Towards Structured Mechanistic Reasoning in Single-cell Prediction

Anonymous ACL submission

Abstract

Large language models (LLMs) have recently gained significant attention as a promising approach to accelerate scientific discovery. However, their application in open-ended scientific domains such as biology remains limited, primarily due to the lack of factually grounded and actionable explanations. To address this, we introduce a structured explanation formalism for virtual cells that represents biological reasoning as mechanistic action graphs, enabling systematic verification and falsification. Building upon this, we propose VCR-AGENT, a multi-agent framework that integrates biologically grounded knowledge retrieval with a verifier-based filtering approach to generate and validate mechanistic reasoning autonomously. Using this framework, we release VC-TRACES dataset, which consists of verified mechanistic explanations derived from the Tahoe-100M atlas. Empirically, we demonstrate that training with these explanations improves factual precision and provides a more effective supervision signal for downstream gene expression prediction. These results underscore the importance of reliable mechanistic reasoning for virtual cells, achieved through the synergy of multi-agent and rigorous verification.

1 Introduction

The development of virtual cells, computational models that simulate cellular behavior, promises to advance biological discovery and drug design (Noutahi et al., 2025; Bunne et al., 2024; Adduri et al., 2025). A central goal of these models is to accurately predict cellular responses to perturbations, such as genetic knockouts or drug treatments. To move beyond correlation-based prediction toward actionable insight, virtual cells must also produce mechanistically grounded explanations. However, generating such explanations that are both biologically plausible and reliable remains a critical bottleneck. Large language models (LLMs) have emerged as a potential solution, demonstrating strong reasoning capabilities in domains such as mathematics and programming (Shao et al., 2024; Chen et al., 2021; DeepSeek-AI et al., 2025; OpenAI et al., 2024). These capabilities are acquired by training on vast, high-quality reasoning datasets.

However, directly transferring the reasoning training paradigms from mathematics or coding to scientific discovery is not straightforward. From a data-centric perspective, a key difficulty lies in the curation of large-scale, reliable reasoning datasets. The datasets are typically derived from two sources: (1) high-quality human annotations (Cobbe et al., 2021; Gao et al., 2024; Hendrycks et al., 2021), or (2) large-scale synthetic generation by LLMs (Wang et al., 2023; Moshkov et al., 2025; Guha et al., 2025). Human annotation, while high-quality, is expensive and non-scalable in domains that require specialized expertise, such as biology. Conversely, LLM-generated dataset is often factually unreliable and prone to hallucination, particularly in settings where LLMs lack sufficient domain grounding. This limits their applicability in scientific contexts where factual correctness and reliability are essential.

Beyond data scarcity, a more fundamental challenge lies in the verification of reasoning. In mathematics and programming, reasoning traces can be automatically verified, e.g., code can be executed, and its output is checked against a ground truth. Biological reasoning, in contrast, rarely admits such direct verification as it relies on disjointed knowledge from literature rather than deterministic rules. This inherent ambiguity makes it difficult for the reliable use of LLMs in scientific explanation.

To overcome these challenges, we introduce *structured explanations* for virtual cells, which constrain biological reasoning into explicit mechanistic actions connected by directed dependencies. Intuitively, rather than relying on ambiguous free-form natural language, we treat the explanation as

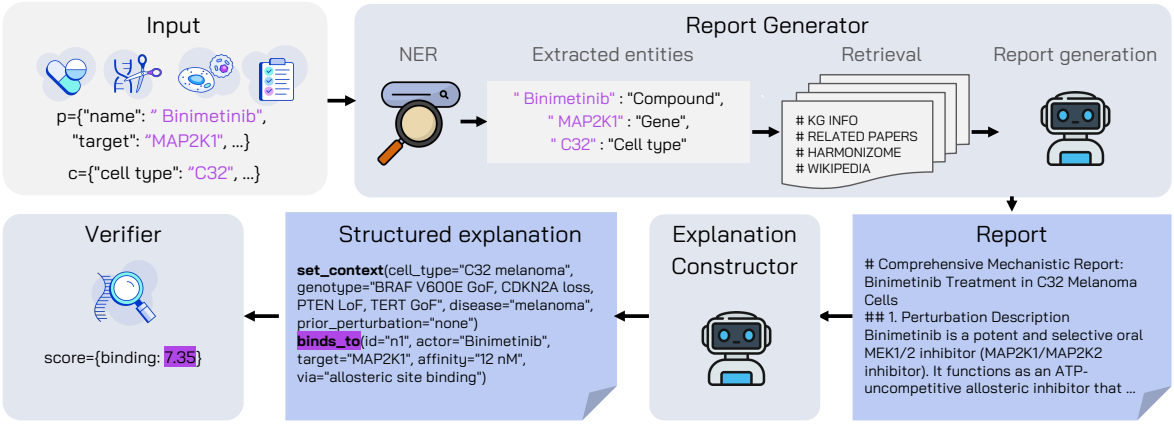


Figure 1: **An Overview of the VCR-AGENT Multi-Agent Framework.** The *Report Generator* accepts the perturbation and cellular context, performing knowledge retrieval and synthesis to produce a biologically grounded report. The *Explanation Constructor* then translates this report into the formal structured mechanistic explanation. This generated structured explanation is subsequently evaluated by the *Verifier* for factual validation and filtering.

a sequence of discrete, biologically-grounded actions. Each action consists of a predefined primitive and domain-specific arguments spanning molecular interactions to phenotype manifestations. By explicitly encoding biological dependencies, such as preconditions or regulatory requirements, these actions are logically connected to form a directed graph. In this representation, nodes constitute the discrete actions, while edges define the mechanistic relationships and dependencies between them. This format effectively restricts the model to a predefined action space, ensuring that every generated explanation is both interpretable and falsifiable. By grounding the reasoning in established biology, our framework produces *mechanistically plausible structures* that provide a rigorous basis for hypothesis generation, while remaining distinct from formal interventional causal discovery.

Building on this formalism, we propose VCR-AGENT, a multi-agent system that orchestrates the generation and validation of structured mechanistic reasoning, as illustrated in Figure 1. Our system decomposes the reasoning process into two specialized modules: (1) a *report generator* that aggregates and summarizes factual biological knowledge from external databases, thereby resolving the issue of factual grounding; and (2) an *explanation constructor* that transforms this retrieved report into the proposed structured format, mitigating the ambiguity of unstructured and unverifiable reasoning. To guarantee scientific reliability, each generated explanation undergoes *verifier-based filtering*, where specialized verifiers evaluate the explanation traces to retain only those that are factually accurate and causally coherent.

We apply our framework to the Tahoe-100M (Zhang et al., 2025) atlas to generate and publicly release a VC-TRACES dataset of structured mechanistic explanation traces. We evaluate this dataset through explanation quality assessment and downstream gene expression prediction under perturbation. Our experiments show that models trained on these verified reasoning traces achieve strong downstream performance.

2 Structured Mechanistic Reasoning for Virtual Cells

Virtual cells must be capable of structured, autonomous reasoning that explains how molecular perturbations lead to observable cellular outcomes. To achieve this, we adopt a structured formulation that serves two primary purposes: (1) it constrains the reasoning space to ensure interpretability and faithfulness, and (2) it enables automatic falsification via biology-grounded verifiers. To align such reasoning processes with the inherent topology of cellular signaling, where information propagates through cascades of mechanistic events, we formalize structured reasoning as the task of inferring a directed acyclic graph (DAG) of mechanistic interactions given a perturbation context. An overview of this problem formulation and an example mechanistic reasoning trace are illustrated in Figure 2a, while the DAG structure is visualized in Figure 2b.

2.1 Problem Formulation

The goal of structured reasoning in virtual cells is to infer how a given perturbation affects the cellular state through a series of mechanistic steps. Formally, given an input $x = (p, c)$, where p de-

(a) **An example of mechanistic reasoning traces.** (b) **An example of DAG.**

Figure 2: **An overview of structured reasoning.** (a) Given an input (p, c) =(Binimetinib, C32), the model generates mechanistic reasoning traces. Blue and light blue indicate the action primitives and the arguments, respectively. The elements within the <dag> tag represent the edge list defining the reasoning graph. (b) An example of DAG. Same color indicates the same action primitive.

notes the *perturbation* (e.g., a chemical compound, genetic knockdown, etc.) and c denotes the *cellular context* (e.g., cell type, disease model, etc.), we aim to generate a reasoning graph $\mathcal{G}$ that captures a chain of mechanistic actions triggered by the perturbation p in the context c.

The output reasoning is denoted as a DAG $\mathcal{G}$:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \quad \text{where } \mathcal{V} = \{n_1, \ldots, n_k\}.$$

Each node $n_i \in \mathcal{A}$ represents a mechanistic action such as binding, modulation, regulation, etc., comprising an action primitive selected from the predefined action space $\mathcal{A}$ (defined in Section 2.2) and its associated arguments. Each directed edge $(n_i, n_j) \in \mathcal{E}$ represents a mechanistic dependency, indicating that the outcome of action n_i enables or influences another action n_j. For example, a ligand–receptor binding (`binds_to`) may precede a downstream signaling modulation (`modulates_pathway_activity`).

Finally, the reasoning model f_θ is defined as

$$f_\theta : x \rightarrow \mathcal{G},$$

where f_θ generates both the mechanistic actions (nodes) and their dependencies (edges). This representation encodes mechanistic plausibility and downstream biological consequences, thereby enhancing the interpretability of the reasoning model's logic while remaining distinct from formal, interventional causal discovery.

2.2 Action Spaces

The *action space* $\mathcal{A}$ defines the set of permissible reasoning actions for a virtual cell. Constraining each reasoning step to a finite and biologically grounded set of high-confidence primitives enables falsifiability, as verifiers can evaluate actions. To ensure systematic consistency, each primitive is parameterized by a specific argument schema, such as assigning an actor and target to a `binds_to` action.

We define twenty action primitives grouped into seven categories: (1) system initialization, (2) metabolic, (3) regulation, (4) functional, (5) interaction, (6) phenotype, and (7) proteostasis. These categories span from molecular interactions to phenotypic manifestations. We provide an overview of defined action primitives in Figure 3 and detailed argument schemas in Appendix A.

To illustrate how an action primitive is defined and parameterized, consider an example action `binds_to`. This action specifies a direct molecular interaction between two biomolecules, such as a drug–target, ligand–receptor, or protein–protein pair. It is parameterized as:

`binds_to`(*id, actor, target, {affinity, unit, residues_actor, residues_target, via, confidence}*).

The *id* argument specifies the node identifier used to connect the actions in the DAG, while *actor* and *target* define the participating entities. Other optional arguments represented in { } are mapped to biological ontologies such as compounds, affinity scores, etc. By leveraging these structured arguments, verifiers can evaluate the reliability of each action with automatic verification against curated databases (Mendez et al., 2019; Ashburner et al., 2000; Croft et al., 2010) and computational tools (Passaro et al., 2025; Love et al., 2014).

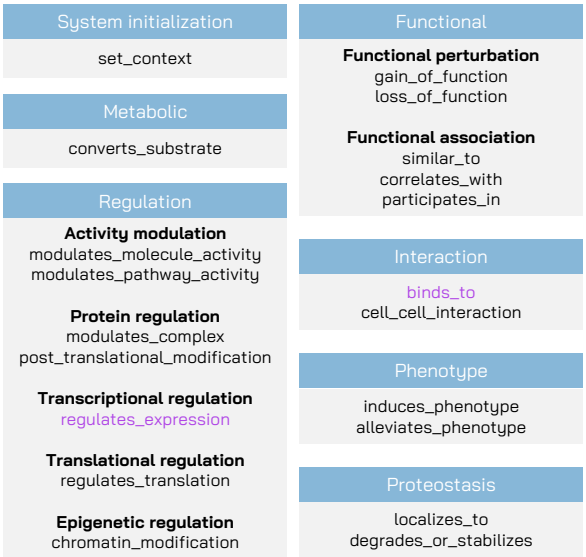
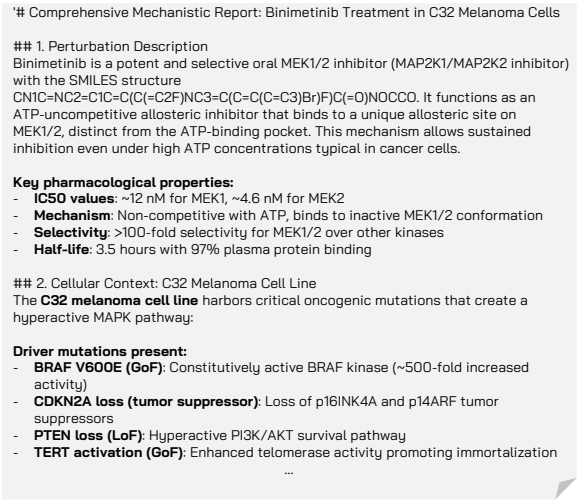


Figure 3: **An overview of action spaces.** The subcategories are represented with **bold** and action primitives with verifier are represented with purple. The argument schemes are in Appendix A.



'# Comprehensive Mechanistic Report: Binimetinib Treatment in C32 Melanoma Cells

1. Perturbation Description
Binimetinib is a potent and selective oral MEK1/2 inhibitor (MAP2K1/MAP2K2 inhibitor) with the SMILES structure CN1C=NC2=C1C=C(C(=C2F)NC3=C(C=C(C=C3)Br)F)C(=O)NOCCO. It functions as an ATP-uncompetitive allosteric inhibitor that binds to a unique allosteric site on MEK1/2, distinct from the ATP-binding pocket. This mechanism allows sustained inhibition even under high ATP concentrations typical in cancer cells.

Key pharmacological properties:
- **IC50 values**: ~12 nM for MEK1, ~4.6 nM for MEK2
- **Mechanism**: Non-competitive with ATP, binds to inactive MEK1/2 conformation
- **Selectivity**: >100-fold selectivity for MEK1/2 over other kinases
- **Half-life**: 3.5 hours with 97% plasma protein binding

2. Cellular Context: C32 Melanoma Cell Line
The **C32 melanoma cell line** harbors critical oncogenic mutations that create a hyperactive MAPK pathway:

Driver mutations present:
- **BRAF V600E (GoF)**: Constitutively active BRAF kinase (~500-fold increased activity)
- **CDKN2A loss (tumor suppressor)**: Loss of p16INK4A and p14ARF tumor suppressors
- **PTEN loss (LoF)**: Hyperactive PI3K/AKT survival pathway
- **TERT activation (GoF)**: Enhanced telomerase activity promoting immortalization

...

Figure 4: **An example of generated report.**

3 VCR-AGENT framework

We introduce VCR-AGENT, our multi-agent system designed to generate structured explanations for virtual cells given input perturbations and cellular contexts. The framework is designed as a three-stage pipeline to ensure factual grounding and structured output, consisting of a *report generator*, *explanation constructor*, and *verifier*. We provide an overview in Figure 1.

First, the report generator is responsible for information retrieval and summarization. It queries comprehensive knowledge bases to gather relevant biological facts about the given perturbation and cellular context. This is then summarized into a comprehensive natural-language report.

Next, the explanation constructor takes this knowledge-grounded report as input, transforming it into the structured reasoning format. By decoupling knowledge acquisition from structured reasoning generation, this enforces knowledge grounding, thereby ensuring both the factual accuracy and structural integrity of the final explanation.

Finally, the verifier filters the explanation trace or partial arguments by comparing against biologically-grounded oracles. This improves the biological and factual accuracy of the generated explanations based on reliable computational tools.

3.1 Report Generator

The report generator operates in a three-step process: (1) entity extraction with biomedical name entity recognition (NER), (2) retrieval with external knowledge bases, and (3) report generation based on the related retrieved information.

First, the entity extraction step identifies relevant biomedical entities from the input perturbation and cellular context. We employ HunFlair2 (Sänger et al., 2024), which extracts the biomedical entities, including chemical compounds, genes, diseases, etc. This NER process simplifies the retrieval by enabling entity-based search instead of relying on complex natural language.

Next, the retrieval step uses these extracted entities to query external knowledge bases to aggregate relevant biological facts. Specifically, these include StarkPrimeKG, a biomedical knowledge graph (Wu et al., 2024b); Harmonizome, a gene-related database (Diamant et al., 2024); PubMed, a biomedical literature database; and Wikipedia. The retrieval process operates as follows:

- **StarkPrimeKG:** We query the entity and its synonyms to search for the matching node in the knowledge graph. If no exact match is found, we identify the most similar node based on the cosine similarity of PubMedBERT (Gu et al., 2021) embeddings. The 1-hop neighbor nodes are then aggregated to provide relevant relational information.

- **Harmonizome:** We query gene entities to enrich gene-specific information. These entities include both the genes encoding protein targets of compound perturbations and the genes targeted by genetic perturbations.

- **PubMed:** We query the database using extracted entities to identify relevant papers, pri-

oritizing those whose abstracts demonstrate the highest similarity to the input entities.

- **Wikipedia:** We query each entity to retrieve the best matching documents from Wikipedia.

We provide the examples of information retrieved from four databases in Appendix B.

Finally, the report generation step summarizes all retrieved information into a single, comprehensive report. We prompt Claude-Sonnet-4 (Anthropic, 2025) with the retrieved information to generate the report. This report provides comprehensive information on the input perturbation and cellular context, which serves as the factual foundation for the explanation constructor. We provide an example report in Figure 4 and the used prompt in Appendix C.

3.2 Explanation Constructor

Next, the explanation constructor generates the structured reasoning, based on the knowledge-grounded report as input. This process enables the verification and falsification of the explanation. We also used Claude 4 (Anthropic, 2025) in the same way as the report generation step, and the used prompt is provided in Appendix C.

3.3 Verifier-based filtering

To improve the biological accuracy of generated explanations, we introduce a verifier-based filtering and quality control pipeline. This prunes reasoning traces and arguments that conflict with established biological facts. The process consists of two stages: per-action verification and filtering.

First, individual actions within a generated trace are evaluated by a corresponding, specialized verifier. Our framework supports an extensible set of verifiers tailored to different action primitives, where coverage is bounded by the availability of reliable computational oracles. We instantiate four verifiers spanning the most frequently occurring action types: drug-target interaction (DTI) for `binds_to`, differential expression (DE) for `regulates_expression`, subcellular localization (LOC) for `localizes_to`, and phenotype (PHENO) for `induces_phenotype` and `alleviates_phenotype`. We provide the details of each verifier in Appendix E.

Next, the resulting verification scores are used to filter out either the entire trace or partial arguments that fail to meet plausibility thresholds. Notably, filtering is conservative by design, i.e., only claims

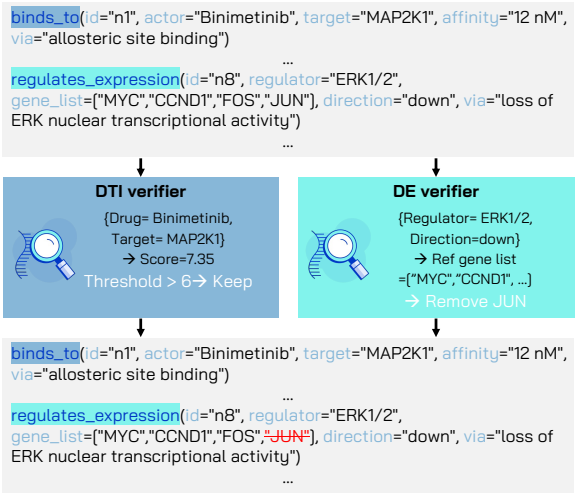


Figure 5: **An example of verifier-based filtering process.** The pipeline processes initial structured explanation (top) through verifiers (middle) to produce filtered output (bottom). Same colors link the action primitive to their corresponding verifiers.

that directly contradict established biological evidence are removed, leaving uncertain or unverifiable claims intact. This avoids false-positive rejection of plausible but uncertain mechanisms, a critical property in biology where ground truth is often context-dependent.

We provide an illustrative example of this verifier-based filtering in Figure 5. For instance, the DTI verifier predicts the physical plausibility of a binding for `binds_to` action with Boltz-2 (Passaro et al., 2025). This process yields a continuous binding probability score. Next, we filter out the traces containing a `binds_to` action with a DTI confidence score below a pre-defined threshold τ.

4 Experiments

In this section, we evaluate both the effectiveness of the VCR-AGENT framework and the quality of the resulting VC-TRACES dataset. Our evaluation is structured into two primary components: (1) an evaluation of explanation quality, and (2) an evaluation of the dataset's utility as a supervision signal for downstream biological tasks. We report additional ablations on retrieval source, backbone LLM, and one-step generation, along with a human evaluation, in Appendix D, and detailed experimental settings in Appendix C.

4.1 Explanation Quality

We first assess the quality of the structured explanations generated by our VCR-AGENT.

	Model	Format		Verifier	
		Validity	Verifiability	Drug-target interaction	Differential expression
Open	DeepSeek-R1-8B	0.003 ± 0.004	1.000 ± 0.081	0.000 ± 0.000	0.000 ± 0.000
	Qwen3-30B	0.860 ± 0.016	0.965 ± 0.007	0.528 ± 0.004	0.272 ± 0.030
	Llama3.3-70B	0.841 ± 0.046	0.887 ± 0.017	0.322 ± 0.034	0.090 ± 0.010
Closed	Claude-Sonnet-4	**1.000** ± 0.000	0.867 ± 0.008	0.657 ± 0.005	0.504 ± 0.064
	VCR-AGENT (Ours)	**1.000** ± 0.000	**0.945** ± 0.011	**0.725** ± 0.017	**0.528** ± 0.060

Table 1: **Explanation quality performance.** The best results are highlighted in **bold**. The standard deviation is computed across cell lines.

Dataset. We derive experimental dataset from the Tahoe-100M (Zhang et al., 2025), extracting a total of 18,950 unique compound perturbation-context pairs. Using these, we constructed the VC-TRACES dataset through VCR-AGENT framework, which transforms the perturbations into reasoning traces. To align with the test split of Tahoe-X1 (Gandhi et al., 2025), we focus our experiment on a subset comprising five cell lines (C32, HOP62, HepG2/C3A, Hs 766T, and PANC-1). Notably, complete VC-TRACES dataset is in https://anonymous.4open.science/r/VC-TRACES_.

Baselines. We compare VCR-AGENT with both open-source and closed-source LLMs. For open-source baselines, we use three models: Qwen3-30B-A3B (Yang et al., 2025), DeepSeek-R1-0528-Qwen3-8B (DeepSeek-AI et al., 2025), and Llama3.3-70B-Instruct (Grattafiori et al., 2024). For the closed-source baseline, we use Claude-Sonnet-4 (Anthropic, 2025), the same base model used within VCR-AGENT for fair comparison.

Metrics. We evaluate performance using two format-based metrics and the two verifier scores. First, regarding format, we report validity, which measures the proportion of traces that are both syntactically correct, i.e., containing proper <explain> and <dag> tags, and structurally valid, meaning all generated action primitives adhere to the definitions in Section 2.2. We also report verifiability, which quantifies the proportion of generated arguments that can be successfully mapped to valid biomedical entities for verification. Finally, the verifier scores are computed with DTI and DE verifiers and averaged to assess the factual correctness of individual mechanistic actions. Notably, the DTI score is computed as the average of the binding score, while the DE score is computed as the proportion of the traces where at least a single DE step included in the trace is correct.

Results. We present the results in Table 1. We observe that VCR-AGENT consistently generates high-quality reasoning traces across both format and verifier-based metrics. These findings demonstrate that our framework generates explanations that are not only structurally valid and verifiable but also more closely aligned with biological reference data than those produced by the baselines.

Crucially, the results reported in Table 1 evaluate the raw generative performance of the models prior to any filtering. While these metrics already demonstrate superior alignment with biological verifiers compared to baselines, our framework further enhances reliability through the verification pipeline: during the construction of VC-TRACES, verifiers excluded 28.2% of faulty DTI claims and refined 87.3% of DE actions to eliminate hallucinations.

4.2 Application: TahoeQA

Next, we evaluate the utility of our VC-TRACES dataset on TahoeQA, a downstream task designed to predict transcriptional responses to chemical compounds using perturbations sourced from Tahoe-100M. This task is inspired by the PerturbQA benchmark (Wu et al., 2024a).

Dataset. We use the same perturbation-context pairs from the Tahoe-100M dataset introduced in Section 4.1. Inspired by PerturbQA, for data labeling, we extracted the top 25 up-regulated, top 25 down-regulated, and 100 random non-regulated genes for each pair. Specifically, we perform differential expression (DE) analysis by fitting a negative binomial-based general linear model to the pseudo-bulked counts, and running a Wald's test (Wald, 1943) to determine if the logFC differs significantly from 0 as implemented in DESeq2 (Love et al., 2014). We define differentially expressed genes following Benjamini and Hochberg (1995) adjusted $p < 0.05$, and the top 25 DE genes are selected based on the magnitude of $\log_2$ fold change.

We select five cell types following the few-shot

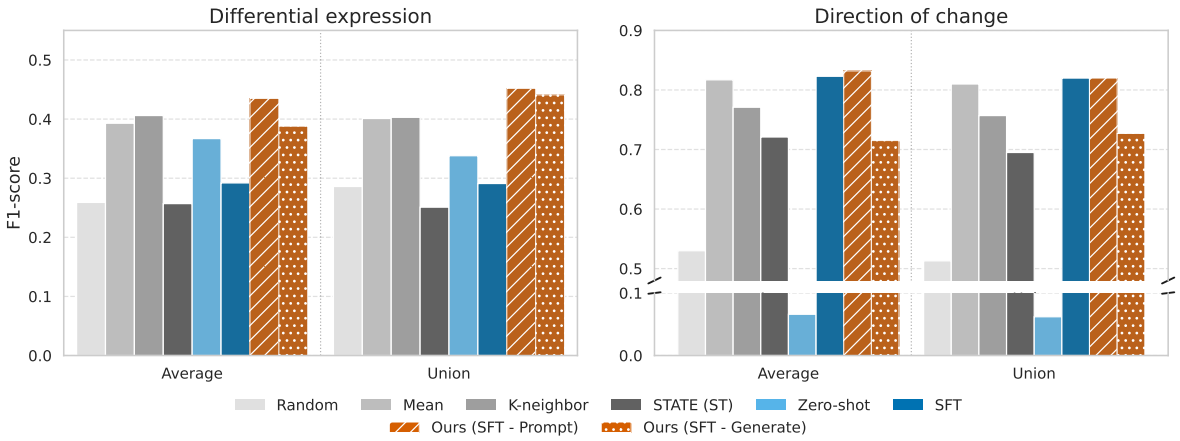


Figure 6: **TahoeQA performance.** Baselines are categorized by model type: statistical and gene foundation models are shown in shades of gray, LLM-based baselines in shades of blue, and our model with structured explanation in brown. Average denotes the mean F1-score across the five individual cell-line test sets while Union denotes the performance on a test set combining all five cell lines.

test split of Tahoe-X1 (Gandhi et al., 2025): C32, HOP62, HepG2/C3A, Hs 766T, PANC-1. The data is split by perturbation to ensure no overlap between training and test perturbations, with 1K test examples randomly selected for evaluation.

Task. Following PerturbQA, the task is a two-fold binary classification: (1) differential expression (DE) and (2) direction of change (DOC). For both, perturbation and cellular context are given as an input question. The DE task predicts whether a perturbation causes differential expression or not, while the DOC task predicts whether the target gene's expression decreases or increases.

Baselines. We compare our trained models against three categories of baselines: (1) statistical models, (2) a transcriptomic foundation model, and (3) LLMs. The statistical baselines include: (1) a random baseline, (2) a mean baseline that predicts labels based on the average gene expression response for a given compound, and (3) a k-nearest-neighbor baseline that performs label classification by aggregating the labels of the k most similar compounds, where similarity is computed using extended-connectivity fingerprints (ECFP) (Rogers and Hahn, 2010). Notably, statistical baselines including mean and k-nearest neighbor have proven to be a strong baseline in differential expression tasks (Kernfeld et al., 2025; Wenkel et al., 2025).

For the transcriptomic foundation model, we include the STATE Transition (ST) model (Adduri et al., 2025) which learns a state-transition function over gene expression from a large corpus of perturbation-response data in Tahoe-100M (Zhang et al., 2025). The model is trained on all available perturbations across cell types and evaluated under our few-shot setting, where test perturbations are held out in the five target cell lines and used only for evaluation. Finally, the LLM baselines include zero-shot prompting and supervised fine-tuning (SFT) without any structured explanations and use the same Qwen3 backbone as our models.

Training. We fine-tune Qwen3-4B-Instruct-2507 (Yang et al., 2025) to assess whether our structured reasoning serves as an effective supervision signal. We train the model using SFT in two configurations: (1) context-augmented prediction (SFT-Prompt), where the model predicts the answer label given the perturbation, cellular context, and the verified structured explanation as input; and (2) generative reasoning (SFT-Generate), where the model is trained to generate the structured explanation followed by the answer, given only the perturbation and cellular context. Notably, we focus on SFT and exclude reinforcement learning (RL) from this study, as reasoning in pure classification settings often suffers from sparse reward signals (Wang et al., 2024; He et al., 2025; Sprague et al., 2025) and leave RL-based optimization for future work.

Metrics. We measure the performance with F1-score due to the label imbalance problem (i.e., 50 positive and 100 negative labels per perturbation-context pair) in the DE task.

Leakage prevention. We confirm that the DE verifier does not introduce test-time label leakage: the empirical overlap between TahoeQA test genes and genes appearing in `regulates_expression`

actions is minimal (0.2% for DE, 0.1% for DOC).

Results. We provide the results in Figure 6 and detailed per-cell-line values in Appendix C. Our experiments demonstrate that structured reasoning consistently enhances predictive accuracy, particularly for the DE task. Conditioning predictions on structured explanations (SFT-Prompt) yields the strongest overall performance. Additionally, our generative model (SFT-Generate), which autonomously constructs the mechanistic reasoning chain, surpasses all baselines in the DE task. This performance gap between our reasoning approaches and standard SFT confirms that training models with biological reasoning provides a more effective supervision signal than the label alone.

Crucially, by leveraging structured reasoning as an inductive bias, our model exhibits superior generalization to novel compounds compared to baselines like STATE, which rely primarily on raw numerical representations. These findings suggest that grounding high-dimensional transcriptomic data in biological reasoning improves performance in sparse-data and out-of-distribution scenarios.

5 Related Work

Reasoning with LLMs. Reasoning of large language models (LLMs) has enhanced the problem-solving capabilities (OpenAI et al., 2024; DeepSeek-AI et al., 2025; Wei et al., 2022). The reasoning models are typically trained through supervised fine-tuning or reinforcement learning, using human-curated (Cobbe et al., 2021; Gao et al., 2024; Hendrycks et al., 2021) reasoning as supervision signals. However, they rely heavily on expensive and domain-limited annotated explanations.

To mitigate data scarcity, LLMs are increasingly leveraged to synthesize reasoning traces (Wang et al., 2023; Moshkov et al., 2025; Guha et al., 2025), yet ensuring factual reliability remains a major challenge. While the reasoning in mathematics and code can be validated via symbolic or programmatic evaluation (Shao et al., 2024; Chen et al., 2021), verification in science, such as biology, is hindered by causal uncertainty and incomplete prior knowledge. As current general-purpose LLMs lack sufficient grounding in specialized scientific contexts, their self-generated explanations frequently lack the reliability.

Reasoning for biology. LLMs are increasingly recognized as powerful tools for scientific discovery, with applications spanning chemistry and biology (Han et al., 2025; Fang et al., 2024; Edwards et al., 2022; Zhang et al., 2024). Yet, limited reasoning capability hinders their ability to answer complex, open-ended scientific questions. Recent efforts to enhance biological reasoning include RL frameworks with soft verifiers (Istrate et al., 2025), SFT on LLM-generated reasoning traces (Phillips et al., 2025), and inference-time compression of gene-centric knowledge graphs (Wu et al., 2025). Despite these advances, they remain largely restricted to gene-centric perturbations and rely on unstructured, free-form natural language rationales.

Prior works face three limitations: (1) Lack of reliable fact discovery: reasoning rely on internal parametric weights without sufficient integration of external knowledge. (2) Structural ambiguity: unstructured rationales prevent the formalization of mechanistic dependencies, which hinders systematic verification. (3) Cross-modal insufficiency: they focus on gene perturbations, neglecting diverse modalities such as drug-induced responses.

To address these challenges, we treat biological reasoning as the construction of structured mechanistic graphs. By replacing free-form natural language with discrete and verifiable actions, VCR-AGENT transforms the LLM into a mechanistic architect. Our framework synthesizes diverse knowledge bases and applies automated verifiers to ensure that every node in the reasoning graph represents an evidence-aligned, falsifiable claim. This enables LLMs not only to predict cellular outcomes but to ground their predictions in structured and reliable reasoning traces.

6 Conclusion

In this work, we tackled the critical bottleneck in reasoning for scientific discovery by defining a structured, falsifiable reasoning format for virtual cells. We introduced VCR-AGENT, a multi-agent system that generates and validates these explanations by separating knowledge retrieval from construction and employing a rigorous verifier pipeline. Our experiments demonstrate that VCR-AGENT produces explanations with factual accuracy and logical coherence. Moreover, this high-quality dataset boosts the performance on the downstream TahoeQA gene expression prediction task. Our framework provides a scalable method for generating grounded and mechanistic reasoning, shedding light on reliable and autonomous virtual cells.

Broader impact

The development of VCR-AGENT addresses a critical bottleneck in the reliability of large language models for scientific tasks: the lack of interpretability and factual grounding in Large Language Models (LLMs) when applied to complex biological systems. By transitioning from unstructured text generation to structured mechanistic graphs, this work provides a framework for the systematic verification and falsification of biological hypotheses.

A primary contribution of this work is the release of VC-TRACES, derived from Tahoe-100M. By providing the research community with a scalable and grounded resource, we lower the barrier to entry for studying complex cellular responses to perturbations. Rather than functioning as a closed-loop solution, this dataset and framework are designed to augment domain experts, allowing them to integrate and refine mechanistic explanations within their specific research pipelines. This facilitates the development of autonomous virtual cells that can navigate cellular events with increased factual precision.

Limitations

Verifier coverage. We instantiate four verifiers (DTI, DE, LOC, PHENO) covering the most frequent action primitives in VC-TRACES, but the remaining 15 primitives in our action space lack dedicated verifiers. Coverage is fundamentally bounded by the availability of reliable computational oracles or curated reference data for each action type and expanding the verifier suite is a community effort.

Boltz-2 as a DTI oracle. The DTI verifier relies on Boltz-2 (Passaro et al., 2025) predictions rather than experimentally measured affinities. This choice is inevitable considering the superior speed of the model but might result in producing noisier scores for novel or under-studied proteins.

Closed-source backbone and reproducibility. Both the report generator and explanation constructor use Claude 4 (Anthropic, 2025). We provide ablations with GPT-4.1 and Gemini-2.5-Flash backbones in Appendix D showing the framework is not tied to a specific model, but exact reproduction of VC-TRACES requires access to the closed-source API. To mitigate this, we release all prompts, retrieval code, and the full VC-TRACES dataset to enable reproducibility without re-generation.

Scope of mechanistic claims. Our framework produces mechanistically plausible reasoning traces grounded in retrieved biological knowledge. These traces are not formal causal discoveries, i.e., edges in the DAG encode mechanistic dependencies as represented in source databases, not causal relationships verified in the target cellular context. We position VCR-AGENT as an explanation-and-hypothesis-generation tool, not a substitute for experimental causal inference.

References

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, and 1 others. 2025. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, pages 2025–06.

Shadab Ahmad, Leonardo Jose da Costa Gonzales, Emily H Bowler-Barnett, Daniel L Rice, Minjoon Kim, Supun Wijerathne, Aurélien Luciani, Swaathi Kandasaamy, Jie Luo, Xavier Watkins, Edd Turner, Maria J Martin, and the UniProt Consortium. 2025. The uniprot website api: facilitating programmatic access to protein knowledge. *Nucleic Acids Research*, 53(W1):W547–W553.

Anthropic. 2025. Introducing claude 4.

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, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, and 1 others. 2024. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

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. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

David Croft, Gavin O'kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, and 1 others. 2010. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Ido Diamant, Daniel J B Clarke, John Erol Evangelista, Nathania Lingam, and Avi Ma'ayan. 2024. Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources. *Nucleic Acids Research*, 53(D1):D1016–D1028.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*.

Shreshth Gandhi, Farnoosh Javadi, Valentine Svensson, Umair Khan, Matthew G. Jones, Johnny Yu, Daniele Merico, Hani Goodarzi, and Nima Alidoust. 2025. Tahoe-x1: Scaling perturbation-trained single-cell foundation models to 3 billion parameters. *bioRxiv*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and 1 others. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *Preprint*, arXiv:2506.04178.

Yang Han, Ziping Wan, Lu Chen, Kai Yu, and Xin Chen. 2025. From generalist to specialist: A survey of large language models for chemistry. In *Proceedings of the 31st International Conference on Computational*

Linguistics, pages 1106–1123, Abu Dhabi, UAE. Association for Computational Linguistics.

Mingqian He, Fei Zhao, Chonggang Lu, Ziyan Liu, Yue Wang, and Haofu Qian. 2025. Gencls++: Pushing the boundaries of generative classification in llms through comprehensive sft and rl studies across diverse datasets. *arXiv preprint arXiv:2504.19898*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M. Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. 2025. rbio1 - training scientific reasoning llms with biological world models as soft verifiers. *bioRxiv*.

Eric Kernfeld, Yunxiao Yang, Joshua Weinstock, A. Little, and Patrick Cahan. 2025. A comparison of computational methods for expression forecasting. *Genome Biology*, 26.

Catherine Kirsanova, Alvis Brazma, Gabriella Rustici, and Ugis Sarkans. 2015. Cellular phenotype database: a repository for systems microscopy data. *Bioinformatics*, 31(16):2736–2740.

Michael I Love, Wolfgang Huber, and Simon Anders. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, and 1 others. 2019. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K. Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, Emmanuel Bengio, Dominique Beaini, Christopher Gibson, Daniel Cohen, and Berton Earnshaw. 2025. Virtual cells: Predict, explain, discover. *Preprint*, arXiv:2505.14613.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. 2025. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*.

Lawrence Phillips, Marc Boubnovski Martell, Aditya Misra, Josefa Lia Stoisser, Cesar A. Prada-Medina, Rory Donovan-Maiye, and Kaspar Märtens. 2025. Synthpert: Enhancing llm biological reasoning via synthetic reasoning traces for cellular perturbation prediction. *Preprint*, arXiv:2509.25346.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

Mario Sänger, Samuele Garda, Xing David Wang, Leon Weber-Genzel, Pia Droop, Benedikt Fuchs, Alan Akbik, and Ulf Leser. 2024. Hunflair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools. *arXiv preprint arXiv:2402.12372*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Peter J Thul and Cecilia Lindskog. 2018. The human protein atlas: a spatial map of the human proteome. *Protein Science*, 27(1):233–244.

Abraham Wald. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding

benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Cian Eastwood, Shawn T Whitfield, Ihab Bendidi, Craig Russell, Liam Hodgson, Yassir El Mesbahi, and 1 others. 2025. Txpert: Leveraging biochemical relationships for out-of-distribution transcriptomic perturbation prediction. *arXiv preprint arXiv:2505.14919*.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. 2024a. Contextualizing biological perturbation experiments through language. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. 2025. Contextualizing biological perturbation experiments through language. In *The Thirteenth International Conference on Learning Representations*.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N. Ioannidis, Karthik Subbian, James Zou, and Jure Leskovec. 2024b. Stark: Benchmarking llm retrieval on textual and relational knowledge bases. In *NeurIPS Datasets and Benchmarks Track*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, and 12 others. 2025. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Kehua Feng, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Tao Huang, Pengju Yan, and 6 others. 2024. Scientific large language models: A survey on biological & chemical domains. *Preprint*, arXiv:2401.14656.

A Details of action primitives

Here, we describe the details of action primitives. The action primitives are grouped into seven categories: system initialization, molecular interactions, metabolic, regulation, functional, interaction, phenotype, and proteostasis. We describe the role of each action below and argument in Table Appx.1.

A.1 System initialization

- `set_context`: Defines the biological background before the new perturbation is applied including the cell model, genotype, disease, and prior treatments.

A.2 Metabolic

- `converts_substrate`: Enzymatic conversion of one chemical entity into another (metabolic reaction or proteolytic processing).

A.3 Regulation

Activity modulation

- `modulates_molecule_activity`: Describes whether increases or decreases the catalytic / signalling activity of a single protein, enzyme, transporter, TF or RNA.

- `modulates_pathway_activity`: Increases or decreases the activity of a named pathway or biological process (e.g. MAPK, autophagy, EMT).

Protein regulation

- `modulates_complex`: Promotes assembly or causes disassembly of a multi-subunit complex; can include stoichiometry changes.

- `post_translational_ modification`: Adds or removes a specific PTM or proteolytic cleavage on a protein site (phospho, ubiquitin, acetyl, etc.)

Transcriptional regulation

- `regulates_expression`: Alters steady-state mRNA level or isoform ratio of one or more genes (bulk, single-cell or signature). Use ‘mechanism‘ to tag cases like alternative splicing.

Table Appx.1: **Argument schema for action spaces.**

Category	Sub-category	Action name	Arguments
system initialization		set_context	set_context({cell_type, genotype, disease, prior_perturbation, extras})
metabolic		converts_substrate	converts_substrate(id, enzyme, substrate, product, {via, confidence})
regulation	activity modulation	modulates_molecule_activity modulates_pathway_activity	modulates_molecule_activity(id, target, direction, {via, confidence}) modulates_pathway_activity(id, pathway, direction, {via, confidence})
	protein regulation	modulates_complex post_translational_modification	modulates_complex(id, members, complex, direction, {stoichiometry, via, confidence}) post_translational_modification(id, protein, mod_type, site, direction, {via, confidence})
	transcriptional regulation	regulates_expression	regulates_expression(id, regulator, gene_list, direction, {mechanism, via, confidence})
	translational regulation	regulates_translation	regulates_translation(id, regulator, rna_id, direction, {mechanism, via, confidence})
	epigenetic regulation	chromatin_modification	chromatin_modification(id, mark, locus, direction, {via, confidence})
functional	functional perturbation	gain_of_function loss_of_function	gain_of_function(id, variant_id, protein, {via, confidence}) loss_of_function(id, variant_id, protein, {via, confidence})
	functional association	similar_to correlates_with participates_in	similar_to(id, entity_a, entity_b, evidence_type, {confidence, via}) correlates_with(id, entity_a, entity_b, evidence_type, {confidence}) participates_in(id, entity, ontology_id, {evidence_type, confidence})
interaction		binds_to cell_cell_interaction	binds_to(id, actor, target, {affinity, unit, residues_actor, residues_target, via, confidence}) cell_cell_interaction(id, sender, receiver, ligand, receptor, outcome, {via, confidence})
phenotype		induces_phenotype alleviates_phenotype	induces_phenotype(id, source, phenotype, {via, confidence, from_state, to_state}) alleviates_phenotype(id, actor, phenotype, {via, confidence, from_state, to_state})
proteostasis		localizes_to degrades_or_stabilizes	localizes_to(id, entity, from_loc, to_loc, {mechanism, via, confidence}) degrades_or_stabilizes(id, regulator, target, direction, {via, confidence})

Translation regulation

- `regulates_translation`: Post-transcriptional control at the ribosome level (e.g., eIF inhibition, uORF usage, IRES activation).

Epigenetic regulation

- `chromatin_modification`: Adds or removes histone/DNA marks at a defined genomic locus, affecting chromatin accessibility.

A.4 Functional

Functional perturbation

- `gain_of_function`: Genetic variant or edit that increases the normal activity of the specified protein.
- `loss_of_function`: Genetic variant, knockdown or knock-out that reduces or abolishes activity of the specified protein.

Functional association

- `similar_to`: Non-causal functional similarity (transcriptomic, phenotypic, structural) between two entities.
- `correlates_with`: Statistical association without established causality (GWAS hit, literature co-mention, co-expression).
- `participates_in`: Links an entity to a GO term, pathway or compartment (background annotation).

A.5 Interaction

- `binds_to`: Describes the direct physical binding of two biomolecules such as drug-target, protein-protein, and ligand-receptor.
- `cell_cell_interaction`: Represents the ligand-receptor signaling from one cell type to another plus the immediate downstream outcome.

A.6 Phenotype

- `induces_phenotype`: Creates or worsens a measurable phenotype or cell-state transition. Optional 'from_state' / 'to_state' capture events like EMT or senescence.
- `alleviates_phenotype`: Reverts or mitigates an abnormal phenotype back toward normal (rescue). Same optional state fields as above.

A.7 Proteostasis

- `localizes_to`: Moves a molecule between compartments. Use 'to="extracellular"' for secretion or 'from="extracellular"' for uptake; add 'mechanism' (e.g. exocytosis, transporter).
- `degrades_or_stabilizes`: Changes protein abundance by altering half-life (ubiquitin-proteasome degradation, PROTAC, chaperone rescue).

B Examples

B.1 Retrieved information

StarkPrimeKG. This is the example of retrieved information from StarkPrimeKG. Note that due to the limited space, we display only partial information of the full retrieved context.

Retrieved information from StarkPrimeKG for (p, c) =(Binimetinib, C32)

KNOWLEDGE GRAPH INFORMATION

- name: MAP2K1
- type: gene/protein
- source: NCBI - details:
- query: MAP2K1
- alias (other gene names): ['CFC3', 'MAPKK1', 'MEK1', 'MEL', 'MKK1', 'PRKMK1']
- genomic_pos (genomic position): 'chr': '15', 'end': 66491656, 'ensemblgene': 'ENSG00000169032', 'start': 66386837, 'strand': 1
- name (gene name): mitogen-activated protein kinase kinase 1
- summary (protein summary text): The protein encoded by this gene is a member of the dual specificity protein kinase family, which acts as a mitogen-activated protein (MAP) kinase kinase. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals. This protein kinase lies upstream of MAP kinases and stimulates the enzymatic activity of MAP kinases upon wide variety of extra- and intracellular signals. As an essential component of MAP kinase signal transduction pathway, this kinase is involved in many cellular processes such as proliferation, differentiation, transcription regulation and development. [provided by RefSeq, Jul 2008].
- relations:
ppi: gene/protein: (EGR1, MYC, GSK3B, PIK3R1, AURKA, CTCF, CHEK1, ETS1, E2F4, UBE2I, UBC, MAPK14, ARRB2, CDKN2A, ...),
target: drug: (Selumetinib, Cobimetinib, Bosutinib, Trametinib, K-252a, 5-Bromo-N-[(2S)-2,3-dihydroxypropoxy]-3,4-difluoro-2-[(2-fluoro-4-iodophenyl)amino]benzamide, (5S)-4,5-difluoro-6-[(2-fluoro-4-iodophenyl)imino]-N-(2-hydroxyethoxy)cyclohexa-1,3-diene-1-carboxamide, 2-[(2-chloro-4-iodophenyl)amino]-N-[(2R)-2,3-dihydroxypropyl]oxy-3,4-difluorobenzamide, ...)
associated_with: disease: (Noonan syndrome, Prader-Willi syndrome, anxiety disorder, cardiofaciocutaneous syndrome, melanoma, cutaneous malignant, susceptibility to, Costello syndrome, LEOPARD syndrome, Noonan syndrome-like disorder with loose anagen hair, familial prostate carcinoma, prostate cancer, hereditary, lung cancer, ovarian cancer, hereditary breast ovarian cancer syndrome, Noonan syndrome with multiple lentigines, melanoma, non-small cell lung carcinoma (disease), lung adenocarcinoma, ovarian neoplasm, hairy cell leukemia, transient ischemic attack (disease), yolk sac tumor, colorectal adenocarcinoma, gonadoblastoma, ovarian mucinous adenocarcinoma, ovarian adenocarcinoma, ...,
interacts_with: cellular_component: (endoplasmic reticulum, late endosome, microtubule organizing center, nucleus, Golgi apparatus, plasma membrane, early endosome, cytosol, mitochondrion, focal adhesion),molecular_function: (protein serine/threonine kinase activity, protein binding, protein kinase activity, transmembrane receptor protein tyrosine kinase activity, protein serine/threonine/tyrosine kinase activity, MAP kinase kinase activity, protein serine/threonine kinase activator activity, protein N-terminus binding, protein threonine kinase activity, protein serine kinase activity, protein C-terminus binding, scaffold protein binding, ATP binding, MAP-kinase scaffold activity),pathway: (MAPK3 (ERK1) activation, ...,

- name: Binimetinib
- type: drug
- source: DrugBank
- details:

- description: Binimetinib, also known as _Mektovi_, is a potent is a potent and selective oral mitogen-activated protein kinase 1/2 (MEK 1/2) inhibitor.
- half_life: The mean (CV%) terminal half-life (t1/2) of binimetinib is 3.5 hours (28.5%).
- protein_binding: Binimetinib is 97% bound to human plasma proteins and the blood-to-plasma ratio is 0.72
- pharmacodynamics: Binimetinib is a MEK inhibitor. MEK is an enzyme that regulates the biosynthesis of the inflammatory cytokines TNF, IL-6 and IL-1. MEK inhibitors interfere with these biosynthetic processes. It is a chemotherapeutic agent that has anti-tumor activity.
- state: Binimetinib is a solid. - atc_1: Binimetinib is anatomically related to antineoplastic and immunomodulating agents. - atc_2: Binimetinib is in the therapeutic group of antineoplastic agents. - atc_3: Binimetinib is pharmacologically related to other antineoplastic agents. - atc_4: The chemical and functional group of is protein kinase inhibitors.
- category: Binimetinib is part of Antineoplastic Agents ; Antineoplastic and Immunomodulating Agents ; Cytochrome P-450 CYP1A2 Substrates ; Cytochrome P-450 CYP1A2 Substrates with a Narrow Therapeutic Index ; Cytochrome P-450 CYP2C19 Substrates ; Cytochrome P-450 CYP2C19 Substrates with a Narrow Therapeutic Index ; Cytochrome P-450 Substrates ; Heterocyclic Compounds, Fused-Ring ; MAP Kinase Kinase 1, antagonists & inhibitors ; MAP Kinase Kinase 2, antagonists & inhibitors ; Narrow Therapeutic Index Drugs ; P-glycoprotein substrates with a Narrow Therapeutic Index ; Protein Kinase Inhibitors ; UGT1A1 Substrates ; UGT1A1 Substrates with a Narrow Therapeutic Index.
- group: Binimetinib is approved and investigational.
- relations:
enzyme: gene/protein: (CYP1A2, CYP2C19, UGT1A1),
target: gene/protein: (IL1B, IL6, TNF, MAP3K1, MAP2K2),
indication: disease: (metastatic melanoma),
synergistic_interaction: drug: (Prednisolone phosphate, Levothyroxine, Diclofenac, Genistein, Resveratrol, Nelfinavir, Nevirapine, Phenytoin, Conjugated estrogens, Morphine, Desogestrel, Valproic acid, Acetaminophen, Amitriptyline, Indomethacin, Olanzapine, Rosiglitazone, Meperidine, Imipramine, Nabumetone, Fluoxetine, Duloxetine, Chlorpromazine, Zidovudine, Ritonavir, Erlotinib, Ciprofloxacin, . . . ,
. . .

Harmonizome. This is the example of retrieved information from Harmonizome. Note that due to the limited space, we display only partial information of the full retrieved context.

Retrieved information from Harmonizome for (p, c) =(Binimetinib, C32)

GENE INFORMATION

TARGET GENE INFORMATION

The target gene of the perturbation is MAP2K1. The gene name is mitogen-activated protein kinase kinase 1 and the gene symbol is MAP2K1.The protein encoded by this gene is a member of the dual specificity protein kinase family, which acts as a mitogen-activated protein (MAP) kinase kinase. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals. This protein kinase lies upstream of MAP kinases and stimulates the enzymatic activity of MAP kinases upon wide variety of extra- and intracellular signals. As an essential component of MAP kinase signal transduction pathway, this kinase is involved in many cellular processes such as proliferation, differentiation, transcription regulation and development. [provided by RefSeq, Jul 2008]The gene synonyms are MKK1, MEK1, CFC3, PRKMK1, MAPKK1.The gene is related to

the proteins MP2K1_HUMAN.

CELL TYPE INFORMATION

The cell type of the context is C32. The related genes are GANAB, CNOT11, HNRNPR, MAP2K1, CASP3, HABP2, NUAK2, KRTCAP2, IL6, BBS9, LARGE1, AKT1, SGMS1, TYR, RARS1, CERS2, NPL, RAN, PXDN, PRDM16. The gene information is as follows:
The gene name is glucosidase, alpha; neutral AB and the gene symbol is GANAB.This gene encodes the alpha subunit of glucosidase II and a member of the glycosyl hydrolase 31 family of proteins. The heterodimeric enzyme glucosidase II plays a role in protein folding and quality control by cleaving glucose residues from immature glycoproteins in the endoplasmic reticulum. Expression of the encoded protein is elevated in lung tumor tissue and in response to UV irradiation. Mutations in this gene cause autosomal-dominant polycystic kidney and liver disease. [provided by RefSeq, Jul 2016]The gene synonyms are GLUII, G2AN, GIIA.The gene is related to the proteins GANAB_HUMAN.
The gene name is CCR4-NOT transcription complex, subunit 11 and the gene symbol is CNOT11.Predicted to be involved in nuclear-transcribed mRNA poly(A) tail shortening. Part of CCR4-NOT complex. [provided by Alliance of Genome Resources, Mar 2025]The gene synonyms are C40, C2ORF29.The gene is related to the proteins CNO11_HUMAN.
The gene name is heterogeneous nuclear ribonucleoprotein R and the gene symbol is HNRNPR.This gene encodes an RNA-binding protein that is a member of the spliceosome C complex, which functions in pre-mRNA processing and transport. The encoded protein also promotes transcription at the c-fos gene. Alternative splicing results in multiple transcript variants. There are pseudogenes for this gene on chromosomes 4, 11, and 10. [provided by RefSeq, Jul 2014] The gene synonyms are HNRPR, HNRNP-R. The gene is related to the proteins HNRPR_HUMAN.
The gene name is mitogen-activated protein kinase kinase 1 and the gene symbol is MAP2K1. The protein encoded by this gene is a member of the dual specificity protein kinase family, which acts as a mitogen-activated protein (MAP) kinase kinase. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals. This protein kinase lies upstream of MAP kinases and stimulates the enzymatic activity of MAP kinases upon wide variety of extra- and intracellular signals. As an essential component of MAP kinase signal transduction pathway, this kinase is involved in many cellular processes such as proliferation, differentiation, transcription regulation and development. [provided by RefSeq, Jul 2008] The gene synonyms are MAPKK1, PRKMK1, MKK1, CFC3, MEK1. The gene is related to the proteins MP2K1_HUMAN.
...

PubMed. This is the example of retrieved information from PubMed. Note that due to the limited space, we display only partial information of the full retrieved context.

% Retrieved information from PubMed for (p, c) =(Binimetinib, C32)

RELATED PAPER LIST

{ “title": "Binimetinib, a novel MEK1/2 inhibitor, exerts anti-leukemic effects under inactive status of PI3Kinase/Akt pathway.", “abstract": "A MEK1/2 inhibitor, binimetinib is promising as a therapeutic agent for malignant melanoma with N-RAS mutation. We examined in vitro effects of binimetinib on 10 human myeloid/lymphoid leukemia cell lines, and found that three of five cell lines with N-RAS mutation and one of five without N-RAS mutation were responsive to treatment with binimetinib. Binimetinib inhibited cell growth mainly by inducing G"
}

{ “title": "Severe Drug-Induced Liver Injury from Combination Encorafenib/Binimetinib.", “abstract": "Encorafenib/binimetinib is a new combination BRAF/MEK inhibitor used in the treatment of advanced or metastatic BRAFV600-mutant melanoma. Though generally tolerated well, mild to moderate aminotransferase elevations are common. However, significant liver injury has not been demonstrated in the literature. Here, we report the first case of severe hepatic injury associated with encorafenib/binimetinib in a 58-year-old gentleman requiring admission and extensive workup. He was successfully treated by withdrawing the combination therapy, and liver function returned to normal range." }

{ “title": "Phase Ib Study of Combination Therapy with MEK Inhibitor Binimetinib and Phosphatidylinositol 3-Kinase Inhibitor Buparlisib in Patients with Advanced Solid Tumors with RAS/RAF Alterations.", “abstract": "This multicenter, open-label, phase Ib study investigated the safety and efficacy of binimetinib (MEK inhibitor) in combination with buparlisib (phosphatidylinositol 3-kinase [PI3K] inhibitor) in patients with advanced solid tumors with RAS/RAF alterations." }

{ “title": "Encorafenib and binimetinib for the treatment of BRAF-mutated metastatic melanoma in the setting of combined hepatic and renal impairment.", “abstract": "Inhibitors of BRAF, a gene coding a protein called B-raf, with or without inhibitors of MEK (MAPK/extracellular signal-regulated kinase) are often used as palliative treatment in BRAF-mutated metastatic melanoma. Recent data show improved progression-free survival with encorafenib with binimetinib, a newer BRAF/MEK inhibitor combination, compared with older agents, but there have been no reports of this treatment in the setting of renal and liver failure. We report a patient with disease-induced transaminitis and renal failure requiring dialysis who was successfully treated with encorafenib and binimetinib. His transaminitis improved and he was able to stop dialysis without any significant adverse effects during treatment, suggesting encorafenib with binimetinib may be used safely and effectively even in patients with end organ damage." }

{ “title": "MEK Inhibitors in the Treatment of Metastatic Melanoma and Solid Tumors.", “abstract": "The mitogen-activated protein kinase (MAPK) cascade is an intracellular signaling pathway involved in the regulation of cellular proliferation and the survival of tumor cells. Several different mutations, involving BRAF or NRAS, exert an oncogenic effect by activating the MAPK pathway, resulting in an increase in cellular proliferation. These mutations have become targets for new therapeutic strategies in melanoma and other cancers. Selective MEK inhibitors have the ability to inhibit growth and induce cell death in BRAF- and NRAS-mutant melanoma cell lines. MEK inhibitor therapy in combination with a BRAF inhibitor is more effective and less toxic than treatment with a BRAF inhibitor alone, and has become the standard of care for patients with BRAF-mutated melanoma. Trametinib was the first MEK inhibitor approved for the treatment of BRAF-mutated metastatic melanoma not previously treated with BRAF inhibitors, and is also approved in combination with the BRAF inhibitor dabrafenib. Furthermore, cobimetinib is another MEK inhibitor approved for the treatment of BRAF-mutated metastatic melanoma in combination with a BRAF inhibitor, vemurafenib. The MEK inhibitor binimetinib in combination with the BRAF inhibitor encorafenib is in clinical development. The addition of an anti-PD-1/PD-L1 agent, such as pembrolizumab, durvalumab or atezolizumab, to combined BRAF and MEK inhibition has shown considerable promise, with several trials ongoing in metastatic melanoma. Binimetinib has also shown efficacy in NRAS-mutated melanoma patients. Future possibilities for MEK inhibitors in advanced melanoma, as well as other solid tumors, include their use in combination

with other targeted therapies (e.g. anti-CDK4/6 inhibitors) and/or various immune-modulating antibodies." } ...

Wikipedia. This is the example of retrieved information from Wikipedia. Note that due to the limited space, we display only partial information of the full retrieved context.

% Retrieved information from Wikipedia for (p, c) =(Binimetinib, C32)

WIKIPEDIA INFORMATION

MAP2K2
Dual specificity mitogen-activated protein kinase kinase 2 is an enzyme that in humans is encoded by the MAP2K2 gene. It is more commonly known as MEK2, but has many alternative names including CFC4, MKK2, MAPKK2 and PRKMK2.
== Function ==
The protein encoded by this gene is a dual specificity protein kinase that belongs to the MAP kinase kinase family. This kinase is known to play a critical role in mitogen growth factor signal transduction. It phosphorylates and thus activates MAPK1/ERK2 and MAPK3/ERK1. The activation of this kinase itself is dependent on the Ser/Thr phosphorylation by MAP kinase kinase kinases. The inhibition or degradation of this kinase is found to be involved in the pathogenesis of Yersinia and anthrax.
== Interactions ==
MAP2K2 has been shown to interact with MAPK3 and ARAF.
== External links ==
GeneReviews/NCBI/NIH/UW entry on Cardiofaciocutaneous Syndrome Overview of all the structural information available in the PDB for UniProt: P36507 (Dual specificity mitogen-activated protein kinase kinase 2) at the PDBe-KB.

Binimetinib
Binimetinib, sold under the brand name Mektovi, is an anti-cancer medication used to treat various cancers. Binimetinib is a selective inhibitor of MEK, a central kinase in the tumor-promoting MAPK pathway. Inappropriate activation of the pathway has been shown to occur in many cancers. In June 2018 it was approved by the FDA in combination with encorafenib for the treatment of patients with unresectable or metastatic BRAF V600E or V600K mutation-positive melanoma. In October 2023, it was approved by the FDA for treatment of NSCLC with a BRAF V600E mutation in combination with encorafenib. It was developed by Array Biopharma.
== Mechanism of action ==
Binimetinib is an orally available inhibitor of mitogen-activated protein kinase kinase (MEK), or more specifically, a MAP2K inhibitor. MEK is part of the RAS pathway, which is involved in cell proliferation and survival. MEK is upregulated in many forms of cancer. Binimetinib, uncompetitive with ATP, binds to and inhibits the activity of MEK1/2 kinase, which has been shown to regulate several key cellular activities including proliferation, survival, and angiogenesis. MEK1/2 are dual-specificity threonine/tyrosine kinases that play key roles in the activation of the RAS/RAF/MEK/ERK pathway and are often upregulated in a variety of tumor cell types. Inhibition of MEK1/2 prevents the activation of MEK1/2 dependent effector proteins and transcription factors, which may result in the inhibition of growth factor-mediated cell signaling. As demonstrated in preclinical studies, this may eventually lead to an inhibition of tumor cell proliferation and an inhibition in production of various inflammatory cytokines including interleukin-1, -6 and tumor necrosis factor.
== Development ==
In 2015, it was in phase III clinical trials for ovarian cancer, BRAF mutant melanoma, and NRAS Q61 mutant melanoma. In December 2015, the company announced that the mutant-NRAS melanoma trial was successful. In the trial, those receiving binimetinib had a median progression-free survival of 2.8 months versus 1.5 months for those on the standard dacarbazine treatment.

NDA submitted Jun 2016, and the FDA should decide by 30 June 2017. In April 2016, it was reported that the phase III trial for low-grade ovarian cancer was terminated due to lack of efficacy. In 2017, the FDA informed Array Biopharma that the phase III trial data was not sufficient and the New Drug Application was withdrawn. In June 2018, it was approved for the treatment of certain melanomas by the U.S. Food and Drug Administration (FDA) in combination with encorafenib. The FDA approved binimetinib based primarily on evidence from one clinical trial (NCT01909453) of 383 patients with BRAF V600 mutation-positive melanoma that was advanced or could not be removed by surgery. The trial was conducted at 162 sites in Europe, North America, and various countries around the world. In October 2023, the US Food and Drug Administration approved encorafenib with binimetinib for adults with metastatic non-small cell lung cancer (NSCLC) with a BRAF V600E mutation, as detected by an FDA-approved test.

MAP2K1
Dual specificity mitogen-activated protein kinase kinase 1 is an enzyme that in humans is encoded by the MAP2K1 gene.
== Function ==
The protein encoded by this gene is a member of the dual-specificity protein kinase family that acts as a mitogen-activated protein (MAP) kinase kinase. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals. This protein kinase lies upstream of MAP kinases and stimulates the enzymatic activity of MAP kinases upon activation by a wide variety of extra- and intracellular signals. As an essential component of the MAP kinase signal transduction pathway, this kinase is involved in many cellular processes such as proliferation, differentiation, transcription regulation and development. MAP2K1 is altered in 1.05% of all human cancers.
== Meiosis ==
The genomes of diploid organisms in natural populations are highly polymorphic for insertions and deletions. During meiosis double-strand breaks (DSBs) that form within such polymorphic regions must be repaired by inter-sister chromatid exchange, rather than by inter-homolog exchange. Molecular-level studies of recombination during budding yeast meiosis have shown that recombination events initiated by DSBs in regions that lack corresponding sequences in the homolog are efficiently repaired by inter-sister chromatid recombination. This recombination occurs with the same timing as inter-homolog recombination, but with reduced (2- to 3-fold) yields of joint molecules. MAP2K1 is also known as MEK1 (see Mitogen-activated protein kinase kinase). MEK1 is a meiotic chromosome-axis-associated kinase that is thought to slow down, but not entirely block, sister chromatid recombination. Loss of MEK1 allows inter-sister DSB repair and also inter-sister Holliday junction intermediates to increase. Despite the normal activity of MEK1 in reducing inter-sister chromatid recombination, such recombination still occurs frequently during normal budding yeast meiosis (although not as frequently as during mitosis), and up to one-third of all recombination events are between sister chromatids.
== Interactions ==
MAP2K1 has been shown to interact with C-Raf, Phosphatidylethanolamine binding protein 1, MAP2K1IP1, GRB10, MAPK3, MAPK8IP3, MAPK1 MP1, and MAP3K1.

B.2 Generated report

Here, we provide an example generated report.

`Retrieved information from Harmonizome for` (p, c) `=(Binimetinib, C32)`

Comprehensive Mechanistic Report: Binimetinib Treatment in C32 Melanoma Cells

1. Perturbation Description

Binimetinib is a potent and selective oral MEK1/2 inhibitor (MAP2K1/MAP2K2 inhibitor) with the SMILES structure CN1C=NC2=C1C=C(C(=C2F)NC3=C(C=C(C=C3)Br)F)C(=O)NOCCO. It functions as an **ATP-uncompetitive allosteric inhibitor** that binds to a unique allosteric site on MEK1/2, distinct from the ATP-binding pocket. This mechanism allows sustained inhibition even under high ATP concentrations typical in cancer cells.
Key pharmacological properties:

- **IC50 values**: ∼12 nM for MEK1, ∼4.6 nM for MEK2
- **Mechanism**: Non-competitive with ATP, binds to inactive MEK1/2 conformation
- **Selectivity**: >100-fold selectivity for MEK1/2 over other kinases
- **Half-life**: 3.5 hours with 97% plasma protein binding

2. Cellular Context: C32 Melanoma Cell Line
The **C32 melanoma cell line** harbors critical oncogenic mutations that create a hyperactive MAPK pathway:
Driver mutations present:

- **BRAF V600E (GoF)**: Constitutively active BRAF kinase (∼500-fold increased activity)
- **CDKN2A loss (tumor suppressor)**: Loss of p16INK4A and p14ARF tumor suppressors
- **PTEN loss (LoF)**: Hyperactive PI3K/AKT survival pathway
- **TERT activation (GoF)**: Enhanced telomerase activity promoting immortalization

This mutational landscape creates **dual pathway hyperactivation**: MAPK (proliferation) and PI3K/AKT (survival), with compromised cell cycle control and DNA damage response.

3. Step-by-Step Mechanistic Analysis
Step 1: Direct Molecular Target Engagement **Relationship: CAUSAL (direct inhibition)**
Binimetinib directly binds to the allosteric site of both MAP2K1 (MEK1) and MAP2K2 (MEK2) in their inactive conformations. The drug:

- Locks MEK1/2 in a catalytically inactive state
- Prevents conformational changes required for kinase activation
- Blocks MEK1/2 phosphorylation by upstream BRAF V600E
- Results in >95% inhibition of MEK1/2 kinase activity at therapeutic concentrations

Step 2: Disruption of MAPK Signaling Cascade **Relationship: CAUSAL (direct pathway blockade)**
MEK1/2 inhibition immediately disrupts the hyperactive MAPK pathway:
BRAF V600E → [BLOCKED] MEK1/2 → [BLOCKED] ERK1/2

- **ERK1/2 phosphorylation**: Dramatic reduction (>90%) in phospho-ERK1/2 (Thr202/Tyr204)
- **ERK nuclear translocation**: Prevented, blocking transcriptional programs
- **Downstream kinase activity**: RSK1/2, MSK1/2 activation blocked
- **Timeline**: ERK dephosphorylation occurs within 1–2 hours of treatment

Step 3: Transcriptional Reprogramming **Relationship: CAUSAL (loss of ERK-dependent transcription)**
ERK1/2 inhibition leads to immediate transcriptional changes:
Early response genes (1–4 hours):

- **c-MYC downregulation**: Loss of ERK-mediated MYC stabilization and transcription

- **Cyclin D1 (CCND1) downregulation**: Reduced ERK/RSK-mediated transcriptional activation
- **c-FOS/c-JUN reduction**: Decreased AP-1 transcriptional activity
- **EGR1 downregulation**: Loss of immediate early response

Cell cycle genes (4–12 hours):

- **CDK4/6 downregulation**: Reduced transcription and protein levels
- **E2F target genes**: Decreased expression due to Rb hypophosphorylation
- **S-phase genes**: MCM2-7, PCNA, DNA polymerase subunits reduced

Step 4: Cell Cycle Arrest Mechanisms **Relationship: CAUSAL (loss of proliferative signals + intact p53-independent mechanisms)**
Despite CDKN2A loss, multiple mechanisms enforce cell cycle arrest:
G1/S checkpoint activation:

- **Rb protein hypophosphorylation**: Due to reduced CDK4/6 activity
- **E2F1 sequestration**: Hypophosphorylated Rb binds and inactivates E2F1
- **S-phase entry blocked**: Reduced origin licensing and DNA replication machinery

Additional arrest mechanisms:

- **p21 upregulation**: ERK-independent, stress-induced CDKN1A transcription
- **p27 stabilization**: Reduced ERK-mediated p27 degradation signals
- **ATM/ATR activation**: Replication stress responses despite compromised DNA damage checkpoints

Step 5: Metabolic Reprogramming **Relationship: CAUSAL (loss of ERK-mediated metabolic control)**
MEK inhibition disrupts cancer cell metabolism:
Glycolytic pathway:

- **Reduced glucose uptake**: Decreased GLUT1 expression and hexokinase activity
- **Glycolytic enzyme downregulation**: PKM2, LDHA expression reduced
- **ATP production decreased**: Shift from aerobic glycolysis toward oxidative phosphorylation

Anabolic pathways:

- **Protein synthesis reduction**: Decreased mTOR signaling and ribosome biogenesis
- **Lipid synthesis impairment**: Reduced SREBP1 activity and fatty acid synthesis
- **Nucleotide synthesis**: Decreased UMPS, CTPS1 expression

Step 6: Compensatory Pathway Activation **Relationship: CAUSAL (feedback mechanisms and bypass signaling)**
Cells attempt to compensate for MAPK blockade:
RTK upregulation:

- **EGFR/PDGFR overexpression**: Transcriptional and post-transcriptional upregulation
- **Insulin/IGF-1R activation**: Enhanced survival signaling
- **Timeline**: Begins 12–24 hours post-treatment

PI3K/AKT hyperactivation:

- **Enhanced AKT phosphorylation**: Loss of ERK-mediated negative feedback on RTKs
- **mTOR complex reactivation**: Compensatory survival signaling
- **FOXO inactivation**: Maintained via AKT despite ERK loss

Step 7: Apoptotic Priming and Cell Death **Relationship: CAUSAL (inadequate survival signaling + metabolic stress)**

Extended MEK inhibition triggers apoptotic cascades:
Mitochondrial dysfunction:

- **BCL-2 family imbalance**: Increased BIM, PUMA; decreased MCL-1, BCL-2
- **Cytochrome c release**: Mitochondrial outer membrane permeabilization
- **ATP depletion**: Metabolic crisis from glycolytic shutdown

Caspase activation:

- **Caspase-9 activation**: Apoptosome formation
- **Caspase-3/7 activation**: Execution phase initiation
- **PARP cleavage**: DNA repair pathway inactivation
- **Timeline**: Detectable at 24–48 hours, maximal at 72–96 hours

Step 8: Phenotypic Resistance Mechanisms **Relationship: CORRELATIVE (long-term adaptation patterns)**
Prolonged treatment selects for resistant subpopulations:
Genetic adaptations:

- **KRAS/NRAS mutations**: Bypass BRAF dependency (15–20% of cases)
- **MEK1/2 mutations**: C121S, P124L/S mutations reduce drug binding
- **RTK amplifications**: EGFR, PDGFR, MET overexpression

Epigenetic changes:

- **EMT activation**: ZEB1, SNAIL upregulation promoting mesenchymal state
- **Stemness programs**: SOX2, NANOG expression in subpopulations
- **Metabolic reprogramming**: Enhanced oxidative phosphorylation capacity

4. Measurable Biological Endpoints
Transcriptomic Signatures (detectable by RNA-seq):
Early response genes (2–6 hours):

- **Downregulated**: MYC, FOS, JUN, EGR1, CCND1, CDK4, CDK6
- **Upregulated**: CDKN1A (p21), GADD45A, DDIT3, ATF3

Cell cycle signatures (12–24 hours):

- **S-phase genes decreased**: MCM2-7, PCNA, RRM1/2, TYMS
- **DNA repair genes**: BRCA1, RAD51, XRCC1 (context-dependent changes)

Metabolic gene signatures (6–24 hours):

- **Glycolytic genes down**: SLC2A1 (GLUT1), HK2, PKM2, LDHA
- **Oxidative metabolism**: Complex I-V subunits, PGC1α (variable)

Apoptotic signatures (24–72 hours):

- **Pro-apoptotic**: BBC3 (PUMA), BCL2L11 (BIM), BAX, BAK1
- **Anti-apoptotic decreased**: MCL1, BCL2L1 (BCL-xL)

Phenomic Signatures (detectable by high-content imaging):
Morphological changes:

- **Cell size reduction**: 20–30% decrease in cell area (12–24 hours)
- **Nuclear morphology**: Condensed chromatin, irregular nuclear shape
- **Cytoplasmic changes**: Reduced cytoplasmic volume, organelle reorganization

Functional measurements:

- **Proliferation rate**: 60–80% reduction in EdU incorporation (24 hours)
- **Mitochondrial potential**: 30–50% reduction in TMRM fluorescence (48 hours)

- **Caspase activity**: 5–10 fold increase in caspase-3/7 activity (48–72 hours)
- **Cell viability**: 40–70% reduction by 96 hours (concentration-dependent)

Cell cycle distribution:

- **G1 accumulation**: 60–80% of cells (vs. 40–50% control)
- **S-phase reduction**: 5–15% of cells (vs. 30–40% control)
- **Sub-G1 population**: 10–30% at 72–96 hours (apoptotic cells)

5. Final Phenotypic Outcome
Binimetinib treatment induces a cytostatic-to-cytotoxic response in C32 melanoma cells. The perturbation initially causes **G1/S cell cycle arrest** through MAPK pathway blockade, leading to **reduced proliferation** and **metabolic stress**. Over time, inadequate survival signaling overcomes the hyperactive PI3K/AKT compensation, resulting in **mitochondrial dysfunction** and **caspase-dependent apoptosis**.
Key measurable outcomes:

- **Transcriptomic**: Immediate early gene suppression, cell cycle gene downregulation, delayed apoptotic gene activation
- **Phenomic**: Progressive cell cycle arrest, reduced cell size, decreased mitochondrial function, increased caspase activity, and ultimate cell death

The response demonstrates how MEK inhibition can overcome oncogene addiction in BRAF-mutant melanoma despite multiple resistance mechanisms, making it a clinically effective therapeutic strategy when combined with BRAF inhibitors.

C Experimental details

We provide the VCR-AGENT code in https://anonymous.4open.science/r/VCR-Agent_.

C.1 Detailed experimental results

Here, we present the detailed numerical results of the experiments discussed in Section 4.2. We provide the results in Table Appx.2.

C.2 Prompts

Report generator. This is the prompt used for report generator.

Prompt for report generator

Biomedical Reasoning Assistant – Report Generation

You are a biomedical reasoning assistant. Your task is to generate a **comprehensive mechanistic report** describing how the specified perturbation affects cellular biology in the given context. The report will later be transformed into a structured explanation, so include **all details necessary for reasoning** that links the perturbation to the phenotype.

—
INSTRUCTIONS
1. **Describe the perturbation in detail**
- Include its type (chemical, genetic, etc.), primary target(s), known binding affinities or potencies, and mechanism of action (e.g., ATP-competitive inhibition, PROTAC-mediated degradation).
2. **Map the full causal chain step by step**
- For example, you can **start from the perturbation** → molecular target → pathway modulation → downstream molecular changes → **phenotypic effect**.
- Explicitly mark whether each step is:
- **Causal** (direct mechanistic or experimental evidence)
- **Correlative** (statistical or inferred association)
3. **Prioritize measurable end nodes (effects)**
- The **final nodes in the chain should, whenever possible, correspond to (or can

Task	Model	C32	HOP62	HepG2/C3A	Hs 766T	PANC-1	Average	Union
Differential expression	Random	0.269	0.254	0.232	0.239	0.301	0.259	0.286
	Mean	0.405	0.419	0.345	0.465	0.330	0.393	0.401
	K-neighbor	0.459	0.419	0.343	0.453	0.355	0.406	0.403
	STATE (ST)	0.273	0.280	0.199	0.219	0.313	0.257	0.251
	Zero-shot	0.363	0.423	0.316	0.332	0.399	0.367	0.338
	SFT	0.344	0.353	0.163	0.363	0.236	0.292	0.291
	Ours (SFT - Prompt)	**0.470**	**0.470**	**0.362**	**0.470**	**0.405**	**0.435**	**0.452**
	Ours (SFT - Generate)	0.412	0.446	0.328	0.424	0.329	0.388	0.441
Direction of change	Random	0.525	0.526	0.517	0.522	0.558	0.530	0.513
	Mean	0.847	0.813	0.802	0.812	0.811	0.817	0.810
	K-neighbor	0.798	0.775	0.761	0.736	0.784	0.771	0.757
	STATE (ST)	0.756	0.685	0.704	0.737	0.721	0.721	0.695
	Zero-shot	0.101	0.063	0.055	0.049	0.061	0.066	0.062
	SFT	0.839	0.810	0.815	**0.822**	0.827	0.823	**0.820**
	Ours (SFT - Prompt)	**0.855**	**0.827**	**0.830**	0.818	**0.832**	**0.832**	**0.820**
	Ours (SFT - Generate)	0.702	0.707	0.704	0.688	0.773	0.715	0.727

Table Appx.2: **TahoeQA performance with total model.** The best results are highlighted in **bold** and the improvement compared to the vanilla SFT is highlighted with teal.

be inferred from) measurable outputs from the assays available**:
- **Transcriptomics**: changes in individual gene expression or gene signatures.
- **Phenomics (imaging/physiology)**: morphological or functional features (e.g., cell size, nuclear morphology, proliferation rate, apoptosis, mitochondrial potential).
- **Only use higher-order functional or disease phenotypes** if these can be clearly linked to transcriptomic or phenomic signatures.
4. **Include associative evidence** and **ontological evidence** when available
- Add correlations, transcriptomic signature similarities, or phenotypic fingerprint associations when direct causality is unclear.
5. **Summarize the final phenotypic outcome**
- Explicitly state whether the perturbation induces, rescues, or exacerbates the measured phenotype and what the phenotype is.
6. **Provide quantitative and qualitative details when you have them**
- Affinities (IC50, Kd), phosphorylation sites, key genes modulated, direction of regulation, morphological metric shifts, etc.
The report should be sufficiently detailed to reconstruct the full reasoning path from **perturbation –> measurable biological effect** and can be used to generate hypotheses with all the details for their falsification.

QUESTION
Q: How does the following perturbation influence the cell in the described context, mechanistically and functionally?
{treatment}

EXPECTED OUTPUT
Write a **detailed mechanistic report** explaining step by step how this perturbation affects the biology in this context, ending with the **measurable transcriptomic or phenomic effects whenever possible**.

Explanation constructor. This is the prompt used for explanation constructor.

Prompt for explanation constructor

Biomedical Reasoning Assistant - Structured Explanation Task

You are a biomedical reasoning assistant. Explain how a specific perturbation alters cellular biology in a given context using both causal mechanisms and associative evidence. Your goal is to translate a mechanistic report into a structured, falsifiable explanation of the perturbation's effect.

Task Overview

You must produce **four outputs**:

1. **Private Reasoning** (<think>) 2. **Mechanism-of-Action Summary** (<answer>) 3. **Structured Explanation** (<explain>) 4. **Causal DAG of Events** (<dag>)
Each step is strictly defined below.

1. Private Reasoning
Wrap your step-by-step biological reasoning inside <think>...</think>.
- Proceed as if you are discovering the answer for the first time. - Simulate the biology: consider binding events, cascades, interactions, and regulatory outcomes, but leverage the reports you have been given to guide your reasoning. - Mark each connection as either causal or correlative. - Indicate when evidence is inferred, indirect, or weak.
—
2. Summary Answer
Wrap a **1-paragraph summary** inside <answer>...</answer>.
- Clearly describe the mechanism of action (MoA). - Use proper biological terminology. - Distinguish between mechanistic (causal) and associative (correlative) findings. - Conclude with the predicted phenotypic or functional outcome.
—
3. Structured Mechanistic Explanation
Wrap your structured actions inside <explain>...</explain>.
- Begin with 'set_context(...)' to define background conditions, including cell type/line, disease model, and any prior perturbation. - Then, **list one primitive per line**, using only **allowed action primitives**. - Each line must follow the **signature** of the primitive action. - Each line must include a unique 'id' (e.g., 'id="n3"') for later referencing in the DAG. - Your last 'explain' step must be a measurable biological output (omics, phenomics, etc.): 'induces_phenotype(...)', 'alleviates_phenotype(...)', 'regulates_expression(...)', 'regulates_translation(...)' - Do not end on upstream mechanistic nodes and always ground the explanation in a testable, measurable biological output. - Do not use the same 'id' for multiple primitives. - Use the 'via' as optional field to reference upstream mechanisms for causal steps or explain the evidence basis for correlative ones.
Example line:
binds_to(id="n1", actor="EW-7197", target="TGFBR1", affinity="0.010 μM", via="ATP site occupancy")
—
4. Causal DAG
Output a directed acyclic graph (DAG) inside <dag>...</dag>.
- The DAG must reference the 'id's declared in the primitives in the <explain> block. - Define directionality (edge types) - 'causal': direct mechanistic consequence - 'correlative': statistical or inferred association without mechanistic support
For example: edge("source_id", "target_id", relation="causal|correlative")
—
Allowed Primitives
Use only the following set of structured primitives:
{action_primitives}
Each primitive has a specific signature and biological scope. Refer to the full list above for descriptions and valid fields.
—
Final Output Format
<think>
...step-by-step causal and correlative reasoning...
</think>
<answer>
...Mechanism-of-action and functional effect summary...
</answer>
<explain>
set_context(cell_type="...", genotype="...", disease="...", prior_perturbation="...")
primitive_1(...) primitive_2(...) ...
</explain>
<dag>
edge("n1", "n2", relation="causal")
edge("n2", "n3", relation="causal")
edge("n3", "n4", relation="correlative") ...
</dag>

Example
Q: How does the following perturbation influence the cell in the described context, mechanistically and functionally?
{ "context": { "perturbation_type": "loss-of-function", "description": "CRISPR knockdown of the LEMD3 gene", "cell_type": "N/A", "disease_model": "Buschke-Ollendorff syndrome" }, "perturbation": { "type": "chemical", "smiles": "CC1=NC(=CC=C1)C2=C(N=C(N2)CNC3=CC=C(=NC=N5)C=C4", "name": "EW-7197", "target": "ALK5", "moa_type": "inhibitor" } }
A: <think>
LEMD3 LoF causes hyper-TGF-β signalling (causal). EW-7197 is an ATP-competitive inhibitor of TGFBR1 (causal). Binding $\rightarrow$ TGFBR1 activity $\downarrow$ (causal) $\rightarrow$ SMAD2/3 phosphorylation $\downarrow$ (causal). Lower p-SMAD2/3 $\rightarrow$ nuclear import $\downarrow$ $\rightarrow$ ECM genes COL1A1 & FN1 $\downarrow$ (causal). High COL1A1/FN1 correlates with fibrotic matrix overproduction (correlative). Therefore EW-7197 is likely to rescue the fibrotic phenotype.
</think>
<answer>
EW-7197 binds the TGF-β type I receptor (TGFBR1) with low-nanomolar affinity, competitively inhibiting its kinase activity. Reduced TGFBR1 signaling lowers SMAD2/3 phosphorylation and nuclear translocation, leading to down-regulation of collagen-I (COL1A1) and fibronectin (FN1) transcripts. Because elevated COL1A1/FN1 expression is associated with excessive extracellular-matrix deposition in LEMD3-deficient fibroblasts, EW-7197 is predicted to rescue the fibrotic-matrix phenotype characteristic of Buschke-Ollendorff syndrome.
</answer>
<explain>
set_context(cell_type="dermal fibroblast", disease_context="Buschke-Ollendorff syndrome", prior_perturbation="LEMD3 CRISPR knockdown") binds_to(id="n1", actor="EW-7197", target="TGFBR1", affinity="0.010 µM", via="ATP site occupancy") modulates_molecule_activity(id="n2", target="TGFBR1", direction="down", via="competitive inhibition") modulates_molecule_activity(id="n3", target="SMAD2/3 complex", direction="down", via="reduced phosphorylation by TGFβ receptor") regulates_expression(id="n4", regulator="SMAD2/3 complex", gene_list=["COL1A1","FN1"], direction="down", via="loss of SMAD2/3 nuclear translocation") alleviates_phenotype(id="n5", actor="EW-7197", phenotype="fibrotic matrix overproduction", via="lower ECM gene expression")
</explain>
<dag>
edge("n1", "n2", relation="causal") edge("n2", "n3", relation="causal") edge("n3", "n4", relation="causal") edge("n4", "n5", relation="correlative")
</dag>

Question
{question}

Original Report
{report}

LLM-judge. This is the prompt used for LLM-judge.

Prompt for LLM-judge

You are an expert biologist. Evaluate the following response for its scientific plausibility, coherence and correctness.

Question:
{question}

Response:
{response}

Assess the explanation on a scale of 0-10 for the following criteria:

1. **"scientific_accuracy"**:
* **Description**: Are the biological claims, pathways, and interactions factually

correct according to current scientific consensus? Are gene/protein names correct? Penalize assertions with low confidence or known inaccuracies.
* **Score**: [0-10]
* **Instruction**: "confidence="low"", "confidence="lost"" should be penalized, the trace should have lower score (0-3), "confidence="correlated"" and "confidence="inferred"" should also be be penalized where the trace should have score 4-5. Unless all steps are fully specified, it shouldn't output a score between 8-10.

2. **"logical_consistency"**:
* **Description**: Does the explanation present a coherent, logical argument? Do the conclusions drawn logically follow from the premises provided within the text?
* **Score**: [0-10]
* **Instruction**: * If there is a loss of function of gene x, it would be wrong if any of the following trace has "binds_to" to x protein. The "logical_consistency" should be penalized. * If the target of "binds_to" is "unknown cellular targets", the following traces would be meaningless. The "logical_consistency" should be penalized. Unless all steps are fully specified, it shouldn't output a score between 8-10.

3. **"mechanistic_clarity"**:
* **Description**: How clearly is the underlying biological mechanism explained? Vague or ambiguous terms should be penalized. Penalize missing actors/targets, unspecified directions, hand-wavy pathways, unlabeled compartments.
* **Score**: [0-10]
* **Instruction**:
* "binds_to": Penalize missing actors/targets. "actor" and "target" should be specific (e.g. "a protein" is not a specified term for proteins and should be penalized. A compound should be specified by name such as "aspirin" or a SMILES syntax. A term like "unknown" or "a drug" should be heavily penalized.).
* "regulates_expression": "gene_list" should be a specific list of genes (e.g., ["FASN","ACC1","SCD1"]) or a well known gene group (e.g., interferon-stimulated genes in interferon signaling pathway). The "direction" must be specified (e.g., "up" or "upregulate", "down" or "downregulate") and terms like "unknown" should be heavily penalized.
* "induces_phenotype"/"alleviates_phenotype": "phenotype" must be specific (e.g., "apoptosis," not "a cellular process").
* "localizes_to": The "entity" must be a specific molecule and the "to_loc" must be a specific compartments (e.g., "nucleus," not "part of the cell").
* "modulates_molecule_activity": The "target" has to be a specified molecule or protein. Ambiguous terms like 'cellular proteins' or 'a molecule' should be heavily penalized.
* "modulates_pathway_activity": The "pathway" has to be a specified and well defined pathway (e.g., MAPK pathway). Any hand-wavy pathway name (e.g., cellular process) should be penalized.
Unless all steps are fully specified, testable, clear, it shouldn't output a score between 8-10.
Your response MUST be a single JSON object containing the numeric scores (0-10) for each criterion. Do not include any other text or explanations. You should only use the following keys: "scientific_accuracy", "logical_consistency", "mechanistic_clarity". Example of a valid response: {{"scientific_accuracy": 8, "logical_consistency": 9, "mechanistic_clarity": 7}}

C.3 Hyperparameters

We report the detailed hyperparameters used for the supervised fine-tuning (SFT) and the subsequent inference generation in Section 4.2.

Training Hyperparameters. The model was fine-tuned with a learning rate of 210^{-4} using a linear scheduler and a warmup ratio of 0.05. To stabilize training, we employed a weight decay of 0.01 and set the gradient accumulation steps to 4. For parameter-efficient fine-tuning, we utilized LoRA with a rank (r) of 64. All experiments were conducted with a fixed random seed of 11 to ensure reproducibility.

Inference Hyperparameters. For generating structured explanations during evaluation and downstream tasks, we adopted specific sampling strategies to balance diversity and precision. We set the temperature to 0.2, combined with nucleus sampling (top-p) of 0.8 and a top-k value of 20.

C.4 Computational resources

We used H100 GPUs for TahoeQA task described in Section 4.2.

D Additional experiments

D.1 Ablation study: Retrieval source

We ablate the contribution of each retrieval source by generating explanation traces using only a single knowledge base at a time, compared to the full combination used by VCR-AGENT. We employ the same metrics as in Table 1. As shown in Table Appx.3, no single source approaches the performance of the full combination. This confirms that each database contributes complementary biological knowledge, with StarkPrimeKG providing relational context, Harmonizome enriching gene-level information, PubMed supplying literature-grounded evidence, and Wikipedia offering broad background knowledge.

D.2 Ablation study: Backbone LLM

We compare the backbone LLM used in VCR-AGENT (Claude) against GPT-4.1. and Gemini-2.5-Flash, keeping the retrieval pipeline and explanation constructor prompts identical. As shown in Table Appx.4, Claude achieves perfect trace validity. On DTI score, Claude substantially outperforms both alternatives. GPT-4.1 achieves a higher DE score, but at the cost of lower verifiability, meaning a large fraction of its outputs are structurally invalid and cannot be reliably evaluated. These results validate Claude as the backbone choice, balancing format adherence with biological accuracy.

D.3 Ablation study: One-step generation

We evaluate whether the two-stage pipeline (report generation followed by explanation construction) is necessary by comparing it against a one-step baseline that directly generates structured explanations from the perturbation input and retrieved information without an intermediate report. As shown in Table Appx.5, the two-step pipeline nearly doubles both DTI and DE scores over one-step generation, while also achieving higher verifiability. This confirms that decoupling knowledge synthesis from structured reasoning generation is critical.

D.4 Ablation study: Given the same retrieved report

To disentangle the contribution of our structured reasoning formalism from the retrieval advantage, we evaluate all baseline models when given the same retrieved report generated by our report generator as input. This isolates whether the perfor-

	Format		Verifier	
Retrieval source	Validity	Verifiability	Drug-target interaction	Differential expression
PubMed	1.000 ± 0.000	0.300 ± 0.099	0.432 ± 0.059	0.435 ± 0.050
StarkPrimeKG	1.000 ± 0.000	0.296 ± 0.089	0.382 ± 0.031	0.449 ± 0.054
Harmonizome	1.000 ± 0.000	0.297 ± 0.092	0.399 ± 0.013	0.457 ± 0.051
Wikipedia	1.000 ± 0.000	0.292 ± 0.091	0.469 ± 0.028	0.485 ± 0.063
All (Ours)	**1.000** ± 0.000	**0.601** ± 0.009	**0.766** ± 0.008	**0.487** ± 0.054

Table Appx.3: **Ablation study on retrieval source.** The best results are highlighted in **bold**. The standard deviation is computed across cell lines.

	Format		Verifier	
Backbone	Validity	Verifiability	Drug-target interaction	Differential expression
GPT-4.1	0.970 ± 0.002	0.299 ± 0.105	0.399 ± 0.063	0.527 ± 0.076
Gemini-2.5-Flash	0.060 ± 0.016	0.963 ± 0.078	0.338 ± 0.036	0.207 ± 0.036
Claude (Ours)	**1.000** ± 0.000	**0.601** ± 0.009	**0.766** ± 0.008	**0.487** ± 0.054

Table Appx.4: **Ablation study on backbone LLM.** The best results are highlighted in **bold**. The standard deviation is computed across cell lines.

mance gap in Table 1 is driven by access to richer context or by the structured explanation pipeline itself. As shown in Table Appx.6, VCR-AGENT maintains the highest validity and verifiability even when all models receive identical retrieved context. Notably, while Qwen3 achieves a higher DE score (0.556 vs. 0.457), this comes at the cost of substantially lower verifiability (0.340 vs. 1.000), meaning that a large fraction of its outputs are structurally invalid and cannot be reliably evaluated or used downstream. Notably, we clarify that the baseline model selection excluding the retrieval in Table 1 is intentional, as the retrieval pipeline is a core component of our framework.

D.5 Ablation study: Verifier-based filtering

In this section, we conduct a component-wise ablation study to evaluate the effectiveness of the verifier-based filtering pipeline proposed in Section 3.3. To evaluate, we compare the quality of explanation traces under four distinct conditions: (1) No Filtering (baseline), (2) DTI-verifier only, (3) DE-verifier only, and (4) the Full Pipeline. For this evaluation, we utilize the complete explanation traces derived from 18,950 perturbation-context pairs within the Tahoe-100M (Zhang et al., 2025) dataset.

Metrics. We assess the quality of the filtered datasets using three LLM-as-judge evaluation metrics:

- Scientific accuracy: Measures whether the mechanistic claims are biologically valid and factually grounded.
- Logical consistency: Evaluates whether each step follows a coherent progression without contradictions.
- Mechanistic clarity: Assesses whether the underlying biological mechanism is articulated clearly.

The detailed prompts are provided in Appendix C.

Results. As presented in Table Appx.7, verifier-based filtering yields modest but consistent improvements in the overall quality of the explanation traces across all metrics. Notably, while explanations refined via verifier-based filtering is designed to improve the mechanistic clarity and factual accuracy, these gains are not always captured by LLM-based evaluators. This limitation stems from the evaluator's lack of the specialized regulatory knowledge such as those from Boltz-2 (Passaro et al., 2025) and Tahoe-100M (Zhang et al., 2025), which are used by our verifiers. Consequently, the evaluator struggles to distinguish between filtered and unfiltered traces as long as both maintain a high degree of surface-level mechanistic plausibility.

	Format		Verifier	
Step	Validity	Verifiability	Drug-target interaction	Differential expression
One-step	0.999 ± 0.002	0.534 ± 0.067	0.329 ± 0.026	0.267 ± 0.051
Two-step (Ours)	**1.000** ± 0.000	**0.601** ± 0.009	**0.766** ± 0.008	**0.487** ± 0.054

Table Appx.5: **Ablation study on one-step generation.** The best results are highlighted in **bold**. The standard deviation is computed across cell lines.

	Format		Verifier	
Model	Validity	Verifiability	Drug-target interaction	Differential expression
DeepSeek	0.046 ± 0.051	0.711 ± 0.269	0.185 ± 0.099	0.246 ± 0.145
Qwen3	0.937 ± 0.055	0.340 ± 0.173	0.081 ± 0.056	0.556 ± 0.152
Llama3.3	0.789 ± 0.105	0.468 ± 0.227	0.079 ± 0.098	0.231 ± 0.146
Two-step (Ours)	**1.000** ± 0.000	**0.601** ± 0.009	**0.766** ± 0.008	**0.487** ± 0.054

Table Appx.6: **Ablation study given the same retrieved report.** The best results are highlighted in **bold**. The standard deviation is computed across cell lines.

	Scientific accuracy	Logical consistency	Mechanistic clarity
Without verifiers	0.641	0.720	**0.725**
+DE	0.642	0.721	**0.725**
+DTI	0.643	0.721	**0.725**
+DE&DTI	**0.645**	**0.724**	**0.725**

Table Appx.7: **Ablation study on verifier-based filtering.** The best results are highlighted in **bold**.

D.6 Human evaluation

To validate that our LLM-based evaluation captures biologically meaningful quality, we conducted a human expert evaluation on a randomized subset of generated explanation traces and measured agreement with LLM-judge scores.

Setup. We randomly sampled 10 explanation traces from VC-TRACES, stratified across the five cell lines used in our experiments. Each trace was independently scored by domain experts with backgrounds in molecular biology and pharmacology, using the same three criteria used in Section D.5: *scientific accuracy*, *logical consistency*, and *mechanistic clarity*. In addition, experts provided a binary plausibility judgment (plausible vs. implausible) for each trace. Annotators were provided with the perturbation-context pair alongside the generated structured explanation, without access to the LLM-judge scores or verifier outputs.

Results. We report the Pearson correlation between human expert ratings and LLM-judge scores in Table Appx.8 and visualize the per-trace agreement in Figure Appx.1. Across all three criteria, we observe strong positive correlation (average $r = 0.72$), confirming that the LLM-judge serves as a reliable proxy for domain expert assessment. We additionally report the average LLM-judge scores across the full set of traces, which indicate consistently high quality across all criteria.

	Scientific accuracy	Logical consistency	Mechanistic clarity
Human–LLM corr. (r)	0.72	0.69	0.76
Average LLM-judge	0.65	0.72	0.73

Table Appx.8: **Human expert evaluation.** Pearson correlation between human expert ratings and LLM-judge scores on a randomized subset of traces, alongside average LLM-judge scores on the full dataset.

As shown in Figure Appx.1, traces judged as plausible by experts (teal) consistently cluster in the upper-right region of the score space, while traces judged as implausible (orange) tend to receive lower scores from both human experts and the LLM-judge. This alignment indicates that the LLM-judge not only correlates with expert scores numerically but also agrees on which traces are biologically sound, providing confidence that it can serve as a scalable proxy for expert review.

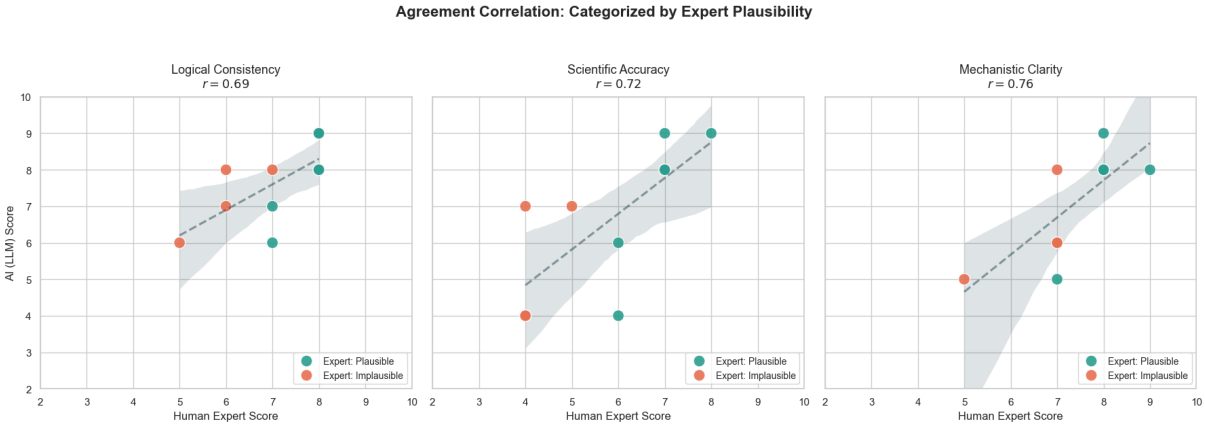


Figure Appx.1: **Agreement between human expert and LLM-judge scores.** Each point represents a single trace. Teal points indicate traces judged as plausible by experts; orange points indicate implausible traces. Dashed lines show the linear regression fit and shaded regions denote the 95% confidence interval. Pearson r is reported per criterion.

E Additional verifiers

In this section, we provide detailed description of additional verifiers.

DE verifier. The DE verifier validates the `regulates_expression` action by checking whether the perturbation regulates the target gene. This verifier queries ground-truth differential expression datasets, such as Tahoe-100M (Zhang et al., 2025), to confirm if the predicted target gene is significantly regulated in the given direction. This identifies hallucinated gene targets, flagging predictions that contradict established biological knowledge. For filtering, the DE verifier prunes factually inconsistent predictions. This step eliminates gene arguments that correspond to incorrectly identified or directionally mismatched gene expression changes.

LOC verifier. The LOC verifier validates the `localizes_to` action by cross-referencing the claimed subcellular localization against curated annotations from UniProt (Ahmad et al., 2025) and the Human Protein Atlas (Thul and Lindskog, 2018). For each `localizes_to` claim, the verifier checks whether the specified entity is annotated to the claimed `from_loc` and `to_loc` compartments.

PHENO verifier. The PHENO verifier assesses the `induces_phenotype` and `alleviates_phenotype` actions by querying phenotypic databases. Specifically, it maps the perturbation to known phenotypic profiles documented in the Cellular Phenotype Database (Kirsanova et al., 2015). The verifier checks whether the claimed phenotype is consistent with documented phenotypic associations for the perturbation or its downstream targets.

We emphasize that the verification suite is designed to be extensible; as reliable computational tools become available for additional action types, new verifiers can be incorporated into the pipeline without modifying the overall architecture.

F Ethical consideration

Despite the improvements in factual precision, we emphasize that the generated reasoning traces are intended for mechanistic plausibility rather than formal causal discovery or direct clinical implementation. The framework's dependency on external knowledge bases (e.g. StarkPrimeKG (Wu et al., 2024b) and Harmonizome (Diamant et al., 2024)) means that inherent biases or incomplete data in these repositories may influence the validity of the reasoning output. Consequently, there is a risk of model hallucination in scientific contexts where external grounding is insufficient or contradictory.

To mitigate these risks, we employ a verifier-based filtering pipeline. While our current implementation includes only the verifiers most critical to our downstream tasks—Drug-Target Interactions and Differential Expression—this represents an extensible foundation rather than an exhaustive solution. We strongly encourage researchers to adopt and expand this verification suite to meet their specific domain requirements. Continued expansion of these capabilities across diverse biological modalities is essential to mitigate the risk of biologically implausible reasoning, which is particularly critical in sensitive biomedical research. Adherence to these rigorous validation standards ensures that autonomous virtual cells serve as reliable and ethical collaborators in the pursuit of scientific truth.