ML Lab Research Proposal
Lookup-Row Dormancy in Cellular Foundation Models

Team: Ethan Wang (solo)
Date: July 22, 2026
Project: InterveneFM: Counterfactual Audit of Per-Gene Perturbation Embeddings
Target Compute: 1x AMD Instinct MI300X GPU, approx. 1 week

---

1. PROBLEM STATEMENT

Cellular foundation models for perturbation prediction, including CPA (Lotfollahi et al., 2023), GEARS (Roohani et al., 2024), and scGPT (Cui et al., 2024), condition their outputs on per-gene embeddings stored as discrete-index lookup tables (nn.Embedding). Ahlmann-Eltze et al. (Nature Methods 2025) showed these models fail to outperform linear baselines. We pinpoint the mechanism: on standard 0/1 splits, held-out test genes receive zero gradient during training, so their embedding rows remain near random initialization. We call this **lookup-row dormancy**. A 30K-parameter MLP (the Conditional Perturbation Generator, CPG) that replaces nn.Embedding with a function of precomputed gene-identity features closes 64% of the gap on K562.

2. PROPOSED APPROACH

We run a counterfactual audit: at inference, we replace only the per-gene embedding with five alternatives (learned, mean-over-training, zero, random, population-mean) while fixing all other model weights. On Norman CRISPRa (4 seeds), GEARS's embedding adds +0.066 DE-Spearman above the population-mean baseline, but on Replogle K562 CRISPRi (3 seeds) the gap inverts to -0.161. scGPT shows the same sign inversion. Trained embedding rows shift 2.42x from initialization along a coherent knockdown axis; held-out rows stay within 1.3x and are nearly orthogonal to it.

For AAAI, we extend the audit to additional architectures (SAMS-VAE, AttentionPert) and test dormancy-immune models (Biolord, which uses Gene Ontology features instead of nn.Embedding) as a control. We also swap the CPG's SVD features for pretrained gene embeddings from GenePT (Zou group, 2024) and Geneformer (Theodoris et al., Nature 2023), and extend to combinatorial perturbations where compounding dormancy (both genes in a pair unseen) may produce failures worse than the sum of individual dormancies.

3. DATASET

Dataset	Source	Size	Use
Norman 2019 CRISPRa	Norman et al., Science 2019	91K cells, 105 single + 131 combo perturbations	Positive control (overall gap > 0). 0/2 split for combinatorial dormancy.
Replogle K562 CRISPRi	Replogle et al., Cell 2022	~2M cells, 9,867 targeted genes	Primary dormancy substrate (gap = -0.161). 0/1 split with 80 held-out genes.
Replogle RPE1 CRISPRi	Replogle et al., Cell 2022	~1.5M cells, ~9,800 genes	Secondary substrate (gap = -0.071). Cross-dataset replication.
sci-Plex (K562 subset)	Srivatsan et al., Science 2020	650K cells, 188 compounds x 4 doses	Dose-response dormancy extension (stretch goal).

Preprocessing

- Highly variable gene (HVG) selection with leakage fix: HVGs computed on training cells only, not full dataset.
- Perturbation embeddings: nn.Embedding(n_genes, embed_dim) with default PyTorch initialization.
- CPG features: truncated SVD (64 components) over control-cell expression matrix. GenePT embeddings: 3072-dim PCA-reduced to 64.

4. EVALUATION METRICS

Experiment	Metric	Target / Interpretation
Architecture extension (SAMS-VAE, AttentionPert, Biolord)	DE-Spearman gap (learned - pop_mean) on K562 0/1 split	SAMS-VAE and AttentionPert show dormancy (gap < 0). Biolord (continuous features) does NOT show dormancy. Confirms mechanism is embedding-specific.
Pretrained gene embeddings (GenePT, Geneformer, ESM-2, Gene2Vec)	Gap closure relative to CPA baseline on K562 and RPE1	GenePT closes more than SVD's 64% (target: 70-80%). Identifies optimal gene representation.
Combinatorial dormancy (Norman 0/2 split)	Compounding ratio: gap(0/2) / (2 * gap(0/1))	Ratio > 1.0 indicates multiplicative compounding. CPG should mitigate (both genes have meaningful features).
Conditional-helper replication (K562, RPE1)	Per-tertile gap stratified by pop_mean performance	LOW tertile (strong effects) shows positive gap; HIGH tertile (weak effects) shows negative gap. Architecture-agnostic.

5. COMPUTE REQUIREMENTS

Most experiments are CPU-bound (small models, cached features). GPU compute is needed only for SAMS-VAE and AttentionPert training on the Replogle datasets.

Phase	Hardware	Est. Wall-Clock	Notes
Architecture extension (SAMS-VAE + AttentionPert + Biolord)	1x MI300X	~8 hours	Training 3 models on Norman + K562, 3 seeds each. ~$25-35 equivalent.
Pretrained gene embeddings (6 feature sets x 2 datasets x 3 seeds)	CPU only	~3 hours	CPG training is fast (30K params). Feature extraction is one-time.
Combinatorial dormancy (Norman 0/2 stratification + CPG)	CPU + 1x MI300X	~3 hours	GEARS retraining on 1/2 split needs GPU. CPA/CPG analysis is CPU.
Conditional-helper replication	CPU only	~1 hour	Reanalysis of existing checkpoints. No new training.
Dose-response (sci-Plex, stretch goal)	1x MI300X	~6 hours	CPA on 1 cell line, 50 drugs, 3 seeds. Only if time permits.
Total	CPU + 1x MI300X	~2 days within 1-week window	~$40-65 GPU equivalent. Remaining budget for reruns.

Software Dependencies

• Python 3.10+, PyTorch 2.x with ROCm backend.
• scanpy, anndata for single-cell data handling.
• GEARS, CPA codebases (public GitHub repos).
• GenePT embeddings via OpenAI text-embedding-3-large API (one-time extraction, cached).

6. EXPECTED DELIVERABLES

• Architecture-extended dormancy table: DE-Spearman gap for 5+ architectures, cleanly separating dormancy-vulnerable (nn.Embedding) from dormancy-immune (continuous features) models.
• Gene embedding comparison: gap closure for GenePT, Geneformer, ESM-2, Gene2Vec, scGPT tokens, and SVD+GenePT concatenation, with 3-seed CIs.
• Combinatorial dormancy analysis: compounding ratio across GEARS, CPA, and CPG on Norman 0/2, with 3-way stratification (0/2, 1/2, 2/2).
• Conditional-helper replication: per-tertile gap on K562 and RPE1 for CPA, GEARS, and scGPT, confirming architecture-agnostic conditional structure.
• An AAAI-27 paper draft with all figures and tables generated from result JSONs.

7. RISK & MITIGATION

Risk	Likelihood	Mitigation
Biolord shows dormancy despite using continuous features	Medium	Check Biolord's generalization gap separately. If it fails, the story shifts to "continuous features are necessary but not sufficient." Still informative.
All pretrained embeddings perform comparably to SVD (remaining 36% gap is architectural)	Medium	Report as evidence that the CPG architecture, not the feature, is the bottleneck. Reframe from "better features" to "architectural limits."
Norman has too few combos for reliable 3-way stratification (~44 per stratum)	Medium	With 3 seeds this gives ~132 observations per stratum. Use median split (2 groups) if tertiles are underpowered.
GenePT embeddings degrade on poorly-annotated Replogle genes	Low	Check annotation coverage before running. Also run at 128-dim and 256-dim to verify PCA reduction does not destroy signal.
SAMS-VAE sparsity mask complicates the audit	Low	Ablate both mask and embedding jointly. Set mask to all-ones for counterfactual conditions.