VCWORLD: A BIOLOGICAL WORLD MODEL FOR VIRTUAL CELL SIMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual cell modeling aims to predict cellular responses to perturbations. Existing virtual cell models rely heavily on large-scale single-cell datasets, learning explicit mappings between gene expression and perturbations. Although recent models attempt to incorporate multi-source biological information, their generalization remains constrained by data quality, coverage, and batch effects. More critically, these models often function as black boxes, offering predictions without interpretability or consistency with biological principles, which undermines their credibility in scientific research. To address these challenges, we present VCWorld, a cell-level white-box simulator that integrates structured biological knowledge with the iterative reasoning capabilities of large language models to instantiate a biological world model. VCWorld operates in a data-efficient manner to reproduce perturbation-induced signaling cascades and generates interpretable, stepwise predictions alongside explicit mechanistic hypotheses. In drug perturbation benchmarks, VCWorld achieves state-of-the-art predictive performance, and the inferred mechanistic pathways are consistent with publicly available biological evidence. Our code is publicly available at https://anonymous.4open.science/r/VCWorld-B970.

1 INTRODUCTION

Cells, the fundamental units of life, maintain organismal function and homeostasis through a complex interplay of biochemical processes (Alberts et al., 2002). A central challenge in modern biology and drug discovery is to understand and predict how cells respond to external perturbations, such as drug treatments or genetic edits (Liberali et al., 2015; Lotfollahi et al., 2019). The ability to forecast these cellular state changes *in silico* would not only illuminate the mechanisms of complex diseases but also accelerate the development of novel therapeutics by reducing the time and cost of experimental screening (Del Sol et al., 2010). The concept of the virtual cell, which leverages computational models to simulate cellular behavior, has emerged as a promising paradigm to address this challenge (Bunne et al., 2024).

Recent advances in deep learning, coupled with significant progress in single-cell sequencing technologies, have spurred the development of various virtual cell models (Lopez et al., 2018; Lotfollahi et al., 2023; Adduri et al., 2025; Tang et al., 2025; Klein et al., 2025). These models typically learn an end-to-end mapping from a given perturbation to a corresponding gene expression profile, trained on large-scale perturbation-response datasets. However, prevailing approaches suffer from two critical limitations. First, they are heavily reliant on the scale, quality, and coverage of the training data. The data-hungry nature makes these models expensive to train. More importantly, it also limits their ability to generalize to novel perturbations that are not present in the training data (Li et al., 2024; Ahlmann-Eltze et al., 2024). Second, these models operate as black boxes. While they may yield predictive outputs, they fail to provide clear, verifiable mechanistic explanations for predictions (Hassija et al., 2024; Noutahi et al., 2025). This lack of interpretability severely undermines their trustworthiness and utility in scientific discovery, making them difficult for biologists to rely upon for designing downstream experiments.

We argue that an ideal virtual cell model should not only provide accurate predictions but also be data-efficient, interpretable, and aligned with established biological principles. Rather than relying solely on statistical correlations, it ought to integrate fundamental biological knowledge to capture

and anticipate cellular responses. Moreover, its reasoning process should be transparent, explicitly revealing the causal mechanisms that underlie predictions and grounding them in the well-established frameworks of cell signaling and gene regulation.

To overcome the above challenges, we introduce VCWorld, a cell-level white-box simulator(Figure 1). The core of VCWorld is a biological world model that simulates the dynamic response of a cell to drug perturbations. Instead of relying solely on statistical patterns, VCWorld integrates structured biological knowledge, such as signaling pathways, protein-protein interactions, and gene regulatory networks, with the iterative reasoning capabilities of Large Language Models (LLMs) (Dubey et al., 2024; Guo et al., 2025). This design allows the model to generalize from limited training data by leveraging a vast repository of open-world biological knowledge. Crucially, VCWorld generates a transparent, traceable reasoning path for each prediction, offering a step-by-step mechanistic explanation that culminates in verifiable hypotheses. Furthermore, to facilitate finer-grained modeling, we introduce GeneTAK, a new benchmark derived from the large-scale Tahoe-100M dataset (Zhang et al., 2025). GeneTAK reframes cell-drug observations into gene-centric perturbation responses, mitigating data sparsity and enabling models to focus directly on the nuanced impact of a drug on individual genes. We highlight three main contributions of our study:

- We propose VCWorld, a novel cell-level white-box simulator architected as a biological world model. It combines structured biological knowledge with LLM-based reasoning, demonstrating a superior balance of data efficiency, interpretability, and predictive accuracy that overcomes the limitations of existing black-box models.
- We construct and introduce GeneTAK, a new benchmark that transforms cell-drug perturbation data into single-gene response profiles. This allows for more granular and robust modeling of drug effects.
- We demonstrate that VCWorld achieves state-of-the-art performance on both the differential expression (DE) and directional (DIR) prediction tasks on the GeneTAK benchmark, validating the effectiveness of our approach.

2 Related Work

2.1 Virtual Cell

Virtual cell was originally a computational tool for simulating intracellular biochemical reactions, diffusion, membrane transport, and electro physiological processes (Loew & Schaff, 2001). With the development of AI, the concept has evolved into predictive cell models that integrate large-scale multimodal biological data to forecast, explain, and guide experimental hypotheses (Bunne et al., 2024; Noutahi et al., 2025). Existing research can be broadly categorized into three approaches: data-driven methods combined with prior knowledge, such as GenePT (Chen & Zou, 2024) and GEARS (Roohani et al., 2024); large-scale pretrained foundational models, like scFoundation (Hao et al., 2024) and scGPT (Cui et al., 2024), trained on tens of millions of cells to learn general representations and excel in perturbation prediction; and generative modeling approaches, including CPA (Lotfollahi et al., 2023), STATE (Adduri et al., 2025), and CellFlow (Klein et al., 2025), which capture complex perturbation effects through decoupled latent spaces, state transitions, or flow-based generation. Further efforts, such as CellForge (Tang et al., 2025), aim to automate data analysis, literature review, and model design, pushing virtual cell toward self-directed evolution. Unlike previous approaches that mainly focus on end-to-end black-box models for cell perturbation prediction, our approach leverages a language model with reasoning and retrieval capabilities to construct a framework that emphasizes interpretability while also achieving superior performance.

2.2 LLMs for Biological Reasoning and Prediction

Recent research focus on leveraging Large Language Models (LLMs) for interpretable biological reasoning, aiming to move beyond traditional black-box predictions. A core approach in this area is the application of Chain-of-Thought (CoT) prompting (Wei et al., 2022) to integrate multi-source heterogeneous data. For instance, frameworks like CoTox (Park et al., 2025) and DrugReasoner (Ghaffarzadeh-Esfahani et al., 2025) combine chemical structures, biological pathways, and Gene Ontology (GO) terms to generate interpretable predictions for drug toxicity and approval

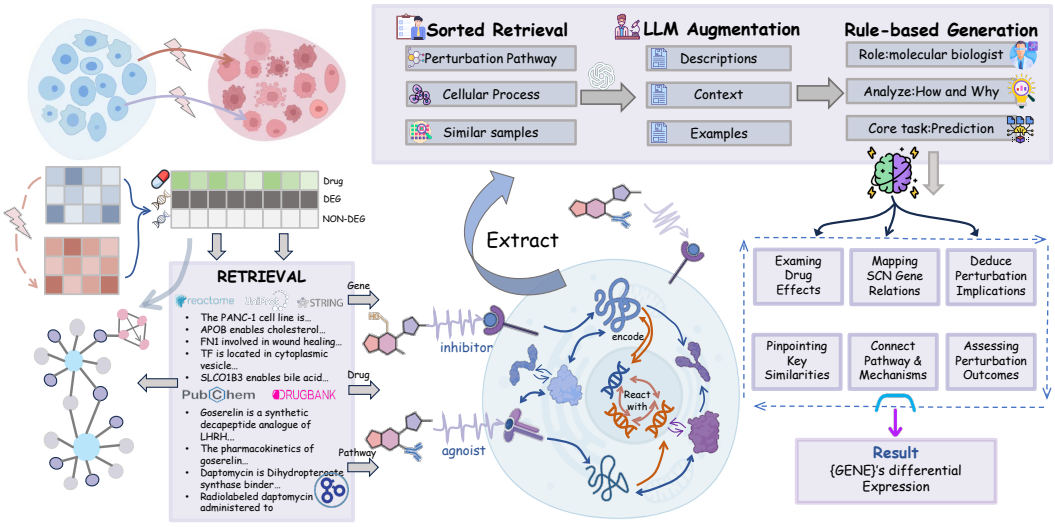

Figure 1: The pipeline of VCWorld. For a given query, VCWorld first retrieves multi-modal biological context from its integrated knowledge base, including pathway information, drug properties, and similar experimental samples. This context is then used to prompt a LLM, which is used to analyze mechanisms and infer how a specific gene will respond. The final output is a prediction for tasks such as differential expression and directional change.

processes. This paradigm has also been extended to fundamental biological contexts, such as the SUMMER framework (Wu et al., 2025), which utilizes retrieval-augmented generation (RAG) to predict perturbation experiments under gene edits. To further enhance the reasoning capabilities of these models, researchers are actively exploring more advanced training methodologies. These studies (Istrate et al., 2025; Hasanaj et al., 2025) collectively demonstrate the strong potential of CoT to effectively integrate multimodal biological data within the virtual cell domain, enabling more nuanced and interpretable predictions compared to foundational models.

2.3 PERT-SEQ DATA REPRESENTATION

Large-scale perturbation datasets (Wu et al., 2024) are foundational for building predictive models of cellular behavior. Recently, scientists have gradually expanded the scope of perturbations. Replogle et al. (2022) enabled predictions of genetic perturbation effects on core cellular functions. Nadig et al. (2024) first introduced the dose effect in the perturbation. Jiang et al. (2025) aided inference of pathway signatures for perturbation-driven changes in disease scenarios. More recently, Wu et al. (2025) released PerturBase, a dedicated database for single-cell perturbation sequencing data, integrating 122 datasets from 46 studies across genetic and chemical perturbations. A landmark in chemical perturbations is Tahoe-100M (Zhang et al., 2025), empowering AI models to predict context-dependent responses to small-molecule drugs across vast cellular diversity. Given the large size and coverage of the Tahoe-100M dataset, we use it as the primary dataset.

3 METHOD

3.1 TASK FORMULATION

The task of predicting single-cell gene expression responses to perturbations is conventionally defined as a regression problem. Given a dataset $D = \{(x_i, p_i, y_i)\}_{i=1}^{N}$, where $x_i, y_i \in \mathbb{R}^d$ are the pre- and post-perturbation expression profiles and p_i is the perturbation, the goal is to learn a mapping function $f : \mathbb{R}^d \times \mathcal{P} \to \mathbb{R}^d$. However, the high dimensionality and sparsity of the data pose significant challenges to achieving accurate predictions at the individual gene level. To this end, we reformulate the problem as a *gene-centric classification task*. The fundamental predictive unit is a triplet (c, p, g),

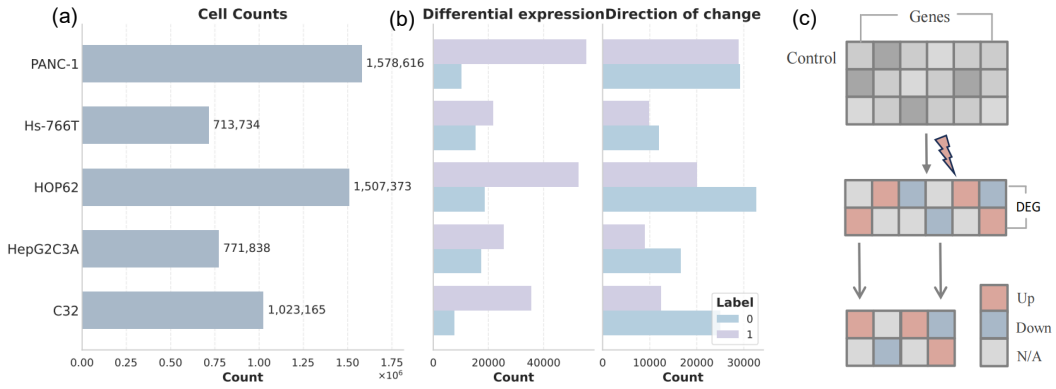

Figure 2: (a) Number of single-cell profiles per cell line. (b) Label distributions for the Differential Expression (DE) and Directional Change (DIR) tasks. (c) Schematic for identifying highly variable genes by comparing perturbed cells against a control group.

which queries the response of a specific gene g to a perturbation p in a given cell lineage c. For each such triplet, we derive a corresponding binary label, let's denote it as $l \in \{0, 1\}$, which represents the outcome for our two fine-grained classification tasks: (i) **Differential Expression (DE):** $l = 1$ if gene g is differentially expressed, and $l = 0$ otherwise. (ii) **Directional Change (DIR):** $l = 1$ for upregulation and $l = 0$ for downregulation. This transformation converts the original dataset D into a new, larger corpus of labeled triplets suitable for training a classifier.

These tasks can be addressed by a new paradigm that shifts from numerical regression to knowledge-based inference. The approach utilizes a reasoning engine, such as a Large Language Model (LLM) L, to process not only the query but also a rich set of biological context (BioContext). This context is dynamically retrieved from external knowledge bases (*e.g.*, signaling pathways, protein-protein interactions). The problem is then transformed into a text-based inference function as follows: $f : L(c, p, g, \texttt{BioContext}) \rightarrow \texttt{Prediction}$.

3.2 GeneTAK

To facilitate the comprehensive and fair assessment of large language models for perturbation prediction, we present GeneTAK, a novel benchmark dataset. GeneTAK is curated from the Tahoe-100M raw expression matrix and comprises 5 distinct cell lines across 348 drug perturbations, designed explicitly to evaluate the generalization capabilities of models in this domain.

Dataset Curation. The GeneTAK expression matrix was first restricted to the 2,000 most highly variable genes, aligning it with the common input dimensions of generative perturbation prediction models. Subsequently, to ensure a diverse and representative selection of cell lines, we performed a Principal Component Analysis (PCA) visualization (Greenacre et al., 2022) (see Appendix A). Guided by this analysis, we selected a final set of five cell lines that includes both a main cluster and several distinct outliers: C32, HOP62, HepG2/C3A, Hs 766T, and PANC-1. Notably, this selection is consistent with the cell lines used in the work of STATE (Adduri et al., 2025), ensuring a degree of comparability and fairness with previous benchmarks.

Then we generated labels for perturbation-gene pairs, denoted as (p, g). The labels were determined by identifying differentially expressed genes (DEGs) for each perturbation using the Wilcoxon signed-rank test (Woolson, 2007). To ensure label quality, we applied strict criteria based on a significant adjusted p-value and consistent expression changes across biological replicates. The resulting dataset of labeled pairs was then split by perturbation into training and test sets at a 3:7 ratio. This specific ratio was intentionally chosen to simulate a challenging few-shot learning scenario, consistent with the evaluation framework used by Adduri et al. (2025) and allowing for a direct comparison of model performance under low-data conditions. The split maintains a similar distribution of DEGs per perturbation in both sets. The final label distribution is shown in Figure 2, with a complete description of all data processing methods provided in Appendix A.

3.3 VCWORLD

Our proposed framework, VCWorld (Figure 1), transforms biological prediction into a multi-stage reasoning process powered by a Large Language Model (LLM). It consists of three key stages: (1) generating rich, symbolic representations for all biological entities; (2) retrieving a support set of causal evidence guided by open-world biological knowledge; and (3) synthesizing all information through a Chain-of-Thought process to produce an interpretable prediction.

3.4 CONSTRUCTION OF THE OPEN-WORLD BIOLOGICAL KNOWLEDGE GRAPH

To construct the open-world biological knowledge graph, we integrated several authoritative databases: PubChem, which provides large-scale compound structures and bioactivity data (Kim et al., 2023); DrugBank, which includes chemical, pharmacological, and clinical annotations of drugs and their target associations (Knox et al., 2024); UniProt, which offers comprehensive protein sequences and functional annotations (Consortium, 2019); Gene Ontology (GO), which defines a unified ontology across molecular function, biological process, and cellular component (Ashburner et al., 2000); Reactome, which curates systematically organized molecular reactions and pathways (Fabregat et al., 2018); STRING, which compiles protein–protein interactions from experiments and predictions (Szklarczyk et al., 2021); and CORUM, which catalogs experimentally validated mammalian protein complexes (Ruepp et al., 2007). Together, these databases cover multi-scale biological knowledge from compounds and drugs to genes, proteins, pathways, and complexes, ensuring the comprehensiveness of the graph.

In the knowledge graph construction process, compounds, drugs, genes, proteins, pathways, and complexes were represented as entity nodes, while interactions, annotations, and hierarchical relationships were encoded as heterogeneous edges. Cross-database integration was achieved through standardized identifiers (*e.g.*, InChIKey, UniProt ID, GO terms), followed by redundancy removal and conflict resolution. The resulting graph systematically represents biomolecules and their interactions, providing a foundation for cellular perturbation prediction.

3.4.1 GENERATIVE NODE FEATURE REPRESENTATION VIA LLMS

Let the open-world biological knowledge be denoted as $G = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, where $\mathcal{V}$ is the set of biological entities (nodes), $\mathcal{R}$ is the set of relation types, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ is the set of triples or edges. The first stage of our framework focuses on defining each biological entity $v \in \mathcal{V}$. To move beyond simple numerical vectors, we generate a rich symbolic representation for each node. For each node v, we first extract its local neighborhood subgraph, $N_k(v)$, defined as the set of all triples within a k-hop distance from v. We then construct a structured prompt P_v, using a template function f_{prompt}:

$$P_v = f_{\text{prompt}}(v, N_k(v)). \tag{1}$$

This function serializes the node's core attributes and its neighborhood triples into a natural language query. Subsequently, our framework employs an LLM to function as a feature generator. The LLM L then processes this prompt to generate a comprehensive textual description(see AppendixC), d_v, serving as the node's initial feature representation:

$$d_v = L(P_v) \tag{2}$$

This process yields a context-aware representation $d_v \in \mathcal{T}$ (where $\mathcal{T}$ is the space of all possible texts) that preserves biological semantics, offering a more expressive alternative to static embeddings.

3.4.2 GRAPH-GUIDED CAUSAL EVIDENCE FRAMEWORK

With rich node representations established, the second stage retrieves relevant experimental cases to form a basis for reasoning. Our training corpus $\mathcal{D}$, consists of M labeled instances derived from the original data, can be formulated as $\mathcal{D} = \{(q_i, l_i)\}_{i=1}^{M}$. Here, each query q_i is a triplet (c_i, p_i, g_i), and l_i is its associated ground-truth binary label ($l_i \in \{0, 1\}$). Given a new query $q_{input} = (c_{\text{input}}, p_{\text{input}}, g_{\text{input}})$, our goal is to construct an evidence support set $S(q_{\text{input}}) \subset \mathcal{D}$.

To ground the LLM's prediction in empirical data, we introduce a structured retrieval mechanism that goes beyond standard Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). The core of this mechanism is to identify and retrieve the most relevant historical cases from our training corpus $\mathcal{D}$ for

given input query q_{input}. To quantify this relevance, we score the similarity between the input query q_{input} and each query $q_i \in \mathcal{D}$. Instead of relying solely on semantic similarity, our method computes a hybrid similarity score, $\text{Sim}(q_{\text{input}}, q_i)$, that also leverages knowledge-graph topology information:

$$\text{Sim}(q_{\text{input}}, q_i) = \alpha \cdot \text{Sim}_{\text{sem}}(d_{q_{\text{input}}}, d_{q_i}) + (1 - \alpha) \cdot \text{Sim}_{\text{struct}}(q_{\text{input}}, q_i) \tag{3}$$

where $\text{Sim}_{\text{sem}}(\cdot)$ is the semantic similarity, calculated as the cosine similarity between the LLM-generated feature descriptions (from Eq. 2), and $\text{Sim}_{\text{struct}}(\cdot)$ is a graph-based structural similarity metric (*i.e.*, path-based similarity). The hyperparameter $\alpha \in [0, 1]$ balances the contribution of the semantic and structural components.

This score allows us to assemble a multifaceted evidence set. Instead of retrieving a single list of similar items, we retrieve two disjoint subsets by searching within predefined outcome groups as follows:

Analogue Cases (S_{analog}): The top-k_a instances from the subset of $\mathcal{D}$ with a positive outcome (*i.e.*, label $l = 1$), ranked by their similarity to the input query.

$$S_{\text{analog}}(q_{\text{input}}) = \underset{q_i \in \{q \in \mathcal{D} \mid l=1\}}{\text{arg top-}k_a} \text{Sim}(q_{\text{input}}, q_i) \tag{4}$$

Contrast Cases (S_{contrast}): The top-k_c instances from the subset of $\mathcal{D}$ with a negative outcome (i.e., label $l = 0$), similarly ranked by similarity.

$$S_{\text{contrast}}(q_{\text{input}}) = \underset{q_i \in \{q \in \mathcal{D} \mid l=0\}}{\text{arg top-}k_c} \text{Sim}(q_{\text{input}}, q_i) \tag{5}$$

The final evidence support set is the union: $S(q_{\text{input}}) = S_{\text{analog}} \cup S_{\text{contrast}}$. This curated set provides a balanced, contextual foundation upon which the final reasoning process is built.

3.4.3 Evidence Synthesis Chain-of-Thought Reasoning

In the final stage, the LLM acts as a computational biologist, which we formulate as a Chain-of-Thought (CoT) reasoning task. A final prompt, P_{CoT}, is synthesized from the symbolic representation of the query, $d_{q_{\text{input}}}$, and the retrieved evidence set, $S(q_{\text{input}})$:

$$P_{\text{CoT}} = f_{\text{CoT\_prompt}}(d_{q_{\text{input}}}, S_{\text{analog}}, S_{\text{contrast}}) \tag{6}$$

The function $f_{\text{CoT\_prompt}}$ formats these inputs using a structured template that first presents the biological query, followed by the lists of analogue and comparison cases as evidence. Finally, it instructs the LLM to provide step-by-step reasoning. (A detailed example is in Appendix A).

The LLM then processes this prompt to generate a text string, O_{final}, we then use a parsing function, f_{parse}, to extract the structured prediction $\hat{l}$ and the textual explanation E from the output:

$$O_{\text{final}} = L(P_{\text{CoT}}) \tag{7}$$

$$(\hat{l}, E) = f_{\text{parse}}(O_{\text{final}}) \tag{8}$$

This process compels the LLM to explicitly integrate qualitative knowledge ($d_{q_{\text{input}}}$) with empirical evidence ($S(q_{\text{input}})$), producing a self-validating and fully interpretable output. In this work, we use Gemeni-2.5 Flash (Comanici et al., 2025) as our reasoning model.

4 Experiment

4.1 Baseline

We benchmarked performance against several baselines using GeneTAK, including a linear model, RANDOM, which is a naive baseline that is assumed to have a statistically accurate prediction of 50% for any binary classification task without any training or prior knowledge; three published deep learning models: scVI (Lopez et al., 2018) is a conditional variational self-encoder that conditionally models perturbation effects in potential space, CPA (Lotfollahi et al., 2023) recognizes novel perturbations by learning disentangled and linearly combined latent embeddings, STATE (Adduri et al., 2025), the current SOTA model that demonstrates the best overall performance in the drug perturbation prediction task.

Table 1: Overall accuracy on two DE and DIR tasks. The best results are shown in bold, and the second-best results are shown with underlines.

Task	Model	C32	HepG2C3A	HOP62	Hs 766T	PANC-1	Average
DE	RANDOM	0.50	0.50	0.50	0.50	0.50	0.50
	scVI	<u>0.66</u>	0.48	<u>0.64</u>	<u>0.61</u>	**0.68**	<u>0.61</u>
	CPA	0.27	0.18	0.21	0.30	0.17	0.23
	STATE	0.16	0.38	0.41	0.08	0.47	0.30
	VCWorld-Llama3	0.35	0.36	0.41	0.36	0.39	0.37
	VCWorld (w/o Biocontext)	0.52	0.54	0.51	0.52	0.50	0.51
	VCWorld (w/o CoT)	0.61	<u>0.57</u>	0.59	0.60	0.56	0.59
	VCWorld	**0.70**	**0.68**	**0.71**	**0.68**	<u>0.61</u>	**0.68**
DIR	RANDOM	0.50	0.50	0.50	0.50	0.50	0.50
	scVI	0.47	0.46	0.41	0.40	0.38	0.42
	CPA	0.44	0.48	0.43	0.46	0.42	0.44
	STATE	0.49	0.51	0.50	0.55	0.51	0.51
	VCWorld-llama3	<u>0.57</u>	0.54	0.55	0.55	0.58	0.56
	VCWorld (w/o Biocontext)	0.52	0.52	0.51	0.52	0.50	0.51
	VCWorld (w/o CoT)	0.55	<u>0.59</u>	<u>0.56</u>	<u>0.58</u>	<u>0.63</u>	<u>0.58</u>
	VCWorld	**0.72**	**0.68**	**0.67**	**0.66**	**0.69**	**0.68**

4.2 METRICS

We evaluate model performance on two primary aspects: differentially expressed gene (DEG) prediction performance and LLM reasoning robustness. For a fair comparison, predictions from baseline models that output continuous expression values (*e.g.*, STATE, scVI) are first converted into binary value. This is achieved by applying the Wilcoxon signed-rank test (Woolson, 2007) to the predicted profiles, mirroring the ground-truth label generation process. All metrics below are calculated on these standardized outputs.

Metric for DEG Prediction Performance. We evaluate DEG prediction performance using four standard classification metrics: Accuracy, Precision, Recall, and the F1 Score. Accuracy provides a global measure of the overall proportion of correctly classified genes. However, given that most genes are not differentially expressed (a significant class imbalance), accuracy alone can be misleading. To provide a more nuanced assessment of performance on the positive class (DEGs), we therefore also report precision, which measures the ability to avoid false positives, and recall, which measures the ability to avoid false negatives. The F1 Score, as the harmonic mean of Precision and Recall, provides a single, balanced summary of these two aspects.

LLM Reasoning Robustness. We define the Abandonment Rate (Q-score) to measure the robustness of the LLM's reasoning process. It is the frequency with which the model abstains from making a prediction for a given query (*e.g.*, by stating it has insufficient information). A lower rate signifies higher robustness. The detailed formulas for all metrics are provided in Appendix B.

4.3 GENE-LEVEL PERFORMANCE

We consider the ability to capture gene-level expression changes as the most important evaluation criterion. This task was split into DE and DIR components for evaluation, and we used Accuracy to evaluate the model's performance in correctly predicting the ground-truth labels. The results are detailed in Table 1.

Our model, VCWorld, achieves an average score of 0.68 on both the DE and DIR tasks, significantly outperforming all baseline models and demonstrating its strong predictive power and generalisability. From the ablation study, we observe the following: (1) LLM reasoning capability is crucial: compared to the Llama3-based version, VCWorld's average score on the DE task improves from 0.37 to 0.68, an

Table 2: Performance comparison of different models on DE task.

Task	Model	Metric	C32	HepG2C3A	HOP62	Hs 766T	PANC-1	Average
DE	STATE	Precision	0.12	0.10	0.18	0.10	0.19	0.13
		Recall	0.15	0.32	0.36	0.09	0.37	0.26
		F1-Score	0.14	0.14	0.24	0.10	0.25	0.17
	CPA	Precision	0.11	0.10	0.14	0.07	0.17	0.12
		Recall	0.02	0.02	0.02	0.01	0.02	0.02
		F1-Score	0.03	0.03	0.04	0.02	0.04	0.03
	scVI	Precision	0.13	0.09	0.18	0.08	0.19	0.13
		Recall	0.48	0.39	0.52	0.38	0.51	0.46
		F1-Score	0.21	0.14	0.27	0.14	0.28	0.21
	VCWorld	Precision	0.60	0.64	0.61	0.57	0.52	0.59
		Recall	0.70	0.68	0.71	0.68	0.61	0.68
		F1-Score	0.65	0.66	0.66	0.62	0.56	0.63

84% performance increase. This shows that model performance is highly dependent on the advanced reasoning capability of the underlying large language model. (2) BioContext is a performance cornerstone: after removing BioContext (VCWorld w/o BioContext), the model's performance on both tasks plummeted to near-random levels (0.51). This confirms that providing relevant biological prior knowledge is crucial to guide the model towards effective reasoning. (3) CoT enhances prediction accuracy: after enabling CoT reasoning, VCWorld's mean score improved by about 15% (from 0.59 to 0.68) compared to the no-CoT version (VCWorld w/o CoT). This indicates that CoT effectively standardises the model's inference path.

As for the capacity boundaries of the baseline models(B.2), we found that: (1) scVI performed well on the DE task (mean score 0.61) and outperformed VCWorld on the PANC-1 dataset, demonstrating its competitiveness in specific scenarios. In contrast, CPA and STATE performed poorly. (2) All traditional baseline models (scVI, CPA, STATE) failed to significantly outperform the random baseline (0.50) in the DIR task. This exposes their inherent limitations in capturing complex inter-gene interactions and pathway-level effects, which likely stems from their reliance on fixed statistical assumptions, contrasting with the flexible reasoning capabilities of our LLM-based framework. In

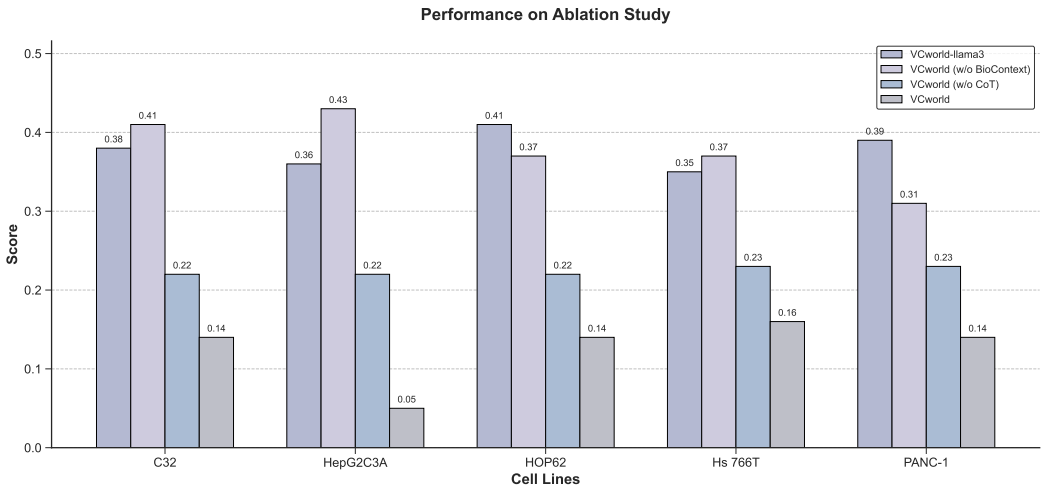

Figure 3: The ablation experiments of VCWorld.

summary, these results not only highlight the superior performance of VCWorld but also systemati-

cally validate our core design principles: leveraging advanced LLM reasoning, grounding predictions in biological context, and structuring the inference process via a Chain-of-Thought.

4.4 Cell-Level and Population-Level Analysis

We believe that examining the collective response of gene populations offers a more accurate representation of cellular changes. To evaluate this, we quantified the number of DEGs predicted across the GeneTAK dataset (see Appendix B).

The results show clear patterns. scVI predicts 3–5 times more DEGs than the ground truth. STATE is less stable, but on average predicts about twice as many DEGs. These results indicate that both models amplify perturbation effects and classify genes as DEGs too aggressively. This explains their high Accuracy but low Precision, Recall, and F1 Score, which limits their reliability. In contrast, CPA consistently underestimates DEG counts, reflecting a limited ability to capture perturbation effects.

To obtain a more balanced view, we further computed Precision, Recall, and the F1 Score (Table 2). The results confirm the above observations: (1) VCWorld had the best overall performance. It has a high and balanced precision (0.59) and recall (0.68), resulting in the highest F1 score (0.63); (2) scVI has a high recall (0.46) but a low precision (0.13), but still the best baseline; (3) STATE has mediocre performance on all metrics and a low overall score, with an F1 score of only 0.17; (4) Both STATE and scVI amplified perturbation effects ,scVI predicts multiple times more DEGs than the ground truth,while STATE predicts about twice as many DEGs.

These findings suggest that deep learning models often lack explicit constraints to regulate DEG counts. They fail to capture the true intensity and scope of perturbation effects. In contrast, the design of VCWorld directly addresses these limitations, which explains its superior performance.

4.5 analysis and ablation study

To evaluate the tendency of models to abstain from answering, we introduced the **Q-score** as a dedicated metric for assessing LLM performance. Figure 3 reports the Q-scores after ablating different components of the model(specific values see AppendixB.3). First, when we replaced VCWorld's reasoning "engine" (Gemini 2.5-flash) with **Llama3-8B**, the model's adherence to prompts dropped sharply: it failed to provide deterministic reasoning results and often abandoned the reasoning process entirely. Second, **VCWorld (w/o BioContext)** performed only on par with Llama3-8B. By contrast, **VCWorld (w/o CoT)** was able to produce answers at a relatively high completion rate. Ultimately, the full **VCWorld** model outperformed the three variants by 40%, 40%, and 11%, respectively. These findings lead us to conclude that: (1) **BioContext** provides indispensable references for reasoning; (2) **CoT** enhances the LLM's ability to tackle the complex task of drug perturbation prediction; and (3) LLMs require sufficient parameter capacity to support such complex reasoning tasks.

5 Conclusion

We demonstrate the efficacy of VCWorld in the task of predicting changes in gene expression following genetic perturbations. Our experiments show that VCWorld far outperforms current state-of-the-art black-box models. However, its main contribution lies in aspects that are crucial for scientific utility. Through extensive ablation studies and case studies, we demonstrate that VCWorld is highly interpretable, with a human-readable, verifiable chain of reasoning for each prediction.

VCWorld still has room for improvement, with future directions including integrating a multi-agent framework to enhance autonomous reasoning and sample retrieval, extending generalization across diverse perturbation types such as gene, pathway, and combinatorial perturbations, and establishing more systematic benchmarks to comprehensively evaluate and compare specialized prediction models.

REFERENCES

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

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines. *BioRxiv*, pp. 2024–09, 2024.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Cell movements and the shaping of the vertebrate body. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

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

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

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Antonio Del Sol, Rudi Balling, Lee Hood, and David Galas. Diseases as network perturbations. *Current opinion in biotechnology*, 21(4):566–571, 2010.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

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

Mohammadreza Ghaffarzadeh-Esfahani, Ali Motahharynia, Nahid Yousefian, Navid Mazrouei, Jafar Ghaisari, and Yousof Gheisari. Drugreasoner: Interpretable drug approval prediction with a reasoning-augmented language model. *arXiv preprint arXiv:2508.18579*, 2025.

Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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

Euxhen Hasanaj, Elijah Cole, Shahin Mohammadi, Sohan Addagudi, Xingyi Zhang, Le Song, and Eric P Xing. Multimodal benchmarking of foundation model representations for cellular perturbation response prediction. *bioRxiv*, pp. 2025–06, 2025.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1): 45–74, 2024.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. rbio1-training scientific reasoning llms with biological world models as soft verifiers. *bioRxiv*, pp. 2025–08, 2025.

Longda Jiang, Carol Dalgarno, Efthymia Papalexi, Isabella Mascio, Hans-Hermann Wessels, Huiyoung Yun, Nika Iremadze, Gila Lithwick-Yanai, Doron Lipson, and Rahul Satija. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *Nature Cell Biology*, pp. 1–13, 2025.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380, 2023.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Lanxiang Li, Yue You, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Yunlin Fu, Jiaming Kong, et al. A systematic comparison of single-cell perturbation response prediction models. *bioRxiv*, pp. 2024–12, 2024.

Prisca Liberali, Berend Snijder, and Lucas Pelkmans. Single-cell and multivariate approaches in genetic perturbation screens. *Nature Reviews Genetics*, 16(1):18–32, 2015.

Leslie M Loew and James C Schaff. The virtual cell: a software environment for computational cell biology. *TRENDS in Biotechnology*, 19(10):401–406, 2001.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

Ajay Nadig, Joseph M Replogle, Angela N Pogson, Steven A McCarroll, Jonathan S Weissman, Elise B Robinson, and Luke J O'Connor. Transcriptome-wide characterization of genetic perturbations. *BioRxiv*, 2024.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, et al. Virtual cells: Predict, explain, discover. *arXiv preprint arXiv:2505.14613*, 2025.

Jueon Park, Yein Park, Minju Song, Soyon Park, Donghyeon Lee, Seungheun Baek, and Jaewoo Kang. Cotox: Chain-of-thought-based molecular toxicity reasoning and prediction. *arXiv preprint arXiv:2508.03159*, 2025.

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

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Andreas Ruepp, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Michael Stransky, Brigitte Waegele, Thorsten Schmidt, Octave Noubibou Doudieu, Volker Stümpflen, et al. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(suppl_1):D646–D650, 2007.

Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

Xiangru Tang, Zhuoyun Yu, Jiapeng Chen, Yan Cui, Daniel Shao, Weixu Wang, Fang Wu, Yuchen Zhuang, Wenqi Shi, Zhi Huang, et al. Cellforge: Agentic design of virtual cell models. *arXiv preprint arXiv:2508.02276*, 2025.

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

Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. *arXiv preprint arXiv:2502.21290*, 2025.

Yan Wu, Esther Wershof, Sebastian M Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. Perturbench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*, 2024.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

A GeneTAK

A.1 DATA PROCESSING

We normalized all gene counts to Log(TP10k+1) values (log-transformed UMI count per 10k), where the count c_{ij} of gene j in cell i is mapped to

$$\log\left(\frac{c_{ij}}{\sum_j c_{ij}} \cdot 10,000 + 1\right). \tag{9}$$

To focus on the most variably expressed genes, the top 2,000 highly variable genes (HVGs) were selected using the Seurat v3 method, and retained in the expression matrix for downstream analyses, with the HVG list exported for reproducibility. To determine differentially expressed genes (DEGs), we ran the Wilcoxon signed-rank test (Wilcoxon, 1945) with Benjamini-Hochberg correction (Benjamini & Hochberg, 2000) between non-targeting control (NTC) cells and perturbed cells, for each perturbation.

We set $\mathrm{FDR}_{\mathrm{threshold}}$ and $\mathrm{logFC}_{\mathrm{threshold}}$ to screen perturbations with phenotypic effects, and $\mathrm{PVAL}_{\mathrm{threshold}}$ to screen negative samples (samples with no significant change in gene expression). Specifically, when

$$\mathrm{FDR}_{\mathrm{threshold}} \leq 0.05 \quad \& \quad \mathrm{logFC}_{\mathrm{threshold}} \geq 0.25$$

(i.e., more than 1.28-fold change in expression), we consider them as samples with phenotypic effects. Based on this, when

$$\mathrm{PVAL}_{\mathrm{neg\_threshold}} \geq 0.1,$$

we consider them as negative samples. In addition, we set $N_{\mathrm{neg\_samples}} = 200$ to limit the number of negative samples, so as to avoid category imbalance.

A.2 Distribution of GeneTAK

To keep consistent with the experimental settings in the state, the dataset of each cell line was split into training and testing sets at a ratio of 30:70.During the experiment, VCWorld only retrieved samples from the training set. Further details regarding the dataset and data split statistics are provided in Table 3 and 4.

Table 3: Number of cells in each cell line.

Cell line	Control	Perturbed
C32	29,651	1,023,165
HepG2C3A	20,451	771,838
HOP62	39,683	1,507,373
Hs 766T	17,802	713,734
PANC-1	43,878	1,578,616

Table 4: GeneTAK statistics for differentially expressed genes.

Cell line	Split	Total	Non-DE	Differentially expressed		
				Total	Up	Down
C32	Train	128 293	55 200	73 093	24 746	48 344
	Test	89 961	38 800	51 161	16 643	34 518
HepG2C3A	Train	102 253	55 200	47 053	16 820	30 233
	Test	72 169	38 800	33 369	11 629	21 740
HOP62	Train	154 969	55 200	99 769	36 615	63 154
	Test	108 777	38 800	69 977	25 472	44 505
Hs 766T	Train	101 707	55 200	46 507	20 774	25 733
	Test	71 891	38 800	33 091	14 420	18 671
PANC-1	Train	163 748	55 200	108 548	54 081	54 467
	Test	115 568	38 800	76 768	38 131	38 637

B Experiment Details

B.1 Metrics

B.1.1 Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total instances.

B.1.2 F1-score

F1-score is the harmonic mean of Precision and Recall, providing a balance between correctly identified positives (Precision) and coverage of actual positives (Recall).

B.1.3 Q-score

$$\text{Q-score} = 1 - \frac{AP}{QP + AP}$$

where AP means the number of answered prompts, QP means the number of prompts without a normal answer.

B.2 Baseline

All baselines were trained on 30% of the 45 cell lines from tahoe-100m and 5 test cell lines, and all of our GeneTAK data was extracted from the remaining 70%, which eliminates the possibility of data leakage, Table5 is the result obtained from the inference of all baseline models on the five cell lines

B.3 Ablation Study of VCWorld

Table 6 shows the responses of all ablation experiments on each cell line, and the Q-score is calculated from this part of the data.

Table 5: Statistics of predicted number of DEGs

Task	Model	Count	C32	HepG2C3A	HOP62	Hs_766T	PANC-1
DE	Truth	Number	128 293	102 253	154 969	101 707	163 748
	STATE	Prediction	157 758	366 238	317 589	86 684	316 434
		Correct	19 488	33 122	55 717	9077	60 640
	CPA	Prediction	21 475	21 204	19 524	19 838	19 253
		Correct	2424	2076	2755	1394	3178
	scVI	Prediction	460 477	466 868	449 047	470 606	435 034
		Correct	60 973	40 108	79 858	38 969	83 801

Table 6: Classification results of different VCWorld models across cell lines.

Task	Model	Metric	C32	HepG2C3A	HOP62	Hs 766T	PANC-1
DE	/	Total	5691	5326	2527	3062	2404
	VCWorld (W/O Biotext)	Answered	3379	3040	1586	1934	1650
		Accurate	1757	1642	809	967	842
	VCWorld (W/O CoT)	Answered	4462	4131	1978	2367	1863
		Accurate	2722	2355	1167	1420	1043
	VCWorld-llama3	Answered	3528	3409	1491	1257	1007
		Accurate	1235	1227	611	453	393
	VCWorld	Answered	4874	5060	2162	2564	2061
		Accurate	3412	3441	1535	1744	1257

B.4 Knowledge Graph-to-Text Pipeline for Prompt Generation

To ensure the factual grounding and mitigate the risk of hallucination in our language model, we developed a deterministic pipeline to programmatically construct input prompts from multiple structured biological knowledge graphs (KGs). This process transforms machine-readable graph data into a human-readable, factual context that guides the model's summarization task. The pipeline consists of three primary stages: data integration and pre-processing, rule-based templated verbalization, and fact aggregation for prompt assembly.

B.4.1 Data Integration and Pre-processing

Our pipeline integrates information from a diverse set of public biological knowledge bases, including Ensembl and UniProt for descriptive annotations, and graph-structured databases such as the Gene Ontology (GO), Reactome, CORUM, BioPlex, and STRING for relational information. Upon ingestion, a pre-processing step is applied to enhance data quality. Notably, we filter out high-degree nodes from relational graphs (e.g., generic GO terms like "protein binding" or highly promiscuous interactors in STRING). This step reduces noise and focuses the downstream context on more specific and informative biological relationships.

B.4.2 Rule-Based Templated Verbalization

The core of our methodology is a rule-based verbalization engine. Recognizing that each knowledge source possesses a unique data schema, we designed a specific template to translate its structured entries into natural language declarative statements.

For instance:

- A relationship from CORUM, representing `(Gene_A, is_member_of, Complex_X, in_cell_line_Y)`, is verbalized as: "Gene A is a member of the Complex X in cell line Y."

- A protein-protein interaction from STRING, representing `(Gene_A, interacts_with, Gene_B)` with associated evidence types `[experimental, database]`, is translated to: "Based on evidence from experimental evidence and database evidence, Gene A may physically interact with Gene B."
- Functional annotations from UniProt are extracted and formatted as statements like: "Functions of Gene A include: [function description]."

This templated approach ensures that the conversion is consistent, deterministic, and preserves the semantic content of the original source.

B.4.3 Fact Aggregation and Prompt Assembly

For each target gene, the declarative statements (facts) generated from all integrated sources are aggregated into a single collection. We use a set data structure to automatically deduplicate identical facts that may arise from data redundancy across different databases. To maintain prompt conciseness and manage cognitive load for the model, we impose a ceiling on the maximum number of facts included, prioritizing information from higher-fidelity sources like UniProt and curated pathway databases.

Finally, this aggregated list of facts is formatted as a bulleted list and injected into a master prompt template under the heading "Domain Knowledge:". This final prompt provides the language model with its role, a precisely defined task, operational constraints, and the crucial, factually grounded context required to generate an accurate and reliable summary.

C Prompts

We provide the prompts used by VCWorld and the model's response at the end of the paper. It clearly demonstrates that VCWorld is interpretable, capable of aligning with real-world biological knowledge, and has a traceable thinking process and final result.

D Ethics Statement

The work reported in this paper involves only computational experiments and publicly available datasets. Therefore, no ethical approval was necessary.

E Reproducibility Statement

The datasets and code for NABench are available at https://anonymous.4open.science/r/VCWorld-B970

F LLM Usage Statement

We use the Large Language Models (LLMs) for grammar checking of the paper to improve its overall readability, while all other work is independently completed by the human authors.

User Prompt

In Hs766T cells, how does Tofacitinib affect the expression of FTH1 ?

Generated Prompt

[Start of Prompt]
You are an expert molecular biologist who studies how small molecule drugs affect gene expression using Perturb-seq. Your goal is to determine: **Does a drug perturbation of Tofacitinib in Hs766T cells result in differential expression of FTH1?**
You are given as input:
Description of the drug (Tofacitinib): description of drug that is to perturb the cell
Description of gene of interest (FTH1): description of gene, the impact on which you wish to infer
Context: description of cell line in which the genes are expressed
Examples: set of experimental observations that describe the impact of small molecule perturbations on related genes, to contextualize your answer
Output: Please extract the most relevant parts of the examples that address these five questions. Be specific.
Which of the observed perturbing drugs are most similar to Tofacitinib (if any, including Tofacitinib itself)?
When using Tofacitinib or similar drugs, what downstream pathways or genes are differentially expressed? Justify your answer with the observed outcomes.
Which of the observed genes of interest are most similar to FTH1 (if any, including FTH1 itself)?
What perturbations of upstream pathways or genes result in differential expression of FTH1 or similar genes (if any)? Justify your answer with the observed outcomes.
Is a perturbation of Tofacitinib in FTH1 likely to result in differential expression of FTH1? For example, if 2) and 4) are unrelated or only indirectly related, it is unlikely we will observe differential expression. On the other hand, if 2) and 4) significantly overlap in specific genes or pathways, we may observe differential expression.
Your final answer should end with one of these three options and nothing else.
No. Perturbation of Tofacitinib does not impact FTH1.
Yes. Perturbation of Tofacitinib results in differential expression of FTH1.
There is insufficient evidence to determine how Perturbation of Tofacitinib affects FTH1.
[End of Prompt]

[Start of Input]
Description of molecule drug (Tofacitinib): Tofacitinib is a reversible Janus kinase (JAK) inhibitor that prevents the phosphorylation and activation of STAT transcription factors. This action blocks the JAK-STAT signaling pathway, which is involved in hematopoiesis and immune cell function. This perturbation may affect the transcription of cytokine-responsive genes that are regulated by the JAK-STAT pathway and are involved in inflammatory responses.
Description of gene of interest (FTH1): FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Context: Hs 766T is a human pancreatic ductal adenocarcinoma (PDAC) cell line, established from the lymph node metastasis of a 73-year-old female patient. It carries hallmark pancreatic cancer alterations including KRAS G12D mutation, TP53 mutation, and CDKN2A inactivation, reflecting the canonical molecular landscape of PDAC. The cells grow adherently with an epithelial morphology, display a doubling time of ~60–70 hours, and are microsatellite stable (MSS). Hs 766T is widely used in studies of pancreatic tumor progression, metastasis, stromal interactions, and therapeutic resistance, and is included in large-scale resources such as the NCI-60 panel for pharmacogenomic profiling.
Examples:
Example 1:
Drug: c-Kit-IN-1
Gene: FTH1

Drug Description: c-Kit-IN-1 acts as an ATP-competitive inhibitor, binding to the kinase domain of the c-Kit receptor tyrosine kinase. This prevents the autophosphorylation of the receptor and the activation of downstream signaling pathways. Perturbation with c-Kit-IN-1 may affect the expression of genes regulated by the c-Kit signaling pathway, such as those in the MAPK and PI3K pathways.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: B) Perturbation of this drug results in differential expression of the gene of interest.
Example 2:
Drug: AZD2858
Gene: FTH1
Drug Description: AZD2858 binds to the bromodomains of BET proteins, preventing them from binding to acetylated histones. This mechanism disrupts the transcriptional activation of key oncogenes, such as MYC. Perturbation with AZD2858 may affect the expression of BET target genes involved in cell proliferation.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: B) Perturbation of this drug results in differential expression of the gene of interest.
Example 3:
Drug: BI-78D3
Gene: FTH1
Drug Description: BI-78D3 is a small molecule inhibitor of the protein-protein interaction between the transcription factor NRF2 and its negative regulator KEAP1. By disrupting this interaction, it modulates the activity of the NRF2 pathway. Perturbation with BI-78D3 may affect the expression of genes regulated by the NRF2 antioxidant response element.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: B) Perturbation of this drug results in differential expression of the gene of interest.
Example 4:
Drug: Vemurafenib
Gene: FTH1
Drug Description: Vemurafenib is a competitive inhibitor of mutated BRAF-serine-threonine kinase, particularly the BRAF V600E mutation. It blocks the downstream MAPK signaling pathway to inhibit tumor growth and trigger apoptosis. This perturbation may affect the expression of genes regulated by the MAPK/ERK pathway, which are involved in cell growth and proliferation.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: A) Perturbation of this drug does not impact the gene of interest.
Example 5:
Drug: Ciclopirox
Gene: FTH1
Drug Description: Ciclopirox is an antifungal agent thought to function by chelating polyvalent metal cations, which inhibits enzymes involved in processes like mitochondrial electron transport. This perturbation may affect the expression of genes involved in cellular metabolism, stress responses, DNA repair, and cell cycle regulation.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: B) Perturbation of this drug results in differential expression of the gene of interest.
Example 6:
Drug: Afatinib
Gene: FTH1
Drug Description: Afatinib is an irreversible ErbB family blocker that covalently binds to the kinase domains of EGFR, HER2, and HER4. This binding irreversibly inhibits tyrosine kinase autophosphorylation, resulting in the downregulation of ErbB signaling. This perturbation may affect the expression of genes downstream of the ErbB signaling pathway that are involved in cell proliferation and survival.

Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: B) Perturbation of this drug results in differential expression of the gene of interest.
Example 7:
Drug: BI-3406
Gene: FTH1
Drug Description: BI-3406 binds to the catalytic domain of SOS1, preventing it from engaging with and activating KRAS. This leads to the inhibition of KRAS signaling and a reduction in the proliferation of KRAS-dependent cancer cells. Perturbation with BI-3406 may affect the expression of genes downstream of the KRAS signaling pathway.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: B) Perturbation of this drug results in differential expression of the gene of interest.
Example 8:
Drug: AZD1390
Gene: FTH1
Drug Description: AZD1390 is a potent inhibitor of Ataxia-Telangiectasia Mutated (ATM) kinase. This action prevents the activation of the DNA damage checkpoint and disrupts DNA repair processes. Perturbation with AZD1390 may affect the expression of genes involved in the DNA damage response and cell cycle checkpoints regulated by ATM kinase.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: A) Perturbation of this drug does not impact the gene of interest.
Example 9:
Drug: Palmatine (chloride)
Gene: FTH1
Drug Description: Palmatine (chloride) has multiple mechanisms of action, including acting as a poison for topoisomerases, which leads to DNA damage. It also inhibits indoleamine 2,3-dioxygenase 1 (IDO-1), an enzyme involved in immune tolerance. Perturbation with Palmatine (chloride) may affect the expression of genes involved in DNA replication and repair, as well as immune response pathways.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: A) Perturbation of this drug does not impact the gene of interest.
Example 10:
Drug: Regorafenib
Gene: FTH1
Drug Description: Regorafenib is a small molecule inhibitor of multiple membrane-bound and intracellular kinases, including VEGFR, KIT, PDGFR, FGFR, and RAF. This broad inhibition targets pathways involved in oncogenesis, tumor angiogenesis, and the maintenance of the tumor microenvironment. This perturbation may affect the expression of genes regulated by these various kinase signaling pathways that are involved in cell proliferation and angiogenesis.
Gene Description: FTH1 is a ferritin heavy chain subunit enabling iron ion binding and sequestering, involved in iron storage and intracellular iron ion homeostasis. Iron overload, iron deficiency, or oxidative stress could impact FTH1 expression to regulate cellular iron levels.
Result: A) Perturbation of this drug does not impact the gene of interest.
[End of Input]

[Start of Output]
1)
2)
3)
4)
5)
[End of Output]

Thoughts

Investigating Tofacitinib's Impact
I'm focusing on Tofacitinib, a JAK inhibitor, and its potential to influence FTH1 expression in Hs766T cells. I've begun by examining Tofacitinib's mechanism, specifically its role in blocking JAK-STAT signaling pathways. This understanding will be crucial in predicting its effect on FTH1, the target gene. **My next step involves researching the link between JAK-STAT and iron metabolism to understand the downstream effects on FTH1 expression.**

Exploring Drug Similarities
I'm now comparing Tofacitinib to the example drugs, focusing on their mechanisms of action. I see c-Kit-IN-1, a tyrosine kinase inhibitor, as potentially relevant due to its kinase-inhibiting nature, though not directly targeting JAK-STAT. The other examples seem quite different, acting on epigenetic regulators or cellular stress responses. I'll need to assess if shared downstream pathways might indirectly link them to FTH1 regulation.

Comparing Drug Kinase Inhibitors
I'm deep into the analysis of kinase inhibitors. Several example drugs share this mechanism with Tofacitinib, though they target distinct kinase families. I've noted c-Kit-IN-1, Vemurafenib, Afatinib, AZD1390 and Regorafenib as particularly relevant, due to their inhibition of kinases involved in various signal transduction pathways. I'm focusing now on identifying potential shared downstream pathways that might connect these drugs to FTH1 regulation, despite their diverse primary targets.

Refining Kinase Connections
Now I'm delving deeper into the potential links between the kinase inhibitors and FTH1. Given the examples, I see that while Tofacitinib directly blocks JAK-STAT, other drugs like c-Kit-IN-1, Vemurafenib, Afatinib, AZD1390, and Regorafenib might indirectly influence FTH1 through shared downstream pathways. I'm focusing on those pathways activated or inhibited by these kinase drugs to evaluate their relevance to FTH1 regulation. This will necessitate cross-referencing published literature.

Identifying JAK-STAT Targets
I'm now focusing on identifying the specific cytokine-responsive genes regulated by the JAK-STAT pathway, which is affected by Tofacitinib. The goal is to determine if any of these target genes might indirectly influence FTH1 expression. While no drugs are directly analogous, some kinase inhibitors (c-Kit-IN-1, Afatinib) showed FTH1 differential expression, highlighting the need to understand potential shared pathways. I'm cross-referencing literature to link Tofacitinib's pathway to FTH1's regulators.

Evaluating Pathway Links
I'm now focusing on how the JAK-STAT pathway, targeted by Tofacitinib, might intersect with FTH1 regulation. Considering the lack of direct JAK-STAT inhibitors in the examples, **I'm examining if inflammatory responses, a key feature of JAK-STAT signaling, could indirectly influence iron homeostasis and FTH1 expression. This will require further literature exploration.**

Linking JAK-STAT to Iron
I'm now focusing on how the JAK-STAT pathway, which is targeted by Tofacitinib, might intersect with iron regulation in Hs766T cells. While I've examined cytokine-responsive genes, I'm now exploring potential connections between inflammatory responses and FTH1 expression, requiring extensive literature review.

Evaluating JAK-STAT Impact
The connection between Tofacitinib's JAK-STAT pathway and FTH1 regulation is under review. I'm focusing on the inflammatory responses driven by this pathway, as these might influence iron homeostasis. I'm also examining the literature regarding STAT transcription factor regulation of iron-related genes to predict FTH1 expression shifts.

Predicting Tofacitinib's Effect
I'm now integrating information on JAK-STAT and iron regulation to predict if Tofacitinib will alter FTH1 in Hs766T. **Given that the JAK-STAT pathway influences inflammatory responses and iron metabolism, it appears plausible that FTH1 expression will shift with Tofacitinib.** This requires further review.

I've determined that Tofacitinib, a JAK-STAT inhibitor, has a plausible link to FTH1 regulation in the Hs766T cells, primarily through the effects of inflammation on iron homeostasis. This requires further scrutiny, as there is currently no direct evidence from the examples.

Analyzing JAK-STAT's Influence
I'm now considering how Tofacitinib, which inhibits the JAK-STAT pathway, interacts with iron regulation. Although there are no direct JAK-STAT inhibitors among the examples, this is a plausible mechanism. FTH1 is known to be regulated by oxidative stress, metal chelation, and the KRAS pathway, all of which are linked.

The link between Tofacitinib, a JAK-STAT inhibitor, and FTH1 in Hs766T cells is clear. The JAK-STAT pathway is involved in inflammation, which impacts iron homeostasis. KRAS, deregulated in this cell line, feeds into the JAK-STAT pathway, which is shown to impact FTH1.

Answer

Which of the observed perturbing drugs are most similar to Tofacitinib (if any, including Tofacitinib itself)?

Tofacitinib is a Janus kinase (JAK) inhibitor that blocks the JAK-STAT signaling pathway. None of the drugs in the provided examples are JAK inhibitors. The most similar class of drugs are other kinase inhibitors, although they target different pathways. These include:
c-Kit-IN-1: A c-Kit receptor tyrosine kinase inhibitor.
Vemurafenib: A BRAF serine-threonine kinase inhibitor.
Afatinib: An ErbB family tyrosine kinase inhibitor.
AZD1390: An ATM kinase inhibitor.
Regorafenib: A multi-kinase inhibitor targeting VEGFR, KIT, PDGFR, FGFR, and RAF.

When using Tofacitinib or similar drugs, what downstream pathways or genes are differentially expressed? Justify your answer with the observed outcomes.

Tofacitinib is described as affecting cytokine-responsive genes regulated by the JAK-STAT pathway. The provided examples do not include drugs that target this pathway, so direct evidence is unavailable. The outcomes for other kinase inhibitors are inconsistent:
Perturbation with c-Kit-IN-1 (affecting MAPK and PI3K pathways) and Afatinib (affecting ErbB signaling) resulted in differential expression of FTH1.
Perturbation with Vemurafenib (affecting MAPK/ERK), AZD1390 (affecting DNA damage response), and Regorafenib (affecting multiple pathways including RAF) did not result in differential expression of FTH1.
This indicates that inhibition of some, but not all, kinase signaling pathways can lead to differential expression of FTH1.

Which of the observed genes of interest are most similar to FTH1 (if any, including FTH1 itself)?
The only observed gene of interest in all provided examples is FTH1 itself.

What perturbations of upstream pathways or genes result in differential expression of FTH1 or similar genes (if any)? Justify your answer with the observed outcomes.

Based on the examples, differential expression of FTH1 results from perturbations of several distinct pathways:
Kinase Signaling: Inhibition of c-Kit (Example 1: c-Kit-IN-1) and ErbB family kinases (Example 6: Afatinib).
KRAS Signaling: Inhibition of the SOS1-KRAS interaction (Example 7: BI-3406). This is particularly relevant as the Hs766T cell line has a KRAS mutation.
Antioxidant/Stress Response: Modulation of the NRF2 pathway (Example 3: BI-78D3), which directly regulates genes involved in oxidative stress, including FTH1.
Metabolic Stress/Ion Homeostasis: Chelation of metal cations (Example 5: Ciclopirox), which directly relates to FTH1's function in iron storage.
Epigenetic Regulation: Inhibition of BET bromodomain proteins (Example 2: AZD2858), which affects broad transcriptional programs.

Is a perturbation of Tofacitinib in FTH1 likely to result in differential expression of FTH1? For example, if 2) and 4) are unrelated or only indirectly related, it is unlikely we will observe differential expression. On the other hand, if 2) and 4) significantly overlap in specific genes or pathways, we may observe differential expression.

The pathways impacted by Tofacitinib (JAK-STAT, inflammation, cytokine signaling) are highly related to pathways shown to regulate FTH1 (stress response, KRAS signaling). There is a well-established biological link between inflammation, cytokine signaling (mediated by JAK-STAT), and the regulation of iron homeostasis, where FTH1 is a key player. Furthermore, in the context of KRAS-mutated pancreatic cancer cells like Hs766T, the oncogenic KRAS pathway is known to drive inflammatory cytokine production, creating a dependency on JAK-STAT signaling. Since the examples show that inhibiting KRAS signaling directly results in differential expression of FTH1 (Example 7), it is highly probable that inhibiting the downstream JAK-STAT pathway with Tofacitinib will also result in differential expression of FTH1.

Yes. Perturbation of Tofacitinib results in differential expression of FTH1.

Prompt

You are an expert molecular biologist who studies how drug affect the cellular reaction using Perturb-seq.

Task: You are writing a brief overview of the small molecule drug{{pert}}, with a focus on its MoA and Targets. You will be provided a set of database entries about the drug. Ensure that your overview remains faithful to this domain knowledge.

Format:
- Write one to two sentences describing the primary MoA and target of {{pert}}.
- Write one sentence describing the potential downstream impact of perturbing drug {{pert}}.

Constraints:
- Maintain a professional tone throughout.
- Do not comment on your own writing.
- Do not add any notes or references. Do not make up additional information.
- Do not discuss the importance or impact of the gene. Focus only on its function.

Domain knowledge:
{{entries}}

Brief overview of drug {{pert}}:

Prompt

You are an expert molecular biologist who studies how genes are related using Perturb-seq.

Task: You are writing a brief overview of the human gene {{name}}, with a focus on its molecular and cellular functions. You will be provided with a set of database entries about the gene. Ensure that your overview remains faithful to this domain knowledge.

Format:
- Write one to two sentences describing the primary molecular and cellular function of gene {{name}}.
- Write one sentence describing what types of perturbations might impact the expression of gene {{name}}. For example, you might consider pathways that are upstream of the gene or compensatory mechanisms.

Constraints:
- Maintain a professional tone throughout.
- Do not comment on your own writing.
- Do not add any notes or references. Do not make up additional information.
- Do not discuss the importance or impact of the gene. Focus only on its function.

Domain knowledge:
{{entries}}

Brief overview of gene {{name}}:

Prompt

You are an expert molecular biologist who studies how drug affect the cellular reactions using Perturb-seq.

Task: You are writing a brief overview of the small molecule drug{{pert}}, with a focus on the downstream effects of perturbing {{name}} .

Inputs: You are provided
- Description of perturbing drug {{pert}}
- Database entries relating {{pert}} to other drugs or targets.

Format:
- Write up to five sentences describing the molecular and cellular impact of perturbing drug{{pert}}.

Constraints:

- Omit the importance or impact of the gene. Focus only on its function.
- Omit all non-specific information and obvious statements, e.g. "this gene is involved in cellular processes."
- Remain faithful to all domain knowledge. Do not make up additional information.
- Maintain a professional tone throughout. Do not comment on your own writing. Do not add any notes or references.

Description of drug {{pert}}: {{summary}}

Relations to other drugs/targets:
{{entries}}

Downstream effects of perturbing drug {{pert}} :

Prompt

You are an expert molecular biologist who studies how genes are related using Perturb-seq.

Task: You are writing a brief overview of the human gene {{name}}, with a focus on molecular and cellular perturbations that may affect the levels of gene {{name}}. For example, you might consider pathways that are upstream of the gene or compensatory mechanisms.

Inputs: You are provided
- Description of gene of interest {{name}}
- Database entries relating {{name}} to other genes or pathways

Format:
- Write up to five sentences describing potential molecular and cellular perturbations that may impact the levels of {{name}}.

Constraints:
- Remain faithful to all domain knowledge. Do not make up additional information.
- Summarize all common aspects succinctly, but point out notable differences within these sets of genes.
- Maintain a professional tone throughout. Do not comment on your own writing. Do not add any notes or references.
- Omit the importance or impact of the gene. Focus only on its function.
- Omit all non-specific information and obvious statements, e.g. "this gene is involved in cellular processes."

Description of gene {{name}}: {{summary}}

Relations to other genes:
{{entries}}

Perturbations that may affect the levels of {{name}}: