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

Istrate *et al.*

Chan Zuckerberg Initative

August 22, 2025

ABSTRACT

Reasoning Models are typically trained against verification mechanisms in formally specified systems such as code or symbolic math. However, in open domains like biology, we do not generally have access to exact rules facilitating formal verification at scale, and oftentimes resolve to testing hypotheses in the lab to assess the validity of a prediction.

Verification by performing real experiments is slow, expensive, and inherently does not scale with computation.

In this work, we show that one can use **world models of biology or other prior knowledge as approximate oracles over biological knowledge to utilize as soft verification to train reasoning systems without the need for additional experimental data.**

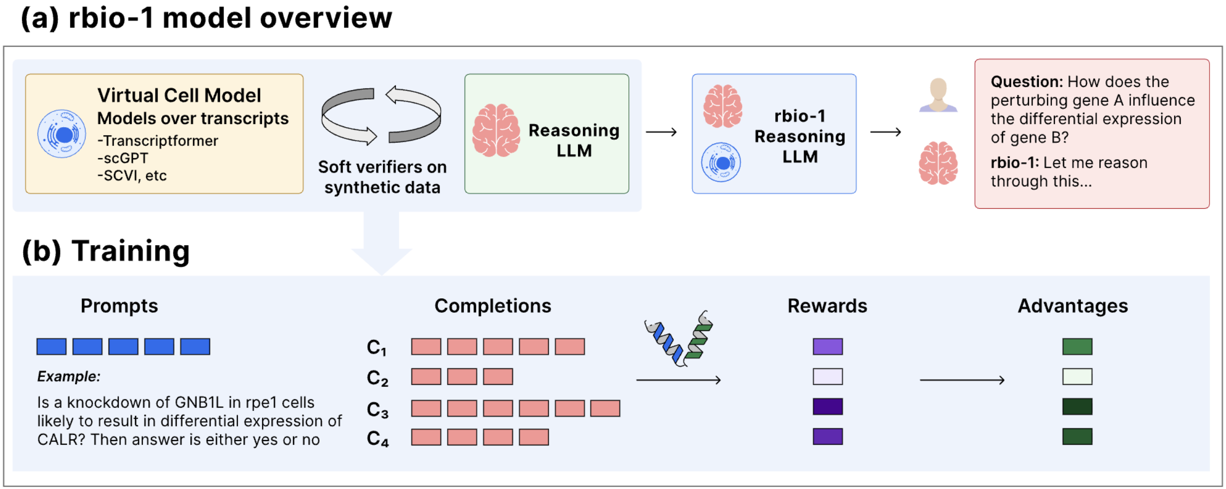
We introduce rbio1, **a reasoning model for biology that is post-trained from a pretrained LLM using reinforcement learning and uses learned models of biology to obtain biological knowledge for verification during training.** We show that soft verification successfully distills biology world models into rbio, at the example of achieving leading performance on perturbation prediction against the PerturbQA benchmark compared to state-of-the-art models; we demonstrate the benefits of compositions of verifiers to learn more general rbio models.

We believe rbio provides a proof of concept that demonstrates that predictions from bio-models can be used to train powerful reasoning models using simulations, rather than experimental data, as a new training paradigm.

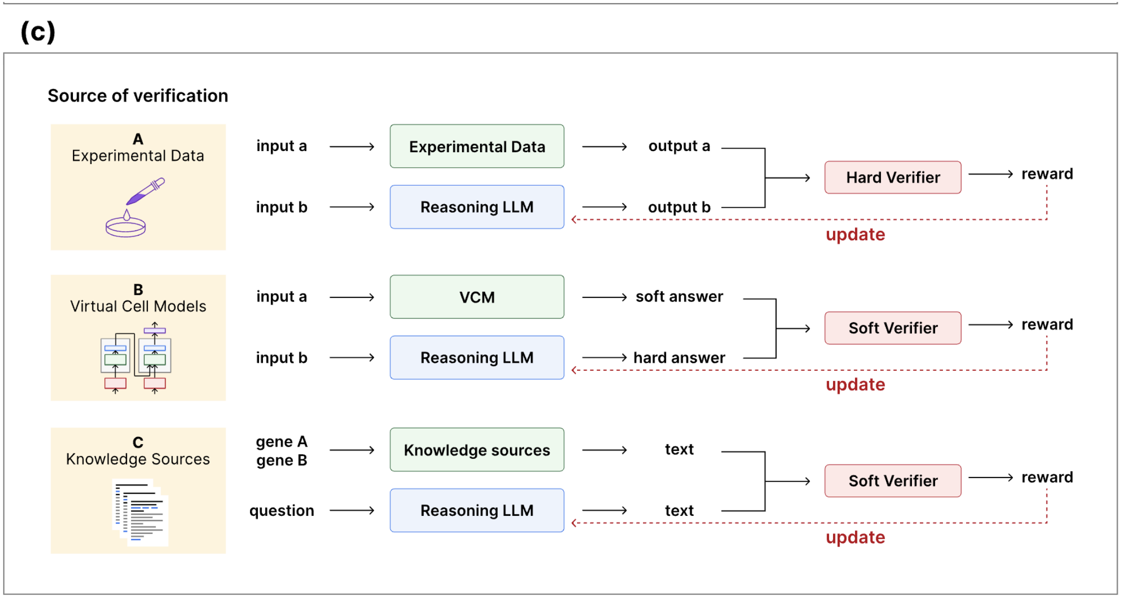
Introduction: Motivation and Core Idea

- **Problem:** Experimental biological supervision does not scale
  - Experiments are expensive, slow, and sparse
  - Many biologically meaningful questions lack labels
- **Limitation of current LLM training**
  - Supervised fine-tuning assumes access to ground truth
  - Fails when labels are unavailable or incomplete
- **Key insight**
  - Biology already exists in multiple forms:
    - Experiments
    - Predictive models (simulations)
    - Curated knowledge
  - These can act as **verifiers**, not labels
- **Core idea of the paper**
  - Replace ground truth with **verification signals**
  - Learn biological reasoning via **reinforcement learning**

The overal structure of rbio1


(a) rbio-1 model overview
Virtual Cell Model
Models over transcripts
-Transcriptformer
-scGPT
-SCVI, etc
Soft verifiers on synthetic data
Reasoning LLM
rbio-1 Reasoning LLM
Question: How does the perturbing gene A influence the differential expression of gene B?
rbio-1: Let me reason through this...
(b) Training
Prompts
Example:
Is a knockdown of GNB1L in rpe1 cells likely to result in differential expression of CALR? Then answer is either yes or no
Completions
C_1
C_2
C_3
C_4
Rewards
Advantages

THE OVERAL STRUCTURE OF RBIO1



- `input a` and `input b`?
- What are the hard/soft verifiers?
- How is the reward computed?

TABLE 1: THE KEY TO UNDERSTAND THE REST OF PAPER

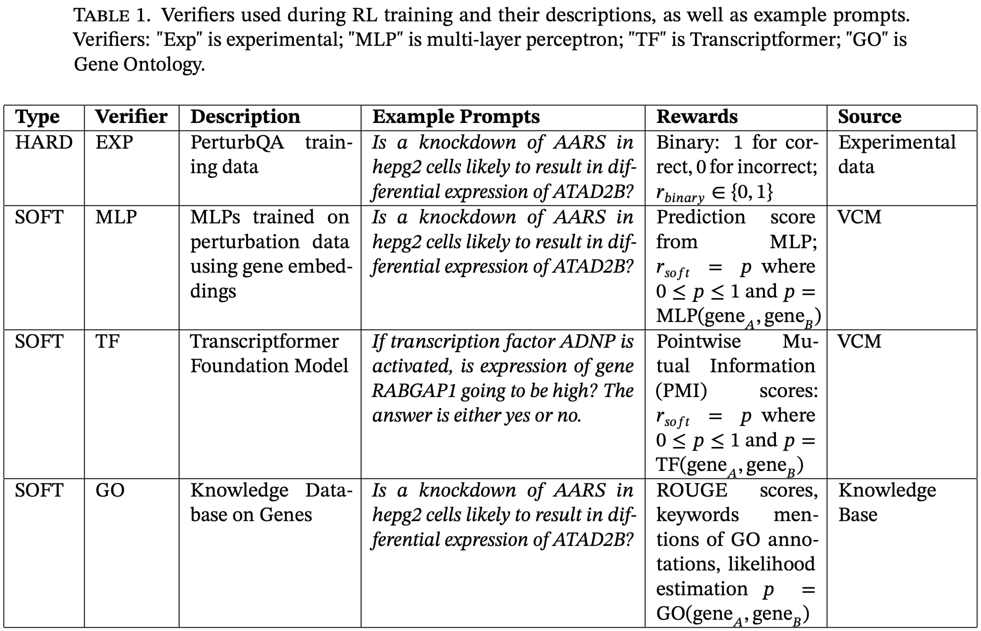
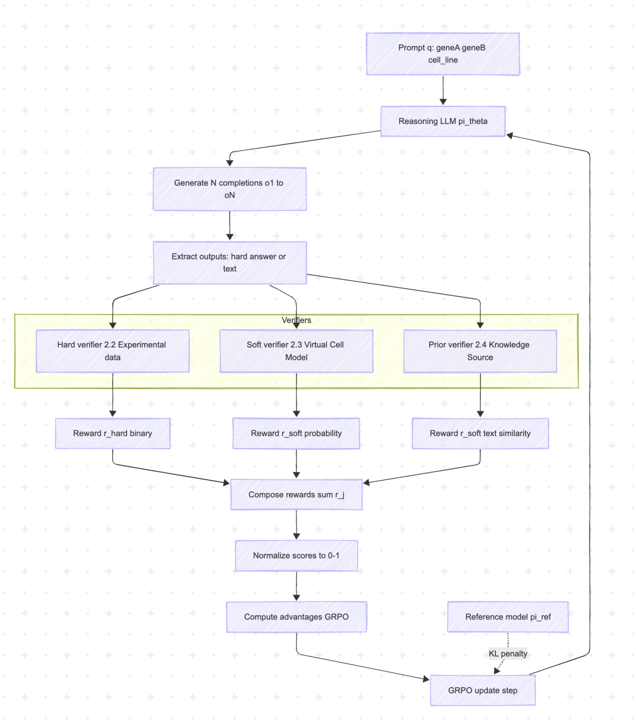


TABLE 1. Verifiers used during RL training and their descriptions, as well as example prompts. Verifiers: "Exp" is experimental; "MLP" is multi-layer perceptron; "TF" is Transcriptformer; "GO" is Gene Ontology.

Type	**Verifier**	**Description**	**Example Prompts**	**Rewards**	**Source**
HARD	EXP	PerturbQA training data	*Is a knockdown of AARS in hepg2 cells likely to result in differential expression of ATAD2B?*	Binary: 1 for correct, 0 for incorrect; $r_{binary} \in \{0, 1\}$	Experimental data
SOFT	MLP	MLPs trained on perturbation data using gene embeddings	*Is a knockdown of AARS in hepg2 cells likely to result in differential expression of ATAD2B?*	Prediction score from MLP; $r_{soft} = p$ where $0 \le p \le 1$ and $p = \text{MLP}(\text{gene}_A, \text{gene}_B)$	VCM
SOFT	TF	Transcriptformer Foundation Model	*If transcription factor ADNP is activated, is expression of gene RABGAP1 going to be high? The answer is either yes or no.*	Pointwise Mutual Information (PMI) scores: $r_{soft} = p$ where $0 \le p \le 1$ and $p = \text{TF}(\text{gene}_A, \text{gene}_B)$	VCM
SOFT	GO	Knowledge Database on Genes	*Is a knockdown of AARS in hepg2 cells likely to result in differential expression of ATAD2B?*	ROUGE scores, keywords mentions of GO annotations, likelihood estimation $p = \text{GO}(\text{gene}_A, \text{gene}_B)$	Knowledge Base

How the training works


Prompt q: geneA geneB cell_line
Reasoning LLM pi_theta
Generate N completions o1 to oN
Extract outputs: hard answer or text
Verifiers
Hard verifier 2.2 Experimental data
Soft verifier 2.3 Virtual Cell Model
Prior verifier 2.4 Knowledge Source
Reward r_hard binary
Reward r_soft probability
Reward r_soft text similarity
Compose rewards sum r_j
Normalize scores to 0-1
Compute advantages GRPO
Reference model pi_ref
KL penalty
GRPO update step

Verifiers for rbio1

Verifier source	**Output form**	**Compared against**	**Reward meaning**	**Strength of truth**	**Why it counts as truth**
Experimental data	Binary label	LLM yes/no	Correctness	⋆ ⋆ ⋆ ⋆ ⋆	Direct observation
Virtual Cell Model	Probability	LLM yes/no	Consistency	⋆ ⋆ ⋆⋆	Learned biological structure
Knowledge Source	Text	LLM text	Alignment	⋆ ⋆ ⋆	Expert consensus

Table. Truth verification signals used in Sections 2.2–2.4

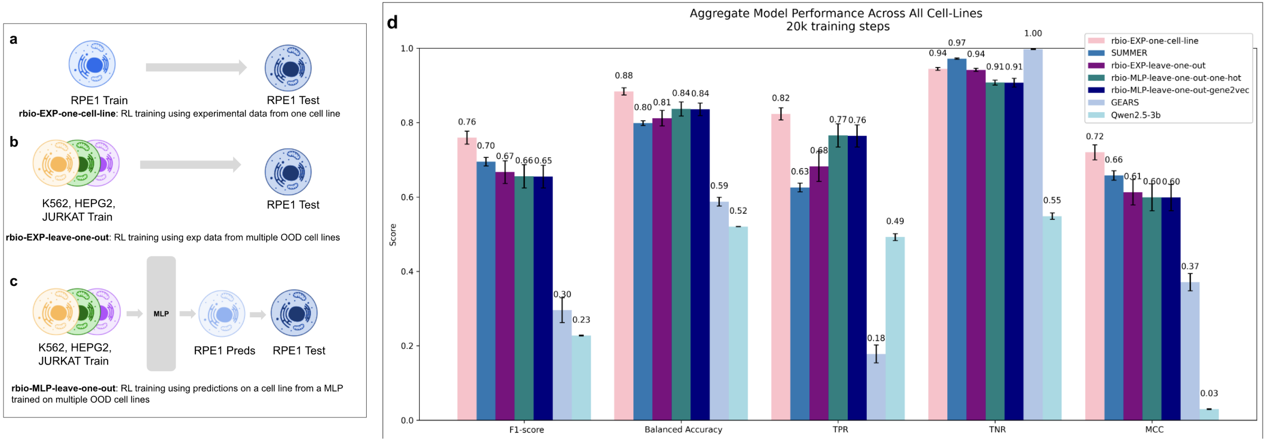
EQUATIONS FOR RBIO1

Mechanism	**Purpose**	**What it stabilizes**	**Failure mode without it**	**One idea / equation**
GRPO (2.1)	Learn from rewards	Policy updates	No learning	$\hat{A}_i = \frac{r_i - \mu}{\sigma}$
Composable rewards (2.5)	Combine signals	Generalization	Overfitting to one truth	$r = \sum_j r_j$
Score normalization (2.6)	Align reward scales	Optimization stability	Reward collapse	Piecewise min–max $\rightarrow [0, 1]$

Metric definitions

$$
\begin{aligned}
Recall\,(TPR) &= \frac{TP}{TP+FN} \qquad (38)\\
TNR &= \frac{TN}{TN+FP} \qquad (39)\\
Precision &= \frac{TP}{TP+FP} \qquad (40)\\
F1\ Score &= \frac{2\cdot Precision\cdot Recall}{Precision+Recall} \qquad (41)\\
Balanced\ Accuracy &= \frac{TPR+TNR}{2} \qquad (42)\\
MCC\ (Matthews\ Correlation\ Coefficient) &= \frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (43)
\end{aligned}
$$

Experimental setup



rbio-EXP-one-cell-line: RL training using experimental data from one cell line

rbio-EXP-leave-one-out: RL training using exp data from multiple OOD cell lines

rbio-MLP-leave-one-out: RL training using predictions on a cell line from a MLP trained on multiple OOD cell lines

Paper Roadmap by Figure

Figure	Sec.	Core question	Training detail (specific point)	Purpose	Conclusion
Fig. 2	3.3	Can soft verification (MLP) match training on experimental labels on-task?	Train rbio using **MLP soft scores (VCM)** instead of EXP labels; leave-one-out cell line logic	Validate **VCM-as-verifier** for perturbation prediction	Soft-verifier training can be competitive with EXP training (esp. recall/TPR tradeoffs)
Fig. 3	3.4	Can off-task VCM biology transfer to perturbation prediction?	Train on **Transcriptformer co-expression** / **PMI** prompts; test on perturbation	Show **transfer** from general biology signals	Off-task biological tutoring improves downstream perturbation behavior
Fig. 4	3.5	Do multiple verifiers improve results vs. single verifiers?	Train with **combinations of TF + GO + MLP** (optionally ESM); prompt mix skewed by dataset density	Demonstrate **composable verification** at scale	Adding diverse verification sources improves generalization/performance
Fig. 5	3.5	What happens in **ablation-style additions**? Any surprises?	Stepwise addition of verifiers; some cases where GO hurts in certain combos	Deepen **composition analysis** beyond Fig. 4	Generally additive gains, with some counterintuitive interactions flagged
Fig. 6	3.6	Does chain-of-thought prompting at inference boost test performance?	**No retraining**; only test-time prompting style changes	Show reasoning helps at inference	CoT improves performance without tools or extra data

Conclusion: What This Paper Shows

- **Biological reasoning can be learned without experiments**
  - Soft verification from models and knowledge is sufficient
  - Experimental data is no longer the only training signal
- **Supervision need not be binary ground truth**
  - Experiments provide hard labels
  - Models and knowledge provide soft, graded signals
  - All can guide learning via verification
- **Compositional verification works**
  - Multiple weak signals outperform any single source
  - Biological supervision is additive, not exclusive
- **Reasoning matters at training and inference**
  - Reinforcement learning aligns reasoning with biology
  - Chain-of-thought improves test-time performance

Rebuttal Overview: Main Criticism Themes

- **Verifier fidelity & noise:** are soft verifiers reliable or do they amplify errors?
- **Generality:** is evidence limited to PerturbQA and closely-coupled rewards?
- **RL vs supervised baseline:** is GRPO actually needed beyond imitation?
- **Clarity & reproducibility:** missing MLP details, unclear equations, missing compute.
- **Metrics/statistical claims:** what is "SOTA" under class imbalance and data fraction?

Author strategy (from official rebuttal):

- "*new controlled-noise, ablation, composition, and cross-domain experiments*"
- "*detailed training, scaling, and reproducibility analyses*" (all in SI)

Reviewer cju1: Verifier Fidelity & Miscalibration

Reviewer cju1 (Weaknesses, verbatim):

> *The approach assumes biology models and GO signals are accurate enough to guide RL, but the paper does not analyze sensitivity to verifier fidelity or miscalibration...*
>
> *There is no analysis of whether RL amplifies verifier errors, no plot of verifier vs model accuracy on held-out perturbations, and no check of verifier disagreement.*

Authors (Response, verbatim highlights):

> *...controlled-noise experiments ... where MLP verifier predictions were progressively randomized... rbio performance decreases smoothly and remains well above the Qwen 2.5-3B baseline until signals are fully random...*
>
> *...confidence–performance analysis ... confirms that recall increases with verifier confidence... indicating that rbio learns a stable biological prior rather than overfitting to verifier noise.*

What they added (SI): controlled-noise (A.4.1), reward ablations (A.4.2), confidence analysis (A.4.3), cross-verifier agreement (A.5).

Takeaway for students: When criticized for “oracle quality”, add *controlled-noise* + *error amplification checks*.

Reviewer cju1: Are Gains Just Generic RL?

Reviewer cju1 (Weakness, verbatim):

> *It is unclear whether generic RL or RLAIF signals (format, helpfulness, self-consistency) would yield similar gains. This weakens the claim that the biology-specific rewards are the true driver of improvement.*

Authors (Response, verbatim highlights):

> *Reward ablations ... show that biological-answer rewards dominate performance gains, while format- or mention-only terms have minimal effect...*
>
> *...we compare ... DeepSeek R1, Qwen Instruct, and OpenAI OSS ... these models fail to reproduce the gains achieved by biology-grounded rewards...*

Takeaway for students: If reviewers suspect "it's just RL regularization", do:

- ablation: domain reward vs format/mention rewards
- stronger external baselines: instruction-tuned / RL reasoning models

Reviewer jVCL: Clarity, Biological Setup, and Ill-Defined Equations

Reviewer jVCL (Weaknesses, verbatim):

> *The exposition needs to be improved strongly: the biological question and setup is not explained...*
>
> *Key details are not included in the paper: how was the tiny MLP trained (input / output / loss?)...*
>
> *Equations are often nonsensical / unhelpful / ill-defined... The paper cannot be reimplemented...*

Authors (Response, verbatim highlights):

> *We now explicitly describe the PerturbQA benchmark in A.1...*
>
> *We added full architecture and hyperparameters ... and ... describe how the MLP's probabilistic outputs serve as rewards during GRPO optimization.*
>
> *We ... re-worked the Methods section for clarity—simplifying notation, improving equation definitions...*

Takeaway for students: Clarity rebuttals must add *reimplementation-critical details*: dataset protocol, model IO/loss, algorithms, and simplified notation.

Reviewer jVCL: Generalization Beyond PerturbQA

Reviewer jVCL (Weakness, verbatim):

> *All experiments are within PERTURBQA... Even a smaller secondary biology reasoning task would make the generality claim more credible.*

Authors (Response, verbatim highlights):

> *...zero-shot disease-state prediction ... shows rbio nearly matches SCVI ... and clearly outperforms Qwen2.5-3B...*
>
> *...demonstrating biological reasoning transfer beyond perturbation tasks.*

Takeaway for students: To rebut “narrow benchmark” critiques:

- add a **cross-domain** experiment (even if smaller)
- make it **zero-shot** to strengthen the generalization claim

Reviewer tnsY: Reproducibility and Code Release

Reviewer tnsY (verbatim):

> *I have some concerns regarding the reproducibility of the results given that the code is not released...*
>
> *...it is the release of the codebase that will lead me to increase my score.*

Authors (Response, verbatim highlights):

> *We have released an anonymous repository ... that reproduces the MLP-verifier experiments and provides an end-to-end example...*
>
> *...clarified the MLP architecture (activation functions, losses, and hyperparameters)...*

Takeaway for students: A reproducibility-focused reviewer can be "won" by:

- minimal end-to-end repo
- architecture + hyperparameters
- algorithms and reward computation details

Reviewer tnsY: Metrics, Significance, and “SOTA” Claims

Reviewer tnsY (verbatim):

> *...claim of state-of-the-art performance ... when the models are trained with 1/5 of the data sample... SUMMER appears to perform better on the TNR metric?*
>
> *...difference in F1-score and the MCC metric does not appear to be statistically significant?*

Authors (Response, verbatim highlights):

> *The PerturbQA datasets are class-imbalanced... identifying true positive perturbations is biologically more important... We therefore emphasize F1, Balanced Accuracy, and MCC...*
>
> *...SUMMER’s higher TNR reflects a different recall–specificity trade-off rather than superior overall accuracy.*

Takeaway for students: When “SOTA” is questioned, rebut by:

- explicitly arguing **which metric matches the scientific objective**
- explaining trade-offs (TPR vs TNR), not just headline numbers

Reviewer AHen: Compute, Scaling, and Baseline Interpretation

Reviewer AHen (verbatim):

> *...computational resources required are not reported...*
>
> *GEARS performs surprisingly poorly... Is there any explanation or analysis regarding this?*

Authors (Response, verbatim highlights):

> *...Training used ... on 8×H100 GPUs for ∼10 days... batch_size=4, learning_rate=5e-6...*
>
> *GEARS's lower F1 arises from an extreme TNR bias—its high specificity (0.997) is offset by low recall...*

Takeaway for students: Always report:

- compute budget + steps + key hyperparameters
- learning curves / scaling plots
- explain surprising baselines (metric mismatch, thresholding effects)

Meta: Why This Rebuttal Worked (Writing Guidance)

- They **matched criticisms with new experiments** (noise, ablations, agreement, cross-domain)
- They **moved from claims to diagnostics** (confidence vs performance, verifier coherence)
- They **fixed reproducibility directly** (repo + algorithms + hyperparams)
- They **clarified evaluation philosophy** (imbalance, objective-aligned metrics)
- They added **stronger baselines** to rule out “generic RL” explanations

Teaching takeaway:

- A strong rebuttal is not persuasion—it is **additional evidence** and **reduced ambiguity**.