

CONTEXTUALIZING BIOLOGICAL PERTURBATION EXPERIMENTS THROUGH LANGUAGE

Menghua Wu^{*†}

Massachusetts Institute of Technology
Cambridge, MA, USA

Russell Littman, Jacob Levine

Biology Research & AI Development, Genentech
South San Francisco, CA, USA

Lin Qiu[†]

Meta AI
Menlo Park, CA, USA

David Richmond, Tommaso Biancalani, Jan-Christian Hütter^{*}

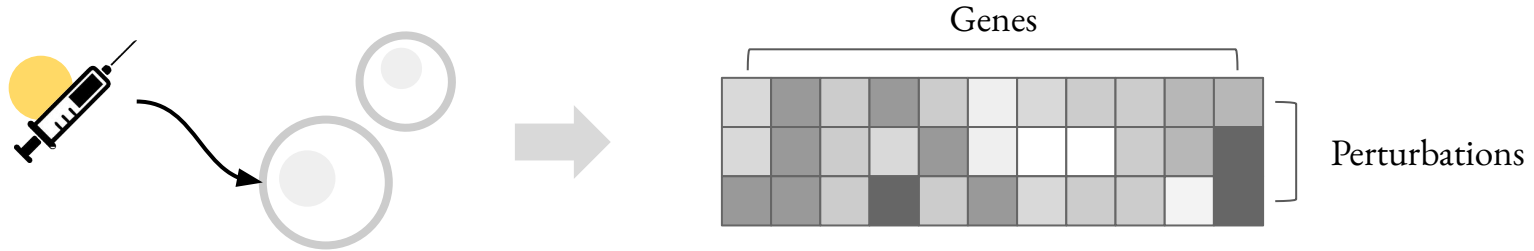
Biology Research & AI Development, Genentech
South San Francisco, CA, USA



Code and data

Genetic perturbation screens

allow biologists to **manipulate** and **measure** biological systems to elucidate their underlying molecular mechanisms.



Single cell

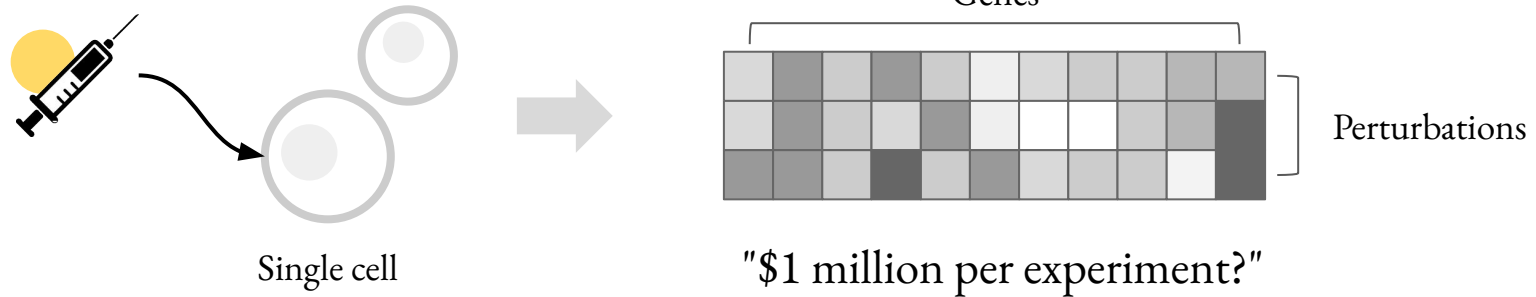
What is the function of this gene?

What happens when we modify the levels of this gene?

Genetic perturbation screens


allow biologists to **manipulate** and **measure** biological systems to elucidate their underlying molecular mechanisms.

are *very* expensive to run and interpret!




Our goals

 Facilitate efficient experimental design → Infer the effects of unseen perturbations

 Reduce human annotation burden → Automatically summarize high-dimensional readouts in context of known biology

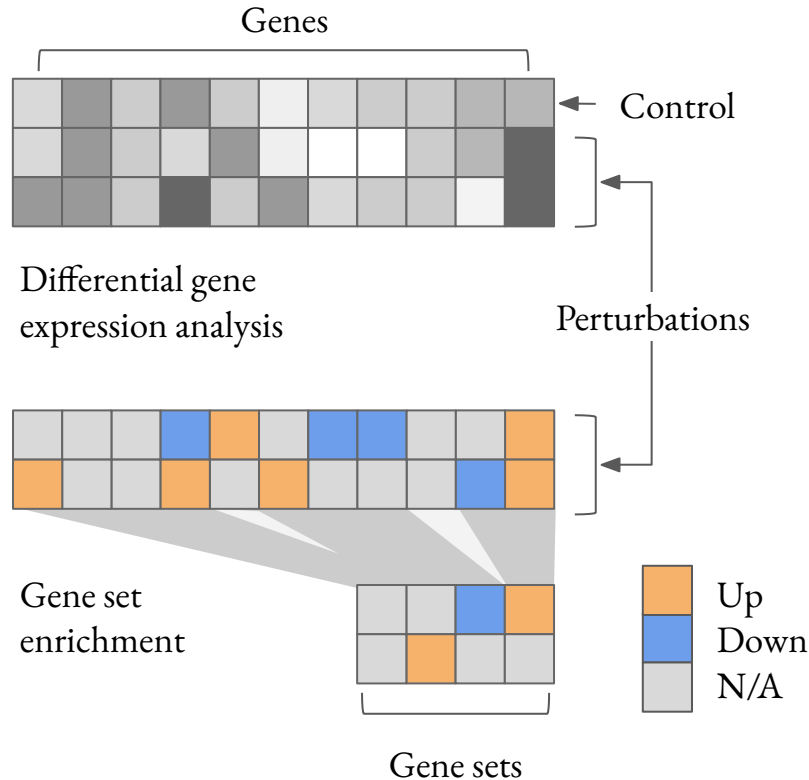
Our contributions

 Claim: Perturbation modeling should reflect downstream analyses.
From individual cells / genes \rightarrow statistical insights.

 PerturbQA: A new benchmark for perturbations + LLM reasoning
for structured biological data / discovering new biology.

 SUMMER: Domain-informed LLM baseline for predicting effects of
unseen perturbations

Discrete insights from transcriptomic data



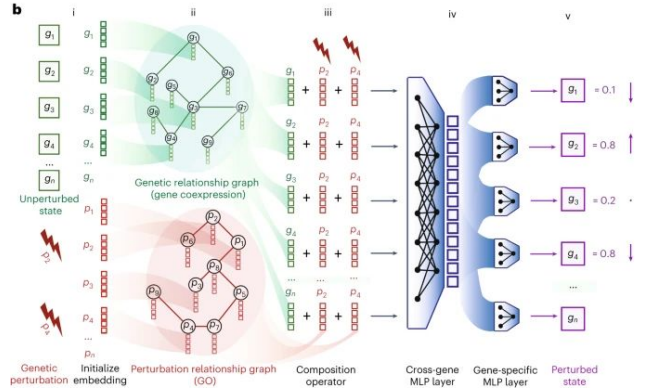
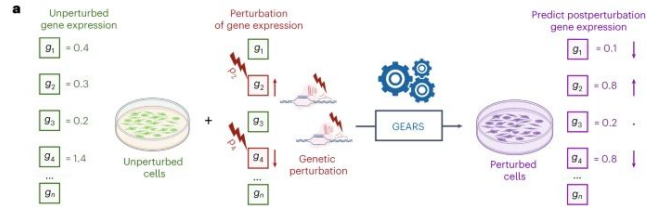
Biologists "read" continuous gene expression through discrete endpoints.

"strong downregulation of NUP62"

"loss of C7orf26 impacted Integrator subunit abundance"

"TMEM242 produced a signature resembling loss of ATP synthase"

Current machine learning perspective




Predict effects of unseen perturbations.

Knowledge graphs relate seen vs. unseen perturbations → This makes sense!

However:

 GNN methods discard rich, textual semantics.

 Most focus on regressing log-fold change of *genes that actually change*.

PerturbQA

Set of real, unsolved tasks related to perturbations, designed as a benchmark for biological reasoning.

Differential
expression

Is a knockdown of **ABCE1** in K562 cells likely to result in ***differential expression*** of **GABARAP**?

Direction of
change


Is a knockdown of **ABCE1** in K562 cells likely to result ***decrease or increase*** of **GABARAP**?


Gene set
enrichment


How are the following genes ***related***, and why do they ***induce similar downstream effects when perturbed***? **CFLAR**, **VIM**, ...


How are the following genes ***related***, and why do they ***respond similarly to perturbation***? **CFLAR**, **VIM**, **CAPG**, ...

Data release

 3 tasks: Differential expression, direction of change, gene set summarization

 DE/Dir: 4 datasets ("cell lines") at individual gene level; 1 dataset at gene set level

 2 sets of human annotations regarding gene cluster function

 Harmonized knowledge graphs with all nodes/edges mapped to text, for biological context

Database	Information
UniProt	Gene
Ensembl	Gene
Gene Ontology	Gene, relations
CORUM	Relations
STRING	Relations
Reactome	Relations
BioPlex	Relations

Data release



Differential expression and direction of change



Datasets can be loaded as follows.

```
from pertqa import load_de, load_dir

# options: "k562" "rpe1" "hepg2" "jurkat" "k562_set"
data_de = load_de("k562")
# train/test splits
X_train = data_de["train"]
X_test = data_de["test"]

data_dir = load_dir("k562")
```


Results

 Existing methods perform poorly on PerturbQA 

	Model	K562	RPE1	HepG2	Jurkat	K562-Set
Differential expression	PHYSICAL	0.53	0.52	0.52	0.54	0.55
	GAT	0.55 \pm .02	0.57 \pm .02	0.57 \pm .02	0.55 \pm .03	0.54 \pm .01
	GEARS	0.54 \pm .01	0.50 \pm .01	0.48 \pm .02	0.51 \pm .01	0.49 \pm .01
	SCGPT	0.52 \pm .00	0.52 \pm .00	0.48 \pm .00	0.51 \pm .00	0.52 \pm .00
	GENEPT-GENE	0.57 \pm .02	0.54 \pm .00	0.55 \pm .02	0.55 \pm .01	0.58 \pm .01
	GENEPT-PROT	0.57 \pm .01	0.56 \pm .00	0.54 \pm .01	0.55 \pm .01	0.58 \pm .01
	LLM (No CoT)	0.52 \pm .01	0.51 \pm .00	0.51 \pm .01	0.52 \pm .00	0.50 \pm .00
	LLM (No retrieval)	0.51 \pm .01	0.48 \pm .00	0.49 \pm .01	0.49 \pm .01	0.50 \pm .01
	Retrieval (No LLM)	0.58 \pm .02	0.58 \pm .01	0.55 \pm .00	0.55 \pm .01	0.64 \pm .00
Direction of change	GAT	0.58 \pm .06	0.60 \pm .04	0.64 \pm .05	0.59 \pm .04	0.53 \pm .03
	GEARS	0.64 \pm .01	0.60 \pm .01	0.52 \pm .01	0.51 \pm .01	0.59 \pm .02
	SCGPT	0.48 \pm .00	0.53 \pm .00	0.51 \pm .00	0.51 \pm .00	0.54 \pm .00
	GENEPT-GENE	0.53 \pm .05	0.57 \pm .03	0.58 \pm .03	0.57 \pm .02	0.56 \pm .02
	GENEPT-PROT	0.57 \pm .01	0.57 \pm .02	0.55 \pm .01	0.58 \pm .03	0.57 \pm .02
	LLM (No CoT)	0.50 \pm .01	0.49 \pm .00	0.49 \pm .00	0.50 \pm .01	0.50 \pm .01
	LLM (No retrieval)	0.49 \pm .04	0.52 \pm .03	0.51 \pm .06	0.53 \pm .05	0.45 \pm .18
	Retrieval (No LLM)	0.50 \pm .00	0.50 \pm .00	0.50 \pm .00	0.50 \pm .00	0.50 \pm .00

Macro AUROC (mean over downstream genes)

Results

 Existing methods perform poorly on PerturbQA ❌

 Naively applying LLMs also performs poorly ❌

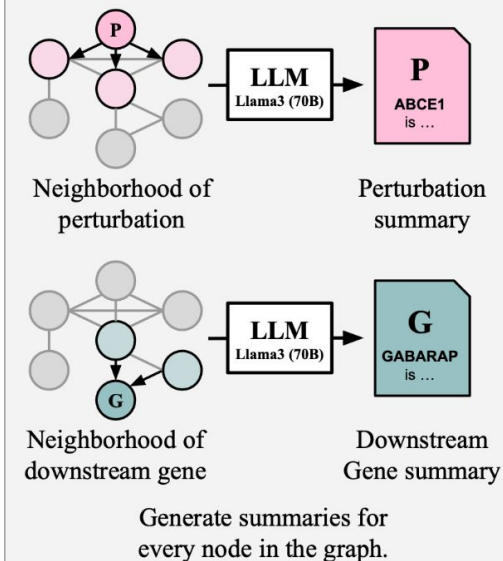
	Model	K562	RPE1	HepG2	Jurkat	K562-Set
Differential expression	PHYSICAL	0.53	0.52	0.52	0.54	0.55
	GAT	0.55±.02	0.57±.02	0.57±.02	0.55±.03	0.54±.01
	GEARS	0.54±.01	0.50±.01	0.48±.02	0.51±.01	0.49±.01
	SCGPT	0.52±.00	0.52±.00	0.48±.00	0.51±.00	0.52±.00
	GENEPT-GENE	0.57±.02	0.54±.00	0.55±.02	0.55±.01	0.58±.01
	GENEPT-PROT	0.57±.01	0.56±.00	0.54±.01	0.55±.01	0.58±.01
	LLM (No CoT)	0.52±.01	0.51±.00	0.51±.01	0.52±.00	0.50±.00
	LLM (No retrieval)	0.51±.01	0.48±.00	0.49±.01	0.49±.01	0.50±.01
	Retrieval (No LLM)	0.58±.02	0.58 ±.01	0.55±.00	0.55±.01	0.64 ±.00
Direction of change	GAT	0.58±.06	0.60±.04	0.64±.05	0.59±.04	0.53±.03
	GEARS	0.64 ±.01	0.60±.01	0.52±.01	0.51±.01	0.59±.02
	SCGPT	0.48±.00	0.53±.00	0.51±.00	0.51±.00	0.54±.00
	GENEPT-GENE	0.53±.05	0.57±.03	0.58±.03	0.57±.02	0.56±.02
	GENEPT-PROT	0.57±.01	0.57±.02	0.55±.01	0.58±.03	0.57±.02
	LLM (No CoT)	0.50±.01	0.49±.00	0.49±.00	0.50±.01	0.50±.01
	LLM (No retrieval)	0.49±.04	0.52±.03	0.51±.06	0.53±.05	0.45±.18
	Retrieval (No LLM)	0.50±.00	0.50±.00	0.50±.00	0.50±.00	0.50±.00

Macro AUROC (mean over downstream genes)

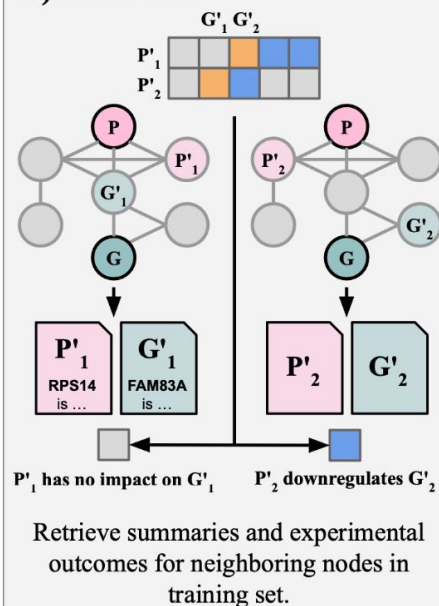
Domain-informed proof of concept

"SUMMER" – Summarize, Retrieve, Answer

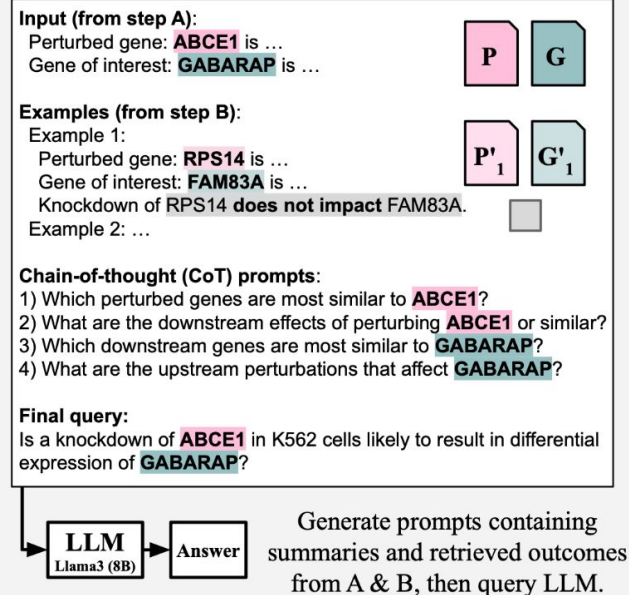
A) SUMMARIZE





B) RETRIEVE





C) ANSWER





Results

 Simple reasoning template +
retrieving experimental
outcomes = does ok 

 Still a long way to go! 

	Model	K562	RPE1	HepG2	Jurkat	K562-Set
Differential expression	PHYSICAL	0.53	0.52	0.52	0.54	0.55
	GAT	0.55 \pm .02	0.57 \pm .02	0.57 \pm .02	0.55 \pm .03	0.54 \pm .01
	GEARS	0.54 \pm .01	0.50 \pm .01	0.48 \pm .02	0.51 \pm .01	0.49 \pm .01
	SCGPT	0.52 \pm .00	0.52 \pm .00	0.48 \pm .00	0.51 \pm .00	0.52 \pm .00
	GENEPT-GENE	0.57 \pm .02	0.54 \pm .00	0.55 \pm .02	0.55 \pm .01	0.58 \pm .01
	GENEPT-PROT	0.57 \pm .01	0.56 \pm .00	0.54 \pm .01	0.55 \pm .01	0.58 \pm .01
	LLM (No CoT)	0.52 \pm .01	0.51 \pm .00	0.51 \pm .01	0.52 \pm .00	0.50 \pm .00
	LLM (No retrieval)	0.51 \pm .01	0.48 \pm .00	0.49 \pm .01	0.49 \pm .01	0.50 \pm .01
	Retrieval (No LLM)	0.58 \pm .02	0.58 \pm .01	0.55 \pm .00	0.55 \pm .01	0.64 \pm .00
	SUMMER	0.60 \pm .00	0.58 \pm .00	0.61 \pm .00	0.58 \pm .00	0.61 \pm .00
	GAT	0.58 \pm .06	0.60 \pm .04	0.64 \pm .05	0.59 \pm .04	0.53 \pm .03
	GEARS	0.64 \pm .01	0.60 \pm .01	0.52 \pm .01	0.51 \pm .01	0.59 \pm .02
	SCGPT	0.48 \pm .00	0.53 \pm .00	0.51 \pm .00	0.51 \pm .00	0.54 \pm .00
	GENEPT-GENE	0.53 \pm .05	0.57 \pm .03	0.58 \pm .03	0.57 \pm .02	0.56 \pm .02
Direction of change	GENEPT-PROT	0.57 \pm .01	0.57 \pm .02	0.55 \pm .01	0.58 \pm .03	0.57 \pm .02
	LLM (No CoT)	0.50 \pm .01	0.49 \pm .00	0.49 \pm .00	0.50 \pm .01	0.50 \pm .01
	LLM (No retrieval)	0.49 \pm .04	0.52 \pm .03	0.51 \pm .06	0.53 \pm .05	0.45 \pm .18
	Retrieval (No LLM)	0.50 \pm .00	0.50 \pm .00	0.50 \pm .00	0.50 \pm .00	0.50 \pm .00
	SUMMER	0.62 \pm .01	0.64 \pm .01	0.65 \pm .00	0.66 \pm .01	0.69 \pm .01

Results (gene set summarization)

 LLMs easily summarize gene sets 

Genes	Annotation	Gene sets (combined)	SUMMER (8b)
SLC25A5, EIF4B, SLC25A3, RPL3, ..., RPL41	translation	cytosol, RNA binding, cytoplasm, membrane, GTP hydrolysis	Ribosomal and Translation Regulation Gene Set The gene set is comprised of genes involved in protein synthesis, translation, and ribosomal function, with a focus on the regulation of protein synthesis and the assembly of ribosomal subunits.
ZC3H13, CBLL1, METTL14, METTL3, PSMG1, RBM15	m6A mRNA methylation	nucleus, nucleoplasm, RNA N6-methyladenosine methyltransferase complex, mRNA processing	M6A Methylation Complex-Associated Genes. The gene set is composed of genes involved in the regulation of N6-methyladenosine (m6A) methylation of RNAs, influencing mRNA splicing and RNA processing. These genes are associated with the WMM complex and interact with each other to modulate gene expression.
CMTR2, RBM14, RBM4, RBM4, UNCX, WDFY3	unknown	no significant sets	RNA Processing and Regulation Gene Set. The gene set is composed of genes involved in RNA processing and regulation, including mRNA cap modification, alternative splicing, and RNA-binding activities. These genes converge on pathways related to mRNA stability, translation, and cellular differentiation.

Results (gene set summarization)

Genes	Annotation	Gene sets (combined)	SUMMER (8b)
SLC25A5, EIF4B, SLC25A3, RPL3, ..., RPL41	translation	cytosol, RNA binding, cytoplasm, membrane, GTP hydrolysis	Ribosomal and Translation Regulation Gene Set The gene set is comprised of genes involved in protein synthesis, translation, and ribosomal function, with a focus on the regulation of protein synthesis and the assembly of ribosomal subunits.
ZC3H13, CBLL1, METTL14, METTL3, PSMG1, RBM15	m6A mRNA methylation	nucleus, nucleoplasm, RNA N6-methyladenosine methyltransferase complex, mRNA processing	M6A Methylation Complex-Associated Genes. The gene set is composed of genes involved in the regulation of N6-methyladenosine (m6A) methylation of RNAs, influencing mRNA splicing and RNA processing. These genes are associated with the WMM complex and interact with each other to modulate gene expression.
CMTR2, RBM14, RBM4, RBM4, UNCX, WDFY3	unknown	no significant sets	RNA Processing and Regulation Gene Set. The gene set is composed of genes involved in RNA processing and regulation, including mRNA cap modification, alternative splicing, and RNA-binding activities. These genes converge on pathways related to mRNA stability, translation, and cellular differentiation.



LLM equal or better to the classical gene set enrichment results in 92% of cases.



Agrees with the independent annotator in 72% of cases.

Results (gene set enrichment)

Enrichment	Top	Gene clusters				Perturbation clusters			
		$R_{\text{ROUGE1}} \uparrow$	$P_{\text{BERT}} \uparrow$	$R_{\text{BERT}} \uparrow$	$F_{\text{BERT}} \uparrow$	$R_{\text{ROUGE1}} \uparrow$	$P_{\text{BERT}} \uparrow$	$R_{\text{BERT}} \uparrow$	$F_{\text{BERT}} \uparrow$
Gene Ontology	5	0.17	0.64	0.66	0.62	0.38	0.66	0.72	0.68
Gene Ontology	10	0.32	0.60	0.65	0.60	0.60	0.62	0.71	0.65
Reactome	5	0.18	0.60	0.65	0.60	0.49	0.60	0.68	0.62
Reactome	10	0.27	0.54	0.64	0.56	0.59	0.56	0.67	0.60
CORUM	5	0.07	0.63	0.45	0.42	0.45	0.64	0.63	0.60
CORUM	10	0.07	0.61	0.44	0.41	0.47	0.61	0.62	0.58
Combined	5	0.14	0.62	0.65	0.61	0.41	0.63	0.71	0.66
Combined	10	0.27	0.59	0.65	0.59	0.63	0.57	0.69	0.62
SUMMER (8b)	desc	0.57	0.63	0.76	0.69	0.26	0.63	0.75	0.68
SUMMER (8b)	name	0.20	0.74	0.76	0.75	0.12	0.75	0.76	0.75
SUMMER (70b)	desc	0.45	0.63	0.77	0.69	0.59	0.65	0.80	0.72
SUMMER (70b)	name	0.15	0.73	0.76	0.74	0.37	0.77	0.82	0.79



Automatic evaluation metrics

Takeaways

1. Biology has lots of assays, modalities, domain knowledge – but language is a natural means for harmonizing these data.
2. Using LMs for biology requires more "biology" than classic approaches.
3. Modeling perturbations has so much room for creative solutions that leverage prior knowledge and model causal relationships

Come play with our data!