# RBIO1 - TRAINING SCIENTIFIC REASONING LLMS WITH BIOLOGICAL WORLD MODELS AS SOFT VERIFIERS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reasoning models are typically trained against verification mechanisms in formally specified systems such as code or symbolic math. In open domains like biology, however, we lack exact rules to enable large-scale formal verification and instead often rely on lab experiments to test predictions. Such experiments are slow, costly, and cannot scale with computation. In this work, we show that world models of biology or other prior knowledge can serve as approximate oracles for *soft verification*, allowing reasoning systems to be trained without additional experimental data. We presemt two paradigms of training models with approximate verifiers: **RLEMF**: reinforcement learning with experimental model feedback and **RLPK**: reinforcement learning from prior knowledge. Using these paradigms, we introduce **rbio1**, a reasoning model for biology post-trained from a pretrained LLM with reinforcement learning, using learned biological models for verification during training. We demonstrate that soft verification can distill biological world models into **rbio1**, enabling it to achieve state-of-the-art performance on perturbation prediction in the PERTURBQA benchmark. We present rbio1 as a proof of concept that predictions from biological models can train powerful reasoning systems using simulations rather than experimental data, offering a new paradigm for model training.

## 1 INTRODUCTION

Building foundation models suitable for scientific tasks is a task of major interest and has produced numerous successful examples in recent memory Abramson et al. (2024); Cui et al. (2024); Lin et al. (2023). Similarly, large language models (LLMs) have shown groundbreaking potential as parametric representations of the world's knowledge, and have been used across every sector. A key challenge is figuring out how to bridge the quantitative accuracy of models of experimental scientific data, for example in biology, with LLMs such that knowledge from these low-level representations of biological systems may be transferred into more flexible and interactive models, such as conversational LLMs, with the explicit goal of being useful for scientific exploration.

Of great promise on scientific tasks are reasoning models, which aim to extend LLMs toward systems that can perform structured, multi-step inference and use test-time compute to generalize better to a given query. Popular reasoning models like DeepSeek-R1 Guo et al. (2025) and QWEN Team (2024) have shown potential in multiple fields, while specialized reasoning LLMs have been explored in fields such as medicine Fallahpour et al. (2025); Cao et al. (2025) and chemistry Narayanan et al. (2025). In frameworks such as reinforcement learning with human feedback (RLHF) Christiano et al. (2017); Stiennon et al. (2020), and reinforcement learning with verifiable rewards (RLVR) Pan et al. (2023), both experimental data collection with human labels and exact oracles of rewards are used to train language models to align to a reward structure and improve their reasoning capabilities. In domains that are not formally specified like biology, however, experimental data and ground-truth verifiers are scarce: while mathematics and code benefit from exact execution and have symbolically accessible oracles, experiments are costly and slow. This motivates exploring alternative supervision strategies for reasoning for such domains.

To overcome these limitations and further advance the utility of reasoning models for scientific tasks in biology, we propose employing models of biological data to run virtual experiments which can be used as sources of probabilistic -or soft- verification signal. This can be seen as a form of reinforcement learning from AI feedback (RLAIF) Lee et al. (2023) with structural adjustments to map to our scientific setting, where RLHF and RLVR are not tractable. We consider those *soft verifiers*, since they return probabilistic rewards which measure the coherence of a biology-model or of biological prior knowledge to a reasoning trace and its returned answer. Much like with RLVR, we can use this *soft verification* paradigm to generate a broad distribution of verified data limited only by how we can query the biology model at hand. We thus turn a (world) model of biology into a reasoning environment to generate rewards to train reasoning models.

Our work also connects with the concept of virtual cell models (VCM) Bunne et al. (2024); Slepchenko et al. (2003); Loew & Schaff (2001), which envisions building powerful predictive systems of biology that can simulate transitions such as diseased $\rightarrow$ healthy states. Advances in compute and large-scale data have enabled construction of such foundation models in specific modalities-transcriptomics Rosen et al. (2023); Pearce et al. (2025); Bian et al. (2024); Ho et al. (2024); Theodoris et al. (2023), imaging Gupta et al. (2024), proteomics Abramson et al. (2024); Lin et al. (2023), genomics Nguyen et al. (2024), and multimodal models Rizvi et al. (2025); Richard et al. (2024); Levine et al. (2024); Schaefer et al. (2024); Choi et al. (2024); Istrate et al. (2024).

Our approach can be seen as using and aligning such world models of biology into a common representation using language as the bridge. This approach not only aggregates knowledge but also makes it accessible through natural language, allowing experimentalists to interact conversationally with biological models. By distilling biological knowledge into LLMs, we transform experimental insights into human-readable reasoning models. Our motivations are threefold: (i) enable training from biological simulations rather than costly experimental data (ii) aggregate diverse models of biology into a universal space, (iii) democratize access to biological knowledge through dialogue.

**Contributions.** Our work contributes to the design of supervision strategies for reasoning LLMs for scientific use, using biological perturbation prediction -e.g., predicting effects of gene knock-downs on differential expression, as a case study:

1. We propose two new processes for training models with AI-verifiers: **RLEMF**: reinforcement learning with experimental model feedback and **RLPK**: reinforcement learning from prior knowledge - that reward with predictive models, and prior knowledge, respectively.
2. RLEMF-trained models generalize OOD and compete with ablation-models trained on experimental data, achieving new state-of-the-art results on the PerturbQA benchmark
3. We show that mixtures of AI-verifiers can be combined to compose stronger models while drawing from different sources of biological knowledge, even when supervision is off-task.
4. We show that inference-time chain-of-thought prompting further improves reasoning performance, allowing **rbio1** to reach state of the art on the PERTURBQA benchmark without tool use or experimental data at inference, even at a fraction of training data.

In summary, rbio-1 extends standard RL training for reasoning models by incorporating AI-based verification through both predictive biological models of experimental data and curated knowledge sources and provides a general framework of using model simulations to train reasoning models.

## 2 RELATED WORK

Recent reasoning-oriented LLMs-such as OpenAI's o-series, Claude 3.7/4, Gemini 2.5, and DeepSeek-R1-exhibit strong multi-step inference and logical deduction across domains. Their development spans four paradigms: (i) inference-time scaling (e.g., chain-of-thought, self-consistency); (ii) pure RL approaches like DeepSeek-R1-Zero, where traces emerge from accuracy- and format-based rewards; (iii) hybrid supervised finetuning plus RL, as in DeepSeek-R1; and (iv) distillation into smaller backbones such as Qwen Team (2024); Yang et al. (2025) or Llama Guo et al. (2025); Touvron et al. (2023). Despite advances, persistent challenges remain in hallucination, logical consistency, verbosity, and interpretability-issues directly tied to the quality of the rewards.

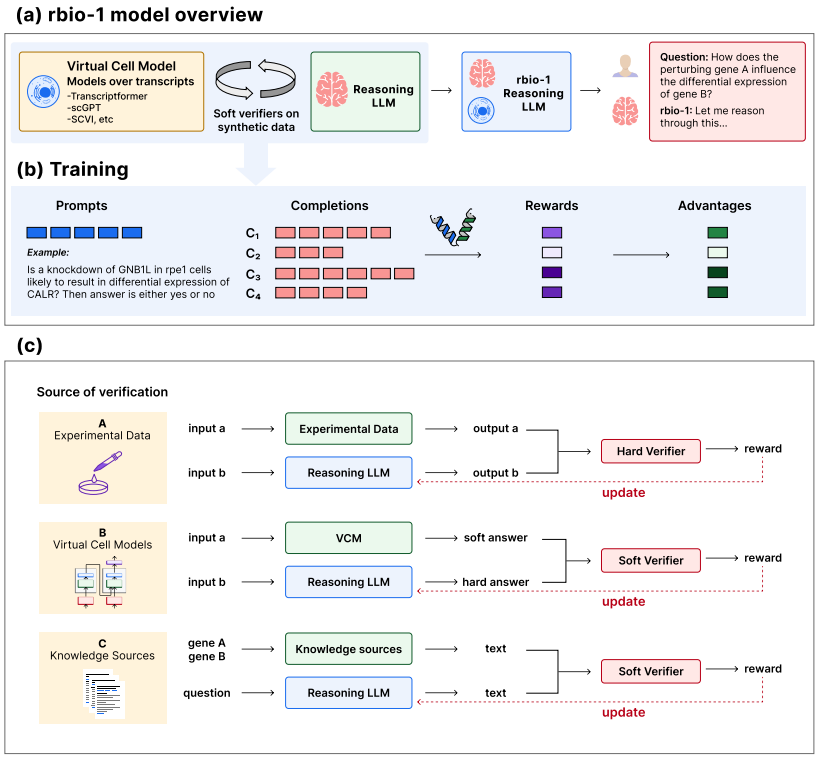


**Figure 1: rbio1 overview.** (a) Distilling VCMs into reasoning LLMs via soft verification. (b) GRPO loop with Virtual Cell Models (VCM) rewards. (c) Soft vs. hard supervision.

Domain-specific reasoning has also been explored. BioReason Fallahpour et al. (2025) combines a genomic encoder with an LLM for disease-pathway inference with interpretable steps, while CellReasoner Cao et al. (2025) frames cell-type annotation explicitly as a reasoning task. Both approaches, however, depend heavily on curated datasets, limiting robustness to noisy or rare populations and motivating richer, more scalable reasoning signals. Our approach differs by using machine learning models of biology directly as reward-generating verifiers. Prior methods integrated external models (e.g., embeddings) into reasoning traces but still evaluated against annotated data. We instead shape rewards themselves with model predictions, showing that biological world models can be distilled into reasoning LLMs -positioning our work within the broader space using AI-rewards.

Wu et al.Wu et al. (2025) propose SUMMER, an inference-time pipeline combining knowledge-graph summaries, retrieval, and chain-of-thought prompting for perturbation prediction. While it outperforms prior methods on PerturbQA, gains are modest, causal directionality remains error-prone, and large models are required even for preprocessing. Unlike SUMMER, our models achieve comparable or better results without experimental data, relying solely on model predictions.

Our work also connects to concurrent research on soft- and AI- verification. In RLAIF Lee et al. (2023) and follow-up work, other LLMs are used as reward mechanisms. Our approach RLEMF 3.3 differs by not requiring an LLM or any text model as an AI-feedback model, and uses models in a different data space of experimental data linked by appropriate prompting techniques and embeddings. Our idea thus builds a bridge between models of experimental data yielding AI-feedback, and the reasoning LLMs learning from that feedback to generate more accurate textual descriptions of valid scientific knowledge. However, we share the approach that model is used to provide a probabilistic verifiable reward. Saad-Falcon et al. (2025) also use LLMs as soft verifiers for other LLMs and combine verifiers. In contrast, we generalize beyond LLMs to arbitrary biological models and combine multiple verifiers as separate reward functions. In a framework closest to our approach RLPK 3.4, Yu et al. (2025) use LLMs to use the reasoning LLM itself to score answers as rewards. In RLPK we do not use answers, but structured databases of prior scientific knowledge.

**Table 1:** Verifiers used during RL training. EXP = experimental data; MLP = multilayer perceptron; GO = Gene Ontology.

| Verifier | Type | Reward Signal | Source |
|---|---|---|---|
| EXP | Hard | Binary $r_i^{hard} \in \{0, 1\}$ | Experimental data |
| MLP | Soft | Probability $r_i^{soft} = p,\ 0 \leq p \leq 1$ | Simulations |
| GO | Soft | ROUGE, keyword, likelihood | Knowledge base |

To our knowledge, we are the first to apply this paradigm to reasoning models for biology, shifting the training signal from experimental data to simulations and broadening the design space of verifiers for reasoning LLMs.

## 3 RBIO1: METHODS

In standard domains, during RL training, verifiers return precise signals-for example, whether code executes or a math solution is correct. In biology, some queries can be validated experimentally (hard verification), but exhaustive lab testing is infeasible due to scale. Consider a biological query related to genetic perturbation, such as: *Is a knockdown of AARS in hepg2 cells likely to result in differential expression of ATAD2B?* with a binary answer: yes/no. During training, the LLM produces completions $o_i$ for query $q$. Rewards can be assigned in three ways that we introduce in the following sections and also showcase in Fig. 1. Table 1 summarizes these verifiers and reward formulations.

### 3.1 REINFORCEMENT LEARNING FOR REASONING

Let $P(Q)$ denote a dataset used for training; $q$ a query sampled from $P(Q)$, $G$ a set of outputs generated during training by the reasoning LLM $\pi_\theta$; $o_i$ a generated sequence of tokens with tokens $o_{i,t}$ in response to $q$; $\pi_{\text{ref}}$ a reference base model from the supervised finetuned LLM; $r_\phi$ a reward model emitting rewards $r_i$; $L_{GRPO}(\theta)$ the surrogate objective and $\beta$ the coefficient for the KL penalty. Given these variables, Group Relative Policy Optimization (GRPO) Mroueh (2025) training maximizes the following objective function, with the goal of increasing the accumulated collective rewards $\{r_{i,\geq t}\}$:

$$J_{GRPO} = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}} \big[L_{GRPO}(\theta)\big]. \tag{1}$$

We use the clipped surrogate objective:

$$L_{GRPO}(\theta) = \frac{1}{|G|}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\frac{\pi_\theta(o_{i,t}|q,o_{i<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i<t})}\hat{A}_{i,t},\ g(\epsilon,\hat{A}_{i,t})\Big) - \beta D_{KL}[\pi_\theta||\pi_{\text{ref}}] \tag{2}$$

$$g(\epsilon,\hat{A}_{i,t}) = \text{clip}\Big(\frac{\pi_\theta(o_{i,t}|q,o_{i<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i<t})},\ 1-\epsilon,\ 1+\epsilon\Big)\,\hat{A}_{i,t} \tag{3}$$

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_1,\ldots,r_G\})}{\text{std}(\{r_1,\ldots,r_G\})} \tag{4}$$

$$D_{KL}[\pi_\theta||\pi_{\text{ref}}] = \frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i<t})}{\pi_\theta(o_{i,t}|q,o_{i<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i<t})}{\pi_\theta(o_{i,t}|q,o_{i<t})} - 1 \tag{5}$$

### 3.2 RBIO-EXP: REINFORCEMENT LEARNING WITH HARD VERIFICATION

In this framework we are most similar to the RLHF scenario, where direct observations of experimental data are translated into language tokens and used to train a reasoning model directly using GRPO. For a task with a binary outcome and a verifier $V$ that emits binary rewards, $r_i$ in Eq. 4 becomes: $\text{r}_i^{hard}(q, o_i) = r_i(q, o_i \mid V) \in \{0, 1\}$

We consider the existence of a broad experimental dataset $D_{EXP}$ that can be used as a source of this reward feedback given a query, where the reward takes the shape of a label about a scientific

fact we can ask the model to reason about. If the outcome of $q$ has been validated experimentally and is available in $D_{EXP}$, we can directly verify $o_i$ and emit a binary reward using $V = \{D_{EXP}\}$:

$$r_i^{hard}(q, o_i) = r_i(q, o_i \mid D_{EXP}) \tag{6}$$

$$r_i^{hard}(q, o_i) = \begin{cases} 1, & o_i = \text{True},\ D_{\text{EXP}}(q) = \text{True}, \\ 1, & o_i = \text{False},\ D_{\text{EXP}}(q) = \text{False}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

### 3.3 RBIO-RLEMF: REINFORCEMENT LEARNING WITH EXPERIMENTAL MODEL FEEDBACK

Similar to the framework of RLAIF, we here propose a related process which utilizes arbitrary other (non-LLM) models as feedback mechanisms for a query, in our example world models of biology defined on experimental data that can be queried appropriately. In the absence of experimental data $D_{EXP}$ for RL-training as explained in Sec. 3.2, predictive models of such data $M$ can act as surrogate verifiers ($V = \{M\}$). If $q$ has not been validated experimentally, we can verify $o_i$ using predictions from a biological model $M$. The reward is $r_i^{soft}(q; o_i) = fn(M(q_i; c_j))$, where $c_j$ denotes the context (e.g., cell line or covariates). The emitted rewards are probabilistic:

$$r_i^{soft}(q, o_i) = r_i(q, o_i \mid V) = p(q, o_i \mid M), \quad 0 \le p \le 1 \tag{8}$$

In the example of the biological application of evaluating perturbation prompts with a model, we consider $M :=$ MLP, and $p(q, o_i \mid M)$ becomes the model's predicted probability for $q$ being true.

### 3.4 RBIO-RLPK: REINFORCEMENT LEARNING FROM PRIOR KNOWLEDGE

Another avenue we propose for injecting knowledge into reasoning models for science is via prior knowledge. Here, given a structured database of prior knowledge, we can query a reasoning model against knowledge in that database and score the model itself against it. Given that knowledge sources-denoted $KS$ are able to act as verifiers $V = \{KS\}$ and emit rewards, we can generate rewards on $o_i$ based on $KS$ using some metric $m$. This setting is outlined in Fig. 1c - C for the case of perturbation prediction. Concretely, we use curated resources such as the Gene Ontology (GO) Aleksander et al. (2023); Ashburner et al. (2000), which provides gene annotations across axes like *molecular processes*, *cellular components*, and *biological processes*. Eq. 4 then becomes:

$$r_i^{soft}(q, o_i) \;=\; r_i(q, o_i \mid V) \;=\; r_m(q, o_i \mid KS), \tag{9}$$

where rewards are not necessarily confined to $[0, 1]$, but depend on the chosen metric.

We experiment with three types of metrics: ROUGE-based scores, keywords-based scores, and likelihood estimations, all of which require querying $KS$ (GO ontology) for prior knowledge on $q$. Assuming we have access to $\{q_j\}$ pieces of prior knowledge on $q$, we accumulate rewards as:

$$q_j^{prior} \mid KS = \texttt{query\_KS}(q_j), \qquad r_i^{soft}(q, o_i) \;=\; \sum_j r_m(q_j^{prior}, o_i \mid KS). \tag{10}$$

**ROUGE-based verifiers.** We request the model to expose the relevant gene facts inside <gene_info> tags - which we refer to as $o_i^{relevant}$ - and compute standard ROUGE-1/2/L F-scores between $q_j^{prior}$ and the extracted $o_i^{relevant}$:

$$r_m \;=\; \sum_j \sum_{X \in \{1,2,L\}} \text{ROUGE}{-}X\big(q_j^{prior},\, o_i^{relevant}\big).$$

**Keywords-based verifiers.** We count normalize overlap of GO keywords present in the reasoning trace:

$$r_m \;=\; \sum_j \text{KWS}\big(q_j^{prior},\, o_i^{relevant}\big), \qquad \text{KWS}(s_1, s_2) \;=\; \frac{|s_1 \cap s_2|}{|s_1|}. \tag{11}$$

**Likelihood-based verifiers.** For likelihood-based verifiers, we use the likelihood of the prior knowledge $\{q_j^{prior}\}$ under our learned policy $\pi_\theta$ to generate rewards under the reasoning model, encouraging higher likelihood for scientifically accurate facts to reduce hallucinations. To account for variability in sequence length, we average over the sequence tokens $y_k$ in $q_j^{prior}$:

$$r_m(q, o_i \mid LL) = \sum_j LL_{\pi_\theta}(q_j^{prior}); \quad LL_{\pi_\theta}\left(q_j^{prior} \mid \pi_\theta\right) = \frac{1}{T}\sum_{k=1}^{T} \log p_{\,\pi_\theta}(y_k \mid y_{<k}) \tag{12}$$

**Normalization.** GRPO uses normalized advantages (Eq. 4), mapping rewards to mean 0 and std 1. When composing multiple verifiers, imbalanced scales can skew updates, with GO-based rewards most affected due to skewed metric distributions. We therefore normalize GO-based rewards to $[0, 1]$ using an Exponential Moving Average (EMA) before policy updates:

$$\tilde{r} \leftarrow (1-\alpha)\tilde{r} + \alpha r_m, \quad \tilde{v} \leftarrow (1-\alpha)\tilde{v} + \alpha(r_m - \tilde{r}_{\text{prev}})(r_m - \tilde{r}), \tag{13}$$

$$\bar{r} = 0.5 + \frac{1}{2z_{\max}} \operatorname{clip}\left(\frac{r_m - \tilde{r}}{\max\left(\sqrt{\tilde{v}+\varepsilon},\, s_{\min}\right)}, -z_{\max},\, z_{\max}\right). \tag{14}$$

This $\bar{r}$ provides a normalized reward signal, used in Eq. 4 to compute the token-level advantages.

### 3.5 Composable Verification for model integration

For all rbio models, we also use formatting rewards $r_{format}$ and mention rewards $r_{mention}$ (e.g., gene mentions). When we train on combinations of verifiers as described in Sec. 4.2, each prompt $q$ can be verified with a different verification source $V_s$. With multiple verifiers $V_k$ emitting rewards $r_{i,k}$, we then have:

$$r_i(q, o_i) = r_{\text{format}} + r_{\text{mention}} + \sum_k \delta_{ks}\, \lambda_k\, r_{i,k}(q, o_i \mid V_k), \qquad \lambda_k \geq 0. \tag{15}$$

Unless otherwise stated, $\lambda_k = 2$. This gives more weight to the variance of the soft verifiers compared to $r_{format}$ and $r_{mention}$ during GRPO updates.

**Use of LLMs**. We used GPT-based tools for minor writing polish and for code assistance in generating plots; all scientific contributions are solely by the authors.

## 4 Experiments

### 4.1 rbio with AI-verification generalizes OOD on perturbation tasks

On PerturbQA Wu et al. (2025) (CRISPRi knockdowns in RPE1, K562, HEPG2, JURKAT), models trained with *soft verifiers* generalize to held-out cell lines, reducing reliance on cell-line–specific experimental data. We first evaluate a 2-layer MLP (64 hidden units) trained on three cell lines and use it to generate predictions on the fourth, which serve as rewards during RL. Gene representations include one-hot, Gene2Vec Du et al. (2019), and ESM Lin et al. (2023). The resulting models, *rbio-MLP-leave-one-out-one-hot* and *rbio-MLP-leave-one-out-gene2vec*, perform comparably to experimental-data–trained rbio models.

We compare to two experimental-data baselines: *rbio-EXP-one-cell-line* (train/test within a cell line; Fig. 2a) and *rbio-EXP-leave-one-out* (train on three cell lines; test on the fourth; Fig. 2b). We also benchmark against SUMMER Wu et al. (2025). As shown in Fig. 2d–e, the soft-verifier models closely match experimental-data models on F1 and MCC, and exceed them in Balanced Accuracy via higher TPR while maintaining similar TNR. Identifying true effects is paramount in perturbation, so higher TPR is valuable even with some F1 trade-off. All rbio variants also outperform GEARS Roohani et al. (2022) and the base Qwen2.5-3B.

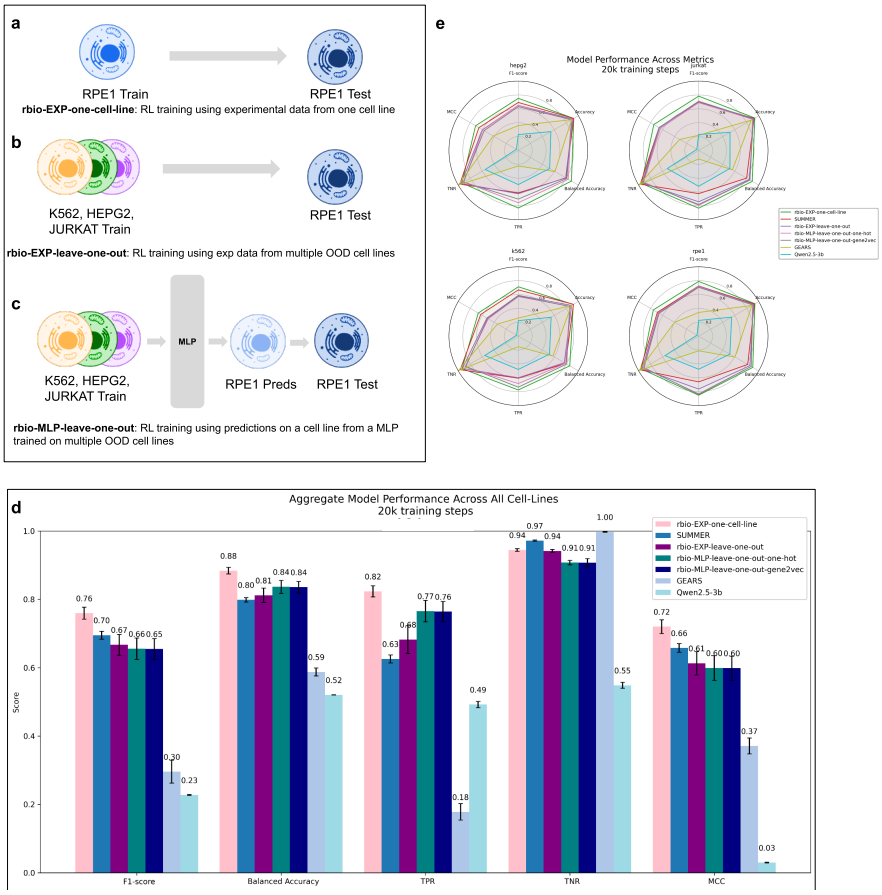


**Figure 2: Model performance for experimental vs. simulation-based soft verification. (a) rbio-EXP-one-cell-line:** trained and tested on the same cell line (in-distribution). **(b) rbio-EXP-leave-one-out:** trained on three cell lines, tested on the held-out one (out-of-distribution). **(c) rbio-MLP-leave-one-out:** trained using MLP predictions on the held-out line (MLP fit on the others). **(d) Aggregate metrics:** computed over four cell lines (K562, RPE1, JURKAT, HEPG2), averaged across 5 runs. **(e) Metrics split by cell line.** Baselines: *SUMMER* (experimental + domain knowledge), *GEARS* (specialized perturbation model),*Qwen2.5-3b* (base reasoning model).

### 4.2 Training rbio on mixtures of AI-verifiers leads to performance gains

We find that combining verifiers improves performance over using them individually. Notably, the order in which models see the verifiers matters, reflecting differences in the knowledge provided. For a pair of verifiers $V_i, V_j$, we evaluate:

1. $V_i$: trained only with $V_i$ , $i \in \{1, 2\}$
2. $V_i \| V_j$: trained sequentially, $V_i$, then $V_j$
3. $V_i \cup V_j$: trained on a random mixture of $V_i$ and $V_j$

We experiment with the following combinations of verifiers:

1. $V_1$ = EXP (hard verifier; experimental data); $V_2$ = MLP (soft verifier; MLP predictions)
2. $V_1$ = EXP (hard verifier; experimental data); $V_2$ = $GO_{all-ll}$ (soft verifier; GO Ontology Knowledge Source, likelihood-based reward)
3. $V_1$ = MLP (soft verifier; MLP predictions); $V_2$ = $GO_{all-ll}$ (soft verifier; GO Ontology Knowledge Source, likelihood-based reward)

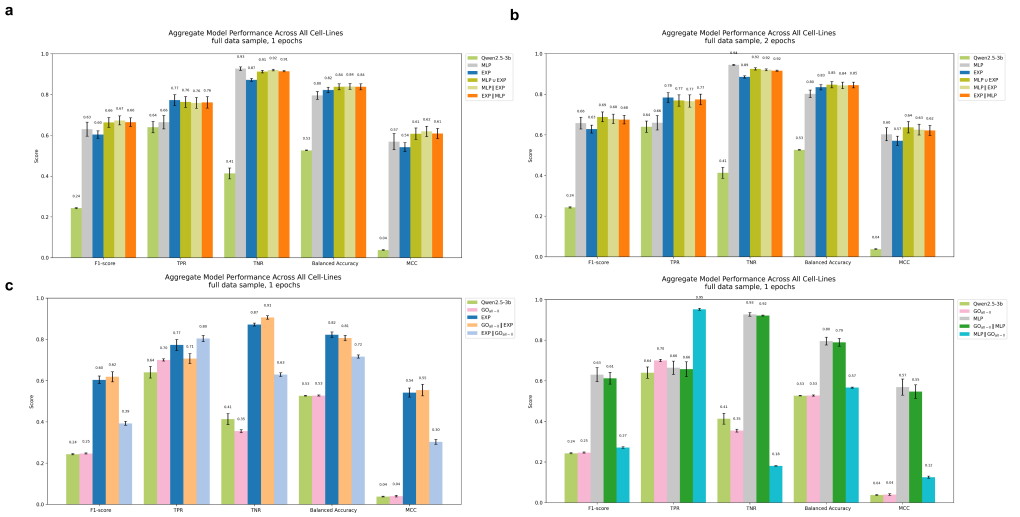


**Figure 3: Model performance for compositions of verifiers**
$V_i \| V_j$ corresponds to training models sequentially, first on $V_i$, then on $V_j$. $V_i \cup V_j$ corresponds to models trained on a random mixture of $\{V_i, V_j\}$; **(a, b)** $MLP$ and $EXP$, trained for 1, and 2 epochs. **(c)**$EXP$ and $GO_{all-ll}$ **(d)**$MLP$ and $GO_{all-ll}$

Note that the training data for each of $V_1, V_2$ is independent of each other - i.e. if $V_1$ is a verifier of experimental data from a dataset $D_1$, emissions from $V_2$ will be on an independent dataset $D_2$ where $D_1 \cap D_2 = \varnothing$. In the case of the $GO_{all-ll}$ the soft verification is the likelihood of the prior knowledge $\{q_j^{prior}\}$ we have under our learned policy $\pi_\theta$ as described in Eq. 12.

As shown in Fig. 3, adding verifiers consistently improves performance over using them individually. For $V_1$ = EXP and $V_2$ = MLP (Fig. 3a,b), all three composition strategies (Sec. 4.2) perform similarly, yet each surpasses the single verifiers, underscoring the complementary value of strong verification sources such as experimental data and models of experimental data.

When mixing knowledge and experimental sources, order becomes critical. In Fig. 3c,d, training first on $GO_{all-ll}$ then on $MLP, EXP$ ($GO_{all-ll} \| MLP$, $GO_{all-ll} \| EXP$) outperforms the reverse. GO-based verification increases TPR (by capturing more positives) - as shown in its performance when evaluated individually compared to baseline - but reduces TNR; subsequent training on experimental data rebalances TNR, improving Balanced Accuracy and MCC. Conversely, starting from EXP or MLP then adding $GO_{all-ll}$ lowers performance, suggesting knowledge sources can dilute experimental signals if applied late. Thus, knowledge is most effective early to guide the model, while stronger experimental signals should refine performance later. This aligns with general training paradigms: start with broader, noisier data (e.g. ontologies) to shape representations, then refine with higher-quality data (e.g. experiments) to maximize performance. The strategy is extendable to multiple verifiers $V_1, V_2, \ldots, V_k$, which could capture different sources of knowledge.

### 4.3 RBIO WITH CHAIN-OF-THOUGHT YIELDS STATE OF THE ART ON PERTURBQA

Adding chain-of-thought (CoT) reasoning at inference improves all rbio variants we tested (Table. 2), surpassing SUMMER as state-of-art performance on the PerturbQA benchmark. The CoT prompt that performed the best was: *'The Biologist will evaluate each step of this problem, using logical reasoning and evidence from the prompt."* Examples of performance increase: *rbio-EXP-all-cell-lines* F1 0.75→0.79, Balanced Accuracy 0.88→0.91, TPR 0.83→0.87; *rbio-MLP-ESM* F1 0.67→0.71, Balanced Accuracy 0.85→0.89, TPR 0.81→0.87. We offer examples of answers and reasoning traces generated by the rbio-models on a perturbation question in Figure 5 in Supplementary material. Shown in Figure 4 are rbio models trained on only one-fifth of the data and tested with and without CoT. Remarkably, adding CoT at inference lets them reach state-of-the-art performance on PerturbQA - with *rbio-MLP* ∪ *EXP-CoT* surpassing SUMMER despite being trained

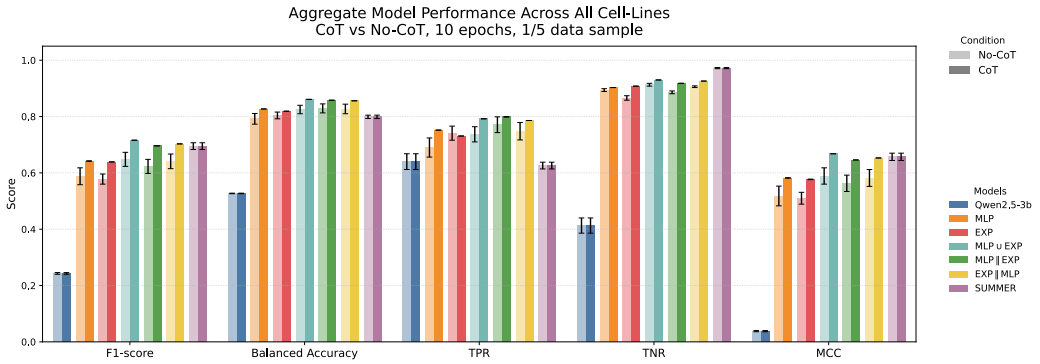


**Figure 4: Effect of chain-of-thought prompting.** Models using CoT achieve state-of-the-art performance on the PerturbQA benchmark.

| Model | F1-score | Balanced Accuracy | TPR | TNR | MCC |
|---|---|---|---|---|---|
| *Models trained on full data size* | | | | | |
| rbio-EXP | 0.750 ± 0.018 | 0.883 ± 0.011 | 0.827 ± 0.018 | 0.939 ± 0.003 | 0.709 ± 0.020 |
| **rbio-EXP-CoT** | **0.786 ± 0.000** | **0.907 ± 0.000** | **0.872 ± 0.000** | 0.943 ± 0.000 | **0.752 ± 0.000** |
| rbio-MLP | 0.669 ± 0.025 | 0.855 ± 0.017 | 0.807 ± 0.030 | 0.902 ± 0.004 | 0.618 ± 0.029 |
| rbio-MLP-CoT | 0.714 ± 0.000 | 0.889 ± 0.000 | 0.873 ± 0.000 | 0.906 ± 0.000 | 0.672 ± 0.000 |
| SUMMER | 0.695 ± 0.012 | 0.799 ± 0.006 | 0.626 ± 0.012 | 0.972 ± 0.002 | 0.657 ± 0.013 |
| Qwen2.5-3b | 0.231 ± 0.002 | 0.522 ± 0.001 | 0.529 ± 0.014 | 0.515 ± 0.013 | 0.032 ± 0.001 |
| GEARS | 0.296 ± 0.033 | 0.588 ± 0.012 | 0.178 ± 0.024 | **0.997 ± 0.001** | 0.371 ± 0.023 |
| *Models trained on 1/15 of full data size* | | | | | |
| rbio-MLP | 0.588 ± 0.030 | 0.792 ± 0.019 | 0.690 ± 0.034 | 0.894 ± 0.005 | 0.518 ± 0.035 |
| rbio-MLP-CoT | 0.642 ± 0.001 | 0.827 ± 0.000 | 0.752 ± 0.000 | 0.903 ± 0.000 | 0.582 ± 0.001 |
| rbio-EXP | 0.578 ± 0.018 | 0.804 ± 0.012 | 0.741 ± 0.025 | 0.866 ± 0.008 | 0.510 ± 0.021 |
| rbio-EXP-CoT | 0.639 ± 0.000 | 0.819 ± 0.000 | 0.731 ± 0.000 | 0.908 ± 0.000 | 0.577 ± 0.001 |
| rbio-MLP ∪ EXP | 0.648 ± 0.025 | 0.825 ± 0.015 | 0.737 ± 0.027 | 0.913 ± 0.005 | 0.589 ± 0.029 |
| **rbio-MLP ∪ EXP-CoT** | **0.716 ± 0.000** | **0.861 ± 0.000** | 0.792 ± 0.000 | 0.930 ± 0.000 | **0.668 ± 0.000** |
| rbio-MLP ‖ EXP | 0.623 ± 0.025 | 0.829 ± 0.016 | 0.771 ± 0.028 | 0.886 ± 0.005 | 0.563 ± 0.029 |
| rbio-MLP ‖ EXP-CoT | 0.696 ± 0.000 | 0.858 ± 0.000 | **0.799 ± 0.001** | **0.918 ± 0.000** | 0.646 ± 0.000 |
| rbio-EXP ‖ MLP | 0.641 ± 0.026 | 0.827 ± 0.017 | 0.748 ± 0.031 | 0.906 ± 0.003 | 0.582 ± 0.030 |
| rbio-EXP ‖ MLP-CoT | 0.703 ± 0.000 | 0.856 ± 0.000 | 0.786 ± 0.000 | 0.926 ± 0.000 | 0.653 ± 0.000 |

**Table 2:** Aggregate performance across datasets on the PerturbQA benchmark. Values are mean ± standard error (SE) over 5 different completions. *rbio-EXP* corresponds to rbio-EXP-all-cell-lines. Comparison to baselines including SUMMER Wu et al. (2025) (current SOTA). Best model in each category bolded.

on a fraction of training data - demonstrating the power of inference-time capabilities and verifier composition in reasoning models.

## 5 Conclusion

We introduce **rbio1**, a suite of reasoning models trained via *soft verification*, where simulations from biological world models provide rewards for reinforcement learning. This approach rivals experimental-data–trained models, especially when combined with chain-of-thought prompting. By leveraging predictive bio-models (e.g., MLPs on gene embeddings) and knowledge sources like the GO Ontology, rbio1 shows that simulations and prior knowledge can substitute for costly experimental supervision. We aim to extend rbio1 across diverse biological models and modalities toward a universal virtual cell system integrating multiple sources into a shared reasoning framework. Beyond biology, soft verification offers a general supervision strategy for reasoning LLMs, enabling scalable training in domains without exact verifiers and raising open questions on verifier design, noisy-signal balancing, and evaluating reasoning quality beyond task accuracy.

## REFERENCES


Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

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.

Haiyang Bian, Yixin Chen, Xiaomin Dong, Chen Li, Minsheng Hao, Sijie Chen, Jinyi Hu, Maosong Sun, Lei Wei, and Xuegong Zhang. scmulan: a multitask generative pre-trained language model for single-cell analysis. In *International Conference on Research in Computational Molecular Biology*, pp. 479–482. Springer, 2024.

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.

Guangshuo Cao, Yi Shen, Jianghong Wu, Haoyu Chao, Ming Chen, and Dijun Chen. Cellreasoner: A reasoning-enhanced large language model for cell type annotation. *bioRxiv*, pp. 2025–05, 2025.

Hongyoon Choi, Jeongbin Park, Sumin Kim, Jiwon Kim, Dongjoo Lee, Sungwoo Bae, Haenara Shin, and Daeseung Lee. Cellama: foundation model for single cell and spatial transcriptomics by cell embedding leveraging language model abilities. *bioRxiv*, pp. 2024–05, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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.

Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(Suppl 1):82, 2019.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ankit Gupta, Zoe Wefers, Konstantin Kahnert, Jan N Hansen, Will Leineweber, Anthony Cesnik, Dan Lu, Ulrika Axelsson, Frederic Ballllosera Navarro, Theofanis Karaletsos, et al. Subcell: Vision foundation models for microscopy capture single-cell biology. *bioRxiv*, pp. 2024–12, 2024.

Nicholas Ho, Caleb N Ellington, Jinyu Hou, Sohan Addagudi, Shentong Mo, Tianhua Tao, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, et al. Scaling dense representations for single cell with transcriptome-scale context. *bioRxiv*, pp. 2024–11, 2024.

Ana-Maria Istrate, Donghui Li, and Theofanis Karaletsos. scgenept: Is language all you need for modeling single-cell perturbations? *bioRxiv*, pp. 2024–10, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Victor Cărune, Abhinav Rastogi, and Sushant Prakash. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pp. 2023–09, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

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.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Siddharth M Narayanan, James D Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G Rodriques, and Andrew D White. Training a scientific reasoning model for chemistry. *arXiv preprint arXiv:2506.17238*, 2025.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.

William Pan, Xuechen Song, Yuchen Zhang, Percy Liang, and Tatsunori B Hashimoto. Reinforcement learning with verifiable rewards from correctness feedback. *arXiv preprint arXiv:2303.17491*, 2023.

James D Pearce, Sara E Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li, et al. A cross-species generative cell atlas across 1.5 billion years of evolution: The transcriptformer single-cell model. *bioRxiv*, pp. 2025–04, 2025.

Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie Lopez, Alexandre Laterre, Maren Lang, et al. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. *bioRxiv*, pp. 2024–04, 2024.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, et al. Scaling large language models for next-generation single-cell analysis. *bioRxiv*, pp. 2025–04, 2025.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Gears: Predicting transcriptional outcomes of novel multi-gene perturbations. *BioRxiv*, pp. 2022–07, 2022.

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

Jon Saad-Falcon, E Kelly Buchanan, Mayee F Chen, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, et al. Shrinking the generation-verification gap with weak verifiers. *arXiv preprint arXiv:2506.18203*, 2025.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Joerg Menche, Eleni M Tomazou, et al. Multimodal learning of transcriptomes and text enables interactive single-cell rna-seq data exploration with natural-language chats. *bioRxiv*, pp. 2024–10, 2024.

Boris M Slepchenko, James C Schaff, Ian Macara, and Leslie M Loew. Quantitative cell biology with the virtual cell. *Trends in cell biology*, 13(11):570–576, 2003.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Caleb Voss, Alec Radford, Dario Amodei, Paul Christiano, and Jan Leike. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, et al. Rlpr: Extrapolating rlvr to general domains without verifiers. *arXiv preprint arXiv:2506.18254*, 2025.

# A Appendix

## A.1 Metrics

We formulate the genetic perturbation prediction task as a question in natural language with a binary answer. Given a pair of genes $gene_A$ and $gene_B$, the model is asked to emit a binary answer - **yes** or **no**. We use four CRISPRi single-gene perturbation knockdown datasets on four cancer cell lines (RPE1, K562, HEPG2, JURKAT), post-processed into natural language queries by PerturbQA Wu et al. (2025). We compute the following metrics:

$$Recall\ (TPR) = \frac{TP}{TP + FN} \tag{16}$$

$$TNR = \frac{TN}{TN + FP} \tag{17}$$

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{19}$$

$$Balanced\ Accuracy = \frac{TPR + TNR}{2} \tag{20}$$

$$MCC\ (Matthews\ Correlation\ Coefficient) = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{21}$$

## A.2 EXAMPLES OF REASONING TRACES

*Is a knockdown of ISCA2 in rpe1 cells likely to result in differential expression of CEP295? The answer is either yes or no*

**system_prompt**
A conversation between User and Biologist. The user asks a question, and the Biologist solves it. The biologist first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>.

**Example answer**

*Knockdown of ISCA2 in RPE1 cells would likely lead to differential expression of CEP295. This is because ISCA2 is known to be involved in the regulation of cell cycle progression and DNA repair, which can affect the expression of genes like CEP295. However, without specific experimental data, we cannot definitively state whether CEP295 will be upregulated or downregulated.*

**system_prompt**
A conversation between User and Biologist. The user asks a question, and the Biologist solves it. The biologist first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>.The Biologist will evaluate each step of this problem, using logical reasoning and evidence from the prompt.

**Example answer using Chain-of-Thought**

*ISCA2 is a protein that is involved in the regulation of cell cycle progression. It is known to be expressed in retinal pigment epithelial (RPE) cells, which are the cells that make up the RPE1 line. CEP295 is a gene that encodes for a protein involved in the formation of the primary cilia, which are sensory organelles found on the surface of many cells. Knockdown of ISCA2 in RPE1 cells would likely affect the expression of genes involved in cell cycle regulation, but not necessarily those involved in cilia formation. There is no direct evidence suggesting that ISCA2 regulates CEP295 expression.*

**system_prompt**
You are an AI model trained as a Biologist through reinforcement learning. I will ask you a question, you will come up with a reasoning process based on what you have learned during training and then you will give me the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. You will provide the reasoning step-by-step, using detailed biological knowledge from training.

**Example answer using self-aware Chain-of-Thought**

*Knockdown of ISCA2 in RPE1 cells would likely lead to differential expression of CEP295. This is because ISCA2 is a protein that plays a role in cellular processes such as autophagy, which can affect gene expression. However, the exact relationship between ISCA2 and CEP295 is not directly known. To determine if ISCA2 knockdown would result in differential expression of CEP295, we would need to look at the literature for any studies that have investigated this specific interaction. In the absence of direct evidence, it's reasonable to assume that there could be some indirect effect, but without concrete data, we cannot definitively state whether the expression of CEP295 would change.*

**Figure 5: — Examples of Reasoning Traces for perturbation prediction answering a question sampled randomly from the test set.** Different example answers for different chain-of-thought techniques.Outputs from a model trained on combinations of soft-verifiers.