovery, and top-list coverage while penalizing associative keyword lists.

Standardized Grading Prompt for Rubric-Based Mechanistic Hypothesis Grading

```
(Continued grading prompt)

(ii) Quality of the final mechanistic answer versus supported
triplets --- 3 points
Evaluate how tightly the final mechanism centers on at least one
supported kinase / transcription-factor / GO biological-process
triplet (including relevant overlapping genes).
- High credit: Integrates a triplet (or closely related homologs)
into a biologically verified, coherent mechanism where the kinase,
TF, and GO process logically interact.
- Partial credit: Correctly identifies components of a triplet or
related pathology, but fails to integrate them into a functional
unit, leaves out the causal link between the TF and the downstream
pathway, or relies on an unsupported functional leap.
- Low/No credit: Names isolated genes with no functional cascade, or
proposes a mechanism unrelated to the supported triplets.

(iii) Quality of the final mechanistic answer versus independent top
lists --- 3 points
Evaluate the breadth and specificity of coverage across the
independent top lists (kinases, TFs, GO processes).
- Award points based on how many top-list entities (or direct
gene-family equivalents) appear within a meaningful mechanistic
context.
- Deduct or withhold credit if listed entities are forced into the
text purely as a list or via broad, non-specific claims (e.g., simply
stating "these genes regulate cell survival" without mechanistic
context).

STUDY SETUP (conditions being hypothesized about; this is not a
reference answer)
Biological context: {biological context}
Control condition: {control condition}
Perturbed condition: {perturbed condition}

----- REFERENCE ENRICHMENT LISTS (grade against these; do not treat
them as the student's text) -----
Supported kinase / transcription-factor / GO biological-process
triplets: {triplets block}
Independent top kinases: {top kinases block}
Independent top transcription factors: {top tfs block}
Independent top GO biological processes: {top go processes block}
----- END REFERENCE ENRICHMENT LISTS -----

TEXT TO GRADE
Reasoning trace: reasoning trace
Final mechanistic answer: final answer

Return ONLY valid JSON with this exact schema (no markdown fences, no
extra keys):
{"reasoning trace score": 0-4, "reasoning trace rationale": "...",
"triplet alignment score": 0-3, "triplet alignment rationale":
"...", "top list coverage score": 0-3, "top list coverage
rationale": "...", "total score": 0-10, "grading summary": "...",
"coherence score": 1-5, "coherence rationale": "..."}
```

Figure S4: The structured prompt template employed for automated rubric-based grading. The judge model receives the experimental condition definitions alongside the hidden reference enrichment lists (triplets, top kinases, top TFs, and top GO terms), scoring the student output across causal step-by-step reasoning, triplet recovery, and top-list coverage while penalizing associative keyword lists.

insulin/cAMP → PI3K/PKA/AMPK → PPARG/PPARA → fatty acid oxidation cascade. The sample's enrichment ranks PRKAA2 (AMPK) as the top kinase, PPARG as the top transcription factor and fatty acid β-oxidation as the second GO process, and PPARG, PIK3R1, ACADL, ACADM and ADIPOQ are all up-regulated. This mirrors established adipogenic signaling: the induction cocktail acts through cAMP-driven CREB activation (Reusch et al., 2000) and insulin-PI3K signaling, whose inhibition blocks adipocyte differentiation (Tomiyama et al., 1995; Xia & Serrero, 1999), and converges on PPARG, the master regulator that is both sufficient and required for adipogenesis (Tontonoz et al., 1994; Rosen et al., 1999). PPARG directly induces adiponectin through a PPAR response element in its promoter (Iwaki et al., 2003) and drives mitochondrial biogenesis and fatty acid oxidation in adipocytes (Wilson-Fritch et al., 2003), linking the cascade to glucose control: adiponectin is an insulin-sensitizing hormone (Yamauchi et al., 2001) and PPARG is the molecular target of antidiabetic thiazolidinediones (Lehmann et al., 1995). One node departs from the literature: pharmacological AMPK activation inhibits rather than promotes adipocyte differentiation (Habinowski & Witters, 2001), so AMPK's enrichment here more likely reflects the metabolic shift toward lipid oxidation than a driver of differentiation. Together with the Brugada syndrome case, this example shows that the model generalizes to unseen diseases across distinct biological systems, proposing cascades that are grounded in each sample's enrichment signal and largely consistent with established mechanisms.