

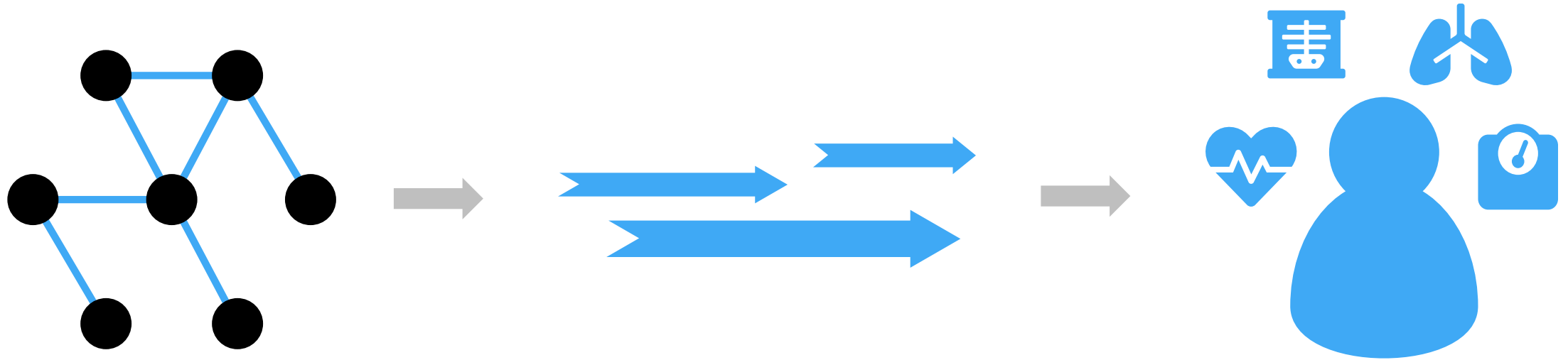
GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations

Caleb Ellington

AI4Bio

Dec. 15th 2022

Gene regulatory networks define gene expression and phenotypes through functional organization

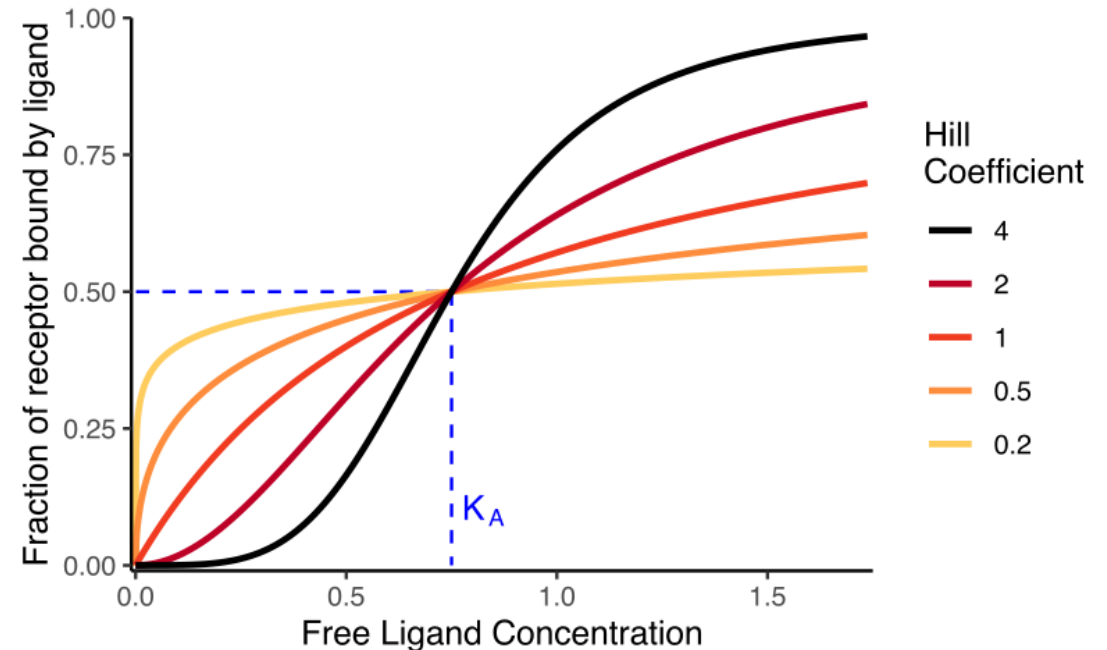


The complexity of gene regulation

- Multiplexed cooperative binding of transcription factors creates highly non-linear and dynamic gene expression patterns
- ~25,000 protein-coding genes, often modulated by cooperative transcription factor binding

Hill coefficient (n) describes how cooperative the ligand binding is

$$\% \text{ bound} = \frac{[\text{Ligand}]^n}{K_d + [\text{Ligand}]^n}$$

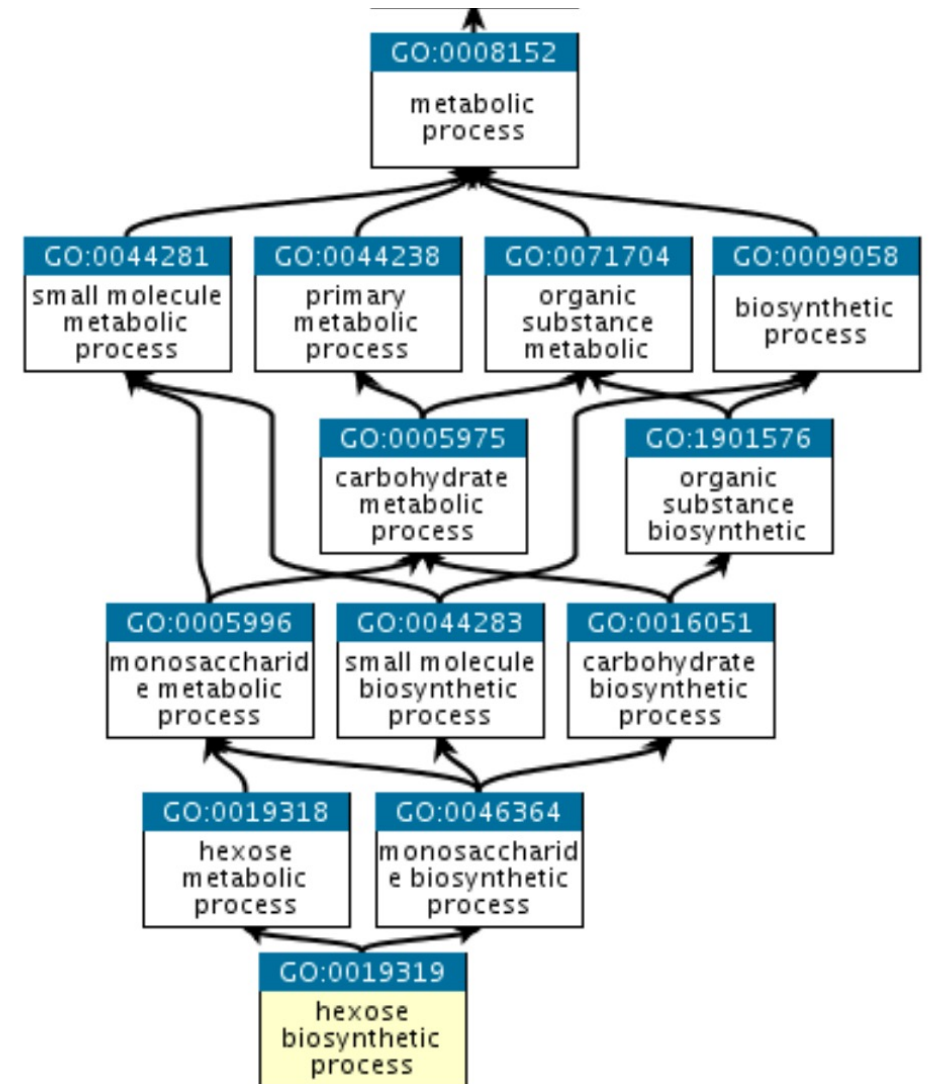


How do we model gene regulatory behavior?

More concretely: how does gene regulation change under experimental conditions? Let's start with gene knockouts as our experimental condition.

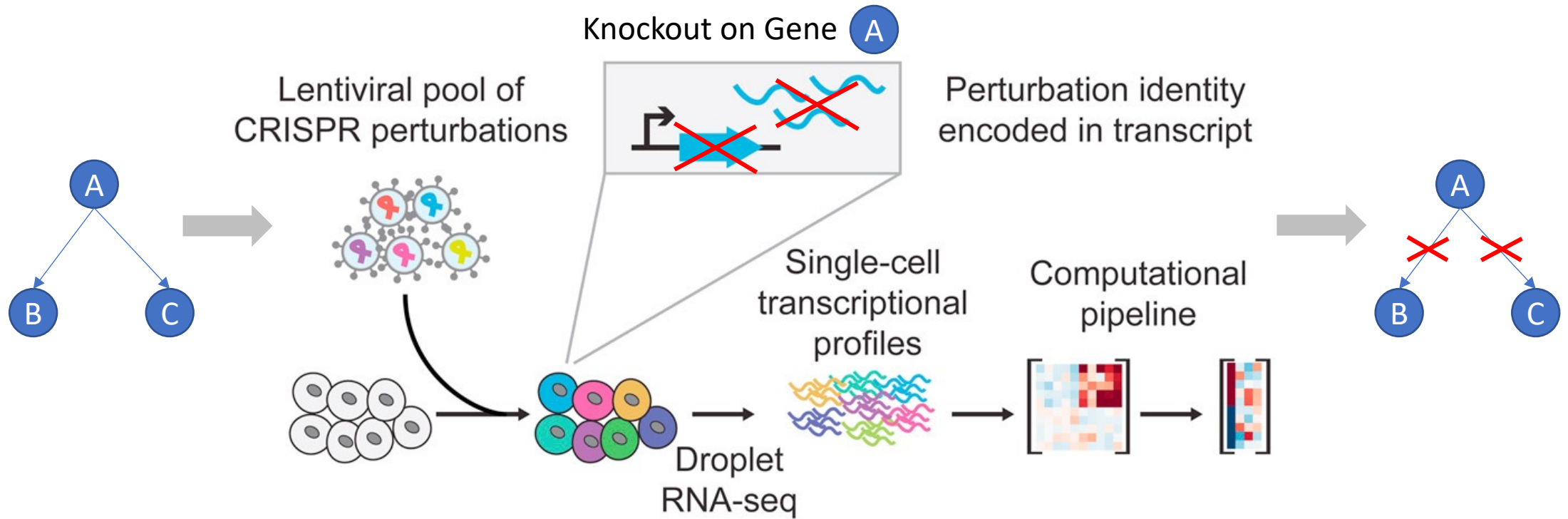
Preliminaries

- Co-expression graph: pairwise correlations between all genes. Can be viewed as a weighted adjacency matrix. Binarizing into an adjacency matrix can be done by thresholding correlation, or selection top N neighbors.
- Gene Ontology: expert-curated hierarchal annotation of genes on functional, pathway, & molecular attributes. Ignoring annotation hierarchy, this can also be viewed as a bipartite graph between genes & annotations.



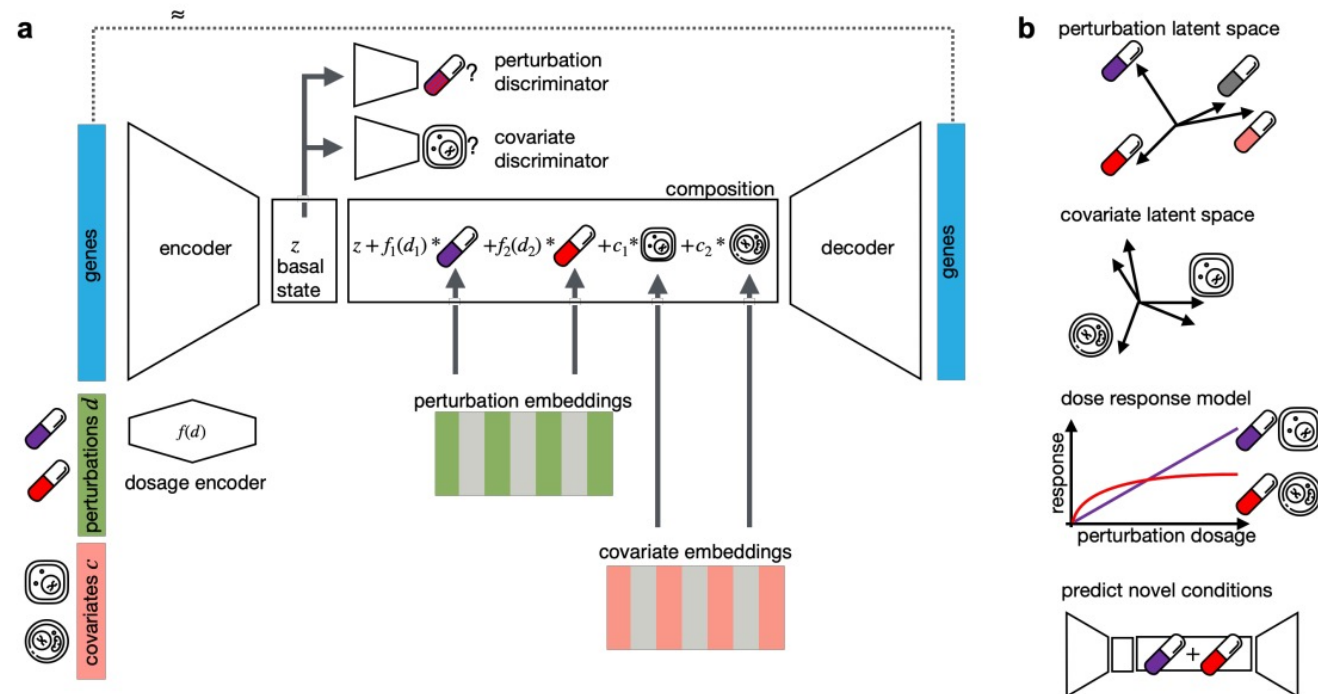
Perturb-seq: Experimental Basis & Limitations

- Low-order do-interventions on gene regulatory graphs

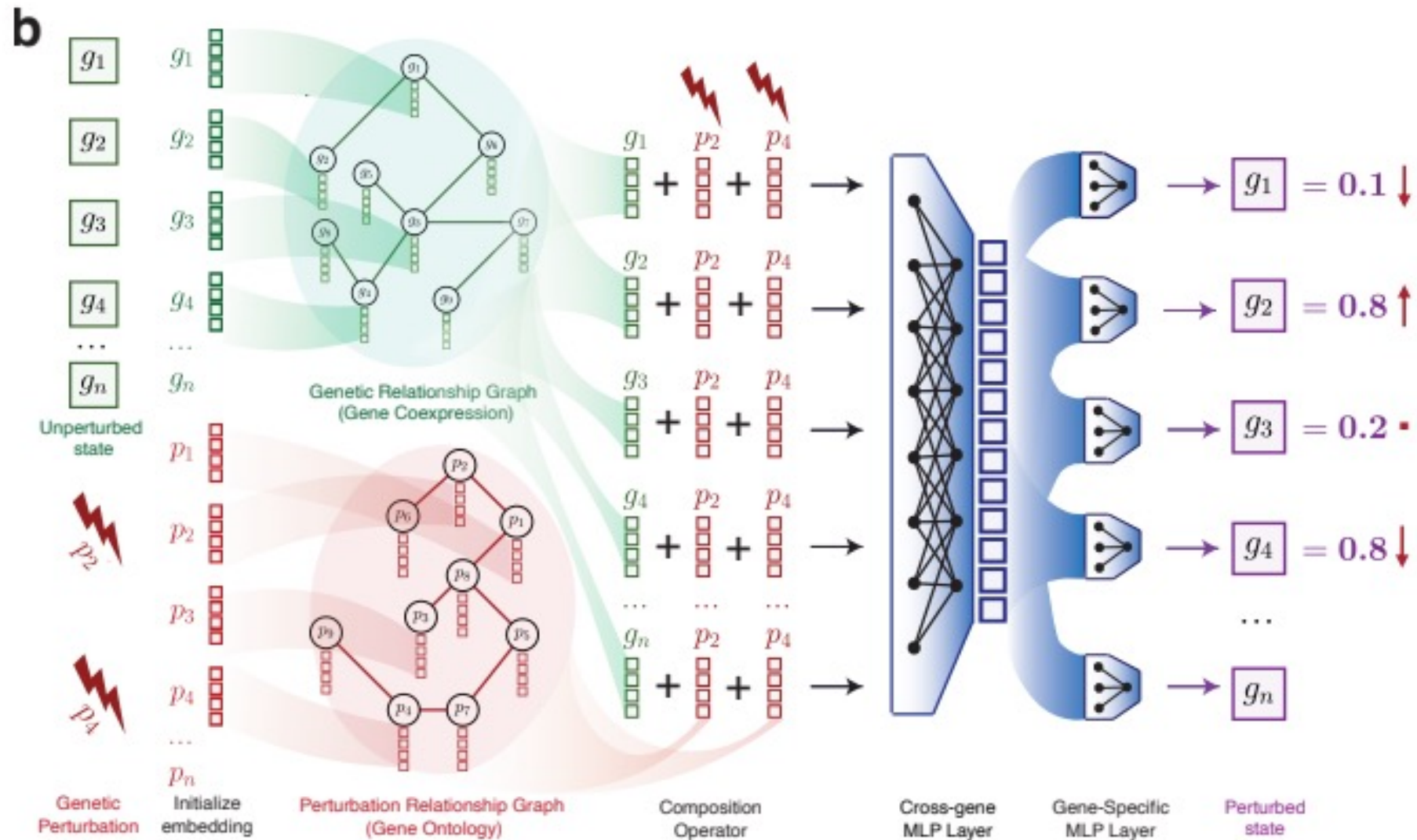


Baseline models

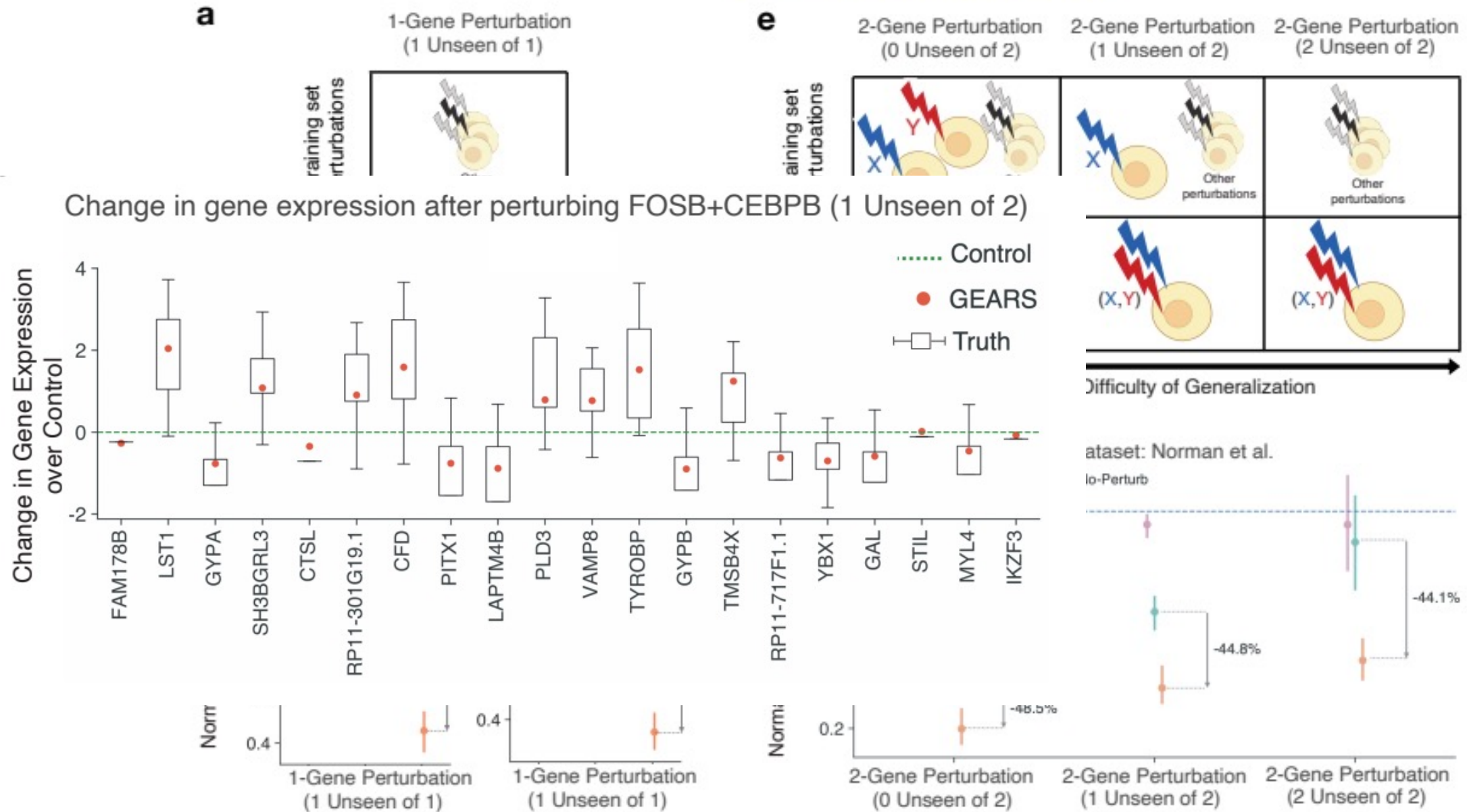
- Linear: 3 linear graph convolutions, using the gene co-expression graph (linear propagation through 3 neighbor hops)
- CPA: Compositional Perturbation Autoencoder



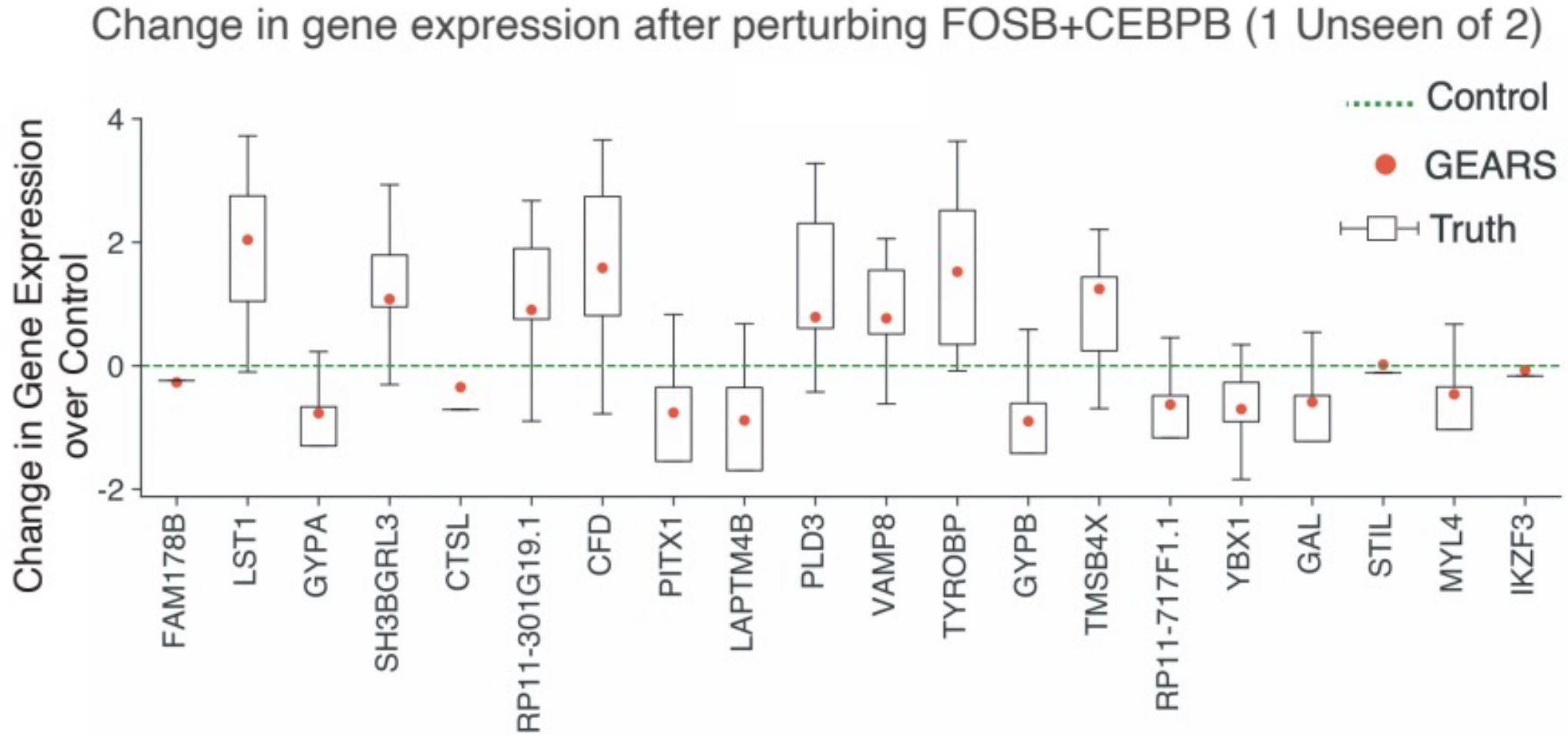
GEARS



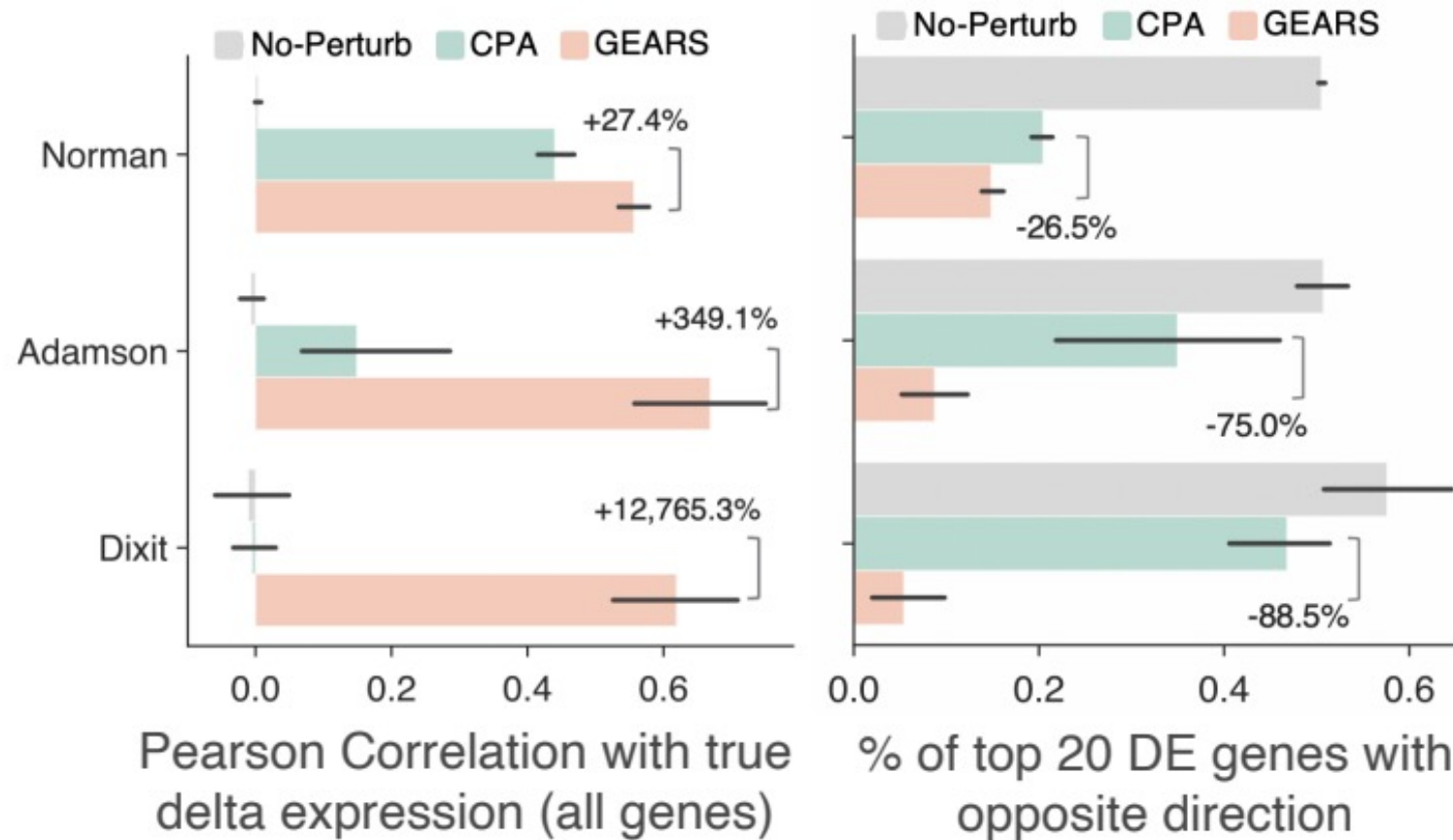
GEARS improves the accuracy of single and multiplexed perturbation predictions



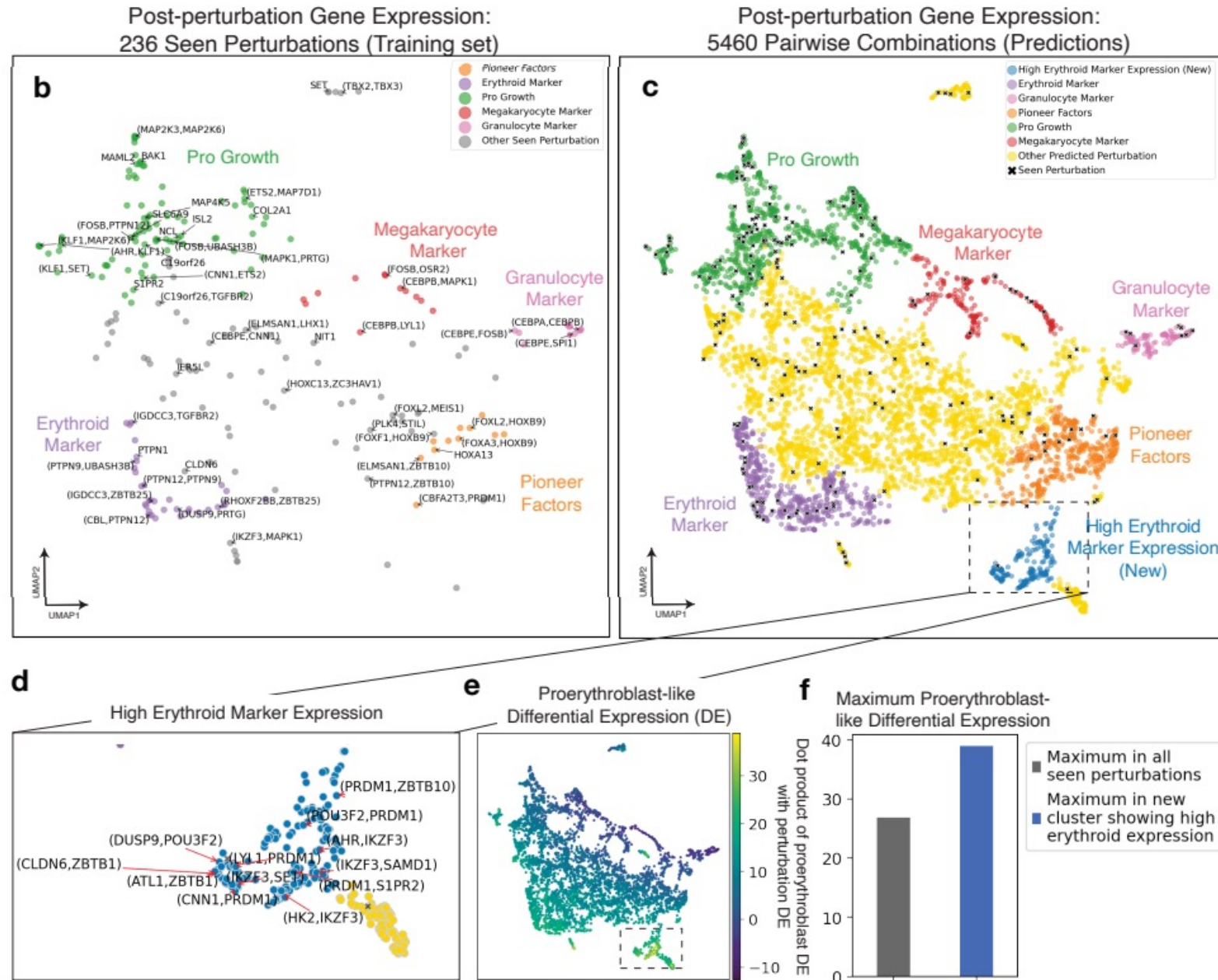
GEARS improves the accuracy of single and multiplexed perturbation predictions



GEARS improves directionality of perturbation predictions

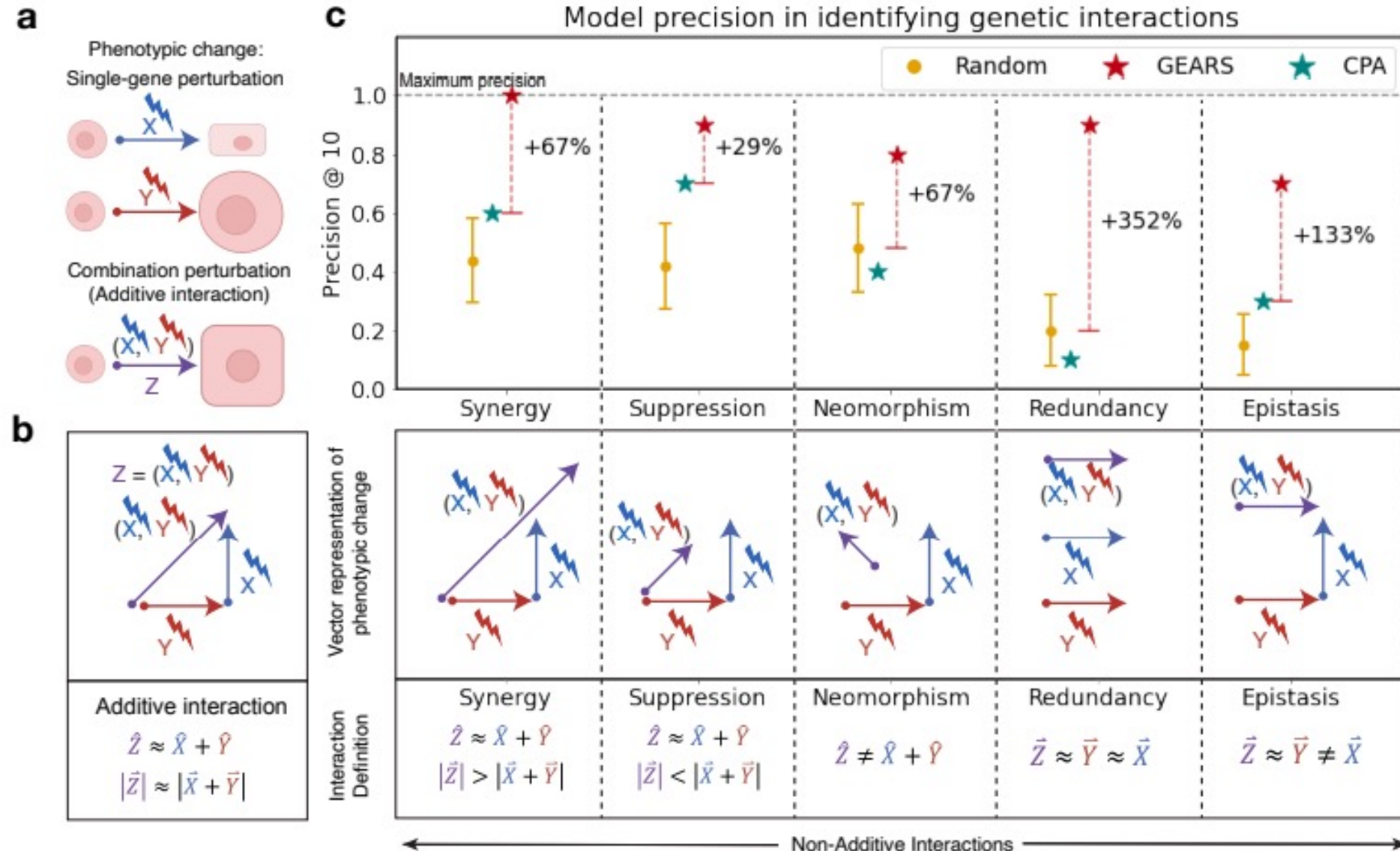


GEARS is sample-efficient and allows inference of unseen states

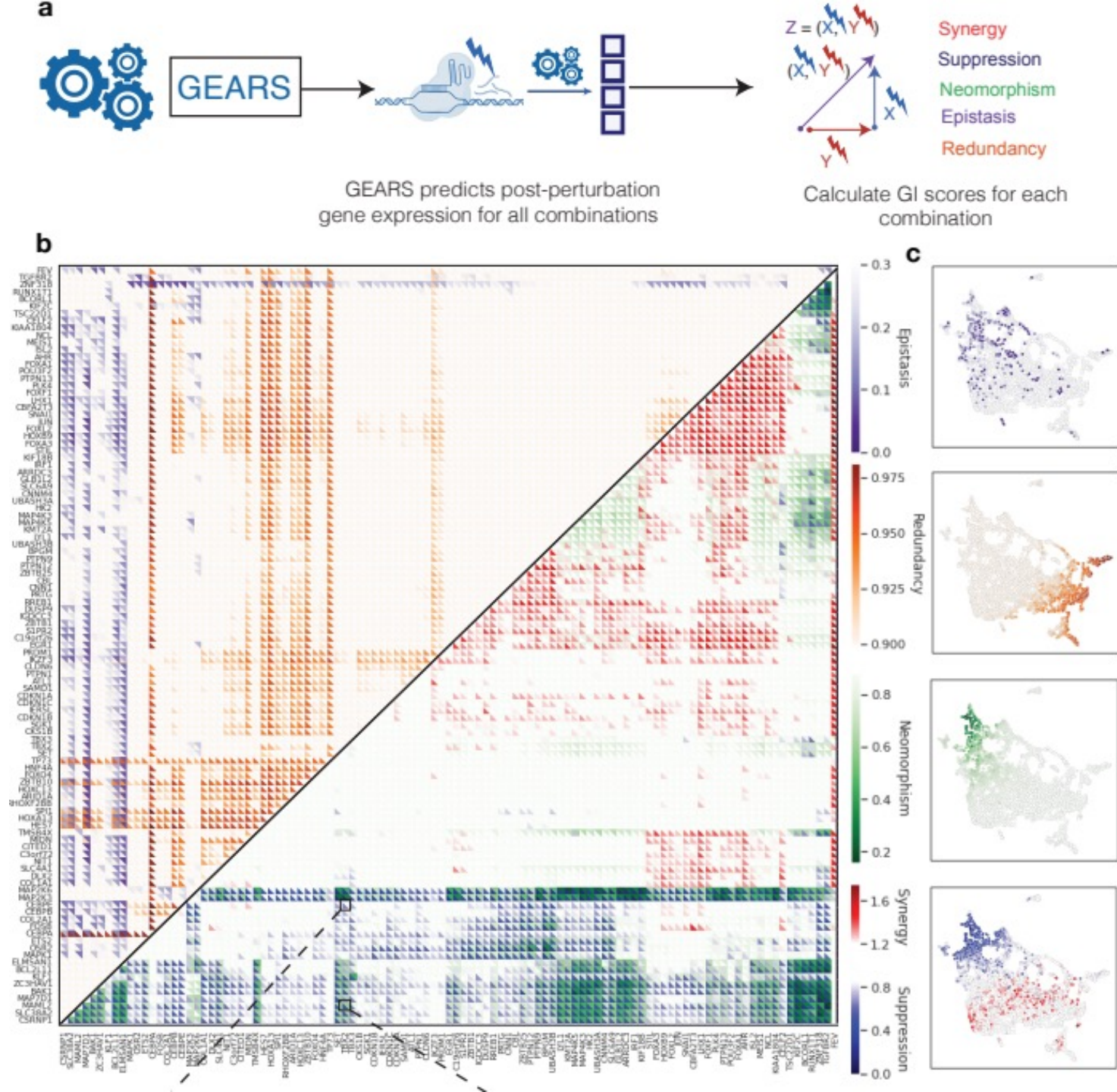


(phenotype speculation,
not experimentally
validated)

GEARS accurately models non-linear gene-gene interactions



GEARS characterizes
multiplexed
interactions at scale



Extra Info

Methods

- 5000 nodes (selected as high variance genes in the training set) + any perturbed genes not in this set
- Perturbation encoder graph is constructed from bipartite gene ontology graph, where neighbors are the top N in terms of Jaccard similarity on their set of annotations.

Losses

$$L_{\text{autofocus}} = \frac{1}{T} \sum_{k=1}^T \frac{1}{T_k} \sum_{l=1}^{T_k} \frac{1}{G} \sum_{u=1}^K (\mathbf{g}_u - \hat{\mathbf{g}}_u)^{(2+\gamma)}$$

$$L_{\text{direction}} = \frac{1}{T} \sum_{k=1}^T \frac{1}{T_k} \sum_{l=1}^{T_k} \frac{1}{G} \sum_{u=1}^K \left(\text{sign}(\mathbf{g}_u - \mathbf{g}_u^{\text{ctrl}}) - \text{sign}(\hat{\mathbf{g}}_u - \mathbf{g}_u^{\text{ctrl}}) \right)^2$$

$$L_{\text{unc}} = \frac{1}{T} \sum_{k=1}^T \frac{1}{T_k} \sum_{l=1}^{T_k} \frac{1}{G} \sum_{u=1}^K \exp(-s_u) (\mathbf{g}_u - \hat{\mathbf{g}}_u)^{(2+\gamma)}$$