OMICSLM: A MULTIMODAL LARGE LANGUAGE MODEL FOR MULTI-SAMPLE OMICS REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can reason over biological knowledge, but they are not naturally equipped to work with quantitative omics measurements. Existing approaches either convert expression profiles into long textual representations, often losing quantitative information, or connect omics encoders to language models for relatively narrow tasks. We introduce OmicsLM, a 9B multimodal language model that represents each bulk or single-cell transcriptome as a compact continuous embedding that can be interleaved with natural-language instructions and with other samples in the same context. An important contribution is the construction of a training corpus that connects experimental measurements with biological knowledge and task supervision. We combine templates linking transcriptomic profiles to annotations and experimental outcomes, reasoning examples connecting evidence across sources, and LLM-generated supervision with task-specific quality checks. The instruction-tuning corpus spans 157 biological and general-purpose task groups. Training includes continued pretraining on 20 billion tokens and supervised fine-tuning on 5.21 million conversations. We further introduce GEO-OmicsQA, a publication-disjoint benchmark of 3,000 questions requiring models to interpret and compare expression profiles from real GEO studies. OmicsLM achieves the best results among the evaluated models on three of four held-out annotation tasks and on all six GEO-OmicsQA variants, and the best average performance among compared methods on PerturbQA perturbation-response prediction. Ablations show a clear advantage for continuous omics representations over a text-only gene-list representation. Together, these results show that quantitative transcriptomic measurements and biological language can be integrated within a general conversational model, enabling both predictive tasks and language-guided interpretation of experimental samples.

1 INTRODUCTION

The advent of high-throughput sequencing technologies has revolutionized our understanding of disease progression, tissue development, and cellular heterogeneity by providing quantitative measurements of biological molecules at both bulk and single-cell resolutions (Regev et al., 2017; Eberwine et al., 2014). Transcriptomic profiles are among the most informative of these measurements: a single RNA-seq assay quantifies the expression of thousands of genes, reflecting cell identity and state as well as responses to disease, genetic perturbation, and drug treatment. Public atlases and repositories now provide such profiles across tissues, cell types, diseases, and experimental conditions (Regev et al., 2017; GTEx Consortium, 2020; The Cancer Genome Atlas Research Network et al., 2013; Clough et al., 2024), making expression a natural interface between experimental measurements and biological knowledge. Concurrently, large language models (LLMs) have demonstrated unprecedented capabilities in reasoning, knowledge retrieval, and instruction following (Brown et al., 2020; OpenAI, 2023). However, a significant gap remains between the quantitative, high-dimensional nature of omics data and the text-centric input space of modern LLMs.

Existing biological foundation models tackle parts of this gap: specialized single-cell models learn representations for downstream prediction (Cui et al., 2024; Theodoris et al., 2023; Hao et al., 2024; Rosen et al., 2026), while text-based biomedical LLMs serialize expression as gene-name sequences (Levine et al., 2024; Rizvi et al., 2025; Choi et al., 2026), requiring long inputs and, for rank-

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.

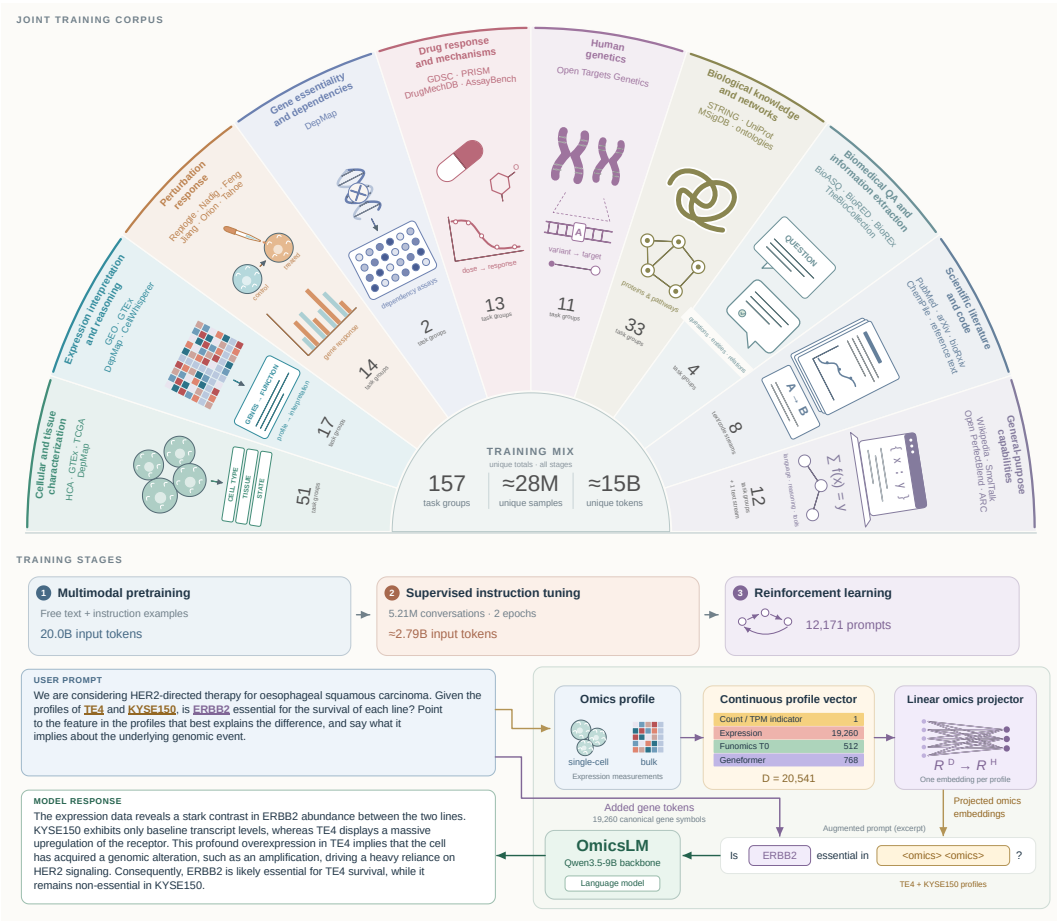


Figure 1: **OmicsLM overview.** *Top:* the joint training corpus spans ten capability groups. Sector counts regroup the 157 instruction-tuning task groups in Table 17; scientific text/code streams and the Wikipedia text stream are counted separately. The centre gives their total and the unique samples and tokens seen across all training stages: pretraining counts only records actually sampled (large text sources such as arXiv were only partly used), and examples repeated across epochs or stages are counted once. *Middle:* the three training stages comprise multimodal pretraining, supervised instruction tuning, and reinforcement learning. *Bottom:* single-cell or bulk expression profiles form continuous vectors projected into the language-model input at `<omics>` positions, alongside text and added gene tokens. The example conversation at lower left connects to the architecture at lower right; highlighted cell names are augmented with profile embeddings. Sector areas are schematic.

only representations, discarding expression magnitudes. Multimodal models connect quantitative transcriptomic representations to language (Schaefer et al., 2025; Fang et al., 2025).

Many biological questions are comparative: control versus perturbation, patient versus model, sensitive versus resistant, or one cellular state versus another. Long textual representations consume context as more profiles are added, and interfaces designed primarily around individual profiles may require adaptation for joint comparisons. These constraints can make multi-sample analysis cumbersome. We ask whether quantitative transcriptomes can instead be represented compactly inside a language model, allowing several experimental samples to be analyzed jointly with biological context and natural-language instructions. Learning to make these comparisons also requires suitable supervision connecting experimental measurements to biological knowledge, task instructions, and examples of interpreting evidence across samples.

In this work, we introduce **OmicsLM**, a multimodal language model trained for diverse omics tasks (Figure 1). We adopt the standard modality-projection approach used in vision-language models (Liu

et al., 2023a; Bai et al., 2023) and find it effective for transcriptomic inputs. A linear layer maps normalized expression values together with pretrained omics embeddings into the input embedding space of an LLM. Each profile occupies a compact representation that can be interleaved with natural-language instructions and other profiles, enabling direct comparisons across experimental samples. We also augment the vocabulary with explicit gene tokens to represent biological nomenclature efficiently.

To help the model interpret omics data in its biological context, we devote substantial effort to constructing the training mixture: templates linking experimental measurements to curated biological knowledge, reasoning examples connecting evidence across sources, and LLM-generated supervision. It draws on expression atlases, perturbation screens, human genetics, and scientific literature. We organise the biological and general-purpose supervision used for instruction tuning into 157 task groups by biological objective and input/output format, bundling related cohorts, context variants, and prompt variants (Table 17). Training proceeds through 20 billion input tokens of continued pretraining, supervised fine-tuning on 5.21 million conversations (3.49 million with omics inputs), and reinforcement learning.

For context, C2S-Scale reports 12 pretraining task formulations (Rizvi et al., 2025), InstructCell three biological tasks (Fang et al., 2025), and CellWhisperer a generated pool of 106,610 chats (Schaefer et al., 2025). Our 157 task groups and 5.21 million SFT conversation records include both biological and general-purpose supervision. Although counting conventions differ, our corpus substantially expands the volume and diversity of supervision used for omics–language training. Sections 3.2 and 4 describe the training stages and data construction.

Our main contributions are:

- **OmicsLM**, a conversational model, fusing continuous transcriptomics with language, supporting several samples in context, evaluated on annotation, perturbation prediction, and biological question answering.
- **GEO-OmicsQA**, a benchmark for language-guided interpretation of one or more transcriptomic profiles, which we are publicly releasing.
- **Large-scale construction of omics–language training data**, transforming heterogeneous experimental measurements, curated knowledge, and scientific literature into supervision for interpreting biological profiles. We combine data harmonization, task design, and evidence-based filtering of generated reasoning to support a training procedure spanning pretraining, instruction tuning, and reinforcement learning.

Evaluation datasets, model checkpoints, and reproduction scripts are described in Appendix A.1.

2 Related Work

General-purpose Gemini (Google DeepMind, 2026b;d), GPT (OpenAI, 2026), and Qwen (Qwen Team, 2026b;c) models offer broad knowledge and reasoning, but dedicated training on quantitative omics measurements paired with biological tasks is not established in their published training descriptions.

Omics Foundation Models scGPT (Cui et al., 2024), Geneformer (Theodoris et al., 2023; Chen et al., 2026), and scFoundation (Hao et al., 2024) learn gene and cell representations through self-supervised transcriptomic pretraining. Funomics extends representation learning across bulk and single-cell expression (Powalski et al., 2026), while GenePT derives embeddings from textual gene descriptions (Chen & Zou, 2025). These features support downstream prediction and can serve as inputs to instruction-following generative models.

Text-Based Biomedical LLMs Cell2Sentence (C2S) (Levine et al., 2024) and C2S-Scale (Rizvi et al., 2025) transform expression matrices into "cell sentences"—plain-text strings of gene names ordered by descending abundance. Portraying Gene Language (PGL) (Choi et al., 2026) explores transcriptome generation from metadata using gene-expression token sequences. Cell-o1 (Fang et al., 2026) similarly represents each cell with ranked top-expressed genes and donor metadata, training a 7B LLM with supervised fine-tuning and reinforcement learning for batch-level annotation. While

compatible with standard language-model architectures, rank-only representations omit explicit expression magnitudes, and longer gene lists consume context that could otherwise accommodate additional samples or instructions. Beyond expression serialization, CellHermes (Gao et al., 2025) converts transcriptomic and protein-interaction data into templated question–answer pairs for joint training. For perturbation prediction, SUMMER (Wu et al., 2025b) combines biological knowledge summaries with retrieval of experimental outcomes, and SynthPert (Phillips et al., 2025) trains on synthetic mechanistic rationales conditioned on known outcomes and filtered by a critic.

Multimodal Omics-Language Alignment CellWhisperer (Schaefer et al., 2025) combines contrastive transcriptome–text alignment with a generative chat model that projects transcriptome embeddings into the LLM token space. It uses paired profiles and descriptions for alignment and generated conversations for chat supervision. InstructCell (Fang et al., 2025) pairs quantitative cell inputs with language instructions, constructing supervision from biological attributes and LLM-generated instruction–response templates. Related interfaces include scMMGPT's cross-modal alignment and generation objectives (Yang et al., 2026) and Cell2Text's adaptation of gene-level embeddings for cell-description generation (Kharouiche et al., 2025). AROMA (Wang et al., 2026) combines retrieved text with graph and protein representations for perturbation reasoning. We build on these interfaces with a training mixture covering diverse tasks and comparative interpretation of multiple samples.

Biological Question Answering Benchmarks CellVerse (Zhang et al., 2025a), SOAR (Liu et al., 2025), and scBench (Workman et al., 2026) cover cell-type annotation, perturbation analysis, scientific QA, and practical single-cell workflows. SC-Arena (Zhao et al., 2026) extends natural-language evaluation to cell captioning, generation, perturbation prediction, and scientific QA; PerturbQA (Wu et al., 2025b) tests experimental perturbation outcomes and gene-set interpretation. GEO-OmicsQA complements these benchmarks with publication-disjoint evaluation of language-guided interpretation across real expression profiles.

3 METHOD

The OmicsLM architecture is designed to bridge the modality gap between continuous omics data and discrete natural language. By treating single-cell and bulk transcriptomes as a continuous modality analogous to visual inputs in vision-language models (VLMs), we enable a pretrained LLM to condition on high-dimensional biological states. In this section, we describe the model architecture and training procedure; the training corpus is described in Section 4, and the GEO-OmicsQA benchmark in Section 5.2.

3.1 MODEL ARCHITECTURE

Composite Omics Representations OmicsLM represents each expression profile as a continuous vector. We align all expression profiles to a fixed human gene panel of $G = 19{,}260$ protein-coding and mitochondrial genes derived from GENCODE annotations (Mudge et al., 2025). For each sample, the default input vector concatenates four feature groups: a one-dimensional input-scale indicator, the normalized expression vector, a 512-dimensional Funomics T0 embedding (Powalski et al., 2026), and a 768-dimensional embedding from Geneformer-V2-104M (Chen et al., 2026), yielding a 20,541-dimensional omics vector. This combination gives the language model access to gene-level expression magnitudes alongside representations learned from transcriptomic data. The Funomics and Geneformer embeddings are precomputed with frozen encoders, which are not updated during language-model training. The input-scale indicator distinguishes count-derived single-cell profiles from transcripts-per-million (TPM)-normalized bulk RNA-seq profiles, aligning the expression values with their source modality.

The feature groups are normalized before training to make their scales comparable. Normalization statistics are computed from the training set. Expression values are first library-size normalized to 10^6, log-transformed with $\log(1 + x)$, centered per gene, and scaled by a single standard deviation computed over the full expression range. The input-scale indicator and learned embedding features are centered and scaled per dimension.

Cross-Modal Alignment via Linear Projection We use the text backbone of Qwen3.5-9B (Qwen Team, 2026a) as the language model. To connect the omics vector to this backbone, OmicsLM adds a trainable affine projection

$$\mathcal{P}_\theta(\mathbf{v}) = W\mathbf{v} + \mathbf{b}, \qquad \mathbf{v} \in \mathbb{R}^D, \quad W \in \mathbb{R}^{H \times D},$$

where $D = 20{,}541$ is the dimensionality of the full omics input vector and H is the LLM token-embedding size. The projection uses a bias term, is initialized with Xavier-uniform weights scaled by gain 0.01 and zero bias, and is trained jointly with the language model.

OmicsLM reuses the multimodal placeholder mechanism provided by the Qwen3-VL template (Bai et al., 2025), with the omics projection supplying the modality embeddings. We denote each profile placeholder schematically as `<omics>`. In the model input, each profile occupies one projected embedding $\mathcal{P}_\theta(\mathbf{v}_j)$ bracketed by two modality delimiter tokens. This lets omics profiles enter the LLM as continuous token embeddings while preserving the surrounding natural-language prompt. Formally, for text tokens x_i and omics vectors $\mathbf{v}_j$, the input embedding sequence, omitting the delimiters for clarity, becomes

$$[E_{\text{txt}}(x_1), \ldots, \mathcal{P}_\theta(\mathbf{v}_1), \ldots, E_{\text{txt}}(x_k), \mathcal{P}_\theta(\mathbf{v}_2), \ldots].$$

Because each profile appears at a designated position in the input sequence, a prompt can interleave multiple biological samples with natural-language instructions while preserving most of the context window for the task description and generated answer.

Gene-Aware Vocabulary Augmentation Standard LLM tokenizers are optimized for common human languages and often fragment biological nomenclature. A gene symbol such as `TP53` or `BRCA1` may be split into several subword pieces, making gene-level generation and matching unnecessarily brittle. We therefore augment the base tokenizer with the same canonical human gene symbols used for the 19,260-gene expression panel. For a gene symbol g that originally tokenizes to subword IDs $T(g)$, the default initialization sets the new token embedding to

$$E_{\text{new}}(g) = \frac{1}{|T(g)|} \sum_{t \in T(g)} E_{\text{old}}(t).$$

This initializes gene tokens from the base model's subword embeddings and allows canonical gene names to be read and generated as atomic tokens. The added gene embeddings are trainable together with the rest of the language model. Encoding each of the 19,260 gene symbols independently with the unmodified Qwen3.5 tokenizer, without leading whitespace or special tokens, requires an average of 3.72 tokens per gene. Vocabulary augmentation reduces this to one token per gene, shortening prompts and target sequences that mention many genes.

3.2 Training Procedure

We jointly update the language backbone, token embeddings, and omics projection while keeping the omics encoders frozen. The three stages below use the data described in Section 4; optimization settings and compute requirements are in Appendix A.7.

Stage 1: Multimodal Pretraining Starting from the augmented pretrained model, we train on 20 billion input tokens, interleaving continued pretraining (CPT) on free text with supervised fine-tuning (SFT) on instruction examples. This stage connects biological knowledge with quantitative omics inputs through alignment and prediction tasks. The causal language-modeling objective supervises text in free-text records and assistant responses in conversational records. Source sampling weights balance exposure across datasets (Appendix A.7.1).

Stage 2: Supervised Instruction Tuning Starting from Stage 1, we concentrate training on 5.21 million biological and general-purpose conversations, including direct answers, reasoning traces, and multi-turn interactions. We minimize negative log-likelihood on selected assistant responses; other messages and excluded assistant turns remain context without contributing to the loss. Appendix A.7.2 details the corpus and supervision.

Stage 3: Reinforcement Learning We refine the SFT model with group relative policy optimization (GRPO; Shao et al., 2024) in VERL (Sheng et al., 2025), optimizing sampled responses against deterministic task verifiers. Difficulty filtering with the SFT initializer selects 12,171 prompts that it can solve occasionally, excluding unsuccessful and nearly saturated groups. Rewards primarily measure task outcomes, with smaller terms for response structure and evidence agreement where ledgers are available. A KL penalty to the frozen SFT reference limits policy drift. Appendix A.7.3 gives filtering coverage, reward definitions, optimization settings, and exploratory small-model results, which include regressions as well as gains.

4 Training Data

The training mixture connects single-cell and bulk measurements with biological knowledge and reasoning, complemented by scientific text and general-purpose instructions. We group subsets into ten broad corpus categories by their primary purpose; one source may contribute to several categories. These categories describe the overall training mixture, whereas Table 17 provides a finer grouping of Stage 2 supervision into 157 task groups. Source preparation and inventories are documented in Appendix A.2, with representative tasks in Appendix A.10.

- **Cellular characterization:** Predicting cell type, tissue, lineage, cell-cycle phase, malignancy, clinical attributes, and cancer-model properties from expression profiles.
- **Expression interpretation and reasoning:** Identifying expressed or silent genes, comparing samples, describing profiles, interpreting pathway and interaction-partner activity, and answering profile-grounded questions.
- **Perturbation response:** Predicting differential expression, response direction and affected genes; identifying interventions; comparing doses; and transferring responses across cellular contexts.
- **Gene essentiality and functional screens:** Predicting context-specific essentiality and identifying or ranking gene hits in phenotypic screens.
- **Drug response and mechanisms:** Predicting sensitivity and quantitative response endpoints, identifying targets, and explaining drug mechanisms and cellular responses.
- **Human genetics:** Assessing gene–disease evidence, classifying variant consequences, and prioritizing candidate genes at GWAS loci.
- **Biological knowledge and networks:** Cell-line genomic facts, protein properties, pathway and interaction-graph reasoning, gene-set membership, and cell and disease ontologies.
- **Biomedical QA and information extraction:** Text-only biomedical questions, answers, reasoning traces, and entity-relation extraction.
- **Scientific and biomedical text and code:** Literature, clinical-trial records, drug labels, research-project descriptions, and biomedical code.
- **General-purpose capabilities:** General text, instruction following, mathematics, coding, multi-turn dialogue, system-instruction following, and tool use.

We use contextual cell-line detection to link mentions in text with DepMap reference profiles, interleaving biological text and quantitative measurements in the same input sequence (Appendix A.3).

Constructing several of these datasets required multistage integration of experimental measurements, curated biological knowledge, and literature evidence, together with quality checks on the generated explanations. The drug mechanism-of-action and marker-grounded reasoning datasets illustrate this process; their construction and filtering procedures are detailed in Appendices A.4 and A.5.

Benchmark exclusions. We apply benchmark-specific exclusions during training-data construction. GEO-OmicsQA test studies and their associated publications are withheld from training; Tabula Sapiens is reserved for evaluation; and 5,683 Human Diseases samples overlapping with the GEO training collection are removed. For PerturbQA, held-out cell-line–perturbation pairs are excluded from genetic training screens and derived task templates. Appendices A.2, A.6.1, and A.6.4 describe the scope of these exclusions.

5 RESULTS

We evaluate OmicsLM across three complementary biological capabilities: identifying samples, interpreting experimental comparisons, and predicting perturbation responses. First, zero-shot annotation tests whether the model can recover cell type, tissue, and disease identity from expression profiles in held-out datasets. Second, GEO-OmicsQA evaluates language-guided interpretation of real experimental profiles, including comparisons across multiple samples. Third, PerturbQA tests whether the same model can predict the direction of gene-expression changes following a perturbation from an unperturbed control profile and biological context. We then use ablations to examine how input representations, backbone size, and gene tokenization contribute to performance.

5.1 ZERO-SHOT ANNOTATION

We first test whether OmicsLM can identify the biological properties of samples from their expression profiles in annotation datasets unseen during training. We use Tabula Sapiens, which is excluded entirely from training, and Human Diseases, whose overlapping samples were removed from the GEO training corpus (Table 5a). The tasks cover cell type and tissue of origin in Tabula Sapiens, and disease subtype and tissue in Human Diseases.

We compare with CellWhisperer, C2S-Scale-27B, and Cell-o1 using macro AUROC. For OmicsLM, C2S-Scale, and Cell-o1, candidate labels are supplied in the prompt and scored by length-normalized constrained likelihood (Appendix A.6.2); C2S-Scale and Cell-o1 receive the 1,000 highest-expressed genes as text, while OmicsLM receives omics embeddings. We evaluate CellWhisperer's Mistral-7B chat model using its transcriptome embeddings and constrained likelihood scoring with task-specific answer prefixes. We also retain its published Geneformer–BioBERT embedding-model results, denoted CellWhisperer (CLIP).

CellWhisperer (LM) achieves the highest macro AUROC on Tabula Sapiens cell type top-20, while OmicsLM leads on Tabula Sapiens tissue and both Human Diseases tasks (Table 1). The transcriptome-embedding models outperform both ranked-gene baselines. C2S-Scale's poor Human Diseases disease-subtype result may partly reflect task mismatch: its reported pretraining objectives include cell-type and tissue annotation, but not disease-subtype classification (Rizvi et al., 2025).

Table 1: **Tabula Sapiens and Human Diseases classification zero-shot benchmarks.** Entries are macro AUROC. CellWhisperer (CLIP) reports published embedding-model results; CellWhisperer (LM) is our evaluation of its released chat model. All language-model rows use constrained likelihood scoring over the candidate labels.

Model	Tabula Sapiens		Human Diseases	
	Cell type top-20	Tissue	Disease subtype	Tissue
OmicsLM	0.926	**0.767**	**0.906**	**0.984**
CellWhisperer (LM) (Schaefer et al., 2025)	**0.948**	0.760	0.839	0.932
CellWhisperer (CLIP) (Schaefer et al., 2025)	0.942	0.747	0.819	0.872
C2S-Scale-27B (Rizvi et al., 2025)	0.684	0.592	0.543	0.631
Cell-o1 (Fang et al., 2026)	0.851	0.720	0.706	0.820

5.2 GEO-OMICSQA

We next move from assigning sample labels to interpreting experimental profiles through natural-language questions, including comparisons across samples. We introduce **GEO-OmicsQA** to evaluate this capability using one or more expression profiles and natural-language context. GEO (Clough et al., 2024) is a large, community-deposited archive of real-world studies from independent laboratories, spanning tissues, primary cells, cell lines, diseases, and genetic or chemical perturbations, each with its own experimental design and unstructured, study-specific metadata (Appendix A.2). GEO-OmicsQA inherits this heterogeneity: its 3,000 questions come from 235 studies and 183 publications, drawn from a held-out pool of 879 studies and 793 publications, and each is generated from its study's publication, so it refers to the samples, conditions, and findings of that specific study. Questions range from yes/no checks of reported findings to open-ended explanations of expression differences

between conditions, and they are clustered by study, with up to 57 questions from a single study. The benchmark comprises 1,000 balanced binary questions, 1,000 simple free-text questions, and 1,000 contextual free-text questions with auxiliary literature context. Each family has 500 standard and 500 minimal prompts. Test studies are publication-disjoint from the training data; a subset was reviewed by biological experts for question–answer validity, sample grounding, and biological coherence. Construction details and evidence annotations are provided in Appendix A.6.1.

Minimal prompts reduce textual scaffolding while retaining the same task and answer space. The standard and minimal subsets are not fully paired, so score differences may also reflect question composition.

We run nine language baselines with ranked lists of up to 1,000 genes per sample, adapting the count to each prompt's token budget; OmicsLM receives matched omics embeddings. We omit CellWhisperer from this benchmark because adapting its released single-profile chat interface to these multi-sample prompts is outside the scope of this comparison. Binary answers use balanced yes/no accuracy and free-text answers use mean DeepEval judge score (Appendix A.6.3). Input budgets, decoding settings, and C2S-Scale's yes/no logit scoring are detailed in Appendix A.6.1.

OmicsLM achieves the highest score across all six prompt variants (Table 2), with aggregate binary accuracy of 0.767 and a free-text judge score of 0.651, compared with 0.662 and 0.613 for Gemini 3.8 Flash. Although ranked gene lists are not a native quantitative input to general-purpose LLMs, several models make effective use of this representation: Flash leads the baselines on every variant, followed by Pro and the GPT models. The very low C2S-Scale free-text score reflects the released checkpoint's failure to answer questions under this protocol (Appendix A.6.1).

Table 2: **GEO-OmicsQA benchmark.** Aggregate columns average over their corresponding variants. Binary QA reports balanced *yes*/*no* accuracy, while free-text reports Gemini-judge score (full DeepEval judge details are in Appendix A.6.3).

	Aggregate		By prompt variant					
Model	Binary QA	Free text	QA std.	QA min.	Simple min.	Simple std.	Context min.	Context std.
Ours								
OmicsLM	**0.767**	**0.651**	**0.752**	**0.782**	**0.560**	**0.672**	**0.638**	**0.734**
General LLMs								
Gemini 3.8 Flash (Google DeepMind, 2026d)	0.662	0.613	0.712	0.612	0.512	0.636	0.607	0.696
Gemini 3.1 Pro Preview (Google DeepMind, 2026b)	0.640	0.536	0.690	0.590	0.442	0.544	0.546	0.612
GPT-6 Sol (OpenAI, 2026)	0.606	0.510	0.632	0.580	0.438	0.525	0.517	0.561
GPT-6 Luna (OpenAI, 2026)	0.575	0.479	0.588	0.562	0.411	0.474	0.495	0.537
Qwen3.5-9B (Qwen Team, 2026b)	0.534	0.360	0.542	0.526	0.319	0.355	0.369	0.395
Qwen3.8-27B (Qwen Team, 2026c)	0.546	0.403	0.560	0.532	0.351	0.397	0.408	0.457
Biology-specialized models								
TxGemma-27B (Wang et al., 2025)	0.518	0.289	0.532	0.504	0.222	0.282	0.307	0.344
Cell-o1 (Fang et al., 2026)	0.522	0.415	0.524	0.520	0.330	0.431	0.402	0.496
C2S-Scale-27B (Rizvi et al., 2025)	0.521	0.028	0.514	0.528	0.023	0.024	0.033	0.031

5.3 Perturbation Prediction

Finally, we evaluate prediction of a perturbation response from an unperturbed profile and biological context. On PerturbQA (Wu et al., 2025b), we predict whether a readout gene increases or decreases after CRISPRi of a target gene, conditional on the readout being differentially expressed. We use the official direction-of-change TEST split: 16,208 questions across four cell lines. OmicsLM receives the unperturbed control profile and public gene-pair notes, and generates an Increase/Decrease answer.

We rescore released SUMMER, retrieval, GenePT, and GEARS (Roohani et al., 2024) predictions and reproduce AROMA on the same test identities. SynthPert is included only as a published reference because its trained checkpoint is unavailable. Table 3 defines the aggregation and metrics; Appendix A.6.4 details scoring and reproduction checks.

OmicsLM attains 0.7635 mean AUROC, exceeding released SUMMER by 0.1088, with higher AUROC in every cell line (Table 3). Its direction macro-F1 is 0.8761, compared with SUMMER's diagnostic 0.7942. Detailed results are in Appendix A.6.4. These descriptive comparisons involve systems with different inputs, training data, and inference budgets; they do not isolate omics conditioning or establish statistical significance. We make no unseen-cell-line or globally unseen-target generalization claim.

Table 3: **PerturbQA direction-of-change prediction** (Wu et al., 2025b). AUROC is averaged over readout genes within each cell and then equally over cells; MF1 is cell-averaged direction macro-F1. All locally scored rows use the same 16,208 TEST questions; OmicsLM and AROMA average three sampled answer scores per question. Bold denotes the best of these scores. SynthPert (paper) is a published reference with unverified test coverage; its mean averages rounded paper values. For SUMMER, retrieval, GenePT, and GEARS, macro-F1 uses released scores thresholded at 0.5; the original decision rules are unverified. — denotes unavailable F1.

	AUROC					
Model	K562	RPE1	HepG2	Jurkat	Mean	MF1
OmicsLM	**.7939**	**.7173**	**.7679**	**.7750**	**.7635**	**.8761**
SUMMER (Wu et al., 2025b)	.6379	.6426	.6634	.6748	.6547	.7942
Retrieval without LLM (Wu et al., 2025b)	.6913	.6331	.6380	.6098	.6431	.7875
GenePT protein (Chen & Zou, 2025)	.5583	.5857	.5558	.5986	.5746	.7447
GenePT gene (Chen & Zou, 2025)	.4891	.5891	.6112	.5828	.5681	.7128
GEARS (Roohani et al., 2024)	.6660	.6064	.5195	.5181	.5775	.4648
AROMA (Wang et al., 2026)	.5645	.5683	.5987	.5838	.5788	.6934
SynthPert (paper) (Phillips et al., 2025)	.65	.73	.72	.65	.6875	—

Table 4: **Selected ablation results.** Overall is the unweighted mean of seven benchmark scores on a 0–100 scale (higher is better). The reference configuration uses Qwen3.5-0.8B, the full omics input, gene tokens, and Pretraining → SFT. Scores average three independent training runs; both 9B configurations report a single run. † denotes the final OmicsLM model. Bold denotes the best score in each block.

Backbone size		*Omics input*		*Training stages*		*Reinforcement learning*	
0.8B	77.6	Full input	**77.6**	Pretraining → SFT	**77.6**	9B	81.9
2B	79.9	No expression	76.2	Pretraining only	75.3	9B + RL†	**82.4**
9B	**81.9**	Cell2Sentence-like	71.5	SFT only	73.4		

5.4 Ablations

We examine the effects of backbone size, omics inputs, and training stages (Table 4). Appendix A.8 reports per-task scores, variability across runs, and additional input and gene-token ablations.

The ablations support three conclusions about building a shared model for omics and language. First, compact continuous profiles outperform the tested ranked-gene representation, supporting their use as an interface to experimental measurements. Second, combining pretraining with instruction tuning yields stronger aggregate performance than either stage alone, supporting the staged training recipe. Third, larger backbones improve biological question answering and perturbation prediction more consistently than sample annotation, suggesting that additional capacity benefits these capabilities differently. Reinforcement learning provides a smaller further improvement. The resulting picture is one of complementary contributions from the input representation, training procedure, and model capacity.

6 Conclusion

OmicsLM demonstrates that a conversational model can support both quantitative transcriptomic prediction and language-guided interpretation of experimental samples. Its performance across annotation, perturbation prediction, and biological question answering shows that these capabilities can coexist within a shared model. An established projection interface proves effective in this setting, with the broader contribution lying in how experimental measurements, biological knowledge, and task supervision are brought together during training. GEO-OmicsQA makes interpretation across samples an explicit target for evaluation. These findings support further development of omics-informed language models, while leaving open how training-data composition shapes transfer to new biological problems.

AI USE STATEMENT

We used generative AI for synthetic data generation (including mechanistic hypotheses in training examples), data cleaning and reformatting, method implementation, and automated answer evaluation. AI was not used for conceptual frameworks, experiment design, interpretation of reported results, research hypotheses, or translation; theoretical models, mathematical claims, and proofs are not applicable. Additionally, AI assisted coding, manuscript drafting, editing and review, literature searches, and reference formatting. Generated data underwent the checks described in the corresponding appendices, including expert review of a GEO-OmicsQA subset; automated judging assesses reference agreement, not independent biological validity. AI-assisted code was reviewed and tested, and authors reviewed manuscript changes. The authors take responsibility for all content, code, data, and results, including AI-assisted work.

REPRODUCIBILITY STATEMENT

Appendix A.1 describes the anonymous reproducibility bundle, including evaluation datasets, precomputed input embeddings, the model checkpoint, and evaluation scripts. Data preparation, training settings, and evaluation protocols are documented in the appendices.

REFERENCES

Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. OpenCodeReasoning: Advancing data distillation for competitive coding. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2504.01943.

Arc Institute. Virtual Cell Challenge: H1 hESC training data, 2025a. URL https://github.com/ArcInstitute/arc-virtual-cell-atlas/blob/main/virtual-cell-challenge/README.md. 2025 challenge; H1 hESC training data.

Arc Institute. pdex: Parallel differential expression for single-cell perturbation sequencing. https://github.com/ArcInstitute/pdex, 2025b. Version 0.1.24. MIT License.

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. doi: 10.1038/75556.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Amos Bairoch. The Cellosaurus, a cell-line knowledge resource. *Journal of Biomolecular Techniques*, 29(2):25–38, 2018. doi: 10.7171/jbt.18-2902-002.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Katy Börner, Sarah A Teichmann, Ellen M Quardokus, James C Gee, Kristen Browne, David Osumi-Sutherland, Bruce W Herr, Andreas Bueckle, Hrishikesh Paul, Muzlifah Haniffa, et al. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. *Nature Cell Biology*, 23(11): 1117–1128, 2021. doi: 10.1038/s41556-021-00788-6.

Broad DepMap. DepMap 24Q4 Public, 2024. URL https://doi.org/10.25452/figshare.plus.27993248.v1.

Tom B 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. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Seth Carbon and Chris Mungall. Gene Ontology data archive, 2025. URL https://doi.org/10.5281/zenodo.15611813. Data release 2025-06-01.

Chan Zuckerberg Initiative. CZ CELLxGENE Discover Census, 2025. URL https://chanzuckerberg.github.io/cellxgene-census/. LTS data release 2025-01-30.

H. Chen, M. S. Venkatesh, J. Gómez Ortega, S. V. Mahesh, T. Nandi, R. Madduri, K. Pelka, and C. V. Theodoris. Scaling and quantization of large-scale foundation model enables resource-efficient predictions in network biology. *Nature Computational Science*, 6(5):450–463, 2026. doi: 10.1038/s43588-026-00972-4.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117784–117812. Curran Associates, Inc., 2024. doi: 10.52202/079017-3741. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d5a58d198afa370a3dff0e1ca4fe1802-Abstract-Conference.html.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14552–14573. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.751. URL https://doi.org/10.18653/v1/2025.findings-acl.751.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6. URL https://doi.org/10.1038/s41551-024-01284-6.

Hongyoon Choi, Haenara Shin, Dongjoo Lee, and Daeseung Lee. Generative single-cell transcriptomics via large language models. *bioRxiv*, 2026. doi: 10.64898/2026.01.12.699186. URL https://www.biorxiv.org/content/10.64898/2026.01.12.699186v1. Preprint.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Emily Clough, Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Research*, 52(D1):D138–D144, 2024. doi: 10.1093/nar/gkad965.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Steven M. Corsello, Rohith T. Nagari, Ryan D. Spangler, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, 2020. doi: 10.1038/s43018-019-0018-6. URL https://doi.org/10.1038/s43018-019-0018-6.

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*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Edward De Brouwer, Carl Edwards, Alexander Wu, Jenna Collier, Graham Heimberg, Xiner Li, Meena Subramaniam, Ehsan Hajiramezanali, David Richmond, Jan-Christian Hütter, Sara Mostafavi, and Gabriele Scalia. AssayBench: An assay-level virtual cell benchmark for LLMs and agents. *arXiv preprint arXiv:2605.10876*, 2026. URL https://arxiv.org/abs/2605.10876.

Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, Ceri E. Van Slyke, Nicole A. Vasilevsky, Melissa A. Haendel, Judith A. Blake, and Christopher J. Mungall. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7. URL https://doi.org/10.1186/s13326-016-0088-7.

James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature Methods*, 11(1):25–27, 2014. doi: 10.1038/nmeth.2769.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html.

Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. BioReason: Incentivizing multimodal biological reasoning within a DNA-LLM model. In *Advances in Neural Information Processing Systems*, volume 38, pp. 123695–123722, 2025. doi: 10.52202/085713-3728. URL https://arxiv.org/abs/2505.23579.

Adibvafa Fallahpour, Arman Seyed-Ahmadi, Parsa Idehpour, Omar Ibrahim, Benedict M. H. Choi, Purav Gupta, Jack Naimer, Kevin Zhu, Abhinav Adduri, Arnav Shah, Shihao Ma, Talu Güloglu, Nuo Liu, Haotian Cui, Arihant Jain, Max de Castro, Amirfaham Fallahpour, Antonio Cembellin-Prieto, John S. Stiles, Filip Nemčko, Alexander A. Nevue, Hyungseok C. Moon, Lucas Sosnick, Olivia Markham, Haonan Duan, Michelle Y. Y. Lee, Andrea F. M. Salvador, Chris J. Maddison, Christoph A. Thaiss, Chiara Ricci-Tam, Brian S. Plosky, Dave P. Burke, Patrick D. Hsu, Hani Goodarzi, and Bo Wang. BioReason-Pro: Advancing protein function prediction with multimodal biological reasoning. *bioRxiv*, 2026. doi: 10.64898/2026.03.19.712954. URL https://doi.org/10.64898/2026.03.19.712954. Preprint.

Yin Fang, Xinle Deng, Kangwei Liu, Ningyu Zhang, Jingyang Qian, Penghui Yang, Xiaohui Fan, and Huajun Chen. A multi-modal AI copilot for single-cell analysis with instruction following. *arXiv preprint arXiv:2501.08187*, 2025. URL https://arxiv.org/abs/2501.08187v2.

Yin Fang, Qiao Jin, Guangzhi Xiong, Bowen Jin, Xianrui Zhong, Siru Ouyang, Yifan Yang, Aidong Zhang, Jiawei Han, and Zhiyong Lu. Cell-o1: Training LLMs to solve single-cell reasoning puzzles with reinforcement learning. *Bioinformatics*, 42(5):btag208, 2026. doi: 10.1093/bioinformatics/btag208. URL https://doi.org/10.1093/bioinformatics/btag208.

Claudia Feng, Elin Madli Peets, Yan Zhou, et al. A genome-scale single-cell CRISPRi map of trans gene regulation across human pluripotent stem cell lines. *Cell Genomics*, 6(2):101076, 2026. doi: 10.1016/j.xgen.2025.101076. URL https://doi.org/10.1016/j.xgen.2025.101076.

Yicheng Gao, Weixu Wang, Yuheng Zhao, Kejing Dong, Caihua Shan, Weizhong Zheng, Till Richter, Zekai Li, Siming Chen, Fabian J. Theis, and Qi Liu. Language may be all omics needs: Harmonizing multimodal data for omics understanding with CellHermes. *bioRxiv*, 2025. doi: 10.1101/2025.11.07.687322. URL https://doi.org/10.1101/2025.11.07.687322. Preprint.

Avishai Gavish, Michael Tyler, Alissa C. Greenwald, Rouven Hoefflin, Dor Simkin, Roi Tschernichovsky, Noam Galili Darnell, Einav Somech, Chaya Barbolin, Tomer Antman, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature*, 618(7965):598–606, 2023. doi: 10.1038/s41586-023-06130-4. URL https://doi.org/10.1038/s41586-023-06130-4.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Gene Ontology Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292.

Maya Ghoussaini, Edward Mountjoy, Miguel Carmona, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49(D1):D1311–D1320, 2021. doi: 10.1093/nar/gkaa840. URL https://doi.org/10.1093/nar/gkaa840.

Adriana Carolina Gonzalez-Cavazos, Anna Tanska, Michael Mayers, Denise Carvalho-Silva, Brindha Sridharan, Patrick A. Rewers, Umasri Sankarlal, Lakshmanan Jagannathan, and Andrew I. Su. DrugMechDB: A curated database of drug mechanisms. *Scientific Data*, 10(1):632, 2023. doi: 10.1038/s41597-023-02534-z. URL https://doi.org/10.1038/s41597-023-02534-z.

Google DeepMind. Gemini 3.1 Flash-Lite model card, March 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/.

Google DeepMind. Gemini 3.1 Pro model card, February 2026b. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.5 Flash-Lite model card, July 2026c. URL https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/.

Google DeepMind. Gemini 3.8 Flash, 2026d. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash.

GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

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. doi: 10.1038/s41592-024-02305-7.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, 51(D1): D870–D876, 2023. doi: 10.1093/nar/gkac947.

Ann C Huang, Tsung-Han S Hsieh, Jiang Zhu, Jackson Michuda, Ashton Teng, Soohong Kim, Elizabeth M Rumsey, Sharon K Lam, Ikenna Anigbogu, Philip Wright, Mohamed Ameen, Kwontae You, Christopher J Graves, Hyunsung John Kim, Adam J Litterman, Rene V Sit, Alex Blocker, and Ci Chu. X-Atlas/Orion: Genome-wide Perturb-seq datasets via a scalable fix-cryopreserve platform for training dose-dependent biological foundation models. *bioRxiv*, 2025. doi: 10.1101/2025.06.11.659105. URL https://doi.org/10.1101/2025.06.11.659105. Preprint.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 369–383. PMLR, 2026. URL https://proceedings.mlr.press/v297/huang26a.html.

Longda Jiang, Carol Dalgarno, Efthymia Papalexi, et al. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *Nature Cell Biology*, 27(3): 505–517, 2025. doi: 10.1038/s41556-025-01622-z. URL https://doi.org/10.1038/s41556-025-01622-z.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1): D672–D677, 2025. doi: 10.1093/nar/gkae909.

Max Karlsson, Cheng Zhang, Loren Méar, Wen Zhong, Andreas Digre, Borbala Katona, Evelina Sjöstedt, Lynn Butler, Jacob Odeberg, Philip Dusart, et al. A single–cell type transcriptomics map of human tissues. *Science Advances*, 7(31):eabh2169, 2021. doi: 10.1126/sciadv.abh2169.

Oussama Kharouiche, Aris Markogiannakis, Xiao Fei, Michail Chatzianastasis, and Michalis Vazirgiannis. Cell2Text: Multimodal LLM for generating single-cell descriptions from RNA-Seq data. *arXiv preprint arXiv:2509.24840*, 2025. URL https://arxiv.org/abs/2509.24840v1.

Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, et al. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*, volume 1, pp. 141–146, 2017.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. BioASQ-QA: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023. doi: 10.1038/s41597-023-02068-4. URL https://doi.org/10.1038/s41597-023-02068-4.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Maxime Labonne. Open-PerfectBlend, 2024. URL https://huggingface.co/datasets/mlabonne/open-perfectblend.

Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, and Avi Ma'ayan. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, 9(1):1366, 2018. doi: 10.1038/s41467-018-03751-6.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. BioREx: Improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487, 2023. doi: 10.1016/j.jbi.2023.104487. URL https://doi.org/10.1016/j.jbi.2023.104487.

Daniel Levine, Syed A Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David Van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024. URL https://proceedings.mlr.press/v235/levine24a.html.

Guangpeng Li, Yue You, Yunlin Fu, Weige Zhou, Feng Tang, Jiaming Kong, and Luyi Tian. RVQ-Alpha: Bridging single-cell transcriptomics and large language models via hierarchical discrete tokenization and fact-aware reinforcement learning. *bioRxiv*, 2026. doi: 10.64898/2026.04.20.719773. URL https://doi.org/10.64898/2026.04.20.719773. Preprint.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023. doi: 10.52202/075280-2264. URL https://arxiv.org/abs/2303.17760.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015. doi: 10.1016/j.cels.2015.12.004. URL https://doi.org/10.1016/j.cels.2015.12.004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023a. doi: 10.52202/075280-1516.

J. Liu, S. Xu, Y. Wu, and J. Zhang. Single-Cell Omics Arena: evaluation of large language models for automatic cell-type annotations on single-cell omics data via RNA-seq bridging. *Briefings in Bioinformatics*, 26(6):bbaf622, 2025. doi: 10.1093/bib/bbaf622. URL https://doi.org/10.1093/bib/bbaf622.

Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416.e11, 2018. doi: 10.1016/j.cell.2018.02.052.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.153.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. BioRED: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022. doi: 10.1093/bib/bbac282. URL https://doi.org/10.1093/bib/bbac282.

James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8):1112–1118, 2010. doi: 10.1093/bioinformatics/btq099. URL https://doi.org/10.1093/bioinformatics/btq099.

José L. McFaline-Figueroa, Sanjay Srivatsan, Andrew J. Hill, et al. Multiplex single-cell chemical genomics reveals the kinase dependence of the response to targeted therapy. *Cell Genomics*, 4(2): 100487, 2024. doi: 10.1016/j.xgen.2023.100487. URL https://doi.org/10.1016/j.xgen.2023.100487.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024. doi: 10.1093/nar/gkad1025.

MiniCPM Team. UltraData-SFT-Agent-2609, 2026a. URL https://huggingface.co/datasets/openbmb/UltraData-SFT-Agent-2609.

MiniCPM Team. UltraData-RL-2609. Hugging Face dataset, 2026b. URL https://huggingface.co/datasets/openbmb/UltraData-RL-2609.

Adrian Mirza, Nawaf Alampara, Martiño Ríos-García, Mohamed Abdelalim, Jack Butler, Bethany Connolly, Tunca Dogan, Marianna Nezhurina, Bünyamin Şen, Santosh Tirunagari, Mark Worrall, Adamo Young, Philippe Schwaller, Michael Martin Pieler, and Kevin Maik Jablonka. ChemPile: A 250 GB diverse and curated dataset for chemical foundation models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, pp. 61976–62024, 2025. doi: 10.52202/085713-1861. URL https://openreview.net/forum?id=l8Mz3dcK1S.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with OpenMathReasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025. URL https://arxiv.org/abs/2504.16891.

Jonathan M Mudge, Sílvia Carbonell-Sala, Mark Diekhans, et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research*, 53(D1):D966–D975, 2025. doi: 10.1093/nar/gkae1078.

Ajay Nadig, Joseph M Replogle, Angela N Pogson, et al. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI. Introducing GPT-6 Sol and Luna, 2026. URL https://openai.com/index/introducing-gpt-6-sol-and-luna/.

Clare Pacini, Emma Duncan, Emanuel Gonçalves, James Gilbert, Shriram Bhosle, Stuart Horswell, Emre Karakoc, Howard Lightfoot, Ed Curry, Francesc Muyas, Monsif Bouaboula, Chandra Sekhar Pedamallu, Isidro Cortes-Ciriano, Fiona M. Behan, Lykourgos-Panagiotis Zalmas, Andrew Barthorpe, Hayley Francies, Steve Rowley, Jack Pollard, Pedro Beltrao, Leopold Parts, Francesco Iorio, and Mathew J. Garnett. A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell*, 42(2):301–316.e9, 2024. doi: 10.1016/j.ccell.2023.12.016. URL https://doi.org/10.1016/j.ccell.2023.12.016.

Oana Palasca, Alberto Santos, Christian Stolte, Jan Gorodkin, and Lars Juhl Jensen. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database*, 2018:bay003, 2018. doi: 10.1093/database/bay003.

Zinaida Perova, Mauricio Martinez, Tushar Mandloi, Federico Lopez Gomez, Csaba Halmagyi, Alex Follette, Jeremy Mason, Steven Newhauser, Dale A Begley, Debra M Krupke, Carol Bult, Helen Parkinson, and Tudor Groza. PDCM Finder: an open global research platform for patient-derived cancer models. *Nucleic Acids Research*, 51(D1):D1360–D1366, 2023. doi: 10.1093/nar/gkac1021. URL https://doi.org/10.1093/nar/gkac1021.

Lawrence Phillips, Marc Boubnovski Martell, Aditya Misra, Josefa Lia Stoisser, Cesar A. Prada-Medina, Rory Donovan-Maiye, and Kaspar Märtens. SynthPert: Enhancing LLM biological reasoning via synthetic reasoning traces for cellular perturbation prediction. *arXiv preprint arXiv:2509.25346*, 2025. URL https://arxiv.org/abs/2509.25346v1. Version 1; Table 2 supplies the published direction-of-change reference values.

Rafał Powalski, Albert Roethel, Joanna Krawczyk, Przemysław Pietrzak, Łukasz Smoliński, Maciej Sypetkowski, Paweł Biernat, Dariusz Plewczyński, and Tomasz Jetka. Simultaneous learning from bulk and single-cell expression data with perceiver-based models. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations (MLGenX)*, 2026. URL https://openreview.net/forum?id=exRCSDdksX.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.5-9B model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Qwen Team. Qwen3.8-27B model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B.

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. doi: 10.7554/eLife.27041.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

S. A. Rizvi, D. Levine, A. Patel, S. Zhang, E. Wang, C. J. Perry, I. Vrkic, N. M. Constante, Z. Fu, S. He, D. Zhang, C. Tang, Z. Lyu, R. Darji, C. Li, E. Sun, D. Jeong, L. Zhao, J. Kwan, D. Braun, B. Hafler, H. Chung, R. Dhodapkar, P. Jaeger, B. Perozzi, J. Ishizuka, S. Azizi, and D. van Dijk. Scaling large language models for next-generation single-cell analysis. *bioRxiv*, 2025. doi: 10.1101/2025.04.14.648850. URL https://www.biorxiv.org/content/10.1101/2025.04.14.648850.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embedding provides a foundation model for cell biology. *Nature*, 656(8126):183–191, 2026. doi: 10.1038/s41586-026-10689-z. URL https://doi.org/10.1038/s41586-026-10689-z.

M. Schaefer, P. Peneder, D. Malzl, S. D. Lombardo, M. Peycheva, J. Burton, A. Hakobyan, V. Sharma, T. Krausgruber, C. Sin, J. Menche, E. M. Tomazou, and C. Bock. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02857-9.

Ruth L. Seal, Bryony Braschi, Kristian Gray, Tamsin E. M. Jones, Susan Tweedie, Liora Haim-Vilmovsky, and Elspeth A. Bruford. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*, 51(D1):D1003–D1009, 2023. doi: 10.1093/nar/gkac888. URL https://doi.org/10.1093/nar/gkac888.

Hyunjin Seo, Hyeon Hwang, Gyubok Lee, Jay Shin, Jimin Park, Taesoo Kim, Sanghoon Lee, Hongjoon Ahn, Sungjun Han, and Sangwon Jung. TheBioCollection: Unified pre-training scale LLM corpus for biology. *arXiv preprint arXiv:2607.08803*, 2026. URL https://arxiv.org/abs/2607.08803.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys '25)*. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://doi.org/10.1126/science.aax6234.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and Yu Rong. ReasonMed: A 370k multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26446–26467. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1344. URL https://aclanthology.org/2025.emnlp-main.1344/.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Disentangling reasoning and knowledge in medical large language models. *arXiv preprint arXiv:2505.11462*, 2025. URL https://arxiv.org/abs/2505.11462.

The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. doi: 10.1038/ng.2764.

The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

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, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015. doi: 10.1126/science.1260419.

Deepak R. Unni, Sierra A. T. Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J. Harry Caufield, et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, 15(8):1848–1855, 2022. doi: 10.1111/cts.13302. URL https://doi.org/10.1111/cts.13302.

Nicole A Vasilevsky, Nicolas A Matentzoglu, Sabrina Toro, Joseph E Flack, Harshad Hegde, Deepak R Unni, Gioconda F Alyea, Joanna S Amberger, Larry Babb, James P Balhoff, et al. Mondo: Unifying diseases for the world, by the world. *medRxiv*, 2022. doi: 10.1101/2022.04.13.22273750. URL https://doi.org/10.1101/2022.04.13.22273750. Preprint.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. TxGemma: Efficient and agentic LLMs for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025. URL https://arxiv.org/abs/2504.06196.

Zhenyu Wang, Geyan Ye, Wei Liu, and Man Tat Alexander Ng. AROMA: Augmented reasoning over a multimodal architecture for virtual cell genetic perturbation modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31855–31881. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1594. URL https://doi.org/10.18653/v1/2026.findings-acl.1594.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. RedPajama: an open dataset for training large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 116462–116492, 2024. doi: 10.52202/079017-3697. URL https://arxiv.org/abs/2411.12372.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-Instruct. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52632–52657. PMLR, 2024. URL https://proceedings.mlr.press/v235/wei24h.html.

Kenny Workman, Zhen Yang, Harihara Muralidharan, Aidan Abdulali, and Hannah Le. scBench: Evaluating AI agents on single-cell RNA-seq analysis. *arXiv preprint arXiv:2602.09063*, 2026. URL https://arxiv.org/abs/2602.09063.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. MedReason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025a. URL https://arxiv.org/abs/2504.00993.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Hütter. Contextualizing biological perturbation experiments through language. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b6e67ae290635d0874c4cb43ba2a2cfb-Abstract-Conference.html.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Hütter. Data distribution for "contextualizing biological perturbation experiments through language", 2025c. URL https://doi.org/10.5281/zenodo.14915313.

Jiaqi Yang, Zhiyuan Liu, Yaorui Shi, Qingchuan Zhang, Changhao Nai, Junfeng Fang, Yahui Long, Yang Zhang, and Xiang Wang. Language-enhanced representation learning for single-cell transcriptomics. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 12609–12620. ACM, 2026. doi: 10.1145/3770855.3819031. URL https://doi.org/10.1145/3770855.3819031.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 2013. doi: 10.1093/nar/gks1111. URL https://doi.org/10.1093/nar/gks1111.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c400474e8a36d0812fdee52739288b12-Abstract-Conference.html.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. CellVerse: Do large language models really understand cell biology? In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025a. doi: 10.52202/085713-2272. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/619cdaf77fd7bda767d0c3a9b1c1d218-Abstract-Datasets_and_Benchmarks_Track.html.

Jesse Zhang, Airol A. Ubas, Richard de Borja, et al. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025b. doi: 10.1101/2025.02.20.639398. URL https://doi.org/10.1101/2025.02.20.639398. Preprint.

Haiteng Zhao, Chang Ma, Fangzhi Xu, Lingpeng Kong, and Zhi-Hong Deng. BioMaze: Benchmarking and enhancing large language models for biological pathway reasoning. *arXiv preprint arXiv:2502.16660*, 2025a. URL https://arxiv.org/abs/2502.16660.

Haiteng Zhao, Chang Ma, Fangzhi Xu, Lingpeng Kong, and Zhi-Hong Deng. BioMaze: all_hsa_graph.graphml. GitHub repository, 2025b. URL https://github.com/zhao-ht/BioMaze/blob/4c22f4da0d7ed544f8495be3fdc260814951fc14/dataset/pathway_graph_env/all_hsa_graph.graphml. KEGG-derived pathway graph; upstream commit 4c22f4d.

Jiahao Zhao, Feng Jiang, Shaowei Qin, Zhonghui Zhang, Junhao Liu, Guibing Guo, Hamid Alinejad-Rokny, and Min Yang. SC-Arena: A natural language benchmark for single-cell reasoning with knowledge-augmented evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html.

A Appendix

A.1 Data Availability

To support reproducibility, we release four components. First, evaluation datasets comprising three processed AnnData files, six GEO-OmicsQA question files, and PerturbQA inputs with control profiles, covering perturbation prediction, tissue classification, disease classification, cell-type annotation, and multi-sample biological question answering. Second, precomputed OmicsLM input embeddings aligned to these evaluation datasets. Third, the released OmicsLM checkpoint. Fourth, evaluation scripts for reproducing the reported benchmark results.

The reproducibility bundle is available during anonymous review at https://huggingface.co/datasets/Authors2027/iclr-reproducibility-bundle. After the anonymity period, the datasets, model checkpoints, and evaluation scripts will be transferred to a permanent

public release with full author attribution. Licences for all released files are documented in the bundle README. All source datasets are publicly available; a complete asset inventory with licence terms is provided in Appendix A.11, and individual source references are provided in Appendix A.2.

A.2 DATASETS

All expression datasets used in this study were aligned to a unified set of 19,260 protein-coding and mitochondrial Ensembl genes derived from GENCODE annotations (Mudge et al., 2025). For single-cell datasets, we retained raw count data prior to normalization, while for bulk expression datasets, we used precomputed transcripts per million (TPM) values.

Table 5a: Summary of expression datasets used for training and evaluation. Expression profiles are aligned to the same set of 19,260 protein-coding and mitochondrial genes. N_{pert} denotes the number of distinct genetic perturbations or, for chemical screens, compound–dose conditions; — indicates an observational study. Pseudobulk denotes scRNA-seq screens aggregated into pseudobulk profiles, which are counted as samples.

Dataset	Samples	Genes	N_{pert}	Modality	Study type
Observational					
Human Cell Atlas (Regev et al., 2017)	6,161,559	19,260	—	scRNA-seq	Observational
Tabula Sapiens (GEO) (The Tabula Sapiens Consortium, 2022)	483,152	19,260	—	scRNA-seq	Observational
GEO (Clough et al., 2024)	615,595	19,260	—	Bulk RNA-seq	Observational
GTEx v10 (GTEx Consortium, 2020)	19,616	19,260	—	Bulk RNA-seq	Observational
TCGA (The Cancer Genome Atlas Research Network et al., 2013)	11,504	19,260	—	Bulk RNA-seq	Observational
Human Diseases (Schaefer et al., 2025)	14,112	19,260	—	Bulk RNA-seq	Observational
DepMap Expression 24Q4 (Broad DepMap, 2024)	1,673	19,260	—	Bulk RNA-seq	Observational
3CA (Gavish et al., 2023)	4,097,846	19,260	—	scRNA-seq	Observational
CancerModels.org (Perova et al., 2023)	6,437	19,260	—	Bulk RNA-seq	Observational
Perturbational					
DepMap Gene Dependency 24Q4 (Broad DepMap, 2024)	1,178	—	17,916	CRISPR	Genetic KO
Replogle et al. (2022)	2,546,823	19,260	9,859	scRNA-seq	Genetic KD
Nadig et al. (2025)	408,230	19,260	2,390	scRNA-seq	Genetic KD
Tahoe-100M (Zhang et al., 2025b)	46,392	19,260	1,129	Pseudobulk	Chemical
sci-Plex3 (Srivatsan et al., 2020)	2,259	19,260	752	Pseudobulk	Chemical
sci-Plex GxE (McFaline-Figueroa et al., 2024)	15,687	19,260	530	Pseudobulk	Genetic KD + Chemical
Jiang et al. (2025)	1,656	19,260	218	Pseudobulk	Genetic KD
Feng et al. (2026)	8,454	19,260	444	Pseudobulk	Genetic KD
Virtual Cell Challenge H1 hESC (Arc Institute, 2025a)	218,268	19,260	150	scRNA-seq	Genetic KD
X-Atlas/Orion (Huang et al., 2025)	36,586	19,260	18,310	Pseudobulk	Genetic KD

Table 5b: Free-text datasets and knowledge sources used for training.

Source	Description
Knowledge graphs and curated databases	
KEGG, via BioMaze (Kanehisa et al., 2025; Zhao et al., 2025a;b)	Knowledge graph: regulatory and metabolic pathways
STRING v12.0 (Szklarczyk et al., 2023)	Knowledge graph: protein–protein interactions
Gene Ontology and MSigDB gene sets (Ashburner et al., 2000; Gene Ontology Consortium, 2026; Liberzon et al., 2015)	Gene sets: gene-program membership
Cellosaurus (Bairoch, 2018)	Database: cell-line annotations

Continued on next page.

Table 5b continued.

Source	Description
UniProt (Swiss-Prot) (The UniProt Consortium, 2025)	Database: protein function and location
Cell Ontology, EFO, and Mondo (Diehl et al., 2016; Malone et al., 2010; Vasilevsky et al., 2022)	Ontology: cell-type and disease hierarchies
DrugMechDB (Gonzalez-Cavazos et al., 2023)	Knowledge graph: drug mechanism paths
Open Targets Genetics (Ghoussaini et al., 2021)	Database: gene–disease and variant evidence
AssayBench (De Brouwer et al., 2026)	Screens: phenotypic CRISPR screen hits
Biomedical question answering, reasoning, and information extraction	
BioASQ (Krithara et al., 2023)	QA: expert biomedical questions
BioRED, BioREx, ChemProt (Luo et al., 2022; Lai et al., 2023; Krallinger et al., 2017)	Relation extraction: entity relations in abstracts
ReasonMed, Medical-R1, MedReason, M23K (Sun et al., 2025; Chen et al., 2025; Wu et al., 2025a; Huang et al., 2026; Thapa et al., 2025)	QA with reasoning: medical questions
TheBioCollection (Seo et al., 2026)	Instructions, free text: biological records as text
CAMEL Biology (Li et al., 2023)	Instructions: biology questions, formatting constraints
BioReason-Pro, BioReason KEGG (Fallahpour et al., 2026; 2025)	Reasoning: protein function, pathways
Scientific and biomedical text and code	
PubMed abstracts	Free text: biomedical abstracts
PubMed Central, incl. biomedically enriched subset	Free text: full-text articles
arXiv	Free text: scientific preprints
bioRxiv, medRxiv	Free text: life-science and medical preprints
AACT, DailyMed, NIH RePORTER	Free text: trials, drug labels, grant abstracts
ChemPile-Code (Mirza et al., 2025)	Code: bio- and chemoinformatics code
General-purpose text and instructions	
Wikipedia (RedPajama) (Weber et al., 2024)	Free text: encyclopedic text
Open PerfectBlend (Labonne, 2024)	Instructions: general conversations
GSM8K, MetaMath, OpenMathReasoning (Cobbe et al., 2021; Yu et al., 2024; Moshkov et al., 2025)	Reasoning: mathematics
Magicoder, OpenCodeReasoning, SmolTalk Self-OSS (Wei et al., 2024; Ahmad et al., 2025; Ben Allal et al., 2025)	Code: code generation and reasoning
ARC, MMLU (Clark et al., 2018; Hendrycks et al., 2021)	Multiple-choice QA: general knowledge
SmolTalk constraints, rewriting, Magpie, multi-turn (Ben Allal et al., 2025)	Instructions: instruction following, dialogue
Hermes, xLAM, ToolACE, When2Call, UltraData-Agent (MiniCPM Team, 2026a)	Tool use: function calls, agent trajectories

A.2.1 Human Cell Atlas

The Human Cell Atlas (HCA) (Regev et al., 2017) was accessed via the CELLxGENE Census (release 2025-01-30) (Chan Zuckerberg Initiative, 2025), from which we downloaded 27 datasets comprising 6,161,559 cells spanning diverse tissues, developmental stages, and donors. HCA forms the primary source of training examples for single-cell tasks.

A.2.2 Tabula Sapiens

Tabula Sapiens (The Tabula Sapiens Consortium, 2022) is a multi-organ atlas covering 24 tissues from multiple donors. It is used exclusively as a held-out evaluation benchmark and is excluded from training. We use the GEO-hosted version of Tabula Sapiens (GSE201048; 483,152 cells), which is the version reported by CellWhisperer (Schaefer et al., 2025), enabling direct comparison. The dataset provides standardized cell-level metadata including cell type, tissue of origin, developmental lineage, cell-cycle phase, and donor sex.

A.2.3 GENE EXPRESSION OMNIBUS (GEO)

Bulk RNA-seq profiles come from the Gene Expression Omnibus (GEO) (Clough et al., 2024). For selected GEO studies, we uniformly re-processed their raw RNA-seq data with the open-source ARCHS4 processing pipeline (Lachmann et al., 2018), available under the Apache License 2.0. We filtered samples by single-cell probability, library selection, and library source, yielding $615{,}595$ human bulk RNA-seq samples with TPM transcript-level expression values. Associated sample metadata, downloaded directly from GEO, is unstructured and highly heterogeneous, posing a significant challenge for standardization at training scale. We addressed this with an LLM-based curation pipeline that (i) generated free-text natural language descriptions for each sample, and (ii) extracted structured annotations including tissue, cell type, cell line, disease, and perturbation target. The generated descriptions and structured annotations provide supervision for profile-description and annotation tasks.

For all GEO-derived tasks, including the GEO-OmicsQA benchmark, we split data at the publication / study level rather than at the sample level. Samples, metadata-derived prompts, and literature-derived QA examples from the same publication are assigned to a single split, preventing validation or test examples from sharing source publications with training examples.

A.2.4 HUMAN DISEASES

The Human Diseases dataset, assembled by Schaefer et al. (2025), is a curated collection of $14{,}112$ bulk RNA-seq samples across 27 tissue types covering a broad spectrum of disease contexts. Each sample carries a tissue-of-origin label and one of 229 free-text disease subtype labels. The $5{,}683$ samples overlapping with our GEO training corpus were removed from GEO prior to training to prevent data leakage.

A.2.5 GENOTYPE-TISSUE EXPRESSION (GTEX)

The Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2020) provides bulk RNA-seq profiles from postmortem donors across 54 tissue sites, designed to characterize tissue-specific patterns of gene expression and regulation. We use the v10 release, comprising 19,616 TPM-normalized samples spanning the full range of profiled tissues, which we use for tissue-of-origin training tasks.

A.2.6 THE CANCER GENOME ATLAS (TCGA)

The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network et al., 2013) is a landmark cancer genomics program that collected tumor and matched normal tissue samples from over 11,000 patients across 33 cancer types. We downloaded TPM-normalized expression data from the GDC Open Access portal, along with clinical metadata from the TCGA Pan-Cancer Clinical Data Resource (Liu et al., 2018), which provides standardized clinical outcome endpoints (overall survival, disease-specific survival, disease-free interval, and progression-free interval) and key demographic and pathological annotations across the full pan-cancer cohort.

A.2.7 DEPMAP

The Cancer Dependency Map (Broad DepMap, 2024; Pacini et al., 2024) is a systematic resource profiling cancer cell lines across multiple molecular modalities, including bulk RNA-seq expression, genome-wide CRISPR knockout screens, DNA mutations, copy number alterations, and proteomics. For each cell line, CRISPR-derived Chronos gene-effect and gene dependency scores quantify the fitness consequence of individual gene knockouts, enabling binary essentiality labels at the (cell line, gene) level. We use the 24Q4 release, spanning diverse tissue lineages and cancer types.

A.2.8 REPLOGLE AND NADIG DATASETS

Replogle et al. (2022) and Nadig et al. (2025) are large-scale Perturb-seq experiments combining CRISPR knockdown with single-cell transcriptomics to measure the genome-wide transcriptional response to thousands of individual gene perturbations. Replogle et al. profiled K562 (chronic myeloid leukemia) and RPE1 (retinal pigment epithelium) cell lines across 9,859 perturbations;

Nadig et al. profiled HEPG2 (hepatocellular carcinoma) and JURKAT (T-cell leukemia) cell lines across 2,390 perturbations. Each dataset includes unperturbed control cells serving as a reference baseline for perturbation prediction tasks.

A.2.9 BIOLOGICAL NETWORKS

Unlike the expression datasets above, pathway graphs, protein–protein interaction (PPI) networks, and gene set collections do not provide transcriptomic profiles; they are used solely to construct natural-language graph-reasoning tasks over biological interaction networks and gene program annotations.

We use the curated biological pathway graph database provided by Zhao et al. (Zhao et al., 2025a;b), derived from KEGG (Kanehisa et al., 2025) (Kyoto Encyclopedia of Genes and Genomes). This graph encodes directed regulatory and metabolic relationships between proteins and small molecules, representing expert-curated knowledge of cellular processes.

STRING (Szklarczyk et al., 2023) is a database of known and predicted protein–protein interactions, integrating evidence from experimental data, co-expression, text mining, and computational prediction. We used STRING v12.0 restricted to human interactions (taxonomy ID 9606). Combined interaction scores were recalculated following the STRING protocol: each per-channel score was prior-corrected as $s^* = (s - p)/(1 - p)$ with prior $p = 0.041$, and database and text-mining channels were excluded from the recombined score to avoid circularity with curated annotations. Four ubiquitous hub proteins (UBB, UBC, RPS27A, UBA52) were removed prior to filtering. We retained edges with a corrected combined score ≥ 700, or ≥ 150 when additionally supported by a database score ≥ 400, yielding a high-confidence interaction set.

The Gene Ontology (GO) (Ashburner et al., 2000; Gene Ontology Consortium, 2026) is a collaborative, continuously updated resource providing structured controlled vocabularies describing gene product functions across species. We downloaded Biological Process annotations from the GO knowledgebase (release `2025-06-01` (Carbon & Mungall, 2025)) and used them to associate genes with high-level functional programs, enabling natural-language tasks grounded in curated gene program membership.

A.2.10 CELLOSAURUS

Cellosaurus (Bairoch, 2018) is a comprehensive knowledge resource for cell lines, providing standardized nomenclature and curated biological annotations. For each cancer cell line in our DepMap catalogue, we retrieved annotations from the Cellosaurus REST API using the corresponding accession identifier (`CVCL_*`), including cell category, disease name with ontology identifier, transformant agent, and associated literature references. These annotations were merged with existing cell-line metadata to enrich natural-language tasks requiring the model to map an expression profile or cell-line name to its biological provenance and disease context.

A.2.11 UNIPROT

UniProt (The UniProt Consortium, 2025) is a comprehensive protein sequence and functional annotation database. We downloaded the manually reviewed Swiss-Prot flat file from the official UniProt FTP server and filtered it to human proteins only. For each protein, we extracted identity fields, functional descriptions, controlled-vocabulary keywords, subcellular localization, and protein class annotations. These were used to construct natural-language tasks requiring the model to answer questions about protein properties given a gene name.

A.2.12 CURATED CANCER CELL ATLAS (3CA)

The Curated Cancer Cell Atlas (3CA) (Gavish et al., 2023) is a collection of tumour single-cell RNA-seq studies with harmonized cell annotations, including the distinction between malignant and non-malignant cells. We use 122 human tumour studies. Five further files were excluded: one with a known technical defect and four that contain mouse data despite a human organism label. From these profiles we construct tasks on cell-type annotation (with and without the cancer type as context), malignant-cell identification, immune status, and developmental lineage, including reasoning variants.

A.2.13 CANCERMODELS.ORG

CancerModels.org (PDCM Finder) (Perova et al., 2023) is an open catalogue of patient-derived cancer models aggregated from multiple providers. We use its bulk RNA-seq TPM profiles, which span four model types: patient-derived xenografts, cell lines, organoids, and the originating patient tumours. The resulting tasks ask for the model type (alone or with the cancer type as context), cancer system, primary site, and histology; whether a profile comes from a patient tumour, a metastasis, a cell line or an organoid; and the cancer system shared by a matched patient–model pair.

A.2.14 CELL AND DISEASE ONTOLOGIES

We derive text-only knowledge tasks from two ontologies. The Cell Ontology (Diehl et al., 2016) provides 3,325 cell-type terms with definitions, synonyms, and paths to the root. For diseases we use the disease branch of the Experimental Factor Ontology (Malone et al., 2010), which integrates Mondo disease identifiers (Vasilevsky et al., 2022) (18,827 terms). Tasks cover subtype (is-a) relations in yes/no and list form, including hard negatives, direct parents, lowest common ancestors, shared classes, ordering by specificity, synonyms, and term definitions.

A.2.15 ASSAYBENCH

AssayBench (De Brouwer et al., 2026) evaluates phenotypic CRISPR screen prediction: given a natural-language description of a screen, a model ranks the genes most likely to be hits. We train on its published training fold, which shares no PubMed identifier with either evaluation fold. Two text-only tasks are derived from it: ranking the 100 strongest hits of a screen, and identifying and ranking hits among a list of candidate genes.

A.2.16 BIOASQ

BioASQ (Krithara et al., 2023) is an expert-curated corpus of biomedical questions with reference answers. We use the Task 11b training set and render its four question types (factoid, list, summary, and yes/no) as text-only question-answering tasks.

A.2.17 DRUGMECHDB

DrugMechDB (Gonzalez-Cavazos et al., 2023) is a curated database of drug mechanisms of action, each represented as a directed graph from a drug through its targets and intermediate biological entities to the treated disease. We use release 2.0.1 and exclude every indication evaluated by the DrugMechCounterfactuals benchmark, retaining 1,622 of 4,842 mechanism graphs. The text-only tasks ask the model to write the mechanism path from drug to disease, to decide whether an edge belongs to a mechanism (in closed and open-world variants), to name a masked intermediate node, and to predict how a perturbed link changes the drug's effect. DrugMechDB mechanisms are also paired with Tahoe-100M chemical perturbation profiles: from control and treated profiles, the model names the compound's mechanistic class or direct targets, or, from an untreated profile, decides whether any direct target of a named drug is expressed in the cells.

A.2.18 DRUG SENSITIVITY SCREENS

We use cancer cell-line drug-response labels from GDSC (Yang et al., 2013) and the PRISM repurposing screens (Corsello et al., 2020). Each example pairs the baseline expression profile of a screened cell line with its name, lineage, and disease, the compound, and, for GDSC, the compound's annotated targets; the model predicts whether the cell line is sensitive or resistant. GDSC labels follow the source's compound-specific GMM/Z-score assignment (48,248 conversations), complemented by instruction-format variants (5,000) and a classification of absolute IC50 values against a threshold (5,964). PRISM labels come from the bimodal single-dose response distribution of the primary screen (69,655) and from cell-specific activity calls in the secondary screen (24,652). Validation and test splits hold out cell lines (leave-cell-line-out, LCO), whole tissues (leave-tissue-out, LTO), and, for GDSC, cell line–compound pairs (leave-pair-out, LPO). The mechanism-augmented variants of these tasks are described in Appendix A.4.

A.2.19 CHEMICAL AND ADDITIONAL GENETIC PERTURBATION SCREENS

Beyond the Replogle and Nadig screens, we use chemical screens from Tahoe-100M (Zhang et al., 2025b) (47 cell-line files) and sci-Plex3 (Srivatsan et al., 2020), the combined genetic–chemical sci-Plex GxE screen (McFaline-Figueroa et al., 2024), and genetic screens from Jiang et al. (2025), Feng et al. (2026), the Virtual Cell Challenge H1 hESC release (Arc Institute, 2025a), and X-Atlas/Orion (Huang et al., 2025) (Table 5a). Single-cell measurements are aggregated into pseudobulk profiles, each resolved to a matched unperturbed control from the same context. A shared builder normalizes profiles, applies quality gates, and subtracts the common response of each context; for chemical screens this response is computed per cell line and dose, because responses at different doses are not interchangeable. Tasks cover response direction of readout genes (three-way and binary), top up- and downregulated genes, perturbation and compound identification, mechanism class and targets, phenotype changes, dose comparisons, and transfer of a knockout response across cytokine contexts. Three percent of perturbations are held out of all tables as an internal test, and the test pairs of PerturbQA, PertReasonQA, and the published chemical benchmarks are removed before training statistics are computed.

A.2.20 HUMAN GENETICS

Human-genetics tasks are derived from an export of Open Targets Genetics (Ghoussaini et al., 2021) built from Open Targets Platform release 26.06. Eleven text-only tasks cover gene–disease evidence as yes/no questions and disease lists, the evidence source and association strength, allelic requirement, rare-variant burden direction, ClinVar clinical significance, gene constraint, prioritization of candidate genes at GWAS loci, coding consequence of a variant, and variant frequency class. Prompts restrict the evidence to quality-filtered GWAS, rare-variant burden, ClinVar, and Gene2Phenotype records. The training split contains 300,000 conversations; separate development (9,087) and validation (16,673) splits are held out.

A.2.21 PERTURBQA TRAINING DATA

Besides serving as a benchmark (Appendix A.6.4), PerturbQA (Wu et al., 2025b) contributes training examples for the K562, RPE1, HepG2, and Jurkat screens: 300,000 direction-of-change questions with minimal prompts that supply the unperturbed control profile and public notes on the perturbation pair, 50,000 differential-expression detection questions, and 5,000 instruction-format variants. Further differential-expression and direction questions are posed on pseudobulk profiles of each cell line, complemented by the extended direction tasks built over all genetic screens (Appendix A.2.19). Benchmark (cell line, perturbed gene) pairs are excluded from all training data.

A.2.22 CELLWHISPERER CONVERSATIONS

We include the profile-linked conversations released with CellWhisperer (Schaefer et al., 2025), in which questions about a profile's cell type, differentiation state, and sample characteristics are answered with reference to the linked expression profile. We use both the main and the fine-tuning conversation sets. These conversations are distinct from the Human Diseases dataset, which is used only for evaluation.

A.2.23 INSTRUCTION AND REASONING DERIVATIVES

These components change the form of supervision rather than introduce new data sources. Omics instruction-following examples (13,437) ask for specific output formats or reformatting of answers to expression-grounded questions, and instruction-format variants of the essentiality, PerturbQA, and GDSC tasks (16,476 in total) rephrase their prompts. Omics dialogue examples (18,222) contain follow-up turns, updates on new evidence, and answer revisions. Text-only reasoning examples come from BioReason-Pro protein-function reasoning (15,000; Fallahpour et al., 2026) and BioReason KEGG pathway reasoning (1,159; Fallahpour et al., 2025), and CAMEL-derived biology answers carry explicit output constraints (19,503).

A.2.24 BIOMEDICAL QA AND INFORMATION EXTRACTION

Medical question–answer and reasoning data come from the CoTMed, ReasonMed, and ResponseMed variants of ReasonMed (Sun et al., 2025), Medical-R1-Distill-Data (Chen et al., 2025), MedReason (Wu et al., 2025a), and M23K (Huang et al., 2026; Thapa et al., 2025). These sources provide both direct answers and generated reasoning traces. We also use the instruction stream of TheBioCollection (Seo et al., 2026), which covers molecular, protein, genomic, cellular, and pathway tasks. The medical exports and TheBioCollection instruction stream are represented as serialized text records. BioASQ, described above, provides biomedical QA conversations, while CAMEL (Li et al., 2023) contributes biology instruction examples.

Relation-extraction datasets pair supplied text with structured biological relationships. BioRED (Luo et al., 2022) provides abstract-level entity and relation extraction; we combine its official training and development sets for training and reserve its test set for evaluation. From BioREx (Lai et al., 2023), we use the DrugProt and CID components of the released training split for relation classification between marked entities. ChemProt (Krallinger et al., 2017) provides chemical–protein relation extraction with supplied entity annotations. We retain its official training, development, and test separation, using only the training split for optimization. These three relation-extraction sources are formatted as conversational tasks, using checksum-pinned source archives or prepared files.

A.2.25 SCIENTIFIC AND BIOMEDICAL TEXT AND CODE

Scientific and biomedical text comes from PubMed abstracts, PubMed Central (PMC) articles, and preprints from arXiv, bioRxiv, and medRxiv. Additional sources include a biomedically enriched PMC subset selected using article-level educational-quality, language, and domain-relevance annotations, and the free-text stream of TheBioCollection (Seo et al., 2026). The latter represents biological records as text, complementing the research literature. Clinical and research context comes from AACT clinical-trial records, DailyMed drug labels, and NIH RePORTER project abstracts.[1] The preparation manifests pin AACT and DailyMed to releases dated 2026-09-01 and NIH RePORTER to the 2026-03-09 release. The bioRxiv and medRxiv full-text XML is rendered as text while retaining figure captions and tables. Where contextual cell-line mentions can be resolved to DepMap profiles, the configured text augmentation inserts references to those profiles (Appendix A.3).

For biomedical code, we select Python files from the CodeParrot component of ChemPile-Code (Mirza et al., 2025). Selection requires imports of supported bioinformatics, cheminformatics, or molecular-simulation libraries and applies syntax and source-license filters. Exact syntax-tree duplicates are removed; related repositories and similar files are grouped for splitting. Groups touching the upstream test split are excluded, and groups touching its validation split are assigned to validation. Code is retained as plain text.

A.2.26 GENERAL-PURPOSE TEXT AND INSTRUCTION DATA

General-purpose data cover encyclopedic text, reasoning, instruction following, conversation, and tool use. The Wikipedia component of RedPajama (Weber et al., 2024) provides free text, while Open PerfectBlend (Labonne, 2024) provides instruction-following conversations. For Open PerfectBlend, we retain nonempty conversations with alternating user and assistant turns, a nonempty source field, and no image placeholders, preserving the original turn contents. Open PerfectBlend uses the filtered upstream training split.

Mathematical reasoning datasets include GSM8K (Cobbe et al., 2021), MetaMath (Yu et al., 2024), and OpenMathReasoning (Moshkov et al., 2025). Coding instructions and reasoning examples come from Magicoder-Evol-Instruct (Wei et al., 2024), OpenCodeReasoning (Ahmad et al., 2025), and SmolTalk's Self-OSS-Instruct component, while ARC (Clark et al., 2018) and MMLU (Hendrycks et al., 2021) provide knowledge questions. Instruction-following and conversational examples include the AutoIF, LMSYS, and UltraChat components selected through Open PerfectBlend, together with SmolTalk constraint-following, rewriting, Magpie, and system-conditioned multi-turn conversations (Ben Allal et al., 2025). Tool-use data cover Hermes, ToolACE, When2Call, and xLAM irrelevance examples, complemented by agent trajectories from UltraData-SFT-Agent-2609 (MiniCPM

[1] Source portals: `https://aact.ctti-clinicaltrials.org/`, `https://dailymed.nlm.nih.gov/dailymed/`, and `https://reporter.nih.gov/exporter`.

Team, 2026a). The retained conversations preserve system and tool messages and any source assistant-turn supervision masks, together with the existing split assignments in the curated instruction corpus.

A.3 Cell-Line Detection and Profile Augmentation in Free Text

We link cell-line mentions in text to reference expression profiles using a dictionary-based detector with contextual rules. Canonical names and unambiguous Cellosaurus synonyms (Bairoch, 2018) are mapped to available DepMap 24Q4 profiles (Broad DepMap, 2024). Name normalization handles common spelling variants, while contextual checks and ambiguity filters help distinguish cell-line mentions from unrelated terms.

We insert an <omics> placeholder after an accepted mention, retaining the original text: for example, "MCF-7 cells" becomes "MCF-7<omics> cells". The placeholder supplies the corresponding profile through the omics projection (Section 3.1), allowing text and profiles to be interleaved within a record. Records without accepted matches remain text-only.

A.4 LLM-Based Mechanism-of-Action Dataset Generation

We built a training dataset that links the mechanism of action of each compound to its measured effect on specific cancer cell lines. Widely used resources largely emphasise one side of this link: drug–target databases and curated mechanism graphs describe mechanisms without experimental context, while sensitivity screens measure response without explaining it. Our dataset connects the two for each drug–cell-line pair. Instead of a single label, each pair comes with a set of explicitly graded mechanistic hypotheses about why that cell line responds as it does. The dataset is organised into three configurations: a compound-level record, a pair-level record for selected drug–cell-line pairs, and a hypothesis-level record with one row per mechanistic hypothesis. In the OmicsLM training mixture, the per-pair hypotheses serve as chain-of-thought rationales for the GDSC drug-sensitivity tasks.

Compound-level records were built for 356 compounds from the GDSC screen (Yang et al., 2013), starting from deterministic sources. Compound names are resolved to canonical identifiers with an explicit refusal threshold, so a compound is left unresolved rather than assigned a confident wrong match. Targets are merged from curated drug–target databases, with per-source provenance kept at every step, and each gene is graded by evidence strength. Genes backed by curated pharmacological or binding evidence become primary targets; genes with association-only evidence are kept as secondary. Targets are then linked to pathways, reactions and metabolites, disease indications and phenotypes, and the compound's measured sensitivity profile. Each compound receives a three-tier confidence label. Tier 1 requires the same gene to be reported by at least two independent strong sources. Tier 2 marks a single-source target or a literature-derived mechanism that passed every verification gate. Tier 3 marks a literature-derived hypothesis that failed at least one gate.

Large language models are used only where curated data runs out, and they never overwrite a fact taken from a database. For compounds with a resolved target, a model turns the established facts into prose without any retrieval. For compounds with no known target, mechanisms are extracted from PubMed articles retrieved through NCBI Entrez, with web search available as a fallback when these are insufficient. Three independent safeguards apply: self-consistency across samples drawn at different temperatures, deterministic verification that every cited PMID was among the retrieved articles and that every gene symbol exists, and an adversarial critic that tries to refute each uncertain claim. The critic caught errors that self-consistency cannot see, because a wrong answer can be reproducible. Among them were a proposed target that the cited sources never mention and downstream effectors reported as direct targets. Every generation step records whether it failed or genuinely found nothing, so a missing mechanism in the dataset reflects a completed search, not a silent failure. All generative steps use Gemini 3.5 Flash-Lite (Google DeepMind, 2026c), selected over Gemini 3.5 Flash and Gemini 3.1 Pro in a pilot comparison on literature extraction, where it achieved full grounding at about one-fourteenth of the cost of Pro. Self-consistency draws three samples at temperatures 0.0, 0.5 and 0.9, and the critic uses the same model.

The main novelty is how the per-pair hypotheses are generated. They are chain-of-thought rationales for an observed outcome: the generator receives the pair's measured response (sensitive or resistant) together with target, pathway and mutation facts, and is asked to explain that response using only

those facts. Each rationale is anchored in an explicit prior: an inhibitor should act on a pathway that is active at baseline in sensitive cell lines and inactive in resistant ones, and for activators the relation is reversed. The prior is checked against the cell line's own transcriptomic pathway activity. Four independent signals—primary pathway, secondary pathway, target mutation and pharmacogenomic association—each produce a separate hypothesis. A hypothesis is labelled supported when the cell line's data match the prior, and hypothetical when the data are ambiguous or the signal is too coarse to confirm it. Pairs whose own data contradict the prior are excluded rather than rationalised, so the dataset favours cases in which the simple mechanistic story holds. The dataset therefore captures a variety of reasoning paths with explicit uncertainty rather than one asserted answer. Mechanisms are also expressed as directed relation chains in the Biolink vocabulary (Unni et al., 2022) used by DrugMechDB (Gonzalez-Cavazos et al., 2023). Every hop traces to a named source, and no edge is added where no source asserts one.

The dataset is guarded against DrugMechCounterfactuals benchmark leakage: mechanism paths reserved for the benchmark are excluded through an explicit allowlist.

A.5 Marker-Grounded Reasoning Data

Most annotation tasks supervise a categorical answer without identifying the expression evidence that supports it. Our marker-grounded reasoning dataset adds this intermediate supervision. Responses connect detected genes to their biological roles, assess relevant alternative interpretations and then state the annotated answer. The examples cover cell type, lineage, immune identity, cell-cycle phase, tissue and donor sex across HCA and 3CA single-cell profiles and GTEx, TCGA and GEO bulk profiles. Cell-type tasks also include groups of cells from the same sample.

Shared design with scCoT-Synth. We adopt the evidence-first format and teacher–student construction of scCoT-Synth in RVQ-Alpha (Li et al., 2026). The teacher sees the source annotation and a set of gene features. In our version these are given as an evidence card with detected markers, gene functions and ontology context. The student receives only the question and expression representation. Predicting the genes cited in the response therefore supplies supervision beyond the final label. RVQ-Alpha uses this design to ground discrete expression tokens. We apply it to continuous omics embeddings across single-cell and bulk inputs, with each canonical gene represented by an atomic token (Section 3.1). Qwen3.8-27B (Qwen Team, 2026c) generates questions without seeing the answer or its marker panel, then writes rationales from the evidence card and the exact question. The final answer is supplied by the source annotation.

Adaptations for marker-grounded annotation. Our construction makes three changes to how evidence is selected and explanations are retained.

- **Reference-defined evidence.** scCoT-Synth supplies its teacher with known markers alongside profile-derived features such as expression-ranked genes. We restrict expression evidence to genes drawn from external marker panels and published programmes, then check their detection in each profile. No marker signatures are learned from our training corpora. Sources include the Human Reference Atlas, CellMarker and Human Protein Atlas (Börner et al., 2021; Hu et al., 2023; Karlsson et al., 2021; Uhlén et al., 2015), TISSUES (Palasca et al., 2018), and UniProt, GO and Reactome (The UniProt Consortium, 2025; Gene Ontology Consortium, 2026; Milacic et al., 2024).
- **Evidence matched to label specificity.** We select examples only when measured markers and context available in the question support the requested distinction. Broad lineage markers alone cannot justify a fine cell state or tissue identity. Unsupported examples are excluded rather than assigned broader labels. This prevents the target annotation from substituting for missing evidence in the explanation.
- **Quality checks used for selection.** For chain-of-thought data, scCoT-Synth logs structural, label-reference and rubric checks but does not filter on them. Only the RVQ code and nomenclature checks can drop a sample. Our grounding checks and separate grading by Gemma-4-31B-it (Gemma Team et al., 2026) determine which rationales are retained. Review checks measured expression claims, biological accuracy, the specificity of the conclusion and confidence in the evidence. Unsupported claims are grounds for rejection

even when the final label is correct. Detection cannot be presented as differential expression, and an unlisted marker cannot be treated as absent.

The dataset thus provides training targets that make the connection between expression, biological knowledge and annotation explicit. These are explanations of source labels, with automated checks on their evidential support.

The first example connects plasma-cell identity to antibody-processing machinery. The second identifies a shared T-cell lineage across six profiles while keeping the conclusion at the level supported by the group. Both are from 3CA. Questions and responses are reproduced verbatim, with consecutive profiles abbreviated as <omics>×N.

Single-cell identity: secretory machinery

Question: What cell type does the single-cell expression profile in <omics> most closely align with? Please provide a concise summary linking specific expression evidence to that identity, placing your final prediction on a separate line following the label "Answer:".

Response: The profile aligns with a plasma cell identity, driven by the co-expression of MZB1 and XBP1, which jointly support the immunoglobulin secretory and unfolded protein response programmes. MZB1 acts as a chaperone promoting IgM assembly, while XBP1 regulates ER stress responses essential for high-volume secretion. Supporting this secretory phenotype, DERL3 participates in ERAD, and SEC11C catalyzes signal peptide cleavage during protein translocation. Additional markers, including CD27, SLAMF7, and PRDM1, are consistent with the plasma cell lineage, though their functions vary from costimulation to transcriptional regulation. This combined evidence identifies the cell as a terminally differentiated B cell.

Answer: plasma cell

Shared cell identity across six profiles

Question: What single cell type identity is best supported by the expression patterns across these separate profiles in <omics>×6, and please place your prediction on the final line beginning with "Answer: "

Response: The co-expression of TCR complex subunits CD3D, CD3E, and CD3G establishes the defining characteristic of this lineage, identifying these cells as T cells. This core identity is reinforced by the detection of PTPRC across all members and the presence of CD8A and CD8B, which function as MHC class I coreceptors in T-cell immune responses. Additional markers such as CD2, CD5, and CD7 further support T-cell lineage through their roles in adhesion and co-stimulation. While cytotoxic markers like GZMA and IFNG are detected, indicating potential effector capacity, the primary evidence firmly anchors the identity at the T-cell level.

Answer: t cell

A.6 Benchmarks

A.6.1 GEO-OmicsQA: Construction and Task Variants

We introduce **GEO-OmicsQA**, a large-scale biological question-answering dataset constructed via automated literature scraping of the Gene Expression Omnibus (GEO) (Clough et al., 2024). Each question contains explicit references to one or more real expression profiles, requiring models to ground their answers in the linked omics data rather than textual priors.

Training–test separation. We exclude all GEO studies associated with GEO-OmicsQA test questions and their associated source publications from our training corpus. These exclusions cover the studies' expression profiles and metadata, the source papers, and training examples derived from these materials. Separation is therefore enforced at the study and publication level, rather than only between individual questions or samples. A direct comparison of GEO series, publication, and sample identifiers found no overlap between the benchmark and the training data. Manual inspection of the GEO profile inputs found no obvious leakage of benchmark questions or reference answers. We release the GEO series, sample, and PubMed identifiers of all benchmark questions, so that developers of other models can remove them from their training data and avoid contamination.

Unlike existing benchmarks that rely on static prompts or artificial single-cell pipeline tasks, GEO-OmicsQA challenges the model to synthesize insights directly from complex, multi-sample omics

profiles in a free-form conversational setting. The evaluation benchmark contains 3,000 held-out examples, split between binary QA and free-text generation. A subset of the held-out test set was manually reviewed by human annotators with biological expertise to check question-answer validity, sample grounding, and biological coherence. In both families, minimal prompts reduce textual scaffolding while retaining the same task and answer space, probing performance with less textual context. The standard and minimal subsets are not fully paired, so score differences may also reflect question composition.

- **Binary QA (1,000 samples).** A perfectly balanced set of yes/no questions asking whether one or more expression profiles support a specific biological claim about a population, phenotype, or experimental condition. These prompts are simple in output space but non-trivial in evidence requirements, since the model must inspect one or more omics profiles and decide whether the stated hypothesis is supported. We evaluate 500 **standard** and 500 **minimal** instantiations of this task.
- **Free-text (2,000 samples).** Open-ended questions that ask for concise biological explanations from supplied omics profiles. We split these into two equal groups by question type: **simple** (1,000), which avoids extra literature context, and **contextual** (1,000), which embeds auxiliary literature context alongside the profiles. Each group contains 500 **standard** and 500 **minimal** prompts, as in the binary QA setting.

Every GEO-OmicsQA question–answer pair includes structured literature evidence to support verification of the reference answer. Each record includes supporting passages, their locations in the source publication, and a short note explaining how they support the reference answer. Together with the GEO and publication identifiers, these annotations make reference answers traceable to reported findings, enabling independent checks of biological claims and their experimental context. Representative examples for each variant are shown in Appendix A.9.

Model inputs and prompting. Text-only baselines receive ranked gene symbols for each sample, with sample identifiers replaced by dummy labels (`sample_1`, `sample_2`, ...); OmicsLM receives `<omics>` placeholders with matched embeddings. Models are instructed to answer binary questions with *yes* or *no*; text-only baselines are asked to answer free-text questions in a few sentences. Each example uses one generation request without tools or code execution. Gene lists omit expression magnitudes and may omit genes mentioned in the question, even with the larger budget.

Dynamic gene budgets. We cap each profile at 1,000 genes, following the setting illustrated in the C2S-Scale model card.[2] For multi-sample prompts, we reduce the per-profile count as needed to fit the context window, consistent with the C2S authors' multi-cell formatting procedure (Rizvi et al., 2025). For a prompt containing P distinct profiles, we initially allocate

$$n = \min\left(1000, \left\lfloor \frac{L - H}{P} \right\rfloor\right)$$

genes per profile, where L is the model's input or context limit and H is the reserved headroom. Because a gene symbol can span several tokens, we then count the complete formatted prompt and shorten all gene prefixes equally until it fits within $L-H$. Counting includes the question, instructions, and chat template where applicable, using provider token-count APIs or the local tokenizer. No question or profile is dropped. Repeated sample identifiers count as one profile. Each of the nine baselines is evaluated on all six 500-example splits.

Generation and scoring settings. Gemini uses temperature 0 and a 10,000-token output cap; GPT retains API-default sampling and the same cap. API reasoning controls are not explicitly overridden. Local models use vLLM with temperature 0; Qwen and Cell-o1 use chat templates with thinking disabled and a 4,096-token output cap. C2S uses its raw prompt format with a 1,000-token free-text cap. For C2S binary questions, we select the higher-probability first-token continuation among `Yes` and `No`; other baselines generate their answers. Binary subsets are class-balanced, so accuracy equals balanced accuracy. Each result is from one run; the comparisons do not control for training exposure or reasoning budget. Prompts are checked against their input budgets. Two Pro, 11 Cell-o1, 10 Qwen

[2] `https://huggingface.co/vandijklab/C2S-Scale-Gemma-2-27B`

Table 6: Selected GEO-OmicsQA input budgets. Gemini limits are input limits; the GPT input limit is separate from its total context. Headroom is the minimum reserved amount, not the actual unused capacity.

Model	Input/context limit	Headroom	Genes/profile
Gemini Flash / Pro	1,048,576	1,000	1,000
GPT Sol / Luna	922,000	10,000	1,000
Qwen 9B / 27B	262,144	4,096	1,000
Cell-o1	32,768	4,096	330–1,000
C2S-Scale-27B	8,192	1,000	80–1,000

9B, and four Qwen 27B responses reached their output caps and remain included in scoring. Judge scores measure agreement with the literature-derived references, not an independent audit of every claim against the supplied profiles.

C2S-Scale output behavior. We evaluate the released `vandijklab/C2S-Scale-Gemma-2-27B` checkpoint without additional QA fine-tuning. Its poor free-text result primarily reflects an output-format failure: it often continues with a list of gene symbols instead of answering the biological question. For example, the first standard simple question asks which of two sorted populations has an arterial versus hematopoietic signature; the response begins `10000001156F18RIK RPL37A RPL37 RPS27 RPL32 RPL39 RPL36 RPS29 RPL41 RPS18` and continues as a gene list. In an audit of all standard simple outputs, 60% of them have at least 80% of their word tokens matching the expression dataset's gene-symbol vocabulary; the median matching fraction is 92.3%. These are output-format diagnostics, not biological correctness measures.

The C2S authors distinguish their released pretrained checkpoints from downstream QA-finetuned models and report that cell-generation training can make the released models produce gene sentences for many prompts.[3] Our result therefore characterizes the released checkpoint under this zero-shot protocol, not the performance of a QA-finetuned C2S model. The binary logit comparison forces a label even when neither word is the preferred natural continuation: it selects *Yes* for 959 of 1,000 questions, with 98.0% Yes recall and 6.2% No recall, yielding 52.1% balanced accuracy.

Sensitivity to expression profiles. Correct profiles yield 76.7% binary accuracy and a mean free-text judge score of 0.65105, compared with 69.5% and 0.61445 for incorrect profiles sampled from the lowest-cosine 5% of eligible profiles. The moderate drop in performance suggests that expression profiles contribute to answering GEO-OmicsQA questions.

A.6.2 Constrained Likelihood Scoring

For closed-set classification tasks, each candidate label is treated as a possible assistant continuation. We encode the prompt once, score each label token sequence under teacher forcing, sum token log probabilities, and normalize by label length. The top-scoring label gives accuracy. To compute macro AUROC, we apply a softmax over the candidate label scores for each example, compute one-vs-rest AUROC for each class with both positive and negative examples, and average over these classes. We report these likelihood-based AUROC values where prior work uses classifier-style metrics, while retaining model-specific continuation formatting.

Ranked-gene annotation baselines. C2S wording follows the released cell-type and tissue templates of Rizvi et al. (2025), with a fixed adaptation for disease subtype. Candidate continuations use a newline, lowercase label, and period for C2S, and bare lowercase labels for Cell-o1. C2S runs use the default Transformers SDPA backend, which omits Gemma-2 attention-logit soft-capping.

[3] Author clarifications: `https://github.com/vandijklab/cell2sentence/issues/15#issuecomment-3397785415` and `https://github.com/vandijklab/cell2sentence/issues/16#issuecomment-3407399888`.

A.6.3 DEEPEVAL JUDGE FOR OPEN-ENDED GENERATION

Open-ended generation tasks, including GEO-OmicsQA simple and contextual tasks, are scored with DeepEval GEval (Liu et al., 2023b) using the Gemini judge model `gemini-3.1-flash-lite` (Google DeepMind, 2026a). Each judged test case is passed to GEval with three fields: the input question, the model's actual output, and the expected reference output. The reported judge score is the mean DeepEval score.

The judge configuration is adapted from the evaluation code as follows:

```
DEEPEVAL_EVALUATION_STEPS = [
    "Read the input question to understand what biological "
    "information is being asked for.",
    "Identify every key biological entity in the expected output: "
    "gene names, cell types or cell lines, tissues, organisms, "
    "diseases, treatments, compounds, and signaling pathways.",
    "Check whether the actual output mentions the same key entities "
    "or their accepted synonyms and aliases. Additional correct "
    "details beyond the expected output should not be penalized.",
    "For directional claims in the expected output, verify that the "
    "actual output states the same direction.",
    "Evaluate whether the biological interpretation and mechanistic "
    "reasoning are consistent with the expected output.",
    "Assess completeness: missing the central finding or conclusion "
    "is a significant deficiency.",
]

DEEPEVAL_RUBRIC = [
    Rubric(score_range=(0, 2),
           expected_outcome="Irrelevant, unanswered, or wrong subject."),
    Rubric(score_range=(3, 4),
           expected_outcome="Correct topic but major factual errors."),
    Rubric(score_range=(5, 6),
           expected_outcome=("Partially correct but incomplete "
                             "or inaccurate.")),
    Rubric(score_range=(7, 8),
           expected_outcome="Main finding and most entities correct."),
    Rubric(score_range=(9, 10),
           expected_outcome=("All key entities, directions, "
                             "and conclusions align.")),
]

geval = GEval(
    name="DeepEval",
    evaluation_steps=DEEPEVAL_EVALUATION_STEPS,
    rubric=DEEPEVAL_RUBRIC,
    evaluation_params=[
        LLMTestCaseParams.INPUT,
        LLMTestCaseParams.ACTUAL_OUTPUT,
        LLMTestCaseParams.EXPECTED_OUTPUT,
    ],
    model=GeminiModel(model="gemini-3.1-flash-lite"),
)
```

A.6.4 PERTURBQA: PROTOCOL AND DETAILED RESULTS

Test population and split scope. We use the released `split=test` rows in the four `<cell>-dir.csv` files at PerturbQA revision `df2fe3a7802f2a46a43876b4f8674734aad3a30a`. Each example is identified by (cell line, perturbed gene, readout gene). Canonicalized identities and gold labels are checked against the official files; duplicate, missing, and extra rows are rejected. There is no new random split. K562 contributes 3,119 rows and 150 readout genes eligible for AUROC; RPE1, 5,365 and 193; HepG2, 3,603 and 166; and Jurkat, 4,121 and 181. We do not evaluate the additional K562-Set dataset or the differential-expression and gene-set-enrichment tasks here.

The data construction policy excludes benchmark (cell line, perturbed gene) pairs from genetic training screens in the same cell line, while allowing the same gene in other cell lines. We apply the same cell-line/perturbation holdout policy to all other training templates derived from the source data used to construct the PerturbQA test set, ensuring that held-out pairs cannot enter training through alternative task formulations. This is a cell/perturbation holdout policy, not a globally unseen-gene split. The present scoring audit verifies test alignment, not the complete training exposure of the evaluated checkpoints. The scoring audit does not establish leakage-free generalization across training records, extended screens, and retrieved pair notes.

Model inputs and inference. Each prompt asks whether knocking down a specified gene increases or decreases the expression of a readout gene in a given cell line, requesting a single-word answer. Inputs also include public notes on the gene pair and the cell line's unperturbed 20,541-dimensional control profile, supplied through the omics projection. We sample three answers per question with temperature 0.6, top-p 0.9, no top-k restriction, an 8,192-token context, and a 4,096-token output cap. Each question/replicate receives a distinct reproducible SHA-256-derived seed; prompt-token hashes agree across runs. All 48,624 outputs were valid and none were truncated.

Three-run answer aggregation. Following SUMMER's methodology (Wu et al., 2025b), we aggregate three sampled answers per question before computing AUROC to enable direct comparison. For each question, OmicsLM maps Increase/Up to 1 and Decrease/Down to 0 and averages the three sampled answer scores. Invalid or truncated answers are excluded from this mean; if all three fail, the fallback is the mean label for that cell/readout gene in official TRAIN, never TEST. This fallback was not needed. AROMA retains its native three-option prompt and maps Up to 1, unchanged (ND) to 0.5, and Down to 0; invalid or truncated answers also receive 0.5. Its score is the arithmetic mean of all three values. We average answers before AUROC, not the three run-level AUROCs, and do not select a best run. Every run is checked against the same complete official test identities and labels.

AUROC. Up is the positive class. Let G_c be the readout genes with both gold classes in cell line c. We compute

$$A_c = \frac{1}{|G_c|} \sum_{g \in G_c} \mathrm{AUROC}(\{(y_{cpg}, s_{cpg}) : p \text{ in TEST}\}), \qquad A = \frac{1}{4} \sum_c A_c.$$

Groups with only one gold class are omitted from AUROC, but not from F1. The implementation checks agreement with the pinned official evaluator. OmicsLM parsing accepts only completed bare labels or final answers following closed thinking blocks. With three valid binary answers, the averaged score takes values 0, 1/3, 2/3, or 1, providing a ranking beyond a single binary answer. Released baseline scores are preserved rather than converted to hard labels for AUROC.

Direction F1. We pool all direction examples within each cell and calculate $F_{1,c,k} = 2\mathrm{TP}_{c,k}/(2\mathrm{TP}_{c,k} + \mathrm{FP}_{c,k} + \mathrm{FN}_{c,k})$ for $k \in \{\mathrm{Up}, \mathrm{Down}\}$. Class F1 is averaged equally across the four cells; macro-F1 is the mean of the two class averages. For OmicsLM and AROMA, the averaged score predicts Up above 0.5 and Down below 0.5; an exact tie predicts neither direction and remains a false negative for its gold class. No test question is removed because of a failed prediction. This population contains no gold unchanged examples, so it cannot measure false positives on unchanged genes or substitute for three-class F1.

Baseline provenance and decision rules. SUMMER, retrieval without LLM, LLM without retrieval, LLM without chain-of-thought, GEARS, GenePT, GAT, and MLP are rescored from the authors' `results.zip` release (Wu et al., 2025b;c); they are not retrained locally. The archive SHA-256 is `d6d3f3861043e5e0f02418c9213ca20f4167dc82bd96ff2ed5af23970ff59df0`. SUMMER's released scores already aggregate sampled answers and are not averaged again. SUMMER and the other scalar-score baselines retain their released `dir_pred` for AUROC. Their diagnostic F1 predicts Up above 0.5 and Down below 0.5; exact ties predict neither. This does not establish the authors' intended hard-decision calibration, especially for GEARS, or recover SUMMER's individual abstentions. For GAT and MLP, AUROC uses the Up logit minus the Down logit, a diagnostic conversion whose equivalence to the original paper is unverified. Their F1 uses native three-class argmax (ND, Down, Up); ND predicts neither direction. GenePT metrics are computed separately for each of three released replicates, then averaged within cells before averaging

Table 7: **Complete PerturbQA direction comparison.** All values are on a 0–1 scale and averaged equally over cells. †: diagnostic threshold-based F1; ‡: diagnostic logit-to-AUROC conversion. These F1 scores are newly computed on the direction subset, not published three-class scores. SynthPert (paper) is a published reference, not a matched reproduction; — denotes unavailable F1.

Model	AUROC	F1 Up	F1 Down	Macro-F1
OmicsLM 9B	0.7635	0.8772	0.8749	0.8761
SUMMER	0.6547	0.7977†	0.7908†	0.7942†
Retrieval without LLM	0.6431	0.7970†	0.7779†	0.7875†
GenePT protein	0.5746	0.7590†	0.7304†	0.7447†
GenePT gene	0.5681	0.7288†	0.6968†	0.7128†
GAT	0.6043‡	0.6963	0.6778	0.6871
MLP	0.4976‡	0.6451	0.6348	0.6400
LLM without retrieval	0.5224	0.5401†	0.6061†	0.5731†
GEARS	0.5775	0.3411†	0.5885†	0.4648†
LLM without CoT	0.4970	0.0994†	0.6354†	0.3674†
AROMA	0.5788	0.7064	0.6803	0.6934
SynthPert (paper) (Phillips et al., 2025)	0.6875	—	—	—

Table 8: **Per-cell PerturbQA results.** Each entry is AUROC / direction macro-F1. Diagnostic conversions and threshold caveats are as in Table 7. GenePT entries average replicate metrics. SynthPert values are published AUROCs with unavailable direction F1.

Model	K562	RPE1	HepG2	Jurkat
OmicsLM 9B	0.7939 / 0.8482	0.7173 / 0.8907	0.7679 / 0.8891	0.7750 / 0.8763
SUMMER	0.6379 / 0.7143	0.6426 / 0.8422	0.6634 / 0.8311	0.6748 / 0.7894
Retrieval without LLM	0.6913 / 0.7164	0.6331 / 0.8478	0.6380 / 0.8239	0.6098 / 0.7618
GenePT protein	0.5583 / 0.6471	0.5857 / 0.8224	0.5558 / 0.7846	0.5986 / 0.7246
GenePT gene	0.4891 / 0.5871	0.5891 / 0.7994	0.6112 / 0.7751	0.5828 / 0.6896
GAT	0.5949 / 0.5381	0.6054 / 0.7885	0.6429 / 0.7613	0.5742 / 0.6603
MLP	0.5223 / 0.4828	0.5136 / 0.7473	0.4953 / 0.7127	0.4594 / 0.6171
LLM without retrieval	0.4984 / 0.4916	0.5223 / 0.6232	0.5311 / 0.5604	0.5376 / 0.6172
GEARS	0.6660 / 0.3433	0.6064 / 0.5941	0.5195 / 0.4553	0.5181 / 0.4665
LLM without CoT	0.5019 / 0.2063	0.4919 / 0.3990	0.4898 / 0.4065	0.5044 / 0.4578
AROMA	0.5645 / 0.6069	0.5683 / 0.7429	0.5987 / 0.7256	0.5838 / 0.6981
SynthPert (paper) (Phillips et al., 2025)	0.65 / —	0.73 / —	0.72 / —	0.65 / —

cells. No replicate is selected by TEST performance. Matched direction predictions for scGPT and STATE are unavailable, so their published three-class scores are not imported into these tables.

SynthPert: published reference only. SynthPert's Table 2 reports direction AUROCs of 0.65, 0.73, 0.72, and 0.65 for K562, RPE1, HepG2, and Jurkat, respectively (Phillips et al., 2025); their arithmetic mean is 0.6875. We cite the specific arXiv v1 table from which these rounded values were taken. Its trained checkpoint is not publicly available, so we report the published results rather than evaluating the model on our matched test set. The tables label them as published references, separate from locally scored results; matched AUROC and direction F1 are unavailable. Its three-class formulation predicts Up, Down, and not differentially expressed jointly, unlike the conditional direction task used here.

AROMA reproduction. We generated AROMA predictions from the authors' released multimodal checkpoint and augmented three-class prompts (Wang et al., 2026), preserving the original chat template and protein and graph feature placeholders. The current three-run evaluation generates predictions only for the 16,208 official direction TEST questions; it does not rerun the full three-class test. Feature lookup preserves case-sensitive gene names; canonicalized identifiers are used separately for test matching. Checkpoint tensors, active adapters, feature mapping, prompt formatting, multimodal injection, and output extraction were audited.

Table 9: **AROMA reproduction on the full three-class test.** Class F1 entries show local reproduction / published value from Table 1 of Wang et al. (2026); published values are rounded to two decimals. Macro-F1 is the local three-class average. These scores are separate from the direction-only comparison.

Cell line	Rows	F1 ND	F1 Down	F1 Up	Macro-F1
K562	23212	0.9594 / 0.96	0.3649 / 0.36	0.6518 / 0.66	0.6587
RPE1	37942	0.9659 / 0.97	0.7257 / 0.73	0.6051 / 0.60	0.7656
HepG2	25749	0.9650 / 0.97	0.6517 / 0.65	0.6431 / 0.65	0.7533
Jurkat	29138	0.9683 / 0.97	0.6948 / 0.69	0.5640 / 0.58	0.7424
Mean macro-F1 (local / mean of rounded paper values)					0.7300 / 0.7325

The answer mapping and aggregation follow the three-run protocol above. AROMA obtains 0.5788 mean direction AUROC and 0.6934 direction macro-F1 (Tables 7 and 8). Inference retains the released sampling settings: temperature 0.7, top-k 20, top-p 0.95, and 512 new tokens, with batch size 32 per GPU and two GPU workers per run. Each replicate/worker uses a distinct SHA-256-derived PyTorch seed and an advancing random stream. These decoding settings differ from OmicsLM's; only the test population and score aggregation are matched. All 48,624 predictions are retained, including 108 invalid or truncated answers scored as 0.5. Runtime package versions differ from the authors' requirements, so this is an independent reproduction, not a bitwise replication.

Separately, an earlier single-run reproduction covered all 116,041 three-class TEST rows with seed 0, including 76 truncated answers and one malformed answer, all retained in scoring. Class F1 is computed within each cell line and macro-F1 averages the three classes and four cell lines equally. Its macro-F1 is 0.7300, compared with 0.7325 obtained by averaging the paper's rounded cell/class values (Table 9). The baseline scoring script rescores full-test predictions when they are supplied; fresh full-test inference remains opt-in. These scores include gold unchanged examples and are distinct from the current three-run direction comparison.

Reproducibility artifacts. The bundle's `perturbqa/README.md` documents inference and CPU rescoring; the scripts save raw answers, run IDs, seed metadata, averaged per-question scores, per-run and per-cell metrics, and asset hashes. CPU rescoring reproduces both models' reported scores, and AUROC agrees with the pinned official evaluator to tolerance 10^{-12}.

A.7 Training Details

We provide optimization, sampling, and supervision details for the three training stages introduced in Section 3.2.

A.7.1 Stage 1: Multimodal Pretraining

This stage jointly performs continued pretraining (CPT) on free-text records and supervised fine-tuning (SFT) on instruction examples, interleaving both throughout training.

Initialization and optimization. We initialize the language backbone from the text-model weights of the official pretrained Qwen3.5-9B checkpoint and augment it with the omics projection and gene vocabulary described in Section 3.1. The language backbone, token embeddings, and omics projection are updated jointly; the omics encoders remain frozen and their outputs are precomputed. We train for 40,700 optimizer updates on 48 NVIDIA H100 80GB GPUs. Each GPU processes one microbatch of up to 10,240 input tokens per update, with gradient accumulation set to one, giving a nominal global batch size of 491,520 input tokens. The measured total is 19.999 billion input tokens. Token counts include masked context and omics input positions and exclude padding. We use 8-bit AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, weight decay 0.1, and gradient-norm clipping at 1.0. The learning rate increases linearly to 2×10^{-5} over 200 updates, then follows a cosine schedule with a minimum of 2×10^{-6}. We use BF16 model weights and gradients, gradient checkpointing, and distributed data parallelism. The recorded training runtime was approximately 43.8 hours, corresponding to 2,103 H100 GPU-hours across 48 GPUs.

Source sampling. We assign a sampling weight to each data subset using source-size estimates and manual adjustments to balance the number of examples, token volume, and coverage of different task types. These weights define target shares of input tokens, including prompt tokens masked from the loss. Table 10 reports the observed input-token share for the completed run using the categories defined in Section 4.

Table 10: Observed Stage 1 input-token shares by training-data category. Counts include masked context and repeated exposure, and exclude padding. Human genetics was not included in this stage. Values are rounded.

Training-data category	Input-token share (%)
Cellular characterization	12.96
Expression interpretation and reasoning	18.93
Perturbation response	7.94
Gene essentiality and functional screens	2.92
Drug response and mechanisms	1.06
Biological knowledge and networks	3.25
Biomedical QA and information extraction	5.45
Scientific and biomedical text and code	40.62
General-purpose capabilities	6.88
Total	100.00

Sequence construction and loss. The maximum sequence length is 10,240 tokens. Longer free-text records are divided into consecutive windows with boundary overlap for next-token prediction. For an overlength conversation, we sample a window containing supervised tokens. Window boundaries preserve complete omics spans. Windows are packed into token-budgeted batches while keeping sequences independent. The first token of each window, omics placeholders, and their delimiters are excluded from the loss. Free-text records otherwise receive next-token supervision throughout, while conversational records receive supervision on assistant responses. The objective is averaged over supervised tokens across devices and accumulation steps.

A.7.2 Stage 2: Supervised Instruction Tuning

Initialization and optimization. We initialize Stage 2 from a Stage 1 checkpoint. The language backbone, token embeddings (including the added gene tokens), and omics projection continue to be updated jointly. The omics encoders remain frozen, and their precomputed outputs are supplied through the same projection as in Stage 1. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, and gradient-norm clipping at 1.0. The learning rate increases linearly to 1.4×10^{-5} over 169 updates (5% of the training budget, rounded down), then follows a cosine schedule to 1.4×10^{-6}. Training uses BF16 computation with FP32 master weights and gradient reduction, gradient checkpointing, and hybrid sharded data parallelism: parameters are sharded across the eight GPUs within each node and replicated across six nodes.

Batching and training budget. We train for two epochs on 48 NVIDIA H100 80GB GPUs, with a global batch of 3,072 conversations per optimizer update (64 per GPU). Each local batch is divided into dynamic microbatches with a budget of 32,768 input tokens per GPU; gradients accumulate over these microbatches before the optimizer update. The number of microbatches depends on sequence lengths. We use padding-free packing with independent attention and recurrent state for each conversation, and a maximum sequence length of 32,768 tokens. All training conversations fit within this limit, so no truncation is required. Dropping the final incomplete batch gives 1,696 updates per epoch and 3,392 updates in total, with 2.788 billion input tokens processed across the two epochs. Token counts include masked context and omics input positions and exclude padding. The training job ran for approximately 12.8 hours, corresponding to 615 H100 GPU-hours, including initialization, validation, and checkpoint writing. The model after the second epoch initializes reinforcement learning.

Training mixture. The corpus contains 5,210,786 training conversations drawn from 104 data components, combining expression-conditioned biological tasks with text-only biological knowl-

edge and general-purpose instructions. Of these conversations, 3,489,641 contain omics inputs and 864,886 contain nonempty reasoning traces; these subsets overlap. The biological examples cover cellular characterization, expression interpretation, perturbation responses, gene essentiality, functional screens, drug sensitivity and mechanisms, human genetics, and biological knowledge and reasoning. Mathematics, coding, general knowledge, instruction following, and tool-use conversations complement the biological examples. The mixture includes direct answers, reasoning followed by a final answer, and multi-turn interactions. Counts refer to conversation records before repeated exposure across epochs; multiple tasks or instruction variants can share the same underlying biological sample. Table 11 accounts for the complete Stage 2 training corpus, grouped by source and task family. Table 17 presents the same corpus as 157 groups by biological objective and input/output format, with group counts summing to 5,210,786 conversations.

Conversation formatting and supervision. We serialize conversations using a training adaptation of the backbone's chat template that preserves reasoning across assistant turns, message order, and omics references. We minimize next-token negative log-likelihood only on assistant turns selected for supervision. System, user, and tool messages, together with explicitly excluded assistant turns, remain available as context but do not contribute to the loss. The corpus contains 3,437 such excluded assistant turns in agent trajectories. On supervised turns, supplied reasoning traces and final answers are both training targets; synthetic empty reasoning prefixes inserted by the formatter are masked. This supports both direct-answer and reasoning examples without making an empty rationale a supervised target.

Table 11: Complete Stage 2 training inventory. Counts cover all 104 components exactly once, before epoch repetition; they are conversation records, not unique biological samples. Marker-grounded examples are included with their source datasets.

Source / task group	Conversations	Tasks and supervision
HCA	1,210,891	Cell type, tissue, phase, lineage, biological role, marker genes, immune status, and sex; direct answers and marker-grounded reasoning.
3CA	303,709	Cell type and lineage, cancer context, immune and malignant status; direct answers and marker-grounded reasoning.
TCGA	172,478	Tissue, cancer type, tumor/normal status, stage, age, sex, and clinical attributes; tissue and sex reasoning.
GTEx	89,999	Tissue and sex prediction, including age/sex-conditioned tissue questions and marker-grounded reasoning.
CancerModels.org	25,261	Model type, cancer system, histology, primary site, metastatic status, and paired-model questions.
DepMap tissue annotation	1,461	Tissue of origin from a cell-line profile with supplied cancer type.
GEO	333,277	Cell line/type, tissue, disease and treatment annotations, perturbation target/type, sample descriptions, description-to-gene prediction, and cell-type/tissue reasoning.
GEO-OmicsQA training data	181,207	Binary and open-ended questions grounded in expression profiles; direct answers and reasoning traces.
Cell-line similarity	85,027	Similar-cell-line retrieval and comparison of paired expression profiles.
CellWhisperer, cell-line QA, expression and pathway tasks	58,048	Profile-linked conversations; cell-line descriptions and facts; top expressed genes in single-cell and pseudobulk profiles; gene-to-pathway mapping.
Pathway activity and composition	4,053	Up/downregulated pathway enrichment from expression and pathway-to-gene queries.
Genetic perturbation screens	97,015	Knockout identification and prediction of top expressed, differentially expressed, upregulated, and downregulated genes.

Continued on next page.

Table 11 continued.

Source / task group	Conversations	Tasks and supervision
Tahoe and sci-Plex chemical screens	109,305	Expressed-gene extraction, post-treatment gene-list prediction, and compound/target identification from control and treated profiles.
PerturbQA	355,000	Direction of change (300,000), differential-expression detection (50,000), and instruction-format variants (5,000).
DepMap essentiality	306,476	Essentiality label and dependency-value prediction (300,000), plus instruction-format variants (6,476).
AssayBench	14,356	Screen-hit gene ranking (1,036) and ranking of supplied candidate genes (13,320).
GDSC sensitivity	59,212	Compound-specific sensitivity labels (48,248), instruction-format variants (5,000), and absolute-IC50-threshold classification (5,964).
PRISM primary and secondary screens	94,307	Cell-specific sensitive/resistant labels from primary (69,655) and secondary (24,652) screens.
Drug-response mechanism examples	1,671	Drug targets/actions (135), context-augmented sensitivity answers (512), and constructed/source reasoning variants (512 each).
DrugMechDB	18,283	Mechanism paths, edge membership and completion, missing nodes, and effects of disrupting mechanism steps.
Open Targets Genetics	300,000	Eleven tasks: gene–disease evidence and disease lists, evidence sources and strength, inheritance, burden direction, clinical significance, gene constraint, GWAS gene prioritization, variant consequence, and frequency class.
KEGG, STRING, and gene sets	587,823	Graph edges, reachability, paths and knockout-path completion, common ancestors, pathway naming/membership, gene-set composition, and enrichment.
Cell-line genomic facts	21,297	Copy-number alterations, cancer mutations, oncogenes, tumor suppressors, and reverse/grouped cell-line queries.
Protein knowledge and reasoning	59,055	Protein properties, functions, and locations (44,055); BioReason-Pro (Fallahpour et al., 2026) protein-function reasoning (15,000).
Cell and disease ontologies	126,858	Definitions, synonyms, is-a relations, parents, common ancestors, and specificity ordering.
BioReason KEGG	1,159	Reasoning about biological pathways and mechanisms from supplied pathway context (Fallahpour et al., 2025).
BioASQ and CAMEL Biology	23,773	Evidence-based biomedical QA (3,773) and biology questions and explanations (20,000).
Medical reasoning	13,109	ReasonMed, MedReason, M23K, and Medical-R1 question–answer reasoning examples.
Omics dialogue and instruction variants	31,659	Biological follow-up conversations, new-evidence updates, and answer revisions (18,222); omics output-format and reformatting instructions (13,437).
Biology formatting constraints	19,503	CAMEL-derived biology answers with casing, prefix/suffix, list, JSON, and other output constraints.
Mathematics	97,717	GSM8K, MetaMath, and OpenMathReasoning solutions with reasoning and final answers.
Coding	89,999	Magicoder-Evol-Instruct, OpenCodeReasoning, and SmolTalk Self-OSS-Instruct code-generation and reasoning examples.
General knowledge	103,100	ARC Easy/Challenge and MMLU auxiliary-training multiple-choice questions.
General instructions and conversation	166,158	Open PerfectBlend AutoIF/LMSYS/UltraChat and SmolTalk constraints, rewriting, Magpie, and system-conditioned multi-turn dialogue.

Continued on next page.

Table 11 continued.

Source / task group	Conversations	Tasks and supervision
Tools and agents	48,540	Hermes, xLAM irrelevance, ToolACE, When2Call, and UltraData-SFT-Agent-2609 tool selection, calls, responses, and agent trajectories.
Total	**5,210,786**	**104 components.**

A.7.3 Stage 3: Reinforcement Learning

Policy initialization. We initialize GRPO from the multimodal Qwen3.5-9B model after two epochs of supervised instruction tuning on a curated corpus of 5,210,786 conversations. SFT starts from an intermediate model obtained during multimodal continued pretraining. The resulting SFT policy provides both the initial actor and the frozen reference policy for RL. GRPO follows SFT directly, without an intervening preference-learning stage. The language backbone, gene-token embeddings, and omics projection remain trainable, while the pretrained omics encoders remain frozen.

Reasoning initialization and response format. We use the preceding CPT and SFT stages as the RL warm start, without an additional RL-specific cold-start SFT run. Multimodal CPT develops the connection between expression profiles, gene tokens, and language. SFT then supplies both direct answers and supervised reasoning: the SFT corpus contains 864,886 conversations with nonempty reasoning traces, alongside instruction, mathematical, biological, and tool-use examples. The SFT formatter preserves reasoning on supervised assistant turns, applies loss only to the selected assistant tokens, and masks synthetic empty reasoning prefixes. System, user, tool, and excluded assistant turns remain context. Consequently, direct-answer examples are not converted into targets that explicitly teach an empty rationale.

RL uses the native thinking-enabled chat template. The prompt supplies the opening `<think>` tag, and a valid continuation supplies exactly one `</think>` followed by a nonempty final answer. The reward adapter restores the opening tag for parsing when it is absent from the generated continuation. Missing answer boundaries, additional chat-role delimiters, or deprecated `<answer>` wrappers receive zero reward. An empty reasoning block is structurally valid but earns no evidence-grounding credit. The verifier scores the final answer separately from the rationale; there is no reward for merely producing a longer chain of thought. These are the RL training settings, not a claim that every benchmark is decoded with thinking enabled.

Exploratory comparison of post-training strategies. Before training the 9B model, we compared SFT, direct GRPO, Kahneman–Tversky Optimization (KTO; Ethayarajh et al., 2024), Noise Contrastive Alignment (NCA; Chen et al., 2024), and preference learning followed by GRPO using Qwen3 1.7B backbones (Table 12). KTO and NCA are alternative alignment methods applied after SFT. We evaluate biological performance on the annotation and GEO-OmicsQA tasks introduced in Sections 5.1 and 5.2. MMLU multiple-choice knowledge questions and a GSM8K mathematical word-problem probe serve as diagnostics of general-capability changes relative to these SFT policies.

Direct GRPO from SFT epoch 2 yields higher observed scores on all eight reported metrics relative to that initializer. GEO-OmicsQA binary accuracy rises from 74.0 to 74.4 and the free-text score from 55.5 to 57.4, while the four annotation AUROCs increase by 0.1–0.4 percentage points. MMLU rises from 34.29 to 54.37 and GSM8K from 72.50 to 75.75. Direct GRPO achieves the highest GEO-OmicsQA scores among the compared workflows, although some preference-learning alternatives achieve higher annotation or general-capability scores. No workflow leads on every metric. For the final 9B model, we use multimodal pretraining followed by two SFT epochs and direct GRPO, without a KTO or NCA stage.

These are descriptive comparisons from single training runs with separately filtered prompt sets and unequal budgets. Direct GRPO starts from SFT epoch 2, whereas KTO and NCA start from SFT epoch 1. The 1.7B models also have a different pretraining lineage from the final 9B model. Consequently, this table does not isolate the effect of the alignment method or establish general-capability retention for the final 9B model or relative to the original pretrained backbone.

Table 12: Exploratory Qwen3 1.7B comparison of biological performance and general capabilities. All values are on a 0–100 scale; higher is better. Tabula Sapiens and Human Diseases report macro AUROC; GEO-OmicsQA reports aggregate binary accuracy and free-text judge score. MMLU and the GSM8K probe use thinking-mode decoding and track general-capability changes relative to SFT. KTO and NCA start from SFT epoch 1. The KTO→GRPO result is an intermediate 100-update checkpoint. Each workflow is a single training run.

	Tabula Sapiens		Human Diseases		GEO-OmicsQA		General capabilities	
Qwen3 1.7B workflow	Cell type top-20	Tissue	Disease subtype	Tissue	Binary QA	Free text	MMLU	GSM8K
SFT epoch 1	93.2	82.1	83.5	96.1	73.1	54.6	34.96	71.25
SFT epoch 2	94.4	81.9	83.2	96.7	74.0	55.5	34.29	72.50
SFT epoch 2 → GRPO	94.8	82.3	83.6	96.8	74.4	57.4	54.37	75.75
KTO	94.2	82.3	82.9	96.6	71.6	55.7	47.78	73.50
NCA	94.9	82.9	83.5	96.9	73.5	56.7	45.69	74.50
NCA → GRPO	95.1	83.0	83.7	96.8	73.0	56.5	57.60	76.25
KTO → GRPO (100)	94.5	82.5	83.4	96.6	72.7	56.1	37.54	74.25

RL sources and task coverage. Table 13 summarizes the 19 reward-task families and their prompt counts. Biological classification and marker-reasoning tasks draw on the Human Cell Atlas (HCA), Curated Cancer Cell Atlas (3CA), GTEx, TCGA, GEO, and CancerModels.org datasets. Gene-list tasks use expression-derived and curated gene sets. Gene-essentiality tasks use DepMap, while drug-sensitivity tasks use GDSC and the PRISM primary and secondary activity screens. Differential-expression and direction tasks use PerturbQA (Wu et al., 2025b), and the four mechanism-of-action tasks use DrugMechDB (Gonzalez-Cavazos et al., 2023). Graph-path questions use a biological graph assembled from STRING resources.

To retain general capabilities, the candidate pool also includes GSM8K training questions (Cobbe et al., 2021), multiple-choice questions from MMLU auxiliary training and ARC training splits (Hendrycks et al., 2021; Clark et al., 2018), and CAMEL biology questions augmented with programmatically checkable instructions (Li et al., 2023). UltraData-RL-2609 contributes only its Math and Knowledge configurations (MiniCPM Team, 2026b); its Code and Long-Context configurations are not used. Agent/tool-use SFT data, Open Targets Genetics, AssayBench, and the UltraInteract preference data should not be conflated with this RL mixture. RL records contain prompts, optional omics inputs, reference answers, and verifier metadata; optimization uses newly sampled responses rather than treating source rationales or filtering completions as supervised targets.

The biological classification family includes GEO binary QA, GEO treatment, cancer-model labels, GTEx tissue, and HCA cell-type, tissue, phase, and related labels. Gene-list tasks include expression-associated genes, enrichment, and perturbation-induced top or downregulated genes. The marker-reasoning collection replaces earlier marker records with 200 prompts from each of 19 training tasks: eight HCA tasks (cell type; cell type with tissue; cell type with phase; population cell type; cell phase; lineage; immune status; tissue), five 3CA tasks (cell type; cell type with cancer; population cell type; immune status; lineage), and two each from GEO (cell type, tissue), GTEx (tissue, sex), and TCGA (tissue, sex).

Data audit and exclusion funnel. We begin with an initial pool of 55,519 RL training prompts. We remove 3,746 superseded marker prompts, 2,841 Gene Ontology prompts whose requested leaf terms were inconsistent with ancestor-expanded targets, and 468 CTRP prompts reserved for evaluation. This leaves 48,464 prompts. A further 6,014 are excluded by overlap, duplicate, or invalid semantic-identity checks, leaving 42,450. Adding the 3,800 accepted marker-reasoning prompts yields the audited pool of 46,250. To obtain these marker prompts, we inspect 4,336 candidates and exclude 530 with more than eight profiles and six with protected held-out profile identities.

The audit checks canonicalized questions, example and profile identities, and task-specific entities: essentiality cell identities, sensitivity cell–drug pairs and cell identities, perturbation cell–knockout pairs, and DrugMechDB mechanism identifiers. Unparseable split identities are quarantined. It also verifies reference-answer compatibility with the task verifier, except for instruction constraints whose targets are executable specifications. All 19,381 distinct referenced expression vectors are loaded and checked for finite values and the expected 20,541-dimensional shape. Prompts allow at most eight profiles and 1,024 tokens; the longest audited prompt has 951 tokens.

Table 13: Audited 9B RL candidate pool and training prompts selected using the SFT initializer. Counts refer to prompts, not the eight sampled completions per prompt.

Task / source	Candidates	Selected
GDSC drug sensitivity	4,779	2,853
DepMap gene essentiality	4,227	565
PerturbQA: differential expression	2,510	334
PerturbQA: direction	2,503	771
PRISM primary activity	1,891	1,202
PRISM secondary activity	1,904	1,288
Biological label classification	3,479	998
Marker reasoning	3,800	60
Gene-list / gene-set prediction	3,001	14
Biological graph paths	5,679	557
DrugMechDB: closed edge	607	245
DrugMechDB: open edge	245	155
DrugMechDB: effect	431	287
DrugMechDB: masked node	620	163
GSM8K mathematics	2,815	917
MMLU / ARC multiple choice	1,892	402
Checkable instruction following	2,868	1,314
UltraData-RL-2609: Math	1,999	26
UltraData-RL-2609: Knowledge	1,000	20
Total	46,250	12,171

The separate RL monitoring split decreases from 2,964 to 2,591 examples after removing 205 superseded marker, 136 inconsistent GO, and 32 CTRP records. Identity checks find no overlap between the resulting RL training and monitoring splits. This is not a claim of universal benchmark decontamination: protection is defined by the registered keys, the graph is a fixed knowledge resource, some RL prompts appeared during SFT, and the monitoring split is not guaranteed to be unseen after SFT. We therefore distinguish monitoring rewards from final benchmark evaluation.

Model-based difficulty filtering. Using the SFT initializer, we sample $G = 8$ responses per prompt with temperature 1.2 and a 2,048-token response limit. Let $c_i \in \{0, 1\}$ denote task-verifier correctness after structural and rejection gates. We retain a prompt if

$$1 \leq \sum_{i=1}^{8} c_i \leq 6. \tag{1}$$

The lower bound retains one-success groups, which are difficult but demonstrably solvable within the sampling budget. Zero-success groups provide no observed correct trajectory, while seven- and eight-success groups are removed to concentrate training on less-saturated problems. Selection uses binary correctness, not the shaped scalar reward; otherwise variation in formatting or rationale length could make an outcome-saturated prompt appear useful. The selected prompt set is frozen before RL, with fresh responses sampled during every optimization step.

Difficulty filtering covers 35,609 of the 46,250 candidate prompts (77.0%; Table 14). Only complete groups of eight sampled responses are considered for selection; incomplete groups and unsampled prompts are excluded. A response that reaches the generation limit still belongs to its completed sampling group, but receives zero if it never produces the required final-answer boundary. The selected task mixture therefore reflects both sampling coverage and verifier correctness. We uniformly shuffle the selected prompts without subsequent task rebalancing.

Checking use of omics inputs. For PRISM primary and secondary activity, we additionally compare correct-profile accuracy with two controls: a matched different-cell profile and no expression profile, retaining the same question, drug, and target. Each task is probed on up to 512 deterministically selected questions. Admission requires at least 128 complete paired probes and an accuracy advantage of at least 0.02 over each control, with a positive lower endpoint of a 95% paired-bootstrap interval (10,000 resamples). Both PRISM tasks pass, so this gate removes no additional selected

Table 14: Prompt selection by model-based difficulty filtering. The three correctness bands partition evaluated prompts; unsampled prompts are reported separately.

Filtering stage	Prompts
Audited candidates	46,250
Evaluated with eight responses	35,609
Excluded: zero correct	11,126
Excluded: seven or eight correct	12,312
Retained: one to six correct	12,171
Unsampled at cutoff	10,641

prompts. The primary-task advantage is 0.044 (interval 0.019–0.070) over the different-cell control and 0.168 (0.143–0.195) over the no-profile control, with 389 pairs; the secondary-task advantages are 0.029 (0.006–0.053) and 0.193 (0.169–0.218), with 408 pairs. These are input-dependence diagnostics for these tasks, not evidence of causal biological reasoning across the entire mixture.

Reward composition. Rewards are computed by deterministic task-specific functions; no learned preference reward model or language-model judge is used. For the shared reward function used by the biological tasks, GSM8K, multiple-choice questions, and instruction constraints, let $O \in [0, 1]$ be the outcome score, $P \in [0, 1]$ the evidence-grounding score, $F \in \{0, 1\}$ the valid-format indicator, and $L \in [0, 1]$ the reasoning-length penalty. After terminal rejection gates, the reward is

$$\widetilde{R} = \mathrm{clip}_{[0,1]}\big(w_O O + w_P P + 0.05F - 0.05L\big), \tag{2}$$

$$R = \begin{cases} \widetilde{R}, & c = 1, \\ \min(\widetilde{R}, 0.25), & c = 0. \end{cases} \tag{3}$$

The default weights are $(w_O, w_P) = (0.75, 0.20)$. Marker-label examples without evidence metadata use $(0.95, 0)$. For the remaining tasks, an empty evidence ledger gives $P = 0$ while retaining the default weights, so the reward before the length penalty is bounded above by 0.80. These task-dependent reward scales are preserved when computing the unnormalized group advantages below.

For a rationale with n whitespace-delimited words and task budget B, $L = \min(\max((n - B)/B, 0), 1)$. Budgets are 400 words for graph paths; 300 for gene sets and GSM8K; 200 for GDSC, essentiality, PerturbQA, PRISM, and instruction constraints; 180 for DrugMechDB; 160 for biological labels and markers; and 120 for general multiple choice. A long rationale can lose at most 0.05 reward. When both outcome and evidence scores are zero, a well-formed answer earns at most the 0.05 format bonus; incorrect answers with partial credit remain capped at 0.25.

The evidence term checks citations against a per-example ledger. Each claim clause must cite known evidence identifiers, and a lexical checker must find agreement with the corresponding stored values; unknown citations, unsupported claims, or uncited claim clauses set $P = 0$. Otherwise, $P = \min(k/2, 1)$, where k is the number of distinct supported facts. Nonempty ledgers occur in GDSC, essentiality, and PerturbQA records. This is a limited rule-based grounding check, not a semantic proof that a reasoning trace is faithful. The shared reward function rejects rationales containing a fixed list of leakage-like strings: `ic50`, `pic50`, `zscore`, `z-score`, `auc_published`, "ground truth," "ground-truth," "the correct answer," "the label is," "prevalence," "is essential in," and "% of cell lines." This conservative lexical gate can also reject legitimate wording; it is not a general leakage detector.

Task-specific outcome verifiers. The outcome score and binary correctness flag are distinct for partially rewarded tasks:

- **Drug sensitivity.** GDSC compares the final JSON label with the reference. A correct label receives $O = 0.5 + 0.5 \tanh(|m|)$; an incorrect label receives zero. For reference potency v in μM, $m = \log_{10}(2/v)$ for sensitive cases and $m = \log_{10}(v/100)$ for resistant cases. This margin is available only to the reward function. PRISM instead requires exactly one lowercase JSON field, `{"label":"sensitive"}` or `{"label":"resistant"}`, and uses binary exact correctness, without additional keys or label synonyms.

- **Essentiality and perturbation.** For essentiality, $O = 0.6 \cdot \mathbf{1}[\hat{y} = y] + 0.4[1 - (z - d)^2]$, with $z \in \{0, 1\}$ obtained from the predicted label and $d \in [0, 1]$ the stored continuous dependency target. Invalid labels receive zero. PerturbQA differential-expression and direction tasks use exact final JSON-label agreement. Incorrect predictions remain subject to the ceiling in Eq. 3.
- **Biological and marker labels.** Normalized exact correctness is weighted by $\min(1/(Kp_y), 4)/4$, where p_y is the recorded class prior and K is the number of classes for the source task. This limits inverse-frequency weighting. Marker answers may include one leading `Answer:` prefix. The correctness flag is the unweighted exact-match result.
- **Gene lists and graph paths.** Gene lists receive $O = 0.92\,\mathrm{IoU} + 0.08u$, where u is the fraction of distinct entries; correctness requires set intersection-over-union of at least 0.5. Graph paths are checked against known edges and relations, knockout constraints, endpoints, and shortest-path length; a valid shortest path or a correct unreachable declaration receives one. Otherwise, partial credit is $\min(0.70v + 0.15e + 0.15s, 0.85)$, where v is the valid-prefix fraction, e indicates that this prefix reaches the target, and s indicates the correct starting node. The incorrect-answer ceiling still applies.
- **Mechanism-of-action tasks.** DrugMechDB closed-edge, open-edge, effect, and masked-node responses use exact entity or label matching after Unicode normalization, case folding, whitespace normalization, and limited surrounding-punctuation removal.
- **General tasks.** MMLU/ARC answers use normalized exact option matching. GSM8K compares the last number in the final answer with the reference, stripping digit-grouping commas and allowing absolute error below 10^{-6}. Instruction-following rewards check one of 11 executable constraints: lowercase, uppercase, no comma, required prefix, required suffix, bullets, sentence count, JSON keys, forbidden letter, keyword occurrence, or maximum word count. These checks measure the specified constraint, not the biological correctness of the underlying CAMEL answer.
- **UltraData.** These two task families use a separate binary reward $R = \mathbf{1}[\hat{y} = y]$ after the common native-answer-boundary check. Math uses a final boxed expression or a strict final answer with symbolic verification through `math_verify`, with one-second parsing and verification timeouts. Knowledge uses normalized exact matching against the gold answer and recorded aliases. These tasks receive no format bonus, evidence-grounding term, reasoning-length penalty, or lexical leakage gate.

GRPO objective and optimization. We use critic-free GRPO (Shao et al., 2024) in VERL (Sheng et al., 2025). At each update, the rollout policy generates eight responses for each of 32 prompts. The sequence-level reward is attached to the final response token, and the group-centered advantage is broadcast across generated tokens:

$$A_i = R_i - \frac{1}{G}\sum_{j=1}^{G} R_j, \qquad G = 8. \tag{4}$$

Unlike the standard deviation-normalized variant of GRPO, our configuration does not divide by the within-group reward standard deviation. For token probability ratio $\rho_{i,t} = \pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\mathrm{old}}(y_{i,t} \mid x, y_{i,<t})$, define

$$s(\rho, A) = \min\{\rho A, \mathrm{clip}(\rho, 0.8, 1.2)A\}, \tag{5}$$

$$s_{\mathrm{dc}}(\rho, A) = \begin{cases} \max\{s(\rho, A), 3A\}, & A < 0, \\ s(\rho, A), & A \geq 0. \end{cases} \tag{6}$$

We minimize the token-averaged loss

$$\mathcal{L} = -\langle s_{\mathrm{dc}}(\rho, A)\rangle + 0.01\langle \widehat{D}_{\mathrm{KL}}(\pi_\theta, \pi_{\mathrm{SFT}})\rangle - 0.002\langle H(\pi_\theta)\rangle, \tag{7}$$

using VERL's low-variance sampled KL estimator and the frozen SFT initializer as reference. KL is an explicit actor-loss term, not part of the task reward or filtering correctness. The entropy term encourages continued exploration. No value network is trained. Table 15 provides the remaining settings.

Table 15: 9B GRPO settings, initialized from the model after two SFT epochs. The optimization budget is distinguished from the realized training duration below.

Setting	Value
Hardware	One node, eight H100 80GB GPUs
Parallelism	Eight-way FSDP; rollout tensor parallelism 1
Precision / memory	BF16 compute; FP32 master weights; CPU offload
Memory controls	Parameter/optimizer offload; gradient checkpointing
Prompts / responses per update	32 / 256 (eight responses per prompt)
PPO minibatch / epochs per rollout	32 prompts / 1
Dynamic microbatch budget	4,096 tokens per GPU
Prompt / response limit	1,024 / 2,048 tokens
Maximum profiles per prompt	8
Sampling temperature / top-p / top-k	1.2 / 1.0 / disabled
Random seed	42
Optimizer	AdamW, $\beta = (0.9, 0.999)$; weight decay 0.01
Learning rate	10^{-6}, constant; no warmup
Gradient-norm clipping	1.0
PPO ratio clip / dual clip	0.2 / 3.0
KL / entropy coefficient	0.01 / 0.002
Group standard-deviation normalization	Disabled
Training budget	2 epochs
Training updates	380 per epoch; 760 budgeted total
Monitoring	2,591 prompts, greedy decoding, every 25 updates

Training procedure and evaluation. The procedure consists of source-data and split verification, validation of omics inputs and reference answers, difficulty filtering using the SFT policy, and on-policy optimization on the fixed set of retained prompts. The trained policy is evaluated using the benchmark protocol. Implementation checks compare rollout and training-policy probabilities, verify finite gradients, and confirm that saved parameters reproduce text-only and omics-conditioned forward passes.

Each epoch drops the final incomplete prompt batch, giving $\lfloor 12{,}171/32 \rfloor = 380$ updates and a two-epoch budget of 760 updates. Monitoring uses greedy decoding with one response per prompt. Training stopped after 724 updates, before completing the budget; the latest retained policy is the checkpoint at update 700.

A.8 Ablation Studies

The ablations support our choice of the full embedding stack, pretraining followed by SFT, and gene tokens initialized from subword embeddings, each of which achieves the best aggregate score within its comparison group. Larger backbones generally perform better, although gains are not uniform across tasks. The final 9B checkpoint improves the overall score from 81.9 to 82.4 after RL, with gains on five of seven metrics and essentially unchanged performance on the remaining two.

A.9 GEO-OmicsQA Examples

Representative question–answer pairs for each GEO-OmicsQA benchmark variant are shown below. The placeholder `<omics>` denotes a transcriptomic expression embedding injected at that position; $\times N$ indicates N replicate sample embeddings.

Binary QA (standard)

Question: Does the addition of PIK3CA*H1047R (samples `<omics>`$\times 3$) further decrease E-cadherin (CDH1) expression compared to ARID1A knockdown alone (samples `<omics>`$\times 3$)?

Answer: No.

Table 16: Ablation study results. The baseline is Qwen3.5-0.8B trained through Pretraining → SFT, with normalized expression, Funomics T0 and Geneformer embeddings, the count indicator, and gene tokens initialized from their mean subword embeddings (Section 3.1). Each block repeats this baseline and varies the named component, except the RL subgroup, which compares the 9B checkpoint after two SFT epochs with the final OmicsLM model (†), reported in Tables 1, 2, and 3. Scores are mean ± standard deviation over three independent training runs; the 9B results are from a single run. Cell2Sentence-like replaces continuous omics inputs with text listing the 1,024 most highly expressed genes, ranked by expression. Scores are scaled to 0–100: macro AUROC for Tabula Sapiens and Human Diseases, mean AUROC for PerturbQA, and binary accuracy/free-text judge score for GEO-OmicsQA. Overall averages the seven benchmark scores. Higher is better; bold marks the best means within each comparison group, including displayed ties.

		Tabula Sapiens		Human Diseases		GEO-OmicsQA		PerturbQA
Model	Overall	Cell type top-20	Tissue	Disease subtype	Tissue	Binary QA	Free text	Mean AUROC
Embedding source								
Full embedding stack	**77.6** ± 0.04	95.3 ± 0.03	75.7 ± 0.08	85.4 ± 0.05	**98.0** ± 0.03	72.6 ± 0.17	**55.2** ± 0.15	**60.7** ± 0.13
No expression embedding	76.2 ± 0.05	**96.7** ± 0.03	70.2 ± 0.07	83.9 ± 0.04	96.3 ± 0.12	**72.9** ± 0.17	54.8 ± 0.20	58.5 ± 0.27
No Funomics T0 embedding	77.2 ± 0.12	95.1 ± 0.04	75.5 ± 0.22	84.2 ± 0.15	**98.0** ± 0.01	72.1 ± 0.47	**55.2** ± 0.20	60.2 ± 0.38
No Geneformer embedding	77.3 ± 0.14	92.9 ± 0.04	**77.5** ± 0.03	**86.3** ± 0.10	**98.0** ± 0.06	71.2 ± 0.38	54.8 ± 0.26	60.3 ± 0.42
Cell2Sentence-like	71.5 ± 0.15	93.6 ± 0.10	75.6 ± 0.15	77.8 ± 0.17	91.2 ± 0.25	62.4 ± 0.42	51.3 ± 0.31	48.8 ± 0.39
Gene tokenization								
Gene tokens	**77.6** ± 0.04	**95.3** ± 0.03	75.7 ± 0.08	**85.4** ± 0.05	**98.0** ± 0.03	**72.6** ± 0.17	55.2 ± 0.15	**60.7** ± 0.13
Random gene initialization	75.6 ± 0.04	**95.3** ± 0.03	74.0 ± 0.15	83.4 ± 0.14	**98.0** ± 0.05	71.9 ± 0.38	55.2 ± 0.11	51.4 ± 0.29
No gene tokens	77.0 ± 0.01	95.2 ± 0.07	**78.9** ± 0.29	82.8 ± 0.09	97.0 ± 0.03	70.7 ± 0.20	**56.9** ± 0.22	57.7 ± 0.24
Stages								
0.8B Pretraining → SFT	**77.6** ± 0.04	**95.3** ± 0.03	75.7 ± 0.08	**85.4** ± 0.05	**98.0** ± 0.03	**72.6** ± 0.17	**55.2** ± 0.15	**60.7** ± 0.13
0.8B Pretraining only	75.3 ± 0.20	95.0 ± 0.39	**80.1** ± 2.13	82.2 ± 1.66	95.9 ± 0.67	66.7 ± 1.85	55.1 ± 0.36	52.2 ± 0.97
0.8B SFT only	73.4 ± 0.28	95.2 ± 0.40	74.3 ± 1.61	81.5 ± 1.91	96.1 ± 0.68	67.2 ± 0.76	49.4 ± 0.49	49.8 ± 0.82
9B Pretraining → SFT	81.9	92.5	75.6	**90.7**	98.2	75.5	**65.1**	75.7
9B Pretraining → SFT → RL†	**82.4**	**92.6**	**76.7**	90.6	**98.4**	**76.7**	**65.1**	**76.4**
Backbone size								
0.8B	77.6 ± 0.04	**95.3** ± 0.03	75.7 ± 0.08	85.4 ± 0.05	98.0 ± 0.03	72.6 ± 0.17	55.2 ± 0.15	60.7 ± 0.13
2B	79.9 ± 0.02	94.1 ± 0.08	**76.0** ± 0.05	85.5 ± 0.16	98.1 ± 0.04	74.9 ± 0.17	60.0 ± 0.29	70.9 ± 0.29
9B	**81.9**	92.5	75.6	**90.7**	**98.2**	**75.5**	**65.1**	**75.7**

Question: Is ABCG1 expression higher in GM-MØ (<omics>×3) while ABCA1 expression is higher in M-MØ (<omics>×3)?

Answer: Yes.

BINARY QA (MINIMAL)

Question: Are Ets-family genes like ETV1, ETV4, and ETV5 unaffected in <omics>×2 compared to <omics>×2?

Answer: No.

Question: Are the epithelial-mesenchymal transition (EMT) markers SNAI1 and MMP3 upregulated in <omics>×3 compared to <omics>×3?

Answer: Yes.

SIMPLE FREE-TEXT (STANDARD)

Question: I am analyzing the transcriptomes of non-hemogenic endothelial cells (<omics>×2) and hemogenic endothelial cells (<omics>×2) isolated from the AGM region of 30- and 32-day-old human embryos based on ACE expression. What does the expression data indicate about the arterial versus venous identity of the hemogenic subset compared to the non-hemogenic subset?

Answer: Both subsets express pan-endothelial markers such as CD34, CDH5, PECAM1, and TEK. However, the hemogenic samples exhibit a clear upregulation of genes classically associated with an arterial fate, including BMX, GJA5, DLL4, CXCR4, and HEY2. Conversely, these hemogenic samples display a lower expression of venous-associated genes like NR2F2 and FLRT2 relative to the non-hemogenic controls. This suggests that the hemogenic subset possesses a distinct arterial gene signature.

Question: Analyze the transcriptomic data for samples <omics>×3 (KBM7 cells treated with the withanolide analog W7) compared to the DMSO control samples <omics>×3. What major metabolic pathways appear to be transcriptionally repressed?

Answer: An evaluation of the gene expression profiles reveals a substantial decrease in the transcript levels of numerous genes central to de novo cholesterol biosynthesis, including HMGCS1, LSS, DHCR7, and SQLE. This coordinated reduction suggests that the compound strongly represses the metabolic pathways responsible for cholesterol production. Consequently, it is evident that sterol metabolism is a primary transcriptional target of this treatment.

SIMPLE FREE-TEXT (MINIMAL)

Question: Examine the EMT process in <omics>×2 versus <omics>×2. What do the epithelial and mesenchymal markers reveal?

Answer: The first group displays a clear upregulation of the master EMT transcription factor SNAI1, accompanied by a strong downregulation of the epithelial marker E-cadherin (CDH1) compared to the second group. Furthermore, there is a modest increase in N-cadherin (CDH2) expression. This coordinated shift from an epithelial to a mesenchymal transcriptional signature strongly supports the conclusion that the perturbation acts as a triggering signal for the initiation of the transition.

Question: I have RNA-seq data from three groups: Group 1 (<omics>×6), Group 2 (<omics>×8), and Group 3 (<omics>×8). Based on this gene expression data, how does Group 3 affect the expression of key myofibroblast markers such as ACTA2, SERPINE1, and ELN compared to Group 2?

Answer: Based on the provided gene expression data, the samples in Group 2 show a pronounced upregulation of myofibroblast markers, including ACTA2, SERPINE1, and ELN, compared to Group 1. In the samples from Group 3, the expression levels of these specific markers are significantly attenuated. This suggests that the treatment in Group 3 effectively inhibits the transcriptional program associated with these markers that was induced in Group 2.

CONTEXTUAL FREE-TEXT (STANDARD)

Question: During perforated-patch current-clamp recordings, we observed that the Venus+ cells display spontaneous action potential firing. What specific voltage-gated channels expressed in the Venus+ dataset (<omics>×4) compared to the Venus- dataset (<omics>×4) provide the molecular basis for this excitability?

Answer: The intrinsic excitability you observed is supported by a robust profile of voltage-gated ion channels in the transcriptomic data. The Venus-positive cells express the voltage-gated sodium channel SCN3A, which is vital for the rapid depolarizing phase of action potentials. Furthermore, they are enriched for specific voltage-gated calcium channels, particularly the P/Q-type CACNA1A and the L-type CACNA1C, which contribute to the action potential duration and trigger intracellular calcium-dependent signaling. Alongside the voltage-gated potassium channel KCNB2 for repolarization, this coordinated suite of ion channels forms the molecular foundation for the spontaneous electrical activity you recorded.

Question: Previous experiments indicate that TIMP1 functions as a cytokine to induce clonal VSMC proliferation. Given this, when comparing the serum-starved control <omics> to the TIMP1-treated sample <omics> from a 65-year-old male, what downstream transcriptional programs support this proliferative state?

Answer: An analysis of the expression data shows an enrichment of cell cycle-related gene sets, including E2F targets and G2/M checkpoint genes, in the TIMP1-treated sample. Concurrently, there is an upregulation of pathways involved in fatty acid metabolism and oxidative phosphorylation. A plausible interpretation is that TIMP1 signaling orchestrates a metabolic reprogramming to meet the increased energetic demands of entering and progressing through the cell cycle.

CONTEXTUAL FREE-TEXT (MINIMAL)

Question: We noted a significant enrichment of oxidative phosphorylation genes (p = 0.02) in our resistant model. Using the RNA-seq data for <omics>×3 compared to <omics>×3, how is this enrichment reflected at the level of individual respiratory chain complex genes?

Answer: The enrichment is distinctly visible across multiple crucial components of the mitochondrial respiratory chain in the first group of samples compared to the second. Specifically, genes encoding subunits of Complex I (NADH dehydrogenase), such as NDUFA1 and NDUFB2, show significant upregulation. Furthermore, multiple subunits associated with Complex IV (cytochrome c oxidase), including COX6C, COX7A2, COX7C, and the assembly factor COX17, are expressed at elevated levels. This coordinated upregulation across different complexes suggests a comprehensive enhancement of the oxidative phosphorylation machinery.

Question: Our experimental rescue showed that knocking down HRK only partially abrogated the apoptosis sensitization seen in samples <omics>×2 compared to <omics>×2. What other pro-apoptotic genes in the expression data might account for the remainder of the sensitization?

Answer: The expression data indicate that multiple parallel pro-apoptotic pathways are activated alongside HRK. For instance, the transcript for the tumor suppressor kinase DAPK1 is highly induced, as is the BH3-only gene BMF. Additionally, direct enhancements to the structural apoptotic machinery, such as the upregulation of BAK1, APAF1, and the executioner CASP7, are clearly evident. These broad transcriptional changes suggest that even without HRK, the collective de-repression of these other sensors and executors creates a highly sensitized environment.

A.10 TASK EXAMPLES

Table 17 shows representative examples and conversation-record counts for 157 instruction task groups, alongside examples from free-text and serialized-text sources. Long questions and answers are truncated. The placeholder <omics> denotes an omics expression embedding injected at that position.

Table 17: Representative training examples by task group and source. Instruction examples are organised into 157 groups by biological objective and input/output format; related cohorts, context variants, and prompt variants are bundled. Each group shows its number of training samples and one question–answer example. Sample counts are conversation records in the released SFT corpus, counted once per record (5,210,786 in total), not unique biological specimens or repeated training exposures. Additional rows illustrate free-text and serialized-text sources without sample counts. Examples are excerpts: long text and lists may be abbreviated, empty thinking tags are omitted, and <omics> denotes an expression input. Free-text records are shown as a single excerpt; when a source serializes a question and its answer into one record, the two parts are separated only for display.

Task	Type	Question	Answer
Single-cell annotation (13 groups; 1,051,316 samples)			
Cell identity with anatomical context 79,976 samples	scRNA-seq	<omics> Given the above expression profile, determine the biological role.	luminal adaptive secretory precursor cell of mammary gland of breast
Cell-cycle phase 99,969 samples	scRNA-seq	<omics> Based on the measured expression data, identify the cell cycle phase of the cell. Possible answers: 'S', 'G2M', 'G1'	G1
Cell-cycle phase verification 99,964 samples	scRNA-seq	Is this sample in the S phase of the cell cycle based on <omics>? Answer with 'No' or 'Yes'.	No
Cell-type annotation (with optional context) 479,838 samples	scRNA-seq	Identify the cell type of <omics> based on the transcriptional profile.	perivascular cell
Immune-cell identification 7,996 samples	scRNA-seq	Given the following expression profile, is this an immune cell? <omics>	No
Cell lineage 39,806 samples	scRNA-seq	Identify the developmental lineage of this sample using the measured expression data. <omics>	endothelial cell of vascular tree, endothelial cell, eukaryotic cell
Biological-sex verification 7,997 samples	scRNA-seq	<omics> Does the above gene expression profile correspond to a male donor? Answer with "No" or "Yes".	No
Single-cell tissue (with optional cell type) 119,803 samples	scRNA-seq	What tissue of origin does the measured expression data suggest? <omics>	breast
Tissue verification 7,999 samples	scRNA-seq	Use <omics> to determine whether the cell is from heart.	No
Tumour cell type (with optional cancer context) 79,968 samples	Tumour scRNA-seq	<omics> Given the measured expression data, identify the cell type.	t cell
Tumour immune-cell identification 8,000 samples	Tumour scRNA-seq	Does <omics> belong to the immune infiltrate rather than the non-immune compartment, based on the expression data? Answer with 'No' or 'Yes'.	Yes
Malignant-cell identification 8,000 samples	Tumour scRNA-seq	Use the transcriptional profile to decide whether <omics> is a cancer cell. Answer with "Yes" or "No".	No
Tumour cell lineage 12,000 samples	Tumour scRNA-seq	Based on the expression data, which lineage terms describe <omics>?	neoplastic cell, abnormal cell, cell
Bulk expression annotation (12 groups; 391,574 samples)			
Cell-line identification 52,191 samples	Bulk RNA-seq	Which cell line does the expression profile support for <omics>?	MSN08
Bulk cell-type annotation 64,122 samples	Bulk RNA-seq	Given <omics>, classify the cell type of it.	cardiac myocytes
GEO tissue annotation 11,812 samples	Bulk RNA-seq	Which tissue best matches it according to <omics>?	embryonic tissue

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Disease or condition annotation 14,792 samples	Bulk RNA-seq	Use the following expression profile to infer the disease condition for this sample. <omics>	Diffuse Intrinsic Pontine Glioma (DIPG)
Perturbation-target identification 69,007 samples	Bulk RNA-seq	Which target was perturbed for <omics> according to the expression data?	Dexamethasone
Perturbation-type annotation 11,837 samples	Bulk RNA-seq	From <omics>, infer the perturbation type for it.	stimulus
Treatment-condition annotation 39,472 samples	Bulk RNA-seq	<omics> From the measured expression data, classify the treatment condition for the sample.	treatment
Expression-to-sample description 11,874 samples	Bulk RNA-seq	Summarize this specimen using <omics>.	Primary pancreatic ductal adenocarcinoma (PDAC) cell line cultured as 3D tumorspheres and treated with Calcitriol (Vitamin D).
Description-to-expressed genes 40,048 samples	Bulk RNA-seq	Using the description, predict the top 200 genes with the highest expression. The sample is described as: H1-NPC cell line exposed to 10uM Ficoll 400 for 10 days, classified as a non-toxic compound.	TUBA1A, EEF1A1, MT-CO1, TMSB10, MT-CO3, ACTG1, . . . (200 genes in total)
GTEx sex verification 14,721 samples	Bulk RNA-seq	<omics> Given the above gene expression profile, was this sample collected from a male donor? Answer with Yes or No.	Yes
GTEx tissue (with optional age or sex) 46,977 samples	Bulk RNA-seq	Given the expression profile, determine the anatomical tissue for <omics>. Possible answers: "Heart - Left Ventricle", "Heart - Atrial Appendage"	Heart - Left Ventricle
GTEx tissue verification 14,721 samples	Bulk RNA-seq	Does <omics> originate from Vagina based on the expression profile?	No
Cancer and clinical annotation (25 groups; 189,757 samples)			
Advanced-stage verification 5,859 samples	Bulk RNA-seq	<omics> Use this expression profile to determine whether this sample is an advanced-stage (III-IV) tumor.	Yes
Age-group annotation 8,825 samples	Bulk RNA-seq	In which age group was the patient when this sample was taken given <omics>?	60-69
Biological-sex annotation 8,873 samples	Bulk RNA-seq	Given the transcriptional profile, is the donor of <omics> male or female?	female
Cancer diagnosis (with optional context) 32,500 samples	Bulk RNA-seq	This sample comes from Ovary. Given the expression data, identify the disease category of <omics>.	Cystic, Mucinous and Serous Neoplasms
Lymph-node involvement 5,671 samples	Bulk RNA-seq	<omics> Use the measured expression data to determine whether this sample has regional lymph node involvement.	No
Distant-metastasis verification 629 samples	Bulk RNA-seq	Does <omics> indicate distant metastasis?	Yes
M stage (with optional context) 9,231 samples	Bulk RNA-seq	Identify the distant metastasis (M) stage for <omics> given the expression data. Options: M0, M1	M0
N stage 5,671 samples	Bulk RNA-seq	<omics> Determine the N (node) stage based on this gene expression profile.	N0
T stage 6,564 samples	Bulk RNA-seq	<omics> Determine the T (tumor) stage for this specimen according to AJCC classification based on the measured expression data.	T3
Overall tumour stage (with optional context) 11,729 samples	Bulk RNA-seq	Based on <omics>, what is the pathological stage of the sample?	Stage IV
TCGA study identification 8,893 samples	Bulk RNA-seq	Based on the transcriptional profile, what TCGA study is <omics> associated with?	TCGA-KIRC

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Primary site (with optional cancer context) 17,767 samples	Bulk RNA-seq	Determine the primary site for the sample using this expression profile. <omics>	Kidney
Primary-site verification 8,892 samples	Bulk RNA-seq	Is the primary site of <omics> Kidney based on the gene expression profile? Answer with "Yes" or "No".	Yes
Tumour versus normal 8,893 samples	Bulk RNA-seq	Does this specimen represent a tumor or normal tissue based on <omics>?	Tumor
Tumour specimen type 8,301 samples	Bulk RNA-seq	Given <omics>, what is the tumor type of the sample?	Primary
Vital status (with optional stage context) 14,737 samples	Bulk RNA-seq	Given the transcriptional profile, determine the vital status for <omics>.	Alive
Cancer organ system 4,353 samples	Cancer-model RNA-seq	Which organ system is the cancer in <omics> derived from, based on the transcriptional profile?	Thoracic Cancer
Matched-pair cancer organ system 1,084 samples	Cancer-model RNA-seq	The two expression profiles below were measured on the same tumour, first in the patient and then in a xenograft established from it. Patient specimen: <omics> Xenograft: <omics> Which organ system is this cancer derived from?	Reproductive System Cancer
Histology (with optional primary site) 8,808 samples	Cancer-model RNA-seq	Based on the expression data, what is the histological subtype of <omics>?	Colon Adenocarcinoma
Metastatic-lesion verification 2,444 samples	Cancer-model RNA-seq	Given the gene expression profile, does <omics> originate from a metastasis?	No
Patient specimen versus model 826 samples	Cancer-model RNA-seq	Does this sample come from a patient tumour rather than a laboratory model, based on the measured expression data? <omics> Answer with 'No' or 'Yes'.	No
Model system (with optional cancer context) 2,860 samples	Cancer-model RNA-seq	What does this gene expression profile reveal about the model system it was measured in? <omics> Options: cell line, organoid, PDX, patient tumour.	organoid
Cell line versus organoid 534 samples	Cancer-model RNA-seq	<omics> Does this expression profile come from a two-dimensional culture rather than an organoid? Answer with 'Yes' or 'No'.	No
Cancer-model primary site 4,352 samples	Cancer-model RNA-seq	Based on the transcriptional profile, what is the site of origin of <omics>?	digestive/gastrointestinal
Cell-line tissue given cancer type 1,461 samples	Cell-line RNA-seq	What tissue of origin does the following gene expression profile suggest for it? Cancer type: Melanoma. <omics>	Skin
Expression interpretation and comparison (7 groups; 362,886 samples)			
GEO binary biological QA 78,828 samples	Bulk RNA-seq	Review the data for <omics> and <omics>. Does the factor depleted in <omics> normally act to prevent the extension of upstream antisense transcripts? Reply with 'Yes' or 'No'.	Yes
GEO binary QA with reasoning 78,788 samples	Bulk RNA-seq	Do the lineage markers CDX1 and CDX2 show robust baseline expression alongside LINC00675 in <omics> (colon epithelial cell line LS180, derived from well-differentiated colorectal cancer), <omics> (colon epithelial cell line C80, derived from well-differentiated colorectal cancer), and <omics> (colon epithelial cell line SW480, derived from well-differentiated colorectal cancer)?Walk through your reasoning before answering "Yes" or "No".	Reasoning: The expression profiles indicate prominent levels of CDX1 and CDX2, which parallel the high abundance of LINC00675, confirming their robust [. . .] Final: Yes

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
GEO open-ended biological QA 23,591 samples	Bulk RNA-seq	I have identified a signature of 42 specific transcripts that differentiate these pre-perfusion allografts destined for delayed graft function (<omics>, <omics>, <omics>) from those with immediate function, after adjusting for reperfusion effects. Based on the expression of these 42 genes in the provided data, what upstream biological processes appear to be driving this impaired physiological resilience?	Evaluating the expression of these 42 signature transcripts reveals a strong enrichment for innate immune system activation and cell-mediated immune responses. The upregulation of transcripts such as UBD, NLRC4, and interferon-gamma inducible targets like CXCL10 and [. . .]
Top-expressed gene ranking 63,995 samples	Expression	Use the expression profile to list the top 62 most expressed genes in <omics>.	MT-RNR1, MT-RNR2, SPARC, AKAP12, VIM, TRIO, . . . (62 genes in total)
Sample and cell-line description / QA 32,657 samples	Expression	<omics> Give a brief description of the sample.	Hippocampal neurons derived from induced pluripotent stem cells (IPSCs) of a fibroblast sample, treated with lithium.
Nearest expression-profile retrieval 45,021 samples	Expression	Given cell lines LN18<omics>, HKA1<omics>, SBC5<omics>, LS513<omics>, KYSE270<omics>, which of them is the most similar to HARA<omics> in terms of expression profile?	KYSE270
Most-similar profile pair 40,006 samples	Expression	Given cell lines DND41<omics>, CFPAC1<omics>, A375SKINCJ1<omics>, TALL1<omics>, SNU182<omics>, NCIH524<omics>, which pair is the most similar in terms of expression profile?	TALL1, DND41
Chemical and genetic perturbation (13 groups; 530,681 samples)			
Compound and target identification 39,985 samples	Chemical perturbation	Untreated HS578T cells: <omics> The same cells after 0.05 uM of a compound: <omics> Which genes does this compound act on, which compound is it, and what is its structure? Answer as "Targets: ...; Molecule: ...; SMILES: ...".	Targets: CYP2J2, HRH1, KCNH1, KCNH2; Molecule: Terfenadine; SMILES: CC(C)(C)C1=CC=C(C=C1)C(CCCN2CCC(CC2)C(C3=CC=CC=C3)(C4=CC=CC=C4)O)O
Downregulated genes after treatment 9,775 samples	Chemical perturbation	If exposure to a compound targeting EGFR (SMILES COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1) at 0.05 μM is applied in <omics> , what would be the top 59 most downregulated genes?	FN1, ABI3BP, PLXNA2, ADAMTS6, KRT81, SUCNR1, . . . (59 genes in total)
Upregulated genes after treatment 9,775 samples	Chemical perturbation	If exposure to a compound targeting EGFR (SMILES COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1) at 0.05 μM is applied in <omics> , what would be the top 59 most upregulated genes?	CLU, FHIP1A, FAM210A, ADM, SOD2, ARRDC3, . . . (59 genes in total)
Top-expressed genes after treatment 9,775 samples	Chemical perturbation	Control expression profile: <omics> Predict the top 103 expressed genes following 0.05 μM exposure to a compound targeting EGFR (SMILES COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1OCCCN1CCOCC1) treatment.	MT-RNR2, MT-RNR1, SPARC, AKAP12, VIM, TRIO, . . . (103 genes in total)
Knocked-out gene identification 46,447 samples	Genetic perturbation	Given the control <omics> <omics> <omics> <omics> <omics> and the perturbed <omics> <omics> <omics> <omics> <omics> , determine the gene whose knockout caused this change.	SLC25A3
Top-expressed genes after knockout 16,801 samples	Genetic perturbation	Control expression profile: <omics> <omics> <omics> <omics> <omics> List the 60 most highly expressed genes after knocking out SLC25A3.	ACTB, CD24, POU5F1, EEF2, ACTG1, HSPA8, . . . (60 genes in total)
Downregulated genes after knockout 16,801 samples	Genetic perturbation	Control expression profile: <omics> <omics> Predict the top 92 downregulated genes following the knockout of SLC25A3.	MT-CO3, SLC25A3, MT-ATP6, MT-ND4, MT-CYB, MT-CO2, . . . (92 genes in total)
Upregulated genes after knockout 16,801 samples	Genetic perturbation	Control expression profile: <omics> <omics> Predict the top 92 upregulated genes following the knockout of SLC25A3.	H1-1, LDB2, ERRFI1, NABP2, COX7A2, TEK, . . . (92 genes in total)
Differentially expressed genes after knockout 165 samples	Genetic perturbation	Control expression profile: <omics> <omics> <omics> <omics> <omics> List the 60 genes most affected by the knockout of SLC25A3.	SLC25A3, MFGE8, HAPLN3, PLAAT3, STC2, PSAT1, . . . (60 genes in total)
Perturbation-response direction with gene notes 300,000 samples	Genetic perturbation	Notes: AAR2 = AAR2 splicing factor; 79 partners incl PRPF8, SNRNP200. RPL7A = ribosomal protein L7a; 567 partners. 1 shared partners. Does RPL7A go up or down when AAR2 is knocked down given <omics>? Answer with Increase or Decrease only.	Decrease

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Differential-expression detection 50,000 samples	Genetic perturbation	Reply Yes or No. Is the expression of DST affected by a knockdown of COPA given <omics>?	Yes
Screen-hit ranking among candidates 13,320 samples	Genetic perturbation	This screen was performed in T.T cells, a Esophageal Squamous Cell Carcinoma Cell Line. [...] A gene is classified as a "hit" if its Knockout significantly decreases cell proliferation. The statistical criterion for significance is: RRA score < -0.5. [...] Genes with low RRA score are ranked most highly. [...] These 23 genes were assayed in this screen and are listed in no particular order: HLTF, SIRT3, TRIM66, SETD5, GTF3C4, CHD6, ... (23 genes in total)	GTF3C4, SETDB1, TTF2, PRMT5, BRD8, HDAC7, HDAC5
Genome-wide screen-hit ranking 1,036 samples	Genetic perturbation	This screen was performed in HAP-1 cells, a Chronic Myelogenous Leukemia Cell Line. [...] A gene is classified as a "hit" if its Knockout significantly increases resistance to virus as measured by increased cell proliferation. The statistical criterion for significance is: FDR < 0.05. [...] Genes with low FDR are ranked most highly. [...] Provide your response as an ordered list of exactly 17 HGNC gene symbols, using the ranking criteria above.	PAPSS1, EXT2, B4GALT7, DPM3, STT3B, EXTL3, ... (17 genes in total)
Cancer genomics (5 groups; 21,297 samples)			
Copy-number-altered genes 14,824 samples	Cell-line genomics	Which cancer genes have a copy number gain in cell line WM35, sorted alphabetically?	AFDN AXIN2 BACH2 CD79B DDX5 DEK ... (16 genes in total)
Cell line from copy-number alterations 1,707 samples	Cell-line genomics	Which cell line is characterized by a given list of cancer genes copy number alterations (loss and gain) CN loss: AMER1 AR ARAF ATRX BCOR BCORL1 BTK CCNB3 CCNQ CDKN2A CDKN2B CRLF2 DDX3X ELF4 ETV6 EZHIP GPC3 KDM5C KDM6A P2RY8 PHF6 PIGA POU3F4 RBM10 SH2D1A SMC1A STAG2 TAF1 TFE3 UBA1 UBE2A XIAP ZMYM3, CN gain: ?	OCIAML5
Cell lines sharing copy-number alterations 840 samples	Cell-line genomics	Which cell lines have a copy number loss in the gene SP140?	BJAB, C32, C4I, C4II, CAL78, EJ1, ... (24 genes in total)
Mutated cancer genes 3,448 samples	Cell-line genomics	Which cancer genes are mutated in cell line WM88<omics>, sorted alphabetically?	BRAF: homozygous
Cell lines carrying a gene mutation 478 samples	Cell-line genomics	Which cell lines have a mutation in the gene EP300?	639V, CAL33, HEC1A, HEC265, JHH5, NCIH1618, ... (12 genes in total)
Pathways, gene sets, and molecular knowledge (19 groups; 653,481 samples)			
Path completion given pathway name 50,000 samples	Graph / gene set	In the context of 'Long patch BER', complete the following biological pathway in the downstream direction: POLB	form complex with MCM7 form complex with PCNA form complex with RFC2 activate FEN1 activate LIG1
Pathway identification from a path 50,000 samples	Graph / gene set	Given the following biological pathway sequence in the downstream direction: NKX3-1 unknown relationship AKT1 inhibit BAD, which pathway or geneset does this represent?	Loss of NKX3-1 to PI3K signaling pathway
Process identification from a gene set 5,565 samples	Graph / gene set	Which biological process is related to the following gene set: GSN, TGFB1, BIK, SH3GLB1, ISL1, MAPK15, ... (126 genes in total)?	Regulation of Protein-Containing Complex Assembly
Gene-set membership retrieval 5,559 samples	Graph / gene set	Which genes are related to Regulation of Protein-Containing Complex Assembly?	GSN, TGFB1, BIK, SH3GLB1, ISL1, MAPK15, ... (126 genes in total)
Gene-set enrichment from genes 26,600 samples	Graph / gene set	Which biological processes are related to the following gene set: FAM114A1, DHCR24, CA2, FEN1, LNP1, KIF12, ... (119 genes in total)?	DNA metabolic process, Chromosome centrometric region, G2m checkpoint, Regulation of CDK activity, Anaphase-promoting complex, E2f targets, Cell Cycle Checkpoints, Mitotic Sister Chromatid Segregation, Mitotic Prometaphase, APC C-mediated Degradation of Cell Cycle [...]

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Directed-edge verification 51,157 samples	Graph / gene set	Is there an 'indirect effect' direct relationship from PRKAG2 to SREBF1?	Reasoning: PRKAG2 has 33 forward neighbour(s): ACACA, ACACB, AKT1S1, CD36, CFTR, CRTC2, CRY1, EEF2K, ELAVL1, FOXO1, FOXO3, FOXO4, ... SREBF1 is among them. Final: Yes
Shortest path after knockout 199,758 samples	Graph / gene set	If the following nodes are knocked out: [[ANK1, CDC42, CEL, PRKCA, PRKCB, TWIST2, VAV3]], what is the shortest path from MIR520E to LIPA in the biological pathways?	Reasoning: Blocked nodes: ANK1, CDC42, CEL, PRKCA, PRKCB, TWIST2, VAV3. Target is LIPA; from MIR520E the shortest route is 9 step(s). From MIR520E: 1 open [...] Final: MIR520E inhibit CD44 bind to IQGAP1 activate MAPK1 activate PRKCG indirect effect PRKACA express HSD17B7 sequential reactions DHCR24 produce DHCR7 produce LIPA
Shortest directed path 30,000 samples	Graph / gene set	How can AKT1S1 reach LIAS through the biological pathways in minimum number of steps?	Reasoning: Target is LIAS; from AKT1S1 the shortest route is 10 step(s). From AKT1S1: 6 open neighbour(s). Of the open ones, MTOR, RPS6KB1 sit 9 step(s) [...] Final: AKT1S1 inhibit MTOR activate PRKCA indirect effect PRKACA activate PPARA express ACOX1 produce ACSS2 produce ACADS produce ACSM1 produce LIPT1 produce LIAS
Directed pathway continuation 70,000 samples	Graph / gene set	Can you provide an example walk from COL1A1 in the biological pathways?	COL1A1 activate CD44 bind to NANOGP8 bind to STAT3 express ESR2 cause expression CTSD
Undirected interaction continuation 40,000 samples	Graph / gene set	Complete the following biological interaction pathway (no direction specified): MRC1	react with FCGR1A unknown relationship with SYK unknown relationship with LYN unknown relationship with INPP5D bind with INPPL1 bind with PIK3R1
Directed reachability 29,996 samples	Graph / gene set	Is CANT1 connected to AK9 through 'produce'?	Yes
Common upstream node and connecting paths 19,795 samples	Graph / gene set	In biological pathways, what is the shortest common upstream node (Lowest Common Ancestor, LCA) for ELK1 and NCAM2? Describe the respective upstream paths from ELK1 to this LCA and from NCAM2 to this LCA.	The upstream path from ELK1 to the LCA is PRNP inhibit HSPA5 inhibit ERN1 activate MAPK8 activate ELK1. The upstream path from NCAM2 to the LCA is PRNP binding/association NCAM2. The LCA is PRNP.
Pathway neighbours 12,169 samples	Graph / gene set	What are the interactions with other proteins for FARSB?	Interactions for FARSB: FARSB unknown relationship with EEF1B2, FARSB unknown relationship with EPRS1, FARSB bind with FARS2, FARSB bind with FARSA, FARSB bind with FDXACB1, FARSB bind with IARS1, FARSB unknown relationship with LARS1, FARSB unknown relationship with [...]
Pathway enrichment from expression 2,668 samples	Graph / gene set	Given the top underexpressed genes in cell line MESSA<omics>, what are the top 20 downregulated pathways?	The following pathways are downregulated: Regulation of fibrinolytic system, PAI, NLRP1 inflammasome signaling pathway, Yersinia YopH to TCR-NFAT signaling pathway, Non-canonical inflammasome signaling pathway, AML1-ETO fusion to PU.1-mediated transcription, HCMV US28 [...]
Protein function 20,000 samples	Protein knowledge	What can be a function of the protein MYOD1?	Acts as a transcriptional activator that promotes transcription of muscle-specific target genes and plays a role in muscle differentiation

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Protein properties 11,456 samples	Protein knowledge	What are the type properties of protein MAP2K6?	MAP2K6 (Mitogen-activated protein kinase kinase 6 ia a enzyme and kinase protein.
Protein subcellular localisation 12,599 samples	Protein knowledge	What are possible locations of protein RFPL4A?	Protein RFPL4A could be located in: centrosome, cytosol.
Protein knowledge with reasoning 15,000 samples	Protein knowledge	Protein: Uncharacterized protein At4g13200, chloroplastic Organism: Arabidopsis thaliana (Mouse-ear cress) Location: Plastid, chloroplast, plastoglobule {ECO:0000269|PubMed:16461379, ECO:0000269|PubMed:22274653}. Domains: - IPR040003: Plastoglobular protein 18-like (family) [1-178] Interaction partners: - Probable plastid-lipid-associated protein 9, chloroplastic Describe this protein's function and name its Gene Ontology leaf terms.	Reasoning: A single sequence-spanning signature is present: IPR040003 (Plastoglobular protein 18-like family) from residues 1–178. The fact that this [. . .] Final: - Functional Summary: A small chloroplast adaptor localized to plastoglobules that anchors to the lipid monolayer and assembles protein complexes at the [. . .]
Disease inference from pathway context 1,159 samples	Pathway knowledge	Chromosome Number: 20 Network Definition of the pathway: PRNP* -> NOTCH1 Genes in the pathway: PRNP; prion protein | NOTCH1; notch receptor 1 Given this context, what is the biological effect of this PRNP allele, specifically what disease does this contribute to?	Reasoning: Step 1: The variant is an insertion of a 24-amino acid octapeptide repeat sequence in the PRNP gene on chromosome 20. Step 2: The normal PRNP [. . .] Final: creutzfeldt-jakob disease
Cell and disease ontologies (9 groups; 126,858 samples)			
Cell type from definition 2,899 samples	Ontology	Which cell type is described by the following definition? "A langerin-positive lymph node dendritic cell that is CD103-positive and CD11b-low."	CD103-positive, langerin-positive lymph node dendritic cell
Cell-type definition generation 2,899 samples	Ontology	Give a definition for the cell type "CD103-positive, langerin-positive lymph node dendritic cell".	A langerin-positive lymph node dendritic cell that is CD103-positive and CD11b-low.
Direct-parent verification 2,962 samples	Ontology	Is "CD24-positive, CD4 single-positive thymocyte" classified immediately under "eukaryotic cell", with no class in between? Answer Yes or No.	No
Cell/disease ancestry verification 62,824 samples	Ontology	Is "superior cervical ganglion TH neuron" a kind of "neural cell"? Answer Yes or No.	Yes
Cell/disease ancestor enumeration 19,819 samples	Ontology	Which cell type categories does "transitional myocyte of anterior internodal tract" fall under? List them separated by commas.	transitional myocyte, transitional myocyte of internodal tract, striated muscle cell, cardiocyte, contractile cell, striated visceral muscle cell, muscle cell, electrically active cell, visceral muscle cell, specialized cardiac myocyte, cardiac muscle cell, [. . .]
Most-specific common cell/disease ancestor 9,000 samples	Ontology	What is the most specific cell type that both "IgD-negative CD38-positive IgG memory B cell" and "group 2 innate lymphoid cell" are? Give one cell type.	lymphocyte
Cell-type specificity ordering 2,862 samples	Ontology	Put these cell types in order of decreasing specificity, separated by " > ": retinal ganglion cell A2 > retinal ganglion cell > alpha retinal ganglion cell (Mmus)	retinal ganglion cell A2 > alpha retinal ganglion cell (Mmus) > retinal ganglion cell
Shared cell/disease class verification 9,000 samples	Ontology	Is "amine precursor uptake and decarboxylation cell" a common cell class of both "superior cervical ganglion TH neuron" and "chromaffin cell"? Answer Yes or No.	No
Cell/disease synonym resolution 14,593 samples	Ontology	What is the standard term for "hippocampus glial cell"?	hippocampal glial cell
Expression annotation with reasoning (9 groups; 504,429 samples)			
Bulk cell type with reasoning 2,619 samples	Bulk RNA-seq	Explain which cell type the overall expression programme in <omics> best supports, then put your single-line prediction after the word "Answer:" on its own final line.	Reasoning: The profile is dominated by a coordinated expression of cardiac contractile and regulatory proteins, including the atrial isoform MYL7, cardiac [. . .] Final: Answer: cardiomyocyte

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Bulk tissue with reasoning 22,219 samples	Bulk RNA-seq	I need a quick label for this bulk sample, so what tissue origin does the overall programme in <omics> best support, and why? Add your concise prediction after "Answer:" on a new line.	Reasoning: The profile is dominated by the surfactant metabolism programme, with high detection of SFTPC, SFTPA1, SFTPA2, SFTPB, SFTPD, NAPSA, and SLC34A2. [. . .] Final: Answer: Lung
Biological sex with reasoning 16,307 samples	Bulk RNA-seq	Which biological sex is most consistent with the expression profile in <omics>, and how strong is the supporting evidence? Put your final prediction on a separate line after the marker "Answer:".	Reasoning: The bulk sample exhibits coherent non-detection of RPS4Y1 and KDM5D, two curated Y-linked markers associated with the male reference category. [. . .] Final: Answer: female
Single-cell type with reasoning (optional context) 253,738 samples	scRNA-seq	Which cell type is best supported by the overall expression programme in <omics>? Explain your reasoning using the key markers, and put your prediction after the 'Answer:' marker on a separate final line.	Reasoning: The profile is defined by the co-detection of CD3D and CD3E, the structural subunits of the T-cell receptor complex that are essential for [. . .] Final: Answer: t cell
Cell-cycle phase with reasoning 30,815 samples	scRNA-seq	Interpret the cell cycle phase of this single cell from <omics> and identify the most informative genes supporting that determination, then place your final classification on a separate line after 'Answer:'.	Reasoning: The detection of CKAP5 and CKS1B provides complementary evidence for a G2M phase. CKAP5 functions as a microtubule-associated protein that [. . .] Final: Answer: G2M
Immune-cell status with reasoning 60,103 samples	scRNA-seq	Is this single cell likely to be immune or non-immune? Briefly note the strongest evidence and any limitations in the <omics> expression. 'Answer:' on the final line.	Reasoning: The profile exhibits a coherent B-cell receptor signalling programme, evidenced by the co-detection of CD19, CD79A, and MS4A1. These genes form [. . .] Final: Answer: Yes
Cell lineage with reasoning 66,788 samples	scRNA-seq	Which expression pattern in <omics> best supports a specific cell lineage, and how does that narrow down to broader categories? Put your final hierarchy on a separate line after "Answer:"	Reasoning: The detection of VIM and COL6A3 provides complementary structural and functional evidence for a fibroblast identity. VIM, a class-III [. . .] Final: Answer: fibroblast, stromal cell, connective tissue cell, eukaryotic cell
Shared cell type across profiles 13,688 samples	scRNA-seq	Which cell type identity is shared by the separate member profiles within this group? The expression data is available at <omics> <omics> <omics>. After a brief explanation, place your final prediction after 'Answer:' on its own line.	Reasoning: The consistent detection of KRT18 across all members, alongside KRT8 and KRT7 in the majority, establishes a shared epithelial identity. These [. . .] Final: Answer: kidney loop of Henle descending limb epithelial cell
Single-cell tissue with reasoning 38,152 samples	scRNA-seq	Explain which tissue assignment is best supported by this single-cell expression profile: <omics>, and put your prediction after "Answer:" on a separate final line.	Reasoning: STC2 and MUCL1 provide the strongest evidence for a breast tissue assignment, as both are detected in this single-cell profile and are [. . .] Final: Answer: breast
Essentiality, drug response, and mechanisms (14 groups; 486,695 samples)			
Gene essentiality and dependency score 300,000 samples	Expression / drug	<omics> This RNA expression profile is from a Synovial Sarcoma cell line (Soft Tissue). Cancer lineage: Soft Tissue > Synovial Sarcoma. Target gene AGA LoF: no. Tissue-distinctive LoF genes: AOX1, ARMCX4, CKMT2, COMMD10, DPY19L2, KPNA1, LARS2, MPDZ, NSD2, OR5M1, PDCD6IP, PDS5A, RGS6, RPGRIP1, SH3RF1, STAG2, SZT2, [. . .] least 0.5 and No otherwise. Truncate the value to one decimal without rounding. Return only a JSON object with "essential_dep" first and "dependancy" second.	{"essential_dep": "No", "dependancy": 0.0}
PRISM secondary drug activity 24,652 samples	Expression / drug	Cell-line expression profile: <omics> Cell line: A2780_OVARY (CVCL_0134). Treatment: drug mesna. Use the PRISM secondary eight-point dose-response screen's released two-component response class. Infer the cell-specific activity from the expression profile. Respond with exactly one JSON object: {"label":"sensitive"} or {"label":"resistant"}.	{"label":"sensitive"}

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
GDSC drug sensitivity 48,248 samples	Expression / drug	<omics> This RNA expression profile is from the A101D cell line (Skin, Melanoma). In the GDSC screen, the compound is CAMPTOTHECIN. Its annotated targets are TOP1. Predict the source's final sensitivity label, assigned by its compound-specific GMM/Z-score procedure and corrections. Is this cell line sensitive or resistant to CAMPTOTHECIN? Answer with sensitive or resistant only.	sensitive
PRISM primary drug sensitivity 69,655 samples	Expression / drug	<omics> This RNA expression profile is from the A375_SKIN cell line (Skin, Melanoma). In the PRISM Primary screen, the compound is proscillaridin-a. The source assigns its label from the compound's bimodal single-dose response distribution. Is this cell line sensitive or resistant to proscillaridin-a? Answer with sensitive or resistant only.	resistant
Drug response with mechanism context 512 samples	Expression / drug	<omics> This RNA expression profile is from the AN3 CA cell line (Uterus, Endometrial Carcinoma). In the GDSC screen, the compound is BRIVANIB. Its annotated targets are VEGFR, PDGFR. Predict the source's final sensitivity label, assigned by its compound-specific GMM/Z-score procedure and corrections. Is this cell [. . .] RTK signaling. Baseline mRNA vs training cells: EMT typical; E2F high; cholesterol high; ABCB1 trace, ABCG2 trace. Answer with sensitive or resistant only.	sensitive
GDSC response at an absolute IC50 threshold 5,964 samples	Expression / drug	<omics> Cell line: K-562 (CVCL_0004) Compound: CHEMBL941 (CHEMBL941) Known facts: - Myeloproliferative Neoplasms [cell.disease] - Myeloid [cell.lineage] A cell line counts as sensitive when the drug's absolute IC50 stays below 2 uM, and resistant when it exceeds 100 uM. Is this cell line sensitive or resistant to CHEMBL941? Answer with a JSON object: {"label": "sensitive"} or {"label": "resistant"}.	{"label": "sensitive"}
Evidence-grounded omics reasoning dialogue 18,222 samples	Expression / drug	<omics> Baseline expression of unperturbed JURKAT cells is shown above. Known facts: - HGNC:10331 (readout) [gene.RPL29.id] Following a knockdown of WARS, is RPL29 upregulated or downregulated? Answer with a JSON object: {"label": "Increase"} or {"label": "Decrease"}.	Reasoning: The readout gene RPL29 is catalogued as HGNC:10331 (readout) [gene.RPL29.id]. The readout is expected to rise once the knocked-down gene is gone. Final: {"label": "Increase"}
Drug targets and molecular actions 135 samples	Expression / drug	Report the documented molecular targets and pharmacological actions of AZ960. Return only JSON with "targets" (HGNC symbols), "actions" (action names), and "mechanisms" (objects with "target" and "action").	{"targets":["JAK2"],"actions":["inhibit"],"mechanisms":[{"target":"JAK2","action":"inhibit"}]}
Mechanism-edge relation classification 2,624 samples	Drug mechanism	Here is how Tinidazole acts on Giardiasis: Tinidazole – positively regulates -> Response to oxygen radical Response to oxygen radical – disrupts -> DNA DNA – in taxon -> Giardia intestinalis Giardia intestinalis – causes -> Giardiasis Someone states: Response to oxygen radical disrupts DNA. How does that claim [. . .] the mechanism relates those two, but the other way round. Unrelated: the mechanism does not link those two directly at all. Answer with the option name only.	Consistent
Mechanism-step membership verification 1,749 samples	Drug mechanism	Someone states: Bacteria takes place in Cell wall. Cefamandole treats Pneumonia. Does the mechanism include that step, in that direction? Answer Yes or No.	No
Missing mechanism entity 3,333 samples	Drug mechanism	One entity has been blanked out of this mechanism of action of reproterol on Asthma: reproterol – increases activity of -> Beta-2 adrenergic receptor Beta-2 adrenergic receptor – positively regulates -> ? ? – negatively correlated with -> Asthma The node shown as ? is a BiologicalProcess. What is it?	relaxation of smooth muscle
Effect of blocking a mechanism step 9,238 samples	Drug mechanism	Indication: Potassium hydrogencarbonate for Hypokalemia. Suppose this step does not take place. Potassium falls as Hypokalemia rises. What does that do to the drug? No Effect: that is not part of how the drug works, so its mechanism is untouched. Partially Blocked: it is part of the mechanism, but another route to the disease remains. Fully Blocked: it is part of the mechanism and every route runs through it. Answer with the option name only.	Fully Blocked

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Drug-to-disease mechanism chain 1,339 samples	Drug mechanism	How does Panobinostat treat Primary cutaneous T-cell lymphoma? Answer as a chain of steps, one per line, written as: source – relation -> target	Panobinostat – decreases activity of -> Histone deacetylase Histone deacetylase – positively regulates -> Histone deacetylation Histone deacetylation – negatively correlated with -> Apoptotic process Apoptotic process – negatively correlated with -> Primary [. . .]
Drug-response reasoning with mechanism context 1,024 samples	Expression / drug	<omics> This RNA expression profile is from the HPAF-II cell line (Pancreas, Pancreatic Adenocarcinoma). In the GDSC screen, the compound is GEMCITABINE. Its annotated targets are Pyrimidine antimetabolite. Predict the source's final sensitivity label, assigned by its compound-specific GMM/Z-score procedure and [. . .] EMT typical; E2F low; cholesterol high; ABCB1 typical, ABCG2 high. Write a brief evidence summary inside <think> tags, then exactly sensitive or resistant.	Reasoning: Unverified source hypothesis: GEMCITABINE targets RRM2 (inhibits); HPAF-II is wild-type for RRM2. Final: resistant
Human genetics and disease associations (11 groups; 300,000 samples)			
Inheritance mode 14,556 samples	Genetics	What inheritance mode does ClinVar record for VPS16 in Splenomegaly? Give the canonical inheritance label.	autosomal recessive
Gene–disease association strength 34,977 samples	Genetics	What is the GWAS association-strength tier for CENPU and injury in Open Targets 26.06? Answer weak, moderate, strong, or very strong.	moderate
Rare-variant burden direction 4,546 samples	Genetics	Is the recorded rare-variant burden effect of CDKN2A on neoplasm risk-increasing or protective? Answer risk or protect.	risk
ClinVar clinical significance 34,977 samples	Genetics	List all clinical significance terms recorded for rs11554628 in DNAJC5 for CLN4B disease by ClinVar in Open Targets 26.06. Return a JSON array.	["benign"]
Gene–disease evidence sources 34,977 samples	Genetics	Use only the quality-filtered GWAS, rare-variant burden, ClinVar, Gene2Phenotype, Genomics England PanelApp, Orphanet, and ClinGen records retained in this benchmark. List all genetic evidence sources recorded for EIF4A3 and asthma in Open Targets 26.06. Return a JSON array of source names.	["GWAS"]
Loss-of-function constraint 14,093 samples	Genetics	Is F7 in the two most loss-of-function-constrained gnomAD sixths? Answer Yes or No.	No
Gene–disease association verification 34,976 samples	Genetics	Use only the quality-filtered GWAS, rare-variant burden, ClinVar, Gene2Phenotype, Genomics England PanelApp, Orphanet, and ClinGen records retained in this benchmark. In Open Targets 26.06, is there recorded human genetic evidence linking AMY2B to dentures? Answer Yes or No.	Yes
Disease retrieval for a gene 21,970 samples	Genetics	List all diseases or traits with Gene2Phenotype evidence for TET3 in Open Targets 26.06. Return a JSON array of names.	["Beck-Fahrner syndrome"]
GWAS effector-gene prioritisation 34,976 samples	Genetics	For the GWAS of Platelet count at 6_36381023_C_T, select the most likely effector gene among these candidates [. . .]. Candidates: [{"features": {[. . .] "distanceFootprintMean": 1.0, [. . .]}, "gene": "ETV7"}, {"features": {[. . .] "distanceFootprintMean": 0.98992, [. . .]}, "gene": "KCTD20"}]. Answer with one candidate gene symbol.	ETV7
Coding-variant consequence 34,976 samples	Genetics	Classify the coding consequence of rs1555734923 for this canonical transcript using the reference/alternate alleles and codon/amino-acid change. Context: {"transcriptId": "ENST00000420124", "codons": "Cga/Tga", "aminoAcidChange": "R2587*", "biotype": "protein_coding", "referenceAllele": "C", "alternateAllele": "T"}. Answer one of: stop_gained, frameshift_variant, stop_lost, start_lost, inframe_insertion, inframe_deletion, missense_variant, synonymous_variant.	stop_gained
Variant population-frequency class 34,976 samples	Genetics	Classify the maximum gnomAD population frequency of rs8592. Answer common, low-frequency, rare, or ultra-rare.	common
Biological and general instruction following (8 groups; 86,298 samples)			

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Constrained expression annotation 13,437 samples	Omics instruction	Given this expression profile, determine the anatomical tissue for it. <omics> Possible answers: "Breast - Mammary Tissue", "Bladder", "Artery - Tibial", [. . .]	Bladder
Constrained essentiality prediction 6,476 samples	Omics instruction	<omics> This RNA expression profile is from a Mature B-Cell Neoplasms cell line (Lymphoid). [. . .] Target gene KCTD10 LoF: no. [. . .] Predict its dependency label and dependency value. [. . .] After reasoning, return one valid JSON object with exactly the key "answer". Preserve the requested conclusion as its value.	{"answer":{"essential_dep":"No", "dependancy":0.1}}
Constrained perturbation prediction 5,000 samples	Omics instruction	[. . .] Cell line: HEK293T. After knocking down PPARA, does ARL15 increase or decrease given <omics>? Answer with Increase or Decrease only. Final output contract for this variant; it overrides any earlier output-format wording: After reasoning, return only the requested conclusion as plain text and do not add commentary after it.	Decrease
Constrained drug-response prediction 5,000 samples	Omics instruction	<omics> This RNA expression profile is from the OVCAR-8 cell line (Ovary/Fallopian Tube, Ovarian Epithelial Tumor). In the GDSC screen, the compound is ELESCLOMOL. Its annotated targets are HSP90. Predict the source's final sensitivity label, assigned by its compound-specific GMM/Z-score procedure and corrections. Is [. . .] contract for this variant; it overrides any earlier output-format wording: After reasoning, begin the conclusion with the exact prefix "FINAL_OMICS_ANSWER: ".	FINAL_OMICS_ANSWER: sensitive
Medical question answering with reasoning 13,109 samples	Biomedical QA	Please answer the following multiple-choice question: The cause of hyperuricemia and gout in glucose-6phosphatase deficiency is :? A. More formation of pentose B. Decreased availability of glucose to tissues C. Increased accumulation of sorbitol D. Impaired degradation of free radicals	Reasoning: Alright, let's think about this. The question is about glucose-6-phosphatase deficiency, which I know causes Von Gierke's disease. This is a [. . .] Final: The chain of thought reasoning concludes that in glucose-6-phosph atase deficiency, the increased formation of pentose through the pentose phosphate pathway is the cause of [. . .]
Open-ended biology instruction following 20,000 samples	Biology QA	What are the common techniques used in identifying a new species, and how can scientists accurately categorize it within the existing taxonomy system?	Identifying a new species and categorizing it within the existing taxonomy system involves several common techniques. These techniques can be broadly classified into morphological, molecular, and ecological approaches. Here's an overview of these techniques: 1. [. . .]
Constrained biology responses 19,503 samples	Biology QA	How does the hypothalamus regulate the secretion of hormones from the anterior pituitary gland? Answer in at most 33 words.	The hypothalamus regulates the secretion of hormones from the anterior pituitary gland through a complex system of hormonal and neural connections. This process involves the following steps: 1.
Biomedical factoid, list, summary, and binary QA 3,773 samples	Biomedical QA	Context: - Here we determine the structure of these novel transcripts, showing that they correspond to circular RNA molecules containing only exons in genomic order. - Thousands of loci in the human and mouse genomes give rise to circular RNA transcripts; at many of these loci, the predominant RNA isoform is a circle. [. . .] we review the breadth of circular RNAs, their biogenesis and metabolism, and their known and anticipated functions. Q: What is the function of circular RNA?	Circular RNAs (circRNAs) are a novel type of RNA that, unlike linear RNAs, form a covalently closed continuous loop and are highly represented in the eukaryotic transcriptome. The biogenesis of circular RNA is an integral, conserved, and regulated feature of the gene [. . .]
General capabilities (12 groups; 505,514 samples)			
Science multiple-choice QA 3,365 samples	Science QA	Which of these is the chemical symbol for magnesium? A. Ma B. Mg C. Mn D. Mu Answer with the letter only.	B

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Multidomain multiple-choice QA 99,735 samples	General QA	A leading question is LEAST likely to be permitted over objection when A. asked on cross-examination of an expert witness. B. asked on direct examination of a young child. C. asked on direct examination of a disinterested eyewitness. D. related to preliminary matters such as the name or occupation of the witness Answer with the letter only.	C
Mathematical reasoning 97,717 samples	Mathematics	Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?	Reasoning: Natalia sold 48/2 = 24 clips in May. Natalia sold 48+24 = 72 clips altogether in April and May. Final: 72
Code instruction following 74,999 samples	Code	Write a function to generate a prime number sequence using Python.	Sure, here is a basic python function to generate a sequence of prime numbers up to a given number. Note that the ‘is_prime‘ helper function is used to check if a number is prime. “‘python def is_prime(n): if n <= 1: return False if n == 2: return True if n % 2 == 0: [. . .]
Code problem solving with reasoning 15,000 samples	Code	Ignoring the air resistance, velocity of a freely falling object v after t seconds and its drop y in t seconds are represented by the following formulas: $ v = 9.8 t $ $ y = 4.9 t^2 $ A person is trying to drop down a glass ball and check whether it will crack. Your task is to write a program to help this [. . .] of datasets is less than or equal to 50. Output For each dataset, print the lowest possible floor where the ball cracks. Example Input 25.4 25.4 Output 8 8	Reasoning: Okay, let’s see. I need to solve this problem where a glass ball is dropped from a certain floor of a building, and I have to find the lowest [. . .] Final: “‘python import math import sys for line in sys.stdin: v = float(line.strip()) h_min = (v ** 2) / 19.6 required_n = (h_min + 5) / 5 print(math.ceil(required_n)) “‘
General constraint following 49,424 samples	Instruction	Can you provide me ideas of how an association of gardening centers may attract young people for apprenticeships or traineeships? Your answer must mention at least one historical event.	To attract young people for apprenticeships or traineeships, an association of gardening centers could consider hosting events that highlight the historical significance of gardening and its evolution over time. For instance, they could organize workshops or seminars [. . .]
General instruction and dialogue 66,734 samples	Dialogue	How do you know if an argument you come across makes sense logically? Can you walk me through the process you would use to break down an argument and understand it?	To determine whether an argument makes sense logically, I follow a process of breaking it down into its core components, analyzing the relationships between them, and evaluating the evidence provided. Here’s how I would walk through this process: When I encounter an [. . .]
Text rewriting 30,000 samples	Text	Hey Sarah, Hope you’re doing well. So, I’ve been getting a bunch of complaints from the users about their browsers being super slow. It’s driving them nuts! I’ve tried a few things but nothing seems to be working. Could you take a look at it? It’s kinda urgent. Cheers, Alex	Dear Sarah, I hope this message finds you well. I am writing to inform you of a significant issue that has been reported by multiple users regarding the unusually slow performance of their browsers. This problem is causing considerable frustration and impacting [. . .]
Tool invocation and dialogue 29,066 samples	Tool use	Hi, can you tell me the current price of Apple’s stock?	<tool_call> {"name": "get_stock_price", "arguments": {"stock_symbol": "AAPL"}} </tool_call>
Irrelevant-tool rejection 5,495 samples	Tool use	Is 49 a perfect square?	None of the provided tools is relevant to this request, so I will not call one.
Tool-use decision making 13,979 samples	Tool use	Find LinkedIn profiles.	To assist you better, could you please specify the name or keyword for the search query?

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
System-guided multiturn dialogue 20,000 samples	Dialogue	System: You are an AI bot that walks users through the process of booking a flight online. User: I've been having some issues with a close friend recently and we ended up having a falling out. I really want to mend our relationship, but I'm not sure how to go about it. Can you provide some advice on how to approach this situation and what steps I can take to repair our friendship?	I'm an AI designed to assist with booking flights and travel information, so my abilities are limited when it comes to providing personal or relationship advice. However, I can suggest you speak with a counselor, a trusted friend or family member, or consider searching [. . .]
Scientific and biomedical text			
PubMed abstract	Free text	Investigation of the carbohydrate metabolism of normal and neoplastic hepatocytes using 2, 6-dichlorophenolindophenol as a probe for NAD(P)H production measured by voltammetry. An electrochemical technique is described for measurement of intracellular NAD(P)H production. This technique involves an auxiliary redox system taken up by the cells which is then measured voltammetrically after reduction by NAD(P)H. The redox syst. . .	
PubMed Central article	Free text	abstract: # Objectives . The general aim was to assess the pattern and trend in cancer mortality among the indigenous people of coastal Chukotka during the period 1961–1990. . # Methods . All cases of cancer deaths of indigenous residents of the Chukotsky district in the north-easternmost coast of Chukotka Autonomous Okrug were copied from personal death certificates. There were a total of 219 cancer deaths during the stu. . .	
arXiv paper	Free text	\{}section{Introduction} Raychaudhuri equation has been introduced by A.Raychaudhuri \{}cite{Ray}. It is a fundamental equation in gravitation and cosmology and has been studied and generalized in many ways, in different cases \{}cite{ehlers,das,ellis,mohajan,hot,dad, kar1,kar2,lobo,k-ts,DDas,Abreu,mohseni1,Alba2,santos,ahmadi,moh,ra1,ra2,kar3}. In Finsler space-time this equation was introduced in \{}cite{st1}, \{}cite{st2}, \{}cite{. . .	
bioRxiv preprint	Free text	Exfoliating Bark Does Not Protect Platanus occidentalis L. From Root-Climbing Lianas Abstract Lianas are structural parasites that depress growth, fertility and survival rates of their hosts, but the magnitude to which they alter these rates differ among host species. We tested the hypothesis that sycamore (Platanus occidentalis L.) would have fewer adventitious root-climbing lianas. We reasoned that because P. occidental. . .	
Clinical trial record (AACT)	Free text	ClinicalTrials.gov trial NCT02232178; AACT snapshot 2026-09-01 A Randomized, Blinded, Inpatient Study to Investigate the PK, Relative Bioavailability and Safety of 2 Doses of XaraColl (200 and 300 mg Bupivacaine HCl) Compared to Bupivacaine HCl Infiltration (150 mg) After Open Laparotomy Hernioplasty Trial registration study type: INTERVENTIONAL; overall status: COMPLETED; phase: PHASE2; enrollment: 64; enrollment type: A. . .	
Drug label (DailyMed)	Free text	Dermacline Label type: HUMAN OTC DRUG LABEL Marketing categories: OTC Monograph Final (legacy) Effective date: 2016-11-08 Source: DailyMed, National Library of Medicine. DailyMed set ID: 2601650a-1870-4fe5-9931-842fb2381c94 Active Ingredient (in each gram) Tetracycline Hydrochloride 30mg Purpose First Aid antibiotic Use First aid to help prevent skin infection in minor cuts, scrapes, and burns. Warnings For esternam use o. . .	
Grant abstract (NIH RePORTER)	Free text	The American Society of Biomechanics is the primary scientific organization for biomechanics in the US. The annual conference this year will be held at the Pennsylvania State University on August 26-29, 2009. The meeting will include presentations on a wide range of topics including aging, biofluids, comparative biomechanics, cellular mechanics, clinical gait analysis, dental, ergonomics, finite element modeling, forensic. . .	
Biomedical code (ChemPile)	Code	!/usr/bin/env python import sys import argparse import os.path Time of script execution and logging module import time import logging import re import itertools from collections import * import random import multiprocessing as mp from Bio.Blast import NCBIXML def main(): args = create_argparser() #Get files to process files_to_process = get_files_to_process(args) sys.stderr.write("Files to process:\{}n\{}t{}\{}n".format("\{}n\{}t"... .	
Encyclopedic text (Wikipedia)	Free text	Harvest gold is a shade of orange and yellow. It was popular with kitchen and other appliances during the 1970s, along with brown, burnt orange, and avocado green. References Shades of yellow	
Biological records and QA			
TheBioCollection record	Free text	The molecule has the SMILES representation: <smiles>CN1C(=O)CC[C@]12CCCN(CC2)C3=NC=CN=C3</smiles> and its IUPAC systematic name is (5S)-1-methyl-9-pyrazin-2-yl-1, 9-diazaspiro[4.6]undecan-2-one. It also has the corresponding InChI: <inchi>InChI=1S/C14H20N4O/c1-17-13(19)3-5-14(17)4-2-9-18(10-6-14)12-11-15-7-8-16-12/h7-8,11H,2-6, 9-10H2,1H3/t14-/m0/s1</inchi>. The compound contains 39 atoms, including 19 heavy atoms, and 41 b. . .	
TheBioCollection instruction	Instruction	You are given a human GRCh38 genomic DNA sequence, taken from the genome forward (+) strand, containing one annotated exon-intron boundary from GENCODE. The cut points are splice sites, with a donor at the start of an intron and an acceptor at its end. The motif fields below. . .	{"splice_sites":[{"site_type": "acceptor","strand":"-","position": 151,"motif_start":151,"motif_end": 152,"subsequence":"CT"}]}

(continued on next page)

(continued from previous page)

Task	Type	Question	Answer
Medical QA (ReasonMed)	Text QA	Please answer the following multiple-choice question: 'Scissors bite' is a term used to describe? A. Anterior posterior crowing with deep bite B. Posterior cross bite C. The type of cross bite where maxillary segments contains full mandibular segments D. A type of anterior c…	<think> Okay, let's dive into this. So, we're talking about 'scissors bite.' I've heard that term before in orthodontics. It has something to do with…
Medical reasoning (Medical-R1)	Reasoning	A 34-year-old man with a history of erratic behavior consistent with manic episodes and who showed improvement with lithium treatment is being considered for discharge from a psychiatric hospital after his symptoms resolve. Given his medical history and treatment response, w…	<think>Okay, so I need to figure out the long-term management strategy for lithium therapy for this 34-year-old man with a history of manic episodes.…
Relation extraction			
Document relations (BioRED)	Relation extraction	Extract the normalized biomedical entities and document-level relations from this PubMed title and abstract. Group unique surface forms that share a database identifier into one entity. Use null as the id only when the text has no normalized identifier. Relations are undirec…	{"entities":[{"id":"D006965", "type":"DiseaseOrPhenoty picFeature","mentions": ["parathyroid hyperplasia"] },{"id":"D009369","type": "DiseaseOrPhenotypicFe…
Chemical–protein relations (ChemProt)	Relation extraction	Extract the chemical–gene/protein relations supported by this article, using the supplied mention IDs. Offsets are zero-based, end-exclusive character positions in the article (title, newline, abstract). GENE-Y and GENE-N both denote gene/protein mentions; Y/N indicates norm…	{"relations":[]}
Marked-entity relation (BioREx)	Relation extraction	Extract the two marked biomedical entities and decide whether the text states a document-level relation between them. Relations are undirected. Entity 1 id: T6 Entity 1 type: ChemicalEntity Entity 2 id: T26 Entity 2 type: GeneOrGeneProduct Relation types: Association, Bind,…	{"entities":[{"id":"T6","type": "ChemicalEntity","mentions": ["indisulam"]},{"id":"T26","type": "GeneOrGeneProduct","mentions": ["hCA VI"]}],"relations":[…
Mechanism and perturbation			
Drug target from profiles (DrugMechDB × Tahoe)	Drug mechanism	the cells profiled before treatment: <omics> and again after an undisclosed compound was applied at 0.05 µM: <omics> What does it bind? Answer with gene symbols only, comma separated.	TYMS
General instructions			
Instruction conversation (Open PerfectBlend)	Instruction	Daniel practices basketball for 15 minutes each day during the school week. He practices twice as long each day on the weekend. How many minutes does he practice during a whole week?	Daniel practices for 15 minutes each day during the school week, which typically consists of 5 days (Monday to Friday). So for the school week, he pra…

A.11 Asset Provenance, Licenses, and Terms of Use

Table 18 summarizes the sources, versions, licenses or terms reported by upstream providers, and uses of third-party assets in this work. Our released artifacts are described in Appendix A.1.

GTEx acknowledgement. The GTEx data used in this work were obtained from GTEx Analysis Release V10 (dbGaP Accession phs000424.v10.p2).

DepMap provenance. DepMap 24Q4 expression and dependency data provide cell-line profiles, gene-essentiality labels, and genomic annotations. The release's stated license and the DepMap portal terms are recorded separately from the licenses of other datasets accessed through the portal, such as PRISM. The raw DepMap files are *not* redistributed in our released artifacts.

Table 18: Third-party assets used in OmicsLM and in evaluation.

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
Expression datasets					
Human Cell Atlas	CellxGene Census 2025-01-30	Chan Zuckerberg Initiative	(Regev et al., 2017; Chan Zuckerberg Initiative, 2025); https://cellxgene.cziscience.com/	CC BY 4.0	Primary training source for single-cell annotation tasks.
Tabula Sapiens	GEO: GSE201048	The Tabula Sapiens Consortium	(The Tabula Sapiens Consortium, 2022); https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201048	CC BY 4.0	Held-out evaluation benchmark (cell-type annotation); not used in training.
GTEx	v10; dbGaP phs000424.v10.p2	NIH Common Fund	(GTEx Consortium, 2020); https://gtexportal.org/	GTEx Portal terms of use	Tissue-of-origin prediction (training). Acknowledgement given in this appendix.
TCGA	GDC Open Access	NCI / NHGRI	(The Cancer Genome Atlas Research Network et al., 2013); https://portal.gdc.cancer.gov/	NIH Genomic Data Sharing Policy	Clinical & bulk prediction tasks (training). Open access data only; no controlled-access data used.
GEO (via ARCHS4 pipeline)	ARCHS4 pipeline Apache 2.0	NCBI / NLM	(Clough et al., 2024; Lachmann et al., 2018); https://www.ncbi.nlm.nih.gov/geo/	NCBI/GEO source terms; ARCHS4 pipeline: Apache 2.0	GEO-OmicsQA benchmark expression profiles and GEO sample-description training tasks.
Human Diseases	—	Schaefer et al. (CellWhisperer)	(Schaefer et al., 2025); https://cellwhisperer.bocklab.org/	CC BY 4.0	Held-out disease-classification evaluation; samples overlapping training removed.
Replogle et al. 2022	scPerturb release	Replogle et al.	(Replogle et al., 2022); https://doi.org/10.25452/figshare.plus.20029387; https://doi.org/10.5281/zenodo.13350497	CC BY 4.0 (Figshare+ and scPerturb depositions)	Perturbation-response training tasks (Perturb-seq).
Nadig et al. 2025	—	Nadig et al.	(Nadig et al., 2025); https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE264667	GEO source terms	Perturbation-response training tasks (Perturb-seq).
DepMap Expression	24Q4	Broad Institute	(Broad DepMap, 2024; Pacini et al., 2024); https://depmap.org/portal/	CC BY 4.0; DepMap portal terms	Cell-line expression profiles for gene essentiality and genomics tasks.
DepMap Gene Dependency	24Q4	Broad Institute	(Broad DepMap, 2024); https://depmap.org/portal/	CC BY 4.0; DepMap portal terms	CRISPR gene dependency labels for essentiality prediction tasks.
X-Atlas / Orion (HCT116)	—	Huang et al.	(Huang et al., 2025); https://doi.org/10.25452/figshare.plus.29190726	**CC BY-NC-SA 4.0** (non-commercial)	Genetic-perturbation response training data. Non-commercial research use only.
Additional expression and perturbation sources					
3CA	Retained study collection	Weizmann Institute / study authors	(Gavish et al., 2023); https://www.weizmann.ac.il/sites/3CA/	Study-specific terms; GEO source terms;	Cancer single-cell characterization and expression tasks.

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
CancerModels.org	Merged TPM snapshot	EMBL-EBI / model providers	(Perova et al., 2023); https://www.cancermodels.org/	EMBL-EBI and provider-specific terms; CC0 1.0	Cancer-model characterization;
Tahoe-100M	Pseudobulk subset	Tahoe Bio	(Zhang et al., 2025b); https://huggingface.co/datasets/tahoebio/Tahoe-100M	CC0 1.0	Chemical-perturbation response and dose comparisons.
sci-Plex3	scPerturb release	Srivatsan et al. / scPerturb	(Srivatsan et al., 2020); https://doi.org/10.5281/zenodo.13350497	CC BY 4.0 (scPerturb deposition)	Chemical-perturbation tasks from pseudobulk profiles.
sci-Plex GxE	GSE225775; GSM7056149	McFaline-Figueroa et al.	(McFaline-Figueroa et al., 2024); https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE225775; https://www.ncbi.nlm.nih.gov/home/about/policies/	NCBI molecular-data usage policy	Genetic–chemical perturbation tasks from pseudobulk profiles.
Jiang et al.	Zenodo 14518762	Jiang et al.	(Jiang et al., 2025); https://doi.org/10.5281/zenodo.14518762	CC BY 4.0 (data deposition)	Genetic-perturbation response tasks.
Feng et al.	Figshare 27989294	Feng et al.	(Feng et al., 2026); https://doi.org/10.6084/m9.figshare.27989294	MIT (count-data deposition)	Genetic-perturbation response tasks.
Virtual Cell Challenge	H1 hESC training data; July 2025 snapshot	Arc Institute	https://github.com/ArcInstitute/arc-virtual-cell-atlas/tree/main/virtual-cell-challenge	CC0 1.0	Genetic-perturbation profiles.
CellWhisperer conversations	Profile-linked subset	Schaefer et al.	(Schaefer et al., 2025); https://cellwhisperer.bocklab.org/	No explicit license; publicly released by the authors as LLM-curated training data (paper data availability statement)	Expression-grounded conversation training; distinct from the held-out Human Diseases dataset. Used for research only; not redistributed.
PerturbQA	Retained training examples	Wu et al.	(Wu et al., 2025b;c); https://openreview.net/forum?id=5WEpbilssv https://github.com/Genentech/PerturbQA/tree/main	CC BY 4.0 (published dataset)	Perturbation detection and response reasoning; released baseline predictions (`results.zip`) rescored for evaluation.
Screens, drug response, and human genetics					
AssayBench	Revision `988daee`	Genentech	(De Brouwer et al., 2026); https://huggingface.co/datasets/Genentech/assaybench	MIT (dataset release)	Ranking and identifying phenotypic-screen hits.
GDSC	Retained response tables	Wellcome Sanger Institute / collaborators	(Yang et al., 2013); https://www.cancerrxgene.org/	GDSC release 8.5 data-use terms	Drug sensitivity and IC50 prediction.
PRISM	Repurposing 19Q4; primary / secondary screens	Broad Institute	(Corsello et al., 2020); https://doi.org/10.6084/m9.figshare.9393293.v4	CC BY 4.0 (19Q4 deposition); DepMap portal terms	Sensitivity and secondary-screen activity prediction.
DrugMechDB	2.0.1	Su laboratory / contributors	(Gonzalez-Cavazos et al., 2023); https://github.com/SuLab/DrugMechDB	CC0 1.0	Mechanism paths and counterfactual reasoning after benchmark exclusions.

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
Open Targets	Platform 26.06; derived v2 export	Open Targets consortium	(Ghoussaini et al., 2021); https://platform-docs.opentargets.org/licence	Platform data: CC0 1.0; retain source attribution	Gene–disease, variant, and GWAS-locus tasks.
Biological databases and ontologies					
GENCODE	—	EMBL-EBI / Wellcome Sanger Institute	(Mudge et al., 2025); https://www.gencodegenes.org/	CC BY 4.0	Defines the 19,260-gene panel.
Gene Ontology	2025-06-01	GO Consortium	(Ashburner et al., 2000; Gene Ontology Consortium, 2026); https://geneontology.org/	CC BY 4.0	Gene-set annotations for pathway reasoning tasks (training).
Cellosaurus	—	SIB Swiss Institute of Bioinformatics	(Bairoch, 2018); https://www.cellosaurus.org/	CC BY 4.0	Cell-line annotations (disease, lineage, transformant) for genomics tasks (training).
UniProt (Swiss-Prot)	—	UniProt Consortium	(The UniProt Consortium, 2025); https://www.uniprot.org/	CC BY 4.0	Protein function, localization, and keyword annotations for protein QA tasks (training).
BioReason-Pro SFT Reasoning Data	Retained SFT subset	WangLab / Fallahpour et al.	(Fallahpour et al., 2026) https://huggingface.co/datasets/wanglab/bioreason-pro-sft-reasoning-data	Apache 2.0 (dataset card)	15,000 protein-function reasoning conversations for Stage 2.
STRING	12.0	SIB / EMBL	(Szklarczyk et al., 2023); https://string-db.org/	CC BY 4.0	Protein–protein interaction graph for network reasoning tasks (training).
Human Protein Atlas	Retained RNA and protein tables	Human Protein Atlas	(Karlsson et al., 2021; Uhlén et al., 2015) https://www.proteinatlas.org/about/licence	CC BY 4.0	Cell, immune and tissue evidence for marker-grounded reasoning data.
TISSUES	Human knowledge associations	JensenLab	(Palasca et al., 2018) https://tissues.jensenlab.org/About	CC BY 4.0	Tissue–gene evidence for marker-grounded reasoning data.
Reactome	Retained pathway programmes	Reactome	(Milacic et al., 2024) https://reactome.org/license	CC0 (pathway data)	Published pathway programmes used in marker evidence cards.
HGNC	Complete gene set	HGNC	(Seal et al., 2023); https://www.genenames.org/about/license/	CC0 1.0	Approved symbols, aliases and chromosome annotations for marker-grounded reasoning data.
Cell Ontology	Retained full OBO release	Cell Ontology contributors	https://obofoundry.org/ontology/cl.html	CC BY 4.0	Cell-type names, synonyms and hierarchy for label normalization.
Pathway and graph-reasoning training sources					
BioMaze (KEGG pathway data)	GraphML; revision 4c22f4d	Zhao et al.	(Zhao et al., 2025a;b); https://github.com/zhao-ht/BioMaze/tree/4c22f4da0d7ed544f8495be3fdc260814951fc14; https://www.kegg.jp/kegg/legal.html	KEGG terms of use	KEGG-derived graph for reasoning training.

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
BioReason KEGG	Training split	WangLab / Fallahpour et al.	(Fallahpour et al., 2025) https://huggingface.co/datasets/wanglab/kegg	Apache 2.0 (dataset card)	1,159 pathway-context reasoning conversations for Stage 2.
MSigDB	Hallmark (H) and GO:BP (C5) gene sets	Broad Institute / UC San Diego	(Liberzon et al., 2015); https://www.gsea-msigdb.org/gsea/msigdb/	CC BY 4.0	Gene sets for pathway-activity tasks and drug-response context programmes (training).
Biomedical QA and information extraction					
BioASQ	Task 11b training set	BioASQ consortium	(Krithara et al., 2023); https://participants-area.bioasq.org/datasets/	CC BY 2.5	Biomedical question answering.
ReasonMed / CoTMed / ResponseMed	Retained text exports	Sun et al.	(Sun et al., 2025); https://huggingface.co/datasets/lingshu-medical-mllm/ReasonMed/tree/f0eb606624bbbd6be88901f400eeec58fb19f700	Apache 2.0 (card covers ReasonMed, CoTMed and ResponseMed)	Medical answers and reasoning traces serialized as text.
Medical-R1-Distill-Data	English export	Freedom Intelligence	(Chen et al., 2025); https://huggingface.co/datasets/FreedomIntelligence/Medical-R1-Distill-Data	Apache 2.0	Medical QA and distilled reasoning.
MedReason	UCSC-VLAA export	Wu et al.	(Wu et al., 2025a); https://huggingface.co/datasets/UCSC-VLAA/MedReason	Apache 2.0 (export card)	Medical reasoning text.
M23K	Text export	UCSC-VLAA	(Huang et al., 2026; Thapa et al., 2025); https://huggingface.co/datasets/zou-lab/BioMed-R1-Train	MIT (dataset release)	Medical reasoning text
TheBioCollection	Revision df09ee3	Trillion Labs	(Seo et al., 2026); https://huggingface.co/datasets/trillionlabs/TheBioCollection	Apache 2.0 (corpus card)	Free-text and instruction streams; retain underlying resource provenance.
CAMEL Biology	Biology subset	CAMEL-AI	(Li et al., 2023); https://huggingface.co/datasets/camel-ai/biology	CC BY-NC 4.0	Biology instruction examples.
BioRED	Checksum-pinned archive	NCBI / original authors	(Luo et al., 2022); https://github.com/ncbi/BioRED	No explicit archive license; article-specific terms	Entity/relation training and held-out evaluation. Restricted to articles whose licenses permit non-commercial reuse.
BioREx: DrugProt / CID	Training components only	NCBI / component authors	(Lai et al., 2023); https://github.com/ncbi/BioREx	DrugProt: CC BY 4.0; CID: public-domain mark; aggregate archive: no explicit license	Relation-classification training; other BioREx components excluded. Restricted to articles whose licenses permit non-commercial reuse.

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
ChemProt	BioCreative VI; BigBio revision `25efb1a`	BioCreative organizers	(Krallinger et al., 2017); `https://huggingface.co/datasets/bigbio/chemprot/tree/25efb1a752ba3775795742f07502f6dd214a18eb`	Public Domain Mark 1.0 (BigBio dataset metadata)	Chemical–protein relations; official train/dev/test separation retained.
Scientific and biomedical text and code					
PubMed abstracts	Retained text snapshot	NLM / article rightsholders	`https://pubmed.ncbi.nlm.nih.gov/`; `https://www.ncbi.nlm.nih.gov/home/about/policies/`	NLM source terms and abstract copyrights	Biomedical text; no blanket public-domain designation for abstracts.
PubMed Central	Original and enriched subsets	NLM / article rightsholders	`https://pmc.ncbi.nlm.nih.gov/tools/openftlist/`	Article-specific terms; enriched subset selected using CC BY / CC0 evidence	Biomedical full text. Restricted to articles whose licenses permit non-commercial reuse.
arXiv	Retained text snapshot	arXiv / article authors	`https://info.arxiv.org/help/license/`	Article-specific licenses	Scientific text. Restricted to articles whose licenses permit non-commercial reuse.
bioRxiv / medRxiv	Full-text XML exports	Preprint authors / servers	`https://www.biorxiv.org/`; `https://www.medrxiv.org/`	Article-specific licenses	Preprint text with captions and tables; source license metadata retained by the exporter.
AACT / ClinicalTrials.gov	2026-09-01	CTTI / NLM / study sponsors	`https://aact.ctti-clinicaltrials.org/`; `https://clinicaltrials.gov/about-site/terms-conditions`	AACT and ClinicalTrials.gov terms	Clinical-trial records rendered as text.
DailyMed	2026-09-01	NLM / product labelers	`https://dailymed.nlm.nih.gov/dailymed/`; `https://www.nlm.nih.gov/web_policies.html`	NLM terms and source-label rights	Drug-label text.
NIH RePORTER	2026-03-09	NIH / grant recipients	`https://reporter.nih.gov/exporter`; `https://www.nih.gov/about-nih/frequently-asked-questions`	NIH website reuse policy	Research-project abstracts used for non-commercial research.
ChemPile-Code	Revision `50f1009`; CodeParrot Python	Jablonka group / code authors	(Mirza et al., 2025); `https://huggingface.co/datasets/jablonkagroup/chempile-code`	Collection: AGPL-3.0; repository-license-label filter: MIT, Apache 2.0, BSD-2/3-Clause	Biomedical Python text; original code, notices, and both license levels retained.
General-purpose text and instruction data					
Wikipedia via RedPajama	Retained Wikipedia export	Wikimedia contributors / Together AI	(Weber et al., 2024); `https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T`; `https://huggingface.co/datasets/legacy-datasets/wikipedia`	CC BY-SA 3.0 (upstream Wikipedia card); GFDL where applicable	General text from the RedPajama Wikipedia slice (2023-03-20 dump).
Open PerfectBlend	Revision `af60f3c` and selected components	Maxime Labonne / source authors	(Labonne, 2024); `https://huggingface.co/datasets/mlabonne/open-perfectblend`	Apache 2.0 collection; source attribution retained	Conversations, including AutoIF, LMSYS, and UltraChat components.

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
GSM8K / MetaMathQA	Selected training examples	OpenAI / MetaMath authors	(Cobbe et al., 2021; Yu et al., 2024); https://huggingface.co/datasets/openai/gsm8k; https://huggingface.co/datasets/meta-math/MetaMathQA	MIT (both dataset cards)	Mathematical reasoning.
OpenMathReasoning	Selected training examples	NVIDIA	(Moshkov et al., 2025); https://huggingface.co/datasets/nvidia/OpenMathReasoning	CC BY 4.0	Mathematical reasoning traces.
ARC / MMLU	Selected training examples	Allen Institute for AI / MMLU authors	(Clark et al., 2018; Hendrycks et al., 2021); https://huggingface.co/datasets/allenai/ai2_arc; https://huggingface.co/datasets/cais/mmlu	ARC: CC BY-SA 4.0; MMLU: MIT (dataset cards)	General knowledge questions.
Magicoder-Evol-Instruct	110K; revision b0079be	Wei et al.	(Wei et al., 2024); https://huggingface.co/datasets/ise-uiuc/Magicoder-Evol-Instruct-110K	Apache 2.0	Coding instructions.
OpenCodeReasoning	Revision 20a1ca1	NVIDIA	(Ahmad et al., 2025); https://huggingface.co/datasets/nvidia/OpenCodeReasoning	CC BY 4.0; original question-source terms	Competitive-programming reasoning traces.
SmolTalk	Revision 5feaf2f; selected components	Hugging Face / source authors	(Ben Allal et al., 2025); https://huggingface.co/datasets/HuggingFaceTB/smoltalk	Constraints, rewriting, Magpie and System-Chats: Apache 2.0; Self-OSS-Instruct: ODC-BY 1.0	Instruction following, rewriting, coding and system-prompt adherence.
Hermes / ToolACE	Selected function-calling examples	Nous Research / ToolACE authors	https://huggingface.co/datasets/NousResearch/hermes-function-calling-v1; https://huggingface.co/datasets/Team-ACE/ToolACE	Apache 2.0 (both dataset cards)	Tool-use conversations.
When2Call	Selected training examples	NVIDIA	https://huggingface.co/datasets/nvidia/When2Call	CC BY 4.0	Tool-use decisions and responses.
xLAM irrelevance	7.5K; revision 34323bf	MadeAgents	https://huggingface.co/datasets/MadeAgents/xlam-irrelevance-7.5k	CC BY 4.0	Recognizing when available tools do not answer a request.
Derived instruction components	Retained SFT export	Original component authors	Appendix A.2.23; parent assets above	CAMEL-derived constraints: CC BY-NC 4.0; omics variants: parent-source terms	Constraints, multi-turn conversations, and reformatted omics instructions. BioReason-Pro is listed separately.
UltraData-SFT-Agent	2609; revision f684cc1	MiniCPM Team	(MiniCPM Team, 2026a); https://huggingface.co/datasets/openbmb/UltraData-SFT-Agent-2609	Apache 2.0 (dataset card)	Agent trajectories and tool interactions.

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
Foundation models and backbones					
Funomics T0	—	Powalski et al.	(Powalski et al., 2026)	Explicit permission from the authors for use in this work	512-dimensional embeddings used directly as part of every omics input vector, obtained from the authors on request.
Geneformer-V2-104M	104M	Chen et al.	(Chen et al., 2026)	Apache 2.0	Pretrained single-cell encoder; embeddings used as input features.
Qwen3.5	0.8B, 2B, 9B	Qwen Team	(Qwen Team, 2026a); `https://huggingface.co/Qwen/Qwen3.5-9B`	Apache 2.0	Pretrained text backbone; jointly trained with the omics projection. Also evaluated as a GEO-OmicsQA baseline. 0.8B and 2B backbones for ablations.
Gemini 3.5 Flash-Lite	API	Google	(Google DeepMind, 2026c); `https://ai.google.dev/gemini-api/terms`	Gemini API terms	Mechanism-of-action data generation (Appendix A.4).
Qwen3.8-27B	27B	Qwen Team	(Qwen Team, 2026c) `https://huggingface.co/Qwen/Qwen3.8-27B`	Apache 2.0	Question and rationale generation for marker-grounded reasoning data. Also evaluated as a GEO-OmicsQA baseline.
Gemma-4-31B-it	31B	Google	(Gemma Team et al., 2026) `https://huggingface.co/google/gemma-4-31B-it`	Apache 2.0	Quality grading of marker-grounded rationales.
Gemini 3.1 Flash-Lite	API	Google	(Google DeepMind, 2026a); `https://ai.google.dev/gemini-api/terms`	Gemini API terms	DeepEval judge model for GEO-OmicsQA free-text scoring.
Baseline models					
CellWhisperer	—	Schaefer et al.	(Schaefer et al., 2025); `https://cellwhisperer.bocklab.org/`	CC BY 4.0	Evaluation baseline: we evaluate the released chat model (LM rows); CLIP rows are taken from the published paper. Profile-linked training conversations are listed separately above.
C2S-Scale-27B	—	Rizvi et al.	(Rizvi et al., 2025); `https://huggingface.co/vandijklab/C2S-Scale-Gemma-2-27B`	CC BY 4.0	Evaluation baseline (zero-shot annotation, GEO-OmicsQA).
Cell-o1	—	Fang et al.	(Fang et al., 2026)	No standard license (model card: `other`); NLM disclaimer	Evaluation baseline (zero-shot annotation, GEO-OmicsQA).
TxGemma-27B	27B	Google	(Wang et al., 2025)	Gemma Terms of Use	Evaluation baseline (GEO-OmicsQA).
Gemini 3.1 Pro / 3.8 Flash	API	Google	(Google DeepMind, 2026b;d); `https://ai.google.dev/`	Google APIs Terms of Service	Evaluation baseline (GEO-OmicsQA).
GPT-6 Sol / Luna	API	OpenAI	(OpenAI, 2026); `https://platform.openai.com/`	OpenAI Terms of Service	Evaluation baseline (GEO-OmicsQA).
AROMA	HF revision `19ea7d7`	Wang et al.	(Wang et al., 2026); `https://huggingface.co/blazerye/AROMA`	No explicit license (model card and repository)	PerturbQA baseline reproduced from the released checkpoint (evaluation only).
Software frameworks					

continued on next page

Table 18 continued from previous page

Asset	Version	Owner / Maintainer	Citation & URL	License / Terms	Use in this work
DeepEval	—	Confident AI	https://github.com/confident-ai/deepeval	Apache 2.0	LLM-as-a-judge framework for GEO-OmicsQA free-text scoring.
PDEX	0.1.24	Arc Institute	(Arc Institute, 2025b); https://github.com/ArcInstitute/pdex	MIT	Parallel differential expression analysis for perturbation-response tasks.
VERL	—	Sheng et al. / contributors	(Sheng et al., 2025); https://github.com/verl-project/verl	Apache 2.0	GRPO reinforcement-learning framework (Stage 3).
vLLM	—	vLLM contributors	(Kwon et al., 2023); https://github.com/vllm-project/vllm	Apache 2.0	Inference engine for local-model evaluation.

A.12 Qualitative Examples of Biological Knowledge and Expression Interpretation

We present nine assessment conversations with OmicsLM. Examples 1–4 illustrate quantitative biochemical recall, mechanistic explanation, and comparative interpretation of expression profiles through the same conversational interface. Examples 5–6 infer a genetic dependency and a cancer-model type from expression profiles, and Examples 7–9 interpret profiles from real GEO studies. Each prompt was run once with greedy decoding, no system message, and a fresh conversation history. The two knowledge questions used text alone; for expression questions, the chat interface resolved the named cell lines or sample identifiers to their corresponding 20,541-feature omics profiles. The prompts below reproduce the user-entered text, before automatic profile insertion. Answers preserve the generated wording, with only LaTeX formatting applied.

These are illustrative assessments rather than an aggregate benchmark. Expression statements were checked against the source data: DepMap 24Q4 values for the cell-line comparisons (Examples 3–5), the CancerModels.org expression matrix for Example 6, and our re-processed profiles of the referenced GEO samples (Appendix A.2) for Examples 7–9. In Examples 3–5 the cell-line names were visible to the model; in Examples 6–9 the sample identifiers were replaced by the profiles, but the question text still describes the experimental context. Because no anonymized or swapped-profile controls were run for this checkpoint, these examples do not isolate the contribution of the omics inputs from prior knowledge of the named cell lines or of the described experiments.

A.12.1 Example 1: Quantitative biochemical recall

Question: In the oxidative pentose phosphate pathway, how many NADPH and CO2 molecules are produced per glucose-6-phosphate converted to ribulose-5-phosphate, and is ATP produced directly? Name the two NADPH-producing enzymes and explain why this pathway supports proliferating cancer cells. Keep the answer under 160 words.

Answer: In the oxidative pentose phosphate pathway, one molecule of glucose-6-phosphate is converted to ribulose-5-phosphate, producing two molecules of NADPH and one molecule of CO2. ATP is not produced directly in this pathway. The two NADPH-producing enzymes are glucose-6-phosphate dehydrogenase (G6PD) and 6-phosphogluconate dehydrogenase (6PGD). This pathway supports proliferating cancer cells by providing NADPH, which is essential for maintaining cellular redox balance and supporting biosynthetic reactions necessary for cell growth and proliferation.

Assessment and significance. The answer supplies all requested quantitative facts: two NADPH and one CO_2 per glucose-6-phosphate, no direct ATP production, and both NADPH-producing enzymes (G6PD and 6PGD). It connects these facts to redox maintenance and biosynthesis, illustrating the ability to turn pathway knowledge into a concise explanation of its relevance to proliferating cells.

A.12.2 Example 2: Connecting a cancer mutation to epigenetic mechanisms

Question: What new enzymatic activity does the cancer-associated IDH1 R132H mutant acquire, which oncometabolite does it produce, and how can that metabolite alter DNA and histone methylation and cellular differentiation? Keep the answer under 150 words.

Answer: The cancer-associated IDH1 R132H mutant acquires a new enzymatic activity that converts 2-oxoglutarate to the oncometabolite 2-hydroxyglutarate (2-HG). This metabolite can inhibit the activity of 2-oxoglutarate-dependent dioxygenases, including histone demethylases and DNA demethylases. As a result, 2-HG can alter DNA and histone methylation patterns, leading to changes in gene expression and cellular differentiation. Specifically, 2-HG can inhibit the activity of histone demethylases, such as KDM5C, which are involved in the removal of methyl groups from histone lysine residues. This inhibition can lead to the accumulation of methylated histones and altered chromatin structure, affecting gene expression and cellular differentiation. Additionally, 2-HG can also inhibit the activity of DNA demethylases, such as TET1 and TET2, which are involved in the removal of methyl groups from DNA. This inhibition can lead to the accumulation of methylated DNA and altered gene expression, further contributing to the oncogenic phenotype.

Assessment and significance. The answer connects the mutant enzyme's conversion of 2-oxoglutarate to 2-HG with inhibition of dioxygenases, increased methylation, and altered gene

expression and differentiation. This mechanistic chain is useful for explaining how a metabolic alteration can affect cell state. The core mechanism is correct. Its description of TET1/TET2 as DNA demethylases is functional shorthand for their role as 5-methylcytosine oxidases in DNA demethylation.

A.12.3 EXAMPLE 3: INTERPRETING LUMINAL EPITHELIAL AND MESENCHYMAL PROGRAMS

Question: Compare the baseline expression profiles of MCF7 and MDAMB231 using ESR1, GATA3, FOXA1, CDH1 and VIM. For each gene, state which profile has higher expression. Then explain what the combined pattern suggests about luminal epithelial versus mesenchymal programs. Describe relative expression rather than claiming absolute absence. Keep the answer under 150 words.

Answer: ESR1, GATA3, and FOXA1 show higher expression in MCF7, while CDH1 is elevated in MCF7 and VIM is elevated in MDAMB231. This pattern indicates that MCF7 maintains a luminal epithelial program characterized by hormone receptor signaling and epithelial adhesion, whereas MDAMB231 exhibits a mesenchymal program marked by intermediate filament expression. The contrast between these profiles reflects distinct cellular identities, with MCF7 retaining differentiated epithelial features and MDAMB231 displaying markers associated with mesenchymal transition.

Assessment and significance. All five expression directions match the recorded DepMap values: ESR1, GATA3, FOXA1, and CDH1 are higher in MCF7, whereas VIM is higher in MDAMB231. The answer integrates the markers into a coherent interpretation of luminal epithelial versus mesenchymal programs and describes relative expression without claiming absolute absence. This is useful for translating gene-level comparisons into interpretable cell-state summaries; the baseline comparison supports a mesenchymal-associated state, not an observed transition over time.

A.12.4 EXAMPLE 4: STRUCTURED COMPARISON OF MYELOID AND EPITHELIAL LINEAGES

Question: Compare the baseline profiles of THP1 and A549. Which has higher expression of each of these genes: LYZ, LST1, TYROBP, FCER1G, SPI1, KRT8 and KRT18? Give a table with gene and higher-expression cell line, followed by one sentence interpreting the lineage difference. Do not discuss treatment effects.

Answer:

gene	higher-expression cell line
LYZ	THP1
LST1	THP1
TYROBP	THP1
FCER1G	THP1
SPI1	THP1
KRT8	A549
KRT18	A549

The data delineates a clear myeloid versus epithelial lineage boundary, with THP1 expressing canonical monocyte markers and A549 expressing simple epithelial keratins.

Assessment and significance. All seven expression directions match the recorded DepMap values, and the response follows the requested table-plus-one-sentence format without introducing treatment effects. It combines five myeloid-associated markers with two epithelial keratins to explain the lineage contrast. This example illustrates a compact, auditable gene-level comparison paired with a biological interpretation, a useful format for reviewing sample identity and communicating expression differences.

Examples 5–9 were run with the same checkpoint and greedy decoding, with the reasoning mode enabled; each response reproduces the generated reasoning followed by the concluding line. Cell-line names were resolved to their DepMap profiles as above. Sample identifiers of CancerModels.org specimens and GEO samples were replaced by the corresponding omics profiles and removed from the text, so the model received no textual cue about the sample beyond the wording of the question. Training exposure differs between the examples: the cell lines in Example 5 are excluded from all gene-essentiality training data, although their profiles appear in other cell-line tasks; the

CancerModels.org specimen in Example 6 is part of the training data for that source; and the studies in Examples 7–9 come from the publication-disjoint test split of GEO-OmicsQA. All expression values quoted in the assessments are TPM from the source matrices.

A.12.5 Example 5: Inferring an oncogene dependency from receptor overexpression

Question: We are considering HER2-directed therapy for oesophageal squamous carcinoma. Given the profiles of TE4 and KYSE150, is ERBB2 essential for the survival of each line? Point to the feature in the profiles that best explains the difference, and say what it implies about the underlying genomic event.

Answer: The expression data reveals a stark contrast in ERBB2 abundance between the two lines. KYSE150 exhibits only baseline transcript levels, whereas TE4 displays a massive upregulation of the receptor. This profound overexpression in TE4 implies that the cell has acquired a genomic alteration, such as an amplification, driving a heavy reliance on HER2 signaling. Consequently, ERBB2 is likely essential for TE4 survival, while it remains non-essential in KYSE150.
Concluding line: Yes

Assessment and significance. Both dependency calls match DepMap 24Q4: ERBB2 is essential in TE4 (dependency probability 1.00, Chronos -2.02) and not in KYSE150 (0.06, Chronos -0.16), and ERBB2 expression differs about 67-fold (2,045 versus 30 TPM). The inferred amplification is supported by the profile itself: neighbouring genes of the 17q12 amplicon are co-elevated only in TE4 (MIEN1 1,524, GRB7 495, PGAP3 463 TPM, versus 15–44 TPM in KYSE150), although the answer does not cite them. Both lines are held out from the essentiality training split. The example links receptor abundance to oncogene addiction and to a probable copy-number event within one tissue type, the reasoning needed to prioritize HER2-directed therapy.

A.12.6 Example 6: Distinguishing a patient tumour from xenograft-derived material

Question: Specimen 216738-123-R comes from the same renal cell carcinoma model. Classify it as a patient tumour, a xenograft, an organoid or a cell line, and justify the call in three bullet points naming immunological cell markers, antigens and other relevant markers

Answer:

- SPP1, HLA-DRA, and C1QB are detected, providing a combined signal of myeloid-associated antigens and complement components that is characteristic of patient-derived tumour material.
- The presence of HLA-DRB1 and C1QA further supports an immune-active microenvironment typical of human patient specimens, rather than the limited transcriptomic scope of cell lines or organoids.
- This specific constellation of markers aligns with the profile of a patient tumour, distinguishing it from the more restricted expression patterns expected in xenografts, organoids, or cell lines.

Concluding line: patient tumour

Assessment and significance. The classification is correct: the specimen is the patient tumour of a CancerModels.org renal cell carcinoma model. Four of the five cited transcripts separate it from a xenograft of the same model (216738-123-R-FEV): HLA-DRA (6,271 versus 111 TPM), HLA-DRB1 (4,555 versus 66), C1QA (613 versus 0.3) and C1QB (505 versus 0.2). These genes mark human antigen-presenting and complement-producing myeloid cells, which are present in patient tissue but absent from xenografts, whose stroma is of mouse origin. SPP1 is not discriminative, as it is equally high in the xenograft (8,231 versus 9,541 TPM) and is therefore expressed by the tumour cells. The prompt names the marker classes to report, which partly guides the answer. The example shows that the model can reason about the tumour microenvironment as well as the malignant compartment, which is relevant to quality control of patient-derived models.

A.12.7 Real Study Examples

Examples 7–9 are adapted from the held-out, publication-disjoint test split of GEO-OmicsQA. Each prompt keeps the original question and appends instructions to describe the observations and

reasoning. The GSM identifiers were replaced by the corresponding profiles, so the sample groups are defined only by their position and by the wording of the question.

A.12.8 Example 7: Host response to LCMV polymerase activity (GSE173072)

Question: We found that the LCMV L protein's ability to activate STAT3 is tied to its RNA-dependent RNA polymerase (RdRp) activity. Looking at the expression of DDIT3 and FOS in the RNA-seq data from samples GSM5259349, GSM5259350, GSM5259351 and GSM5259352, GSM5259353, GSM5259354, what does this suggest about the functional consequences of this RdRp activity on the host genome? State which sample group shows higher expression. Then check whether unfolded protein response genes (HSPA5, XBP1, ATF4, TRIB3, PPP1R15A), immediate-early genes (JUN, EGR1, ATF3), interferon-stimulated genes (ISG15, IFIT1) and STAT3 itself change in the same direction, and explain what kind of host response this pattern indicates. Answer in full sentences.

Answer: The expression data reveals that DDIT3 and FOS are significantly upregulated in the second group of samples compared to the first. This induction is accompanied by a concurrent upregulation of unfolded protein response genes (HSPA5, XBP1, ATF4, TRIB3, PPP1R15A), immediate-early genes (JUN, EGR1, ATF3), and interferon-stimulated genes (ISG15, IFIT1). The coordinated activation of these diverse stress and immune pathways indicates that the RdRp activity of the viral protein triggers a massive, multi-faceted host stress response, likely driven by the aberrant activation of STAT3.
Concluding line: The second group shows higher expression.

Assessment and significance. The prompt does not state which group is infected, so the direction has to be read from the profiles. The second group is the LCMV-infected one, consistent with its interferon response (ISG15 about 690 versus 6 TPM, IFIT1 113 versus 0.3). DDIT3 rises from 8 to 237 TPM and FOS from about 1 to 8. The unfolded protein response is induced (HSPA5 4.3-fold, XBP1 3-fold, TRIB3 5-fold, PPP1R15A 13-fold), as are the immediate-early genes (JUN 10-fold, EGR1 from 0 to 42 TPM, ATF3 from 2 to 291 TPM). ATF4 changes only 1.2-fold, so its inclusion overstates the effect, and the answer does not report STAT3 itself (1.8-fold); attributing the response to STAT3 follows the premise of the question. The integrated reading, an endoplasmic-reticulum stress and interferon response to viral RNA synthesis, matches the data and shows how the model summarizes coordinated changes across several pathways.

A.12.9 Example 8: Resisting a leading question on KDM2B knockdown (GSE81043)

Question: Evaluating the control glioblastoma samples (GSM2141218, GSM2141219) against the KDM2B knockdowns (GSM2141220, GSM2141221), are anti-apoptotic genes like MCL1 globally upregulated by the perturbation? Before answering, describe what you observe for the genes in question in each sample group, then explain the underlying biology step by step. Explain your reasoning, then answer with Yes or No on the last line.

Answer: The expression data reveals that anti-apoptotic genes, including MCL1, are actually downregulated in the knockdown samples compared to the controls. This reduction in survival factors indicates that the perturbation compromises cellular viability rather than promoting it.
Concluding line: No

Assessment and significance. The answer rejects the premise of the question, correctly. The knockdown is effective (KDM2B 13 versus 2 TPM), MCL1 falls from about 180 to 88 TPM and BCL2A1 from 9 to 1. The generalization to anti-apoptotic genes as a class is partly overstated, since BCL2L1, BIRC5 and XIAP are unchanged. The interpretation of reduced viability is supported by pro-apoptotic genes that the answer does not mention: PMAIP1 (NOXA, 20 to 47 TPM), TNFRSF10B (DR5, 112 to 165) and CASP8 (18 to 29). The step-by-step description requested in the prompt is brief. The example shows that the model follows the data rather than the framing of a leading question.

A.12.10 EXAMPLE 9: DNA-REPAIR GENE LOSS AFTER CDK12/13 DEGRADATION (GSE262119)

Question: Based on the RNA-seq data for DMSO control samples (GSM8157881, GSM8157882, GSM8157883) and YJ9069-treated samples (GSM8157884, GSM8157885, GSM8157886), is there a noticeable decrease in the expression of the DNA double-strand break repair genes BRCA1 and Rad51 upon treatment? Before answering, describe what you observe for the genes in question in each sample group, then explain the underlying biology step by step. Explain your reasoning, then answer with Yes or No on the last line.

Answer: The expression data reveals a marked downregulation of BRCA1 and Rad51 in the YJ9069-treated samples compared to the DMSO controls. These genes are critical for homologous recombination, the primary pathway for repairing DNA double-strand breaks. The reduction in their transcript levels indicates that the treatment impairs the cell's ability to mount an effective DNA damage response, thereby compromising genomic stability.
Concluding line: Yes

Assessment and significance. Both genes decrease after treatment: BRCA1 from about 16 to 4 TPM and RAD51 from 11 to 2. Other long DNA-damage-response genes follow the same pattern (ATR 6 to 0, FANCD2 29 to 7, FANCI 61 to 21 TPM), in line with the known role of CDK12/13 in transcriptional elongation of long genes and the resulting homologous-recombination deficiency; the answer states the functional consequence but not this mechanism. Housekeeping genes such as GAPDH and ACTB roughly double in these TPM profiles, reflecting a compositional shift, but the four- to six-fold decrease of BRCA1 and RAD51 exceeds it. The example connects a drug-induced expression change to a DNA-repair phenotype relevant to therapeutic combinations with PARP inhibition.

A.13 Multi-turn Conversation about Epithelial and Mesenchymal Cell States

This five-turn conversation illustrates how OmicsLM connects an initial expression comparison to follow-up questions about cell identity, causal inference, and a hypothetical perturbation experiment. We used the model with greedy decoding, no system message, and a maximum of 16,000 new tokens per turn. The chat interface inserted the two corresponding 20,541-feature omics profiles into the first prompt; conversation history and those profile inputs were retained throughout, without resetting between turns. The user-entered prompts and model responses below are reproduced verbatim, with only LaTeX formatting applied. The observations in the final prompt are hypothetical, not results of a laboratory experiment.

A.13.1 Turn 1: Comparing baseline expression profiles

Question: Compare the baseline expression profiles of MCF7 and MDAMB231 using ESR1, GATA3, FOXA1, CDH1 and VIM. Which profile has higher expression of each gene, and what biological programs does the pattern suggest? Keep your answer under 130 words.

Answer: ESR1, GATA3, FOXA1, and CDH1 are highly expressed in MCF7, while VIM is elevated in MDAMB231. This pattern indicates that MCF7 is a luminal epithelial breast cancer cell line, whereas MDAMB231 is a mesenchymal-like cell line.

A.13.2 Turn 2: Distinguishing cell state from a historical transition

Question: Does that comparison demonstrate that one of these cell lines underwent epithelial-to-mesenchymal transition, or that it arose from the other? Explain what we can and cannot infer from these baseline profiles. Keep your answer under 100 words.

Answer: No, the comparison does not demonstrate that one cell line underwent epithelial-to-mesenchymal transition or arose from the other. The baseline profiles only reflect the current state of the two cell lines, not their evolutionary history.

A.13.3 Turn 3: Connecting adhesion to cell identity

Question: Focusing on the adhesion gene in our comparison: what does its protein do, and would restoring that gene alone in the more mesenchymal line necessarily recreate the luminal identity of the other line? Explain briefly, in under 130 words.

Answer: CDH1 encodes E-cadherin, a calcium-dependent cell adhesion protein that mediates homophilic interactions to maintain epithelial integrity. Restoring CDH1 alone in the mesenchymal line would not necessarily recreate the luminal identity, as other luminal-specific genes like ESR1, GATA3, and FOXA1 are also required.

A.13.4 Turn 4: Outlining a perturbation experiment

Question: Suppose we want to test whether TGF-beta can induce an EMT-like change in the more epithelial line. Propose a small time-course experiment with controls, molecular readouts and one functional readout. Explain how you would avoid mistaking cell growth for migration. Treat the response to TGF-beta as something to test, not assume. Keep your answer under 180 words.

Answer: Treat MCF7 cells with TGF-beta and monitor changes in VIM expression and cell migration over time. Use untreated MCF7 cells as a control. Measure cell growth separately to ensure that any observed changes in migration are not due to differences in cell growth.

A.13.5 TURN 5: INTERPRETING MIXED HYPOTHETICAL RESULTS

Question: Here is a hypothetical result from that experiment: VIM RNA increases, CDH1 RNA is unchanged, junctional E-cadherin staining becomes weaker, and a scratch closes faster, but treated cells also proliferate faster. How would you interpret this mixed result, and which two follow-up measurements would most help resolve it? Keep your answer under 150 words.

Answer: The mixed result suggests that TGF-beta treatment induces partial EMT-like changes, including increased VIM expression and weakened junctional E-cadherin staining, but does not completely abolish CDH1 expression. The faster scratch closure could be due to both increased cell migration and faster cell proliferation. To resolve this, I would measure cell proliferation rates and assess the functional impact of TGF-beta treatment on cell migration in the absence of proliferation.

Assessment and significance. All five expression directions in the first response agree with the cached DepMap source: ESR1, GATA3, FOXA1, and CDH1 are higher in MCF7, whereas VIM is higher in MDAMB231. The follow-ups retain the relevant biological context, distinguish a baseline phenotype from evidence of a historical transition, and connect CDH1 to E-cadherin-mediated adhesion without equating restoration of one gene with recreation of a luminal identity. The final response recognizes that proliferation can contribute to faster scratch closure.