

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 ENHANCING BIOLOGICAL REASONING IN LLMs VIA SYNTHETIC REASONING TRACES FOR CELLULAR PERTURBATION PREDICTION

Anonymous authors

Paper under double-blind review

## ABSTRACT

Predicting cellular responses to genetic perturbations represents a fundamental challenge in systems biology, critical for advancing therapeutic discovery and virtual cell modeling. While large language models (LLMs) show promise for biological reasoning, their application to perturbation prediction remains underexplored due to challenges in adapting them to structured experimental data. We present SynthPert, a novel method that enhances LLM performance through supervised fine-tuning on synthetic reasoning traces generated by frontier models. Using the PerturbQA benchmark, we demonstrate that our approach not only achieves state-of-the-art performance but surpasses the capabilities of the frontier model that generated the training data. Our results reveal three key insights: (1) Synthetic reasoning traces effectively distill biological knowledge even when partially inaccurate, (2) This approach enables cross-cell-type generalization with 87% accuracy on unseen RPE1 cells, and (3) Performance gains persist despite using only 2% of quality-filtered training data. This work shows the effectiveness of synthetic reasoning distillation for enhancing domain-specific reasoning in LLMs.

## 1 INTRODUCTION

Accurately predicting cellular responses to genetic perturbations represents a fundamental challenge in systems biology, with critical applications from drug discovery to virtual cell modeling (Bunne et al., 2024). While deep learning approaches like GEARS (Roohani et al., 2024a) and scGPT (Cui et al., 2024) have advanced perturbation prediction, they struggle with generalization to unseen biological contexts—a key requirement for real-world applicability. Recent breakthroughs in large language models (LLMs) offer new paradigm through their ability to reason over biological knowledge, yet their potential remains underexploited for this task.

The rapid evolution of LLM reasoning capabilities, driven by architectural innovations (Guo et al., 2025) and techniques like chain-of-thought prompting (Wei et al., 2022), has enabled novel scientific applications. Projects like AI Co-Scientist (Gottweis et al., 2024) demonstrate LLMs’ ability to assist hypothesis generation, while benchmarks like LAB-Bench (Laurent et al., 2024) assess practical biological reasoning. However, existing evaluations focus on literature analysis and experimental design, neglecting the core challenge of predicting intervention outcomes in complex cellular systems.

PerturbQA (Wu et al., 2024) addresses this gap by reformulating perturbation experiments into natural language tuples (cell type, perturbation, gene)  $\rightarrow$  {up, down, unperturbed}. While this three-class formulation simplifies the complexity of gene expression dynamics, it provides a practical framework grounded in statistical hypothesis testing for actionable biological predictions. However, current approaches simplify the task through problematic decomposition: first predicting differential expression, then directionality only for perturbed genes. This contrasts with real-world scenarios where researchers lack prior knowledge of perturbation effects, requiring joint three-class prediction.

Our work makes three key contributions:

- **Synthetic Reasoning Distillation:** A novel method enhancing LLMs through fine-tuning on generated chain-of-thought explanations rather than raw experimental data

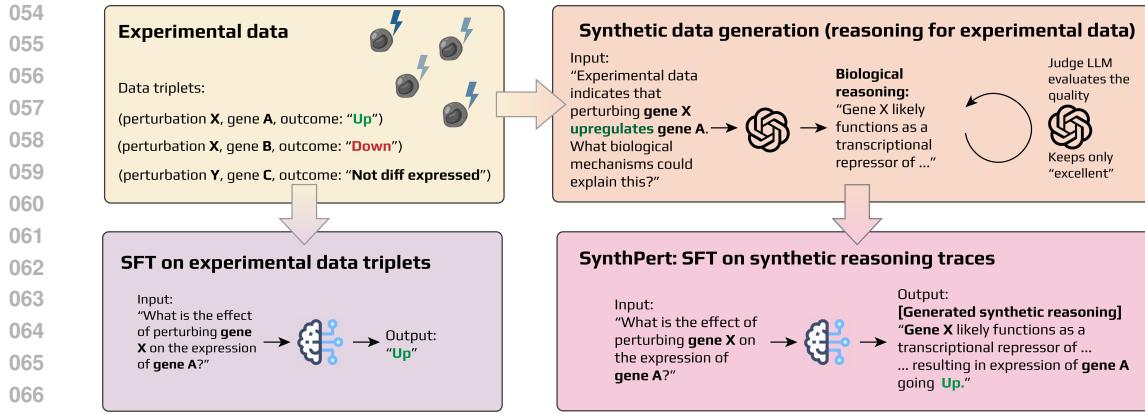


Figure 1: Illustration of the SynthPert workflow. Given experimental perturbation data in the form of “(perturbation, gene, outcome)” data tuples (top left panel), our goal is to create an LLM capable of predicting responses to unseen perturbations. We consider two strategies: (i) SFT on experimental data directly (bottom left panel), and (ii) a synthetic chain-of-thought based supervised fine-tuning. The arrows between panels indicate information flow. In particular, the latter involves experimental data *indirectly*, in the process of creating synthetic reasoning traces for given data tuples, using a frontier LLM. A separate judge LLM evaluates their quality, and keeps only those synthetic explanations that were graded “excellent”. Finally, we fine-tune the base LLM on the generated chain-of-thought explanations.

- **Generalizable Prediction:** State-of-the-art performance (78% AUROC) with successful cross-cell transfer (87% accuracy on unseen RPE1 lineages)
- **Practical Task Formulation:** Direct three-class prediction without artificial task decomposition, better matching biological use cases

Central to our approach is the insight that reasoning structure, rather than factual accuracy, drives biological generalization. By filtering synthetic explanations from frontier models (o4-mini) through quality critics, we enable smaller LLMs to surpass teacher models’ performance—a distillation paradox revealing untapped potential in pretrained biological schemas. Our results challenge conventional wisdom through three findings: (1) Structured explanations improve minority class prediction despite data imbalance, (2) External biological databases degrade performance when forced into reasoning paths, and (3) Practical three-class prediction outperforms decomposed binary tasks, better matching real-world use cases.

We evaluate our approach on the PerturbQA benchmark, demonstrating state-of-the-art performance and strong generalization to unseen cell types. Our analysis reveals insights about the role of reasoning structure in biological prediction tasks and establishes synthetic reasoning distillation as an effective approach for domain-specific LLM enhancement.

## 2 BACKGROUND

### 2.1 CLASSICAL DEEP LEARNING APPROACHES FOR CELLULAR PERTURBATION PREDICTION

Predicting cellular perturbation outcomes has been an active area of method development to infer interventional distributions over gene expression vectors (Lotfollahi et al., 2023; Lopez et al., 2023; Gebregziabher et al., 2024). Of particular interest, is to predict *unseen* pert outcomes, and this requires introduction of some notion of prior knowledge.

**GEARS** Roohani et al. (2024b) proposed GEARS, a graph neural network that incorporates structured biological prior knowledge through gene co-expression graphs and gene ontology networks. This enables generalization to perturbations involving genes absent from training data by exploiting connectivity patterns between seen and unseen genes.

108 **Single-cell foundation models** Single-cell transformer models such as scGPT (Cui et al., 2024)  
 109 take a different approach by pre-training on large-scale gene expression atlases and tokenizing  
 110 gene expression vectors. Despite their sophisticated architecture and a large number of parameters,  
 111 these models often struggle to outperform much simpler baselines for perturbation prediction tasks  
 112 (Ahmann-Eltze & Theis, 2024; Kernfeld et al., 2024), suggesting limitations in their ability to model  
 113 perturbation data effectively.

114

## 115 2.2 LLM-BASED APPROACHES FOR CELLULAR PERTURBATION BIOLOGY

116

117 More recently, large language models have emerged as a promising alternative for modeling perturba-  
 118 tion biology, offering unique capabilities to incorporate biological knowledge and reasoning.

119

120 **LLM-derived gene embeddings** GenePT (Chen & Zou, 2024) proposed an approach that leverages  
 121 LLMs to embed gene or protein summaries from scientific literature and using these embeddings in  
 122 downstream modelling, such as for creating cell embeddings from gene expression data. By generating  
 123 gene embeddings from NCBI text descriptions, GenePT captures rich semantic information about  
 124 gene functions and relationships. These embeddings have shown relatively strong performance when  
 125 combined with ML models such as Gaussian Processes for perturbation prediction (Märtens et al.,  
 126 2024), demonstrating that LLMs can indeed effectively encode biologically relevant prior knowledge  
 127 that generalizes to unseen perturbations.

128

129 **LLM-based reasoning frameworks** SUMMER (Wu et al., 2024) introduced a retrieval-augmented  
 130 generation (RAG) framework for perturbation biology, employing LLMs to: (1) summarize gene  
 131 descriptions, (2) retrieve similar seen perturbations via knowledge graph proximity, and (3) reason  
 132 through structured prompting. This approach outperformed previous methods on the PerturbQA  
 133 benchmark without fine-tuning, thus providing a natural framework we build upon.

134

135 However, SUMMER notably simplified the problem by converting the natural three-way classification  
 136 task into two separate binary classifications. First, they classified genes as differentially expressed or  
 137 not, then predicted directionality (“up” or “down”) only for the subset of genes already identified  
 138 as differentially expressed. While this decomposition proved effective, in practical applications  
 139 researchers typically don’t have advance knowledge of which genes will be differentially expressed,  
 140 making it easier to work with a 3-class prediction model.

141

142 **Limitations of current LLM approaches** While showing promise, current LLM-based methods  
 143 primarily operate at inference time rather than systematically learning biological reasoning patterns.  
 144 GenePT uses LLMs solely as embedding generators, while SUMMER relies on a retrieval-based  
 145 approach that may struggle with perturbation scenarios that lack similar examples in its retrieval  
 146 database. Neither approach fully leverages the reasoning capabilities of LLMs through targeted  
 147 fine-tuning on perturbation biology tasks. This suggests an opportunity to develop models that can  
 148 learn from perturbation data directly while maintaining the interpretability and knowledge integration  
 149 advantages of LLMs—the direction we pursue in this work.

150

## 151 3 METHOD: SYNTHPERT

152

153 In this work, we develop an LLM-based perturbation prediction model that, given a *cell type*,  
 154 *perturbation*, and *gene of interest*, predicts whether the gene is up-regulated, down-regulated, or not  
 155 differentially expressed. Formally, we learn the mapping:

156

$$(cell\ type, perturbation, gene) \mapsto \{\text{“up”}, \text{“down”}, \text{“not differentially expressed”}\}$$

157

158 parameterized by an LLM. Although predictions are made for individual genes, our approach enables  
 159 genome-wide in silico screening through systematic application across all genes.

160

161 Unlike Wu et al. (2024), who predicted directionality (i.e., “up” or “down”) only for genes pre-  
 162 identified as differentially expressed, our approach directly classifies all three possible states simul-  
 163 taneously. This comprehensive formulation better matches real-world biological workflows where  
 164 researchers lack prior knowledge of perturbation effects.

162 3.1 SYNTHETIC TRAINING STRATEGIES  
163164 While existing work uses pretrained LLMs without modification, we explore two complementary  
165 fine-tuning approaches:  
166167 **SFT on Data Tuples (Baseline)** The straightforward approach involves direct supervised fine-  
168 tuning (SFT) on observed (input, output) data tuples. While we consider this primarily as a baseline,  
169 to our knowledge this represents the first application of SFT to perturbation prediction. This method  
170 trains the model to associate inputs with outcomes without explanatory context.  
171172 **SFT on Synthetic Explanations (SynthPert)** Our novel strategy focuses on enhancing biological  
173 reasoning through synthetic data generation. Rather than training directly on experimental tuples, we  
174 use a frontier model to generate mechanistic explanations for observed outcomes, then fine-tune on  
175 these reasoning traces. This indirect approach teaches causal relationships rather than surface-level  
176 associations.  
177178 3.2 SYNTHETIC CHAIN-OF-THOUGHT GENERATION  
179180 As illustrated in Figure 1, our workflow generates biological explanations through two complementary  
181 strategies:  
182183 **Approach 1 (Prediction + Explanation):** The frontier model (OpenAI o4-mini) receives only the cell  
184 type, perturbation, and gene name, requiring it to both predict the outcome and generate supporting  
185 reasoning. We retain only traces where the model correctly predicted the outcome.  
186187 **Approach 2 (Explanation from Outcome):** Here we provide the ground truth outcome alongside  
188 the input tuple, tasking the model with generating a mechanistic rationale. A separate critic model  
189 (also a frontier model) grades explanations on a 5-point scale: ‘excellent’, ‘good’, ‘average’, ‘bad’, or  
190 ‘terrible’. We retain only ‘excellent’ graded explanations.  
191192 Both approaches allow the model to respond with “I don’t know” to reduce hallucinations. Approach 2 produced higher quality explanations (see Supplementary Material) and forms our final  
193 implementation.  
194195 3.3 SUPERVISED FINETUNING IMPLEMENTATION  
196197 We perform SFT using Low-Rank Adaptation (LoRA) (Hu et al., 2022) on a DeepSeek-R1 8B model  
198 (Guo et al., 2025). Despite generating traces for only a subset of training data initially – filtered to a  
199 small fraction via quality control – SYNTHPERT achieves superior performance to models trained on  
200 full datasets, demonstrating remarkable data efficiency.  
201202 For rigorous comparison, we implement a LLM + SFT ON DATA baseline using identical hyperpa-  
203 rameters but training directly on perturbation tuples without explanations. This controlled experiment  
204 isolates the value added by synthetic reasoning traces versus mere exposure to experimental data.  
205206 4 RESULTS  
207208 4.1 EXPERIMENTAL SETUP  
209210 **Datasets** For evaluation, we use the PerturbQA benchmark released by Wu et al. (2024). This is a  
211 relatively comprehensive data source, spanning Perturb-seq experiments across four cell lines: K562,  
212 RPE1, HepG2, and Jurkat, derived from data by Replogle et al. (2022) and Nadig et al. (2024).  
213214 In PerturbQA, for every cell line, we have a number of tuples “(*perturbation, gene, outcome*)”.  
215 While the outcome label is one of {“up”, “down”, “not differentially expressed”}, the benchmark  
216 separates two tasks: 1) differential expression, and 2) direction of change, where the latter task  
217 has been subsetted to those data tuples that are indeed differentially expressed. In total, the dataset  
218 contains  $N = 84,550$  examples for the direction of change task and  $N = 614,479$  for the differential  
219 expression task. These have been split into a 75% train and 25% test sets. A detailed breakdown by  
220 cell line is available in Table 5 in the Appendix.  
221

216 **Implementation details** For all SYNTHPERT results reported in this paper, we supervised fine-tuned  
 217 on synthetic data created via the critic-based method (our second approach described in Section 3.1,  
 218 and as illustrated in Figure 1), as we saw similar results across both methods but this approach is  
 219 more data efficient as the critic filters out low-quality explanations before fine-tuning.  
 220

221 **Baselines** Following the PerturbQA original data splits, we compare our “LLM + SFT ON DATA”  
 222 baseline and SYNTHPERT with the baselines considered by Wu et al. (2024). Specifically, we include  
 223 PHYSICAL, which naively predicts differential expression based on known physical interactions  
 224 between genes from STRINGDB (Szklarczyk et al., 2021), and GAT (Veličković et al., 2018), a  
 225 graph attention network trained on biological knowledge graphs with a ternary classification objective.  
 226 We also compare against state-of-the-art methods previously described in the background section:  
 227 GEARS (Roohani et al., 2024a), scGPT (Cui et al., 2024), and GenePT (Chen & Zou, 2024).  
 228 Additionally, we evaluate against SUMMER (Wu et al., 2024), a RAG-based approach that leverages  
 229 knowledge graphs and LLM reasoning to predict perturbation effects in biological systems.  
 230

231 **Evaluation Protocol** We evaluate our models using two distinct protocols. First, we follow the  
 232 original PerturbQA setup as proposed by Wu et al. (2024), adhering to their data splits and AUROC  
 233 metric for the two separate tasks (differential expression and direction of change), as detailed in  
 234 Section 4.2.

235 Second, we introduce a more unified evaluation approach that directly addresses the three-class  
 236 classification problem {“up”, “down”, “not differentially expressed”}, which better aligns with our  
 237 model’s design. For this three-class evaluation, we report precision, recall, and F1 scores for each  
 238 class, along with overall accuracy, providing a more comprehensive assessment of model performance  
 239 across the imbalanced class distribution. This approach allows for a more comprehensive assessment  
 240 of SYNTHPERT’s performance relative to both the base LLM and existing methods.  
 241

242 Furthermore, alongside the original PerturbQA split, we propose an additional evaluation protocol  
 243 where an entire cell type (RPE1) is held out as a test set. This cross-cell-type evaluation provides a  
 244 more rigorous assessment of generalization capabilities across biological contexts, testing whether  
 245 models can transfer knowledge to previously unseen cellular environments, as detailed in Section 4.3.  
 246

247 **Baseline Rationale** The motivation for including the “LLM + SFT ON DATA” baseline is two-fold.  
 248 First, we aim to quantify the utility of our synthetic chain-of-thought data relative to the information  
 249 content directly present in the observed experimental data. Second, when treated as a three-class  
 250 problem, the data distribution is highly imbalanced—the “not differentially expressed” category  
 251 significantly dominates the other two classes. Therefore, we needed to ensure that SYNTHPERT  
 252 is learning meaningful biological relationships and not simply memorizing the underlying data  
 253 distribution. The “LLM + SFT ON DATA” baseline, trained directly on the experimental data,  
 254 provides a control that would demonstrate performance achievable through distribution learning  
 255 alone.  
 256

257 In Section 4.3, we also include the frontier model OpenAI o4-mini as an additional baseline. This  
 258 is the same model we used for generating synthetic chain-of-thought traces, providing a direct  
 259 comparison between the teacher model and our fine-tuned LLM. This comparison is particularly  
 260 meaningful as it helps evaluate whether our approach can effectively distill and enhance the biological  
 261 reasoning capabilities of the frontier model. Generally, one would not expect a much smaller (8  
 262 billion parameter) model to match or exceed the capabilities of a large frontier model on complex  
 263 reasoning tasks. However, this comparison allows us to test the hypothesis that targeted fine-tuning on  
 264 high-quality synthetic reasoning chains might enable more efficient utilization of parameter capacity  
 265 for specific biological reasoning tasks. The comparison also examines whether frontier models  
 266 contain implicit knowledge about biological mechanisms that can be extracted and refined through  
 267 our synthetic data approach, potentially allowing a smaller model to achieve strong performance on  
 268 this specialized task.  
 269

270 **Model Configuration** In all experiments, we use a DeepSeek-R1 distilled Llama 3.1 8B model  
 271 (Team et al., 2024; Guo et al., 2025). Importantly, we fine-tune the model only once and evaluate it  
 272 on both tasks (differential expression and directionality prediction), demonstrating the versatility of  
 273 our approach.  
 274

270 4.2 DIFFERENTIAL EXPRESSION AND DIRECTION OF CHANGE  
271

272 Table 1 presents the comparison of AUROC values on the PerturbQA Differential Expression task  
273 across all four cell lines. While SUMMER achieves the highest performance among existing methods,  
274 SYNTHPERT substantially and significantly outperforms it across all cell lines. Notably, SYNTHPERT  
275 demonstrates remarkable improvements on the RPE1 cell line (AUROC increase from 0.58 to 0.78)  
276 and Jurkat cell line (increase from 0.58 to 0.79).

277 Table 1: Results on Differential Expression as binary prediction, following the original PerturbQA  
278 evaluation (Wu et al., 2024). AUROC is computed over the predictions associated with each gene,  
279 and averaged over perturbations.

282 Task	283 Model	284 K562	285 RPE1	286 HepG2	287 Jurkat
288 Differential 289 expression	290 PHYSICAL	291 0.53	292 0.52	293 0.52	294 0.54
	295 GAT	296 0.55	297 0.57	298 0.57	299 0.55
	300 GEARS	301 0.54	302 0.50	303 0.48	304 0.51
	305 SCGPT	306 0.52	307 0.52	308 0.48	309 0.51
	310 GENEPT-GENE	311 0.57	312 0.54	313 0.55	314 0.55
	315 GENEPT-PROT	316 0.57	317 0.56	318 0.54	319 0.55
	320 SUMMER	321 0.60	322 0.58	323 0.61	324 0.58
325 LLM + SFT ON DATA		326 0.59	327 0.59	328 0.60	329 0.60
330 SYNTHPERT		331 <b>0.70</b>	332 <b>0.78</b>	333 <b>0.74</b>	334 <b>0.79</b>

293 Importantly, our “LLM + SFT ON DATA” baseline performs significantly worse than SYNTHPERT,  
294 confirming that naive fine-tuning with the training data alone would not have achieved these perfor-  
295 mance gains. This comparison clearly demonstrates the added value of our synthetic chain-of-thought  
296 generation workflow in enhancing the model’s biological reasoning capabilities.

297 Table 2 presents results for the Direction of Change prediction task. SYNTHPERT demonstrates  
298 strong performance on this task as well, despite being fine-tuned on three-class reasoning traces.  
299 For this evaluation, we simply adjust the prompt to inform the model that the gene is differentially  
300 expressed and ask it to predict one of the two directional labels. SYNTHPERT outperforms SUMMER  
301 on three out of four cell lines, while performing comparably on the fourth (Jurkat). Consistent with  
302 our previous findings, the “LLM + SFT ON DATA” baseline performs substantially worse across all  
303 four cell lines, further confirming that extracting meaningful patterns from perturbation data requires  
304 more sophisticated approaches than direct fine-tuning.

305 Table 2: Results on Direction of Change as binary prediction, following the original PerturbQA  
306 evaluation (Wu et al., 2024). AUROC is computed over the predictions associated with each gene,  
307 and averaged over perturbations.

310 Task	311 Model	312 K562	313 RPE1	314 HepG2	315 Jurkat
316 Direction 317 of change	318 GAT	319 0.58	320 0.60	321 0.64	322 0.59
	323 GEARS	324 0.64	325 0.60	326 0.52	327 0.51
	328 SCGPT	329 0.48	330 0.53	331 0.51	332 0.51
	333 GENEPT-GENE	334 0.53	335 0.57	336 0.58	337 0.57
	338 GENEPT-PROT	339 0.57	340 0.57	341 0.55	342 0.58
	343 SUMMER	344 0.62	345 0.64	346 0.65	347 <b>0.66</b>
	348 LLM + SFT ON DATA	349 0.47	350 0.55	351 0.55	352 0.50
353 SYNTHPERT		354 <b>0.65</b>	355 <b>0.73</b>	356 <b>0.72</b>	357 0.65

320 4.3 DIRECT THREE-CLASS PERTURBATION EFFECT PREDICTION  
321

322 Having established the strong performance of SYNTHPERT on the individual binary tasks in the  
323 PerturbQA benchmark, we now evaluate its capabilities on the more challenging and biologically

relevant direct three-class prediction problem. Tables 3 and 4 present detailed performance metrics across all three possible gene expression states.

Table 3 shows that SYNTHPERT significantly outperforms all baselines in terms of overall accuracy (89% versus 15% for the base LLM and 52% for both the SFT baseline and o4-mini). Notably, SYNTHPERT substantially outperforms o4-mini—the very model used to generate its training data—demonstrating that our synthetic fine-tuning approach effectively distills and refines the biological reasoning capabilities of the larger model. While o4-mini achieves the highest recall for upregulated genes (0.62 compared to SYNTHPERT’s 0.14), it suffers from poor precision (0.12), suggesting it tends to over-predict this class. In contrast, SYNTHPERT shows more balanced performance with significantly higher precision across all classes.

Interestingly, SYNTHPERT shows a different precision-recall trade-off compared to the base LLM for upregulated genes. SYNTHPERT achieves the highest precision (0.49 compared to only 0.07 for the base LLM) but lower recall (0.14) compared to both the base LLM (0.32) and o4-mini (0.62). This suggests that SYNTHPERT has learned to be more selective in its predictions, only identifying the most confident cases of upregulation. In contrast, the other models appear to overpredict upregulation, capturing more true positives but at the cost of numerous false positives—a pattern consistent with insufficient understanding of the biological mechanisms governing gene upregulation. Despite this trade-off, SYNTHPERT still achieves a better F1 score for upregulated genes than the base LLM (0.22 vs 0.11), indicating that its precision gains outweigh its recall losses in this balanced metric.

The dramatic improvement in overall accuracy is primarily driven by SYNTHPERT’s exceptional performance in correctly identifying non-differentially expressed genes (F1 score of 0.95 compared to o4-mini’s 0.68 and base LLM’s 0.21), which constitute the majority class in most perturbation datasets. Beyond the majority class, SYNTHPERT also demonstrates substantially better performance on downregulated genes, with an F1 score of 0.53—nearly triple the performance of both o4-mini and the SFT baseline (0.18). This suggests our approach not only distills knowledge from the teacher model but actually enhances it through the focused learning on high-quality synthetic reasoning examples.

Table 3: Performance comparison on direct three-class perturbation effect prediction. All perturbations evaluated are unseen during training. Results show precision, recall, and F1 scores for each class (Not Differentially Expressed, Upregulated, and Downregulated), along with overall accuracy across all models.

Model	Not Diff. Expressed			UP Regulated			DOWN Regulated			Accuracy
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	
Base LLM	0.87	0.12	0.21	0.07	0.32	0.11	0.08	0.38	0.14	0.15
SFT on data	0.90	0.55	0.69	0.08	0.29	0.13	0.12	0.35	0.18	0.52
o4-mini	<b>0.92</b>	0.54	0.68	0.12	<b>0.62</b>	0.20	0.15	0.23	0.18	0.52
SYNTHPERT	<b>0.92</b>	<b>0.97</b>	<b>0.95</b>	<b>0.49</b>	0.14	<b>0.22</b>	<b>0.52</b>	<b>0.53</b>	<b>0.53</b>	<b>0.89</b>

Table 4: Cross-cell-type generalization performance with RPE1 cell line held out. Models are fine-tuned exclusively on data from HepG2, Jurkat, and K562 cell lines, then evaluated on the unseen RPE1 cell line to assess transfer learning capabilities across biological contexts

Model	Not Diff. Expressed			UP Regulated			DOWN Regulated			Accuracy
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	
Base LLM	0.87	0.12	0.22	0.05	<b>0.29</b>	0.09	0.09	0.34	0.14	0.15
SFT on data	0.87	0.43	0.58	0.06	0.24	0.09	0.10	0.19	0.14	0.40
o4-mini	0.91	0.54	0.68	0.09	0.20	0.13	0.13	<b>0.53</b>	0.21	0.52
SYNTHPERT	<b>0.92</b>	<b>0.96</b>	<b>0.94</b>	<b>0.31</b>	0.16	<b>0.21</b>	<b>0.49</b>	0.41	<b>0.45</b>	<b>0.87</b>

Table 4 presents results from our more challenging cross-cell-type generalization experiment, where models trained on HepG2, Jurkat, and K562 cell lines are evaluated on the completely unseen RPE1

378 cell line. SYNTHPERT maintains remarkably strong performance even in this zero-shot transfer setting,  
 379 with only a modest drop in overall accuracy from 89% to 87%. This robust generalization suggests that  
 380 SYNTHPERT has learned transferable biological principles rather than simply memorizing patterns  
 381 specific to individual cell types.

382 Comparing SYNTHPERT to o4-mini in this cross-cell-type scenario reveals particularly interesting  
 383 insights. Despite being fine-tuned on synthetic data generated by o4-mini, SYNTHPERT significantly  
 384 outperforms its teacher model in overall accuracy (87% vs. 52%). This substantial gap demonstrates  
 385 that our synthetic fine-tuning approach not only distills knowledge but enhances generalization  
 386 capabilities across different biological contexts. While o4-mini achieves the highest recall for  
 387 downregulated genes (0.53), surpassing even SYNTHPERT (0.41), its precision is considerably lower  
 388 (0.13 vs. 0.49), resulting in a weaker F1 score (0.21 vs. 0.45). This pattern suggests that o4-mini  
 389 tends to over-predict gene downregulation in novel cellular contexts, while SYNTHPERT makes more  
 390 calibrated predictions.

391 Across all classes, SYNTHPERT demonstrates better balanced performance than the baselines, with  
 392 particularly dramatic improvements in identifying non-differentially expressed genes (F1 score of  
 393 0.94 compared to o4-mini’s 0.68 and the base LLM’s 0.22). For minority classes, SYNTHPERT  
 394 consistently favors precision over recall, making fewer but more reliable predictions—a valuable  
 395 characteristic for practical experimental design.

396 Manual evaluation of reasoning traces by domain experts (detailed in Appendix B) revealed that  
 397 chains-of-thought leading to correct predictions had higher factual accuracy (0.92 vs 0.76 for incorrect  
 398 predictions), and relied more on gene-level functional descriptions rather than high-level pathway  
 399 categorizations. This suggests that SYNTHPERT succeeds by learning mechanistic relationships  
 400 between specific gene functions rather than exploiting superficial biological abstractions.

401 These results collectively demonstrate that SYNTHPERT’s synthetic fine-tuning approach enables  
 402 effective transfer learning across fundamentally different cell types with distinct biological char-  
 403 acteristics. This generalization capability is particularly noteworthy given that RPE1 cells (retinal  
 404 pigment epithelial cells) have significantly different biological functions compared to the training  
 405 cell types (liver cancer cells, leukemia cells, and chronic myelogenous leukemia cells). The model’s  
 406 ability to maintain strong performance across these diverse cellular contexts suggests it has captured  
 407 fundamental regulatory principles that transcend cell-type-specific patterns.

## 410 5 DISCUSSION

411 Our work demonstrates that synthetic reasoning traces enable LLMs to overcome their pretraining  
 412 limitations in biological prediction tasks. SYNTHPERT establishes three key advances in this domain.

413 First, our synthetic reasoning approach proves more effective than direct training on experimental data.  
 414 By fine-tuning on quality-filtered chain-of-thought (CoT) explanations rather than raw experimental  
 415 data, we achieve state-of-the-art PerturbQA performance (78% AUROC) while using only 2% of  
 416 the potential training data. This finding confirms that it’s the structured reasoning patterns that drive  
 417 accurate perturbation prediction. The efficiency gains suggest that targeted reasoning enhancement  
 418 can be more effective than simply increasing dataset size.

419 Second, we demonstrate effective cross-cell-type generalization through synthetic explanations.  
 420 SYNTHPERT’s 87% accuracy on unseen RPE1 cells, compared to 40% for direct supervised fine-  
 421 tuning on experimental data, reveals that synthetic traces encode transferable biological principles  
 422 rather than cell-specific patterns. This suggests the model has learned fundamental mechanisms of  
 423 gene regulation that apply across different cellular contexts.

424 Third, we observe a striking result where SYNTHPERT substantially surpasses its teacher model  
 425 o4-mini (89% vs 52% accuracy) despite being trained on synthetic data generated by that same model.  
 426 This is particularly noteworthy given that the base model initially achieved only 15% accuracy. The  
 427 substantial improvement suggests that targeted fine-tuning on high-quality synthetic reasoning traces  
 428 can unlock latent capabilities in smaller models for domain-specific tasks.

429 This finding supports the Superficial Alignment Hypothesis (Zhou et al., 2023): limited fine-tuning  
 430 data (14k examples) suffices because pretraining on scientific text embeds biological reasoning

432 capabilities that targeted CoT traces activate. This suggests that teaching models to reason through  
 433 biological mechanisms can be highly effective for domain-specific tasks.  
 434

### 435 5.1 LIMITATIONS AND CHALLENGES

437 Three technical limitations emerge from our analysis. First, class imbalance effects create challenges  
 438 similar to those seen in long-tailed classification (ValizadehAslani et al., 2024). The precision-recall  
 439 tradeoff for upregulated genes (0.49 vs 0.14) reflects PerturbQA’s skewed label distribution, where  
 440 85% of genes are non-differentially expressed. However, our LLM + SFT ON DATA baseline, which  
 441 could have benefited from exposure to the class distribution, performed substantially worse than  
 442 SYNTHPERT (52% vs 89% accuracy). To further verify that our results weren’t due to memorizing  
 443 class proportions, we rebalanced the SYNTHPERT training data such that each class represented  
 444 exactly a third of the training data. The resulting model performed roughly the same as the original,  
 445 confirming that performance gains stem from the reasoning structure rather than class distribution  
 446 learning.

447 Second, our attempts to incorporate external biological knowledge were counterproductive. Adding  
 448 EnrichR pathway analysis (Kuleshov et al., 2016) during inference degraded performance ( $\Delta$ AUROC  
 449 = -0.07), likely due to input length saturation (Liu et al., 2023). This suggests that explicit biological  
 450 knowledge injection may interfere with LLMs’ implicit reasoning capacities.

451 Third, validating the biological accuracy of CoT rationales remains challenging. While we provide  
 452 manual verification of reasoning traces for a small sample (Appendix B), this process is extremely  
 453 time-consuming as it requires evaluating each biological claim within the reasoning chain rather than  
 454 just the final prediction. This represents an important consideration for LLMs in scientific discovery  
 455 applications (Ji et al., 2023).

### 457 5.2 FUTURE DIRECTIONS

459 Our findings suggest several promising research directions for enhancing biological reasoning in  
 460 language models. Implementing Reinforcement Learning from Biological Feedback could signifi-  
 461 cantly improve model performance, potentially through approaches like GRPO (Shao et al., 2024)  
 462 with rewards derived from pathway database consistency checks against resources like STRING  
 463 (Szklarczyk et al., 2021). This would allow models to receive direct feedback on the biological  
 464 plausibility of their reasoning chains, potentially reducing hallucinated mechanisms while reinforcing  
 465 scientifically valid explanations.

466 Beyond RL, we see opportunities in Multi-Task Co-Distillation by jointly training models on per-  
 467 turbation prediction and biomedical question answering datasets such as MedQA (Jin et al., 2020),  
 468 enabling knowledge transfer across related biological reasoning tasks. Additionally, the compu-  
 469 tational efficiency of our approach could be further improved by applying recent LoRA variants  
 470 (Hayou et al., 2024) to reduce memory overhead during Chain-of-Thought synthesis, making these  
 471 approaches more accessible to researchers with limited computational resources while potentially  
 472 improving parameter efficiency.

### 474 5.3 CONCLUSION

476 SYNTHPERT establishes synthetic reasoning distillation as an efficient paradigm for biological  
 477 LLMs, achieving state-of-the-art prediction and unprecedented cross-cell generalization. Our results  
 478 challenge the prevailing assumption that biological AI requires massive experimental datasets –  
 479 instead, carefully structured reasoning traces unlock pretrained knowledge through targeted activation.  
 480 For ML practitioners, this work offers a blueprint for domain-specific reasoning enhancement; for  
 481 biologists, a step toward interpretable in silico experimentation.

## 483 REFERENCES

485 Constantin Ahlmann-Eltze and Fabian J Theis. Deep learning for single-cell genomics: models,  
 challenges and opportunities. *Nature Methods*, 21(1):46–57, 2024.

- 486 Christian Bunne, Jacob FV Haim, Simon Mathis, Mohammad Lotfollahi, and Fabian J Theis.  
 487 How to build a virtual cell: A roadmap for ai-powered simulation in biology. *arXiv preprint*  
 488 *arXiv:2403.02165*, 2024.
- 489 Yiqun Chen and James Zou. Genept: a simple but effective foundation model for genes and cells  
 490 built from chatgpt. *bioRxiv*, pp. 2023–10, 2024.
- 492 Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang.  
 493 scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature*  
 494 *Methods*, pp. 1–11, 2024.
- 495 Bereket Gebregziabher, Leon Hetzel, Anna C Schaar, Fabian J Theis, and Francesco Casale. Mod-  
 496 elling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In  
 497 *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- 499 Julian Gottweis, Samuel G Rodrigues, Bo Shopsin, David O’Donovan, David GRG Jones, George M  
 500 Church, and Lucy J Colwell. Towards an ai co-scientist for experimental biology. *arXiv preprint*  
 501 *arXiv:2407.12648*, 2024.
- 502 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
 503 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
 504 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 506 Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models,  
 507 2024. URL <https://arxiv.org/abs/2402.12354>.
- 508 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,  
 509 and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International*  
 510 *Conference on Learning Representations (ICLR)*, 2022.
- 512 Ziwei Ji, Nayeon Lee, and ... Survey of hallucination in natural language generation. *ACM Computing*  
 513 *Surveys*, 2023.
- 514 Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What  
 515 disease does this patient have? A large-scale open domain question answering dataset from medical  
 516 exams. *CoRR*, abs/2009.13081, 2020. URL <https://arxiv.org/abs/2009.13081>.
- 518 Eric Kernfeld, Yunxiao Yang, Joshua S. Weinstock, Alexis Battle, and Patrick Cahan. A sys-  
 519 tematic comparison of computational methods for expression forecasting, October 2024. URL  
 520 <https://www.biorxiv.org/content/10.1101/2023.07.28.551039v2>. Pages:  
 521 2023.07.28.551039 Section: New Results.
- 522 Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan,  
 523 Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann,  
 524 Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma’ayan. Enrichr:  
 525 a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*,  
 526 44(W1):W90–W97, 05 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw377. URL <https://doi.org/10.1093/nar/gkw377>.
- 528 Benjamin Kuznets-Speck, Leon Schwartz, Hanxiao Sun, Madeline E Melzer, Nitu Kumari, Benjamin  
 529 Haley, Ekta Prashnani, Suriyanarayanan Vaikuntanathan, and Yogesh Goyal. Fluctuation structure  
 530 predicts genome-wide perturbation outcomes. *bioRxiv*, pp. 2025–06, 2025.
- 532 C Laurent, NRLZ Anastacio, A Garriga-Alonso, C Bunne, FJ Theis, et al. LAB-Bench: A com-  
 533 prehensive benchmark for language models in biology. *bioRxiv*, pp. 2024–05, 2024.
- 534 Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni,  
 535 and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL  
 536 <https://arxiv.org/abs/2307.03172>.
- 538 Romain Lopez, Mohammad Lotfollahi, L Sole-Boldo, D De Donno, ASRR Al-Rawi, AS Jordan, and  
 539 Fabian J Theis. Learning interoperable representations of single-cell perturbation effects. *Nature*  
*Biotechnology*, 41(6):798–808, 2023.

- 540 Mohammad Lotfollahi, Romain Lopez, F Alexander Wolf, and Fabian J Theis. Predicting cellular  
 541 responses to novel perturbations with generative modeling. *Nature Biotechnology*, 41(6):787–797,  
 542 2023.
- 543
- 544 Kaspar Märtens, Rory Donovan-Maiye, and Jesper Ferkinghoff-Borg. Enhancing generative perturba-  
 545 tion models with llm-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning*  
 546 *for Genomics Explorations*, 2024.
- 547
- 548 Anika Nadig, Joseph M Replogle, Brittnia KYL Chan, Alina Guna, S Adrian Scharenberg, Jeffrey A  
 549 Hussmann, Luke A Gilbert, and Jonathan S Weissman. Transcriptome-wide measurement of  
 550 complex genetic interaction effects in single cells. *Cell*, 187(12):2977–2992, 2024.
- 551
- 552 Joseph M Replogle, Reuben A Saunders, Andrew N Pogson, Jeffrey A Hussmann, Alex Lenail, Alina  
 553 Guna, Lisa Mascibroda, Elana J Wagner, Brittnia KYL Chan, Luke A Gilbert, et al. Mapping  
 554 information-rich genotype–phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):  
 555 2559–2575, 2022.
- 556
- 557 Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel  
 558 multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024a.
- 559
- 560 Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel  
 561 multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024b.
- 562
- 563 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
 564 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of  
 565 mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- 566
- 567 Damian Szkłarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-  
 568 Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. The STRING  
 569 database in 2021: customizable protein–protein networks, and functional characterization of  
 570 user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021.
- 571
- 572 The Llama 3.1 Team, Louis-Philippe Morency, Guillaume Grattafiori, Hakan Celebi, Joanna Lee,  
 573 Maryam Fazel, Nicola Bux, Gido de Jong, Sam Hosseini, et al. The llama 3.1 series of models.  
 574 *arXiv preprint arXiv:2407.19524*, 2024.
- 575
- 576 Taha ValizadehAslani, Yiwen Shi, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and  
 577 Hualou Liang. Two-stage fine-tuning with chatgpt data augmentation for learning class-imbalanced  
 578 data. *Neurocomputing*, 592:127801, 2024. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2024.127801>. URL <https://www.sciencedirect.com/science/article/pii/S0925231224005721>.
- 579
- 580
- 581 Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua  
 582 Bengio. Graph attention networks. In *International Conference on Learning Representations*  
 583 (*ICLR*), 2018.
- 584
- 585 Jason Wei, Yi Tay, Rishi Bommasani, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H.  
 586 Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language  
 587 models. *NeurIPS*, 2022.
- 588
- 589 Zijun Wu, Yusuf Roohani, and Jure Leskovec. Contextualizing perturbation biology with language  
 590 models. *arXiv preprint arXiv:2405.15074*, 2024.
- 591
- 592 Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia  
 593 Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information  
 Processing Systems*, 36:55006–55021, 2023.

594 **A APPENDIX**595 **A DATASET STATISTICS**596 **Table 5: Number of data points across tasks and cell lines in PerturbQA.**

600 <b>Task</b>	<b>HepG2</b>	<b>Jurkat</b>	<b>K562</b>	<b>RPE1</b>	<b>Total</b>
602 Direction of Change	17,860	20,058	19,980	26,652	84,550
603 Differential Expression	126,889	142,822	157,679	187,089	614,479

604 **B BIOLOGICAL FACTUALITY OF CHAINS-OF-THOUGHT**

605 To better understand the utility of SynthPert’s reasoning, we manually evaluated its chain-of-thought  
 606 for a random sample of 10 incorrect and 7 correct predictions. For each prediction, we evaluated the  
 607 factual accuracy of every sentence in the reasoning trace, treating each as a distinct reasoning clause.  
 608 This allowed us to calculate a factuality score for each trace, representing the fraction of correct  
 609 clauses. We found that chains-of-thought leading to correct predictions had an average factuality  
 610 score of 0.92, compared to 0.76 for those leading to incorrect predictions.

611 The o4-mini generated synthetic CoTs created via SynthPert’s synthetic data pipeline (described  
 612 as Approach 2 in section 3.2) summarized gene-level functions quite well for the perturbation and  
 613 target genes. However, when it came to describing the functions of super-pathways (i.e., biological  
 614 processes) to which each gene belonged, the descriptions varied. We found that it was good at  
 615 describing biological processes for well-annotated genes. However, if a gene was not well-annotated,  
 616 it would assume the gap in the literature linking the gene to a pathway was indeed true, as opposed to  
 617 an assumption that needs to be stated.

618 Upon examining the failure and success modes of SynthPert (i.e. SynthPert made incorrect and  
 619 correct predictions, respectively), we noticed an interesting phenomenon: when SynthPert made  
 620 correct predictions; its CoTs reasoned about gene-level functions of both the perturbation and target  
 621 genes and how they may interact. In contrast, the CoTs for incorrect predictions relied heavily on  
 622 super-pathway-level descriptions of each gene (i.e. akin to Gene Ontology’s Biological Process)  
 623 and how those biological processes may interact. This likely arises from a lack of literature for the  
 624 perturbation or target gene in a knockdown study, thus making the model rely on the superpathway-  
 625 level descriptions from gene ontology. Unfortunately, most gene ontologies are annotated by humans,  
 626 creating an artificial bias that may be assigning an inappropriate biological process to a gene or  
 627 incompletely describing a gene’s function. Gene ontology is also affected by the lantern effect, where  
 628 the most well-studied genes have the most literature, annotations, and ontology terms. This makes  
 629 the models suffer from inductive bias when it comes to less well-annotated genes. Conversely,  
 630 this implies that the model works well for perturbation and target gene pairs, where the gene-level  
 631 functions are well known, and the perturbation is not so large that it leads to a new cell state. A  
 632 similar finding was reported in CIPHER (Kuznets-Speck et al., 2025), where linear response theory  
 633 was used to achieve the state-of-the-art prediction performance of perturbation data by using the  
 634 covariance of gene-gene pairs of cell lines in an unperturbed state.

635 Here is an example of a synthetic CoT generated o4-mini in SynthPert’s synthetic data pipeline with  
 636 our factuality rating in parentheses:

637 PFDN2 encodes a prefoldin subunit that functions primarily as a cytosolic chaperone (**TRUE**),  
 638 delivering nascent polypeptides (especially cytoskeletal and other cytosolic proteins) to the TRiC/CCT complex for folding. While prefoldin activity  
 639 can impact global proteostasis and, indirectly, ribosome biogenesis and translation fidelity, it does not directly regulate transcription of mitochondrial outer membrane  
 640 proteins such as VDAC3 (**FALSE**). VDAC3 is encoded in the nucleus, transcribed by factors (e.g., NRF1, PGC-1E) linked to mitochondrial biogenesis (**TRUE**), and  
 641 its post-translational import and  $\beta$ -barrel folding rely on mitochondrial-specific chaperones (SAM complex, small TIMs) rather than prefoldin (**FALSE**). Transcrip-  
 642 tional stress or mild proteostatic imbalance from PFDN2 knockdown in HepG2

648 cells may trigger a general stress response, but there is no established pathway  
 649 connecting PFDN2 depletion to up- or downregulation of VDAC3 mRNA (**TRUE**).  
 650 Empirical CRISPRi screens in this cell context report no significant change in  
 651 VDAC3 expression upon PFDN2 knockdown.  
 652

653 Here is an example of SynthPert correctly predicting down-regulation in the target gene with our  
 654 factuality rating in parentheses:  
 655

656  
 657 NCBP1 encodes the nuclear cap-binding protein 1, a core component of the cap-  
 658 binding complex (CBC) that associates with nascent pre-mRNAs to facilitate 5'-cap  
 659 recognition, co-transcriptional splicing, nuclear export, and protection from decay  
 660 (**TRUE**). In RPE1 cells, which rely on robust ribosome biogenesis to maintain  
 661 high rates of protein synthesis, efficient processing of ribosomal protein mRNAs  
 662 is critical (**TRUE**). Knockdown of NCBP1 via CRISPRi is expected to impair  
 663 cap-dependent mRNA maturation and export globally, but particularly affects  
 664 transcripts with complex splicing or weak cap-proximal elements (**TRUE**). RPL28,  
 665 a nuclear-encoded ribosomal protein gene, depends on proper capping and splicing  
 666 for stable mRNA and export to the cytoplasm (**TRUE**). Disruption of NCBP1  
 667 reduces RPL28 mRNA stability and processing, leading to lower steady-state  
 668 RPL28 transcript levels (**TRUE**). Moreover, compromised ribosome biogenesis  
 669 can trigger a feedback stress response, further downregulating ribosomal protein  
 670 gene expression (**TRUE**). Thus, in single-cell RPE1 context, NCBP1 knockdown  
 671 diminishes RPL28 mRNA abundance.  
 672

673 Here is an example of SynthPert incorrectly predicting a gene to be not differentially expressed when  
 674 it is actually downregulated:  
 675

676 COA5 encodes a mitochondrial complex IV assembly factor (**TRUE**). Its CRISPRi-  
 677 mediated knockdown in RPE1 cells impairs cytochrome c oxidase activity, reducing  
 678 mitochondrial respiration and ATP production (**TRUE**). Energy stress activates  
 679 AMPK and inhibits mTORC1, a central driver of ribosome biogenesis and riboso-  
 680 mal protein (RP) gene transcription (**TRUE**). mTORC1 downregulation decreases  
 681 RNA Pol I/III activity in the nucleolus and reduces RNA Pol II-dependent tran-  
 682 scription of RP genes, including the large ribosomal protein RPLP0. Additionally,  
 683 mitochondrial unfolded protein response (UPRmt) and integrated stress response  
 684 (ISR) trigger transcriptional reprogramming that favors stress-response factors over  
 685 ribosomal protein genes (**FALSE**). Together, these pathways explain a decrease in  
 686 RPLP0 mRNA upon COA5 knockdown in RPE1 cells.  
 687  
 688  
 689

## 690 C PROMPTS 691

692 This section provides the prompts used in the paper, the bold text in the prompts are variables.  
 693

### 694 695 C.1 STANDARD PROMPTS 696

697 These are the prompts we used for both LLM + SFT ON DATA and SYNTHPERT in Table 1 and for all  
 698 the models in Table 3 and Table 4. To generate the results shown in Table 1 for predicting whether a  
 699 gene is differentially expressed, we classified model predictions of “upregulated” or “downregulated”  
 700 as “differentially expressed”.  
 701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

**System Prompt**

You are a molecular and cellular biology expert analyzing gene regulation upon CRISPRi knockdown. First, provide your reasoning process within `<think> </think>` tags. Consider relevant pathways (e.g., cell-type specific biology, ribosome biogenesis, transcription, mitochondrial function, stress response), gene interactions, and cell-specific context. Then, choose one option from the following and place your choice within `<answer> </answer>` tags: 'upregulated', 'downregulated', or 'not differentially expressed'. Example: `<think> [Your reasoning here] </think><answer> [upregulated / downregulated / not differentially expressed] </answer>`

727

728

729

730

731

**User Query**

Analyze the regulatory effect of knocking down the **[gene name]** gene on the **[target gene]** gene in a single-cell **[cell type]** cell line using CRISPR interference.

732

733

734

735

**C.2 DIRECTION OF CHANGE PROMPTS**

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

**User Query**

It is given that the gene in question is differentially expressed, choose one of the following options:

1. upregulated
2. downregulated

Choose ONLY ONE of the options, UPREGULATED OR DOWNREGULATED, and PLACE YOUR CHOICE WITHIN `<answer> </answer>` TAGS. For this question 'not differentially expressed' is NOT an OPTION. Analyze the regulatory effect of knocking down the **[gene name]** gene on the **[target gene]** gene in a single-cell **[cell type]** cell line using CRISPR interference.

753

754

755

**C.3 SYNTHETIC DATA CREATION PROMPTS**

756

757

758

759

We used the following generator and critic prompts in our synthetic data pipeline. The generator user query is the same as above.

756

**Generator System Prompt**

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

You are a molecular and cellular biology expert analyzing and predicting gene regulation upon CRISPRi knockdown. The regulatory effect of knocking down the [gene name] gene on the [target gene] gene in a single-cell [cell type] cell line is given to you [solution]. Please provide detailed reasoning for your solution by considering the following:

1. relevant pathways
2. (e.g., cell-type biology, ribosome biogenesis, transcription, mitochondrial function, stress response),
3. gene interactions, and cell-specific context.

Then, choose one option from the following and place your choice within <answer> </answer> tags: 'upregulated', 'downregulated', 'not differentially expressed', or 'I do not know'. When answering provide a reasoning in regulatory effect such that you use the following template: <think> </think> <answer> [upregulated / downregulated / not differentially expressed / I do not know] </answer> Example of a CORRECT response: <think>Knocking down TF\_A, a known activator of Target\_Gene in this cell type, likely reduces its transcription. Relevant pathways include X and Y.</think><answer>downregulated</answer>

**Critic System Prompt**

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

You are an expert molecular and cellular biology expert acting as a critic.

Your task is to evaluate the reasoning process of another AI model that was asked to predict gene expression changes based on a perturbation.

Focus only on the quality, logical flow, and biological relevance of the provided reasoning (<think> block). Do not judge the final answer, only the steps taken to reach it.

Is the reasoning sound? Does it mention relevant and correct biological concepts (pathways, mechanisms, functions)? Does it logically connect the perturbation to the gene in the given cell type context?

Output your evaluation only in the following format choosing a single value for the evaluation:

<reasoning> [Provide a brief justification for your evaluation here. Explain why the reasoning is excellent, good, average, bad, or terrible.] </reasoning>

<evaluation> [excellent/good/average/bad/terrible] </evaluation>

**Critic User Query**

793

794

795

796

797

798

799

Original User Query: [user query] AI's Reasoning (<think> block): [generated thinking] Critique Task: Evaluate the AI's reasoning based on the criteria mentioned in the system prompt. Output your evaluation and justification in the specified format (<evaluation>...</evaluation><reasoning>...</reasoning>).

**D HYPERPARAMETERS**

800

801

802

803

804

805

806

807

808

809

We trained and evaluated both LLM + SFT ON DATA and SYNTHPERT using unsloth<sup>1</sup>. Table 6 lists all custom hyperparameters used in our experiments.

**E COMPUTE RESOURCES**

Both training and testing was conducted on a single NVIDIA A100 80GB GPU at BF16 precision, using unsloth. Training took 1.5 GPU hours.

<sup>1</sup><https://github.com/unslothai/unsloth>

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
Table 6: Hyperparameter settings for all experiments

820 821 822 823 824 825 826 <b>Hyperparameter</b>	827 828 829 830 831 <b>Value</b>	832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 <b>Description</b>
Learning rate	$1 \times 10^{-4}$	AdamW optimizer
Batch size	4	Per GPU
Warmup steps	5	Linear warmup
Max sequence length	2048	Output truncation
Epochs	50	Training duration

820  
821  
**F CODE AND DATA AVAILABILITY**822  
**F.1 CODE REPOSITORY**823  
Our implementation is publicly available at <https://github.com/qpdcdp/SynthPert>.824  
**F.2 SYNTHETIC DATA**825  
The synthetic data created and used in this work are available at [https://figshare.com/articles/dataset/Synthetic\\_Data\\_used\\_for\\_SynthPert\\_/29135996](https://figshare.com/articles/dataset/Synthetic_Data_used_for_SynthPert_/29135996). It is currently under embargo and will become available upon paper acceptance. The synthetic data used for supervised fine-tuning is provided for both the default split (used to produce results in Tables 2-4) and the RPE1 cell type holdout split (used to produce results in Table 5).