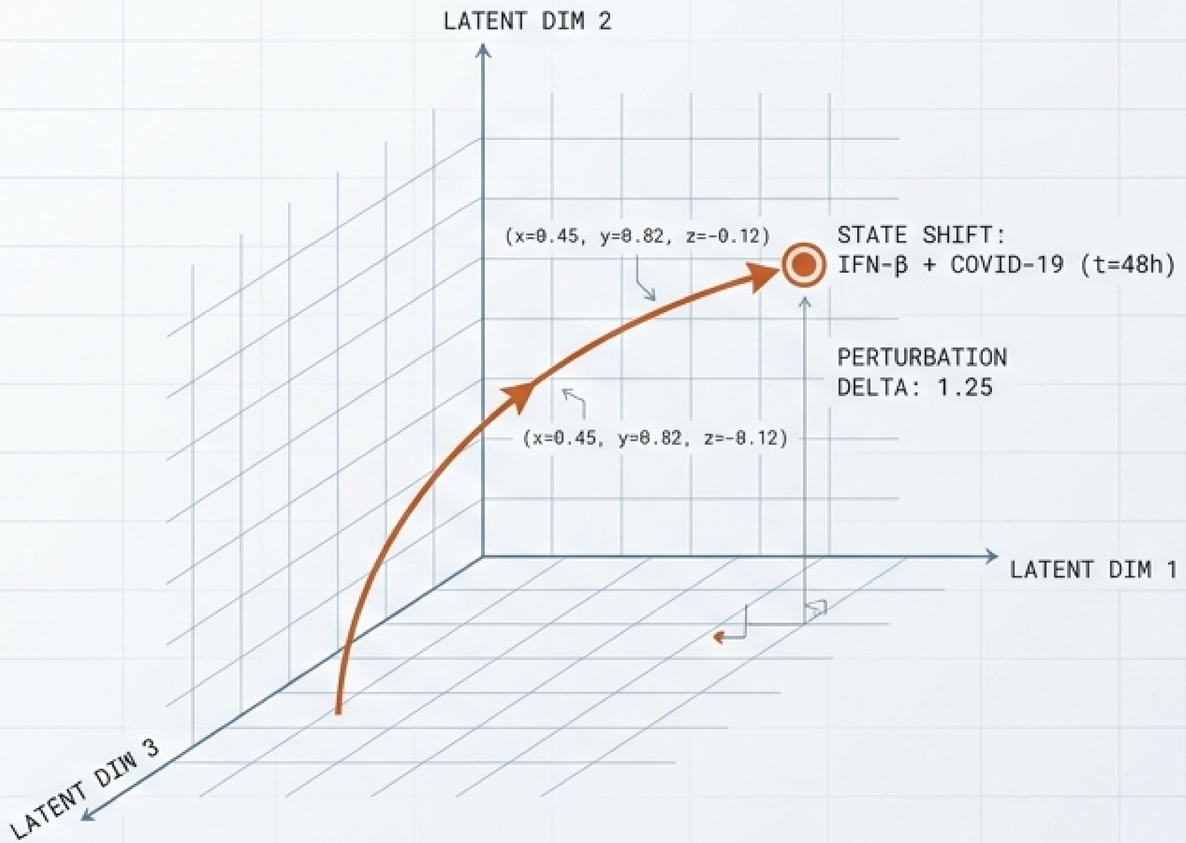


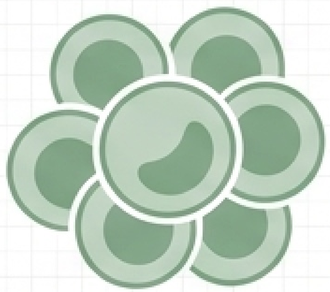
Mini Virtual IFN Immune Cell

A perturbation-aware single-cell virtual cell prototype that learns interferon and COVID immune-state transitions

Shivaprasad Patil

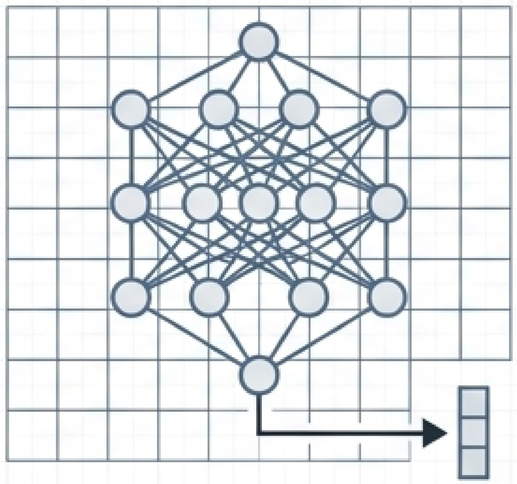
Executive Summary

Input (The Goal)



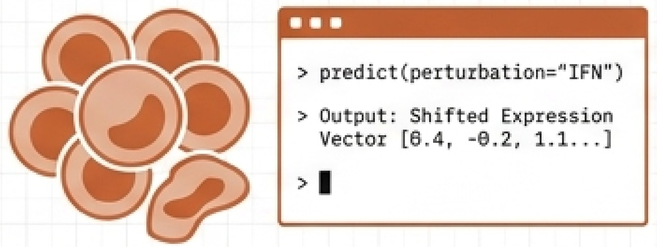
Learn how immune cells transition through latent state spaces under interferon and COVID-19 perturbations.

Engine (The Method)



Combine a foundation-model encoder with a learned latent perturbation vector and magnitude calibration.

Output (The Result)



A lightweight prototype featuring a simple inference API that predicts directionally meaningful, biologically interpretable gene expression shifts.

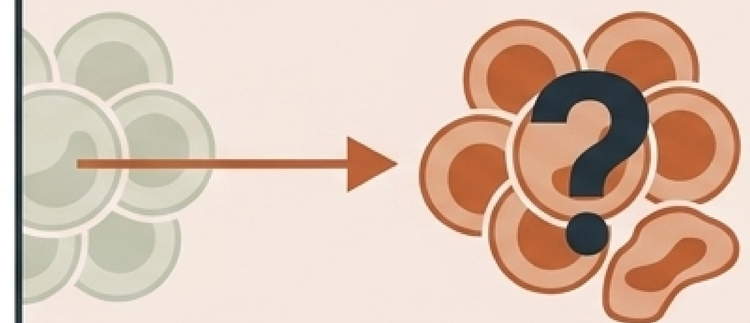
The Core Scientific Question

Observed Baseline

```
CELL_ID: C_001
STATE: Baseline
EXP_PROFILE:
  GENE_A: 1.24
  GENE_B: 0.85
  GENE_C: 2.10
  ...
```

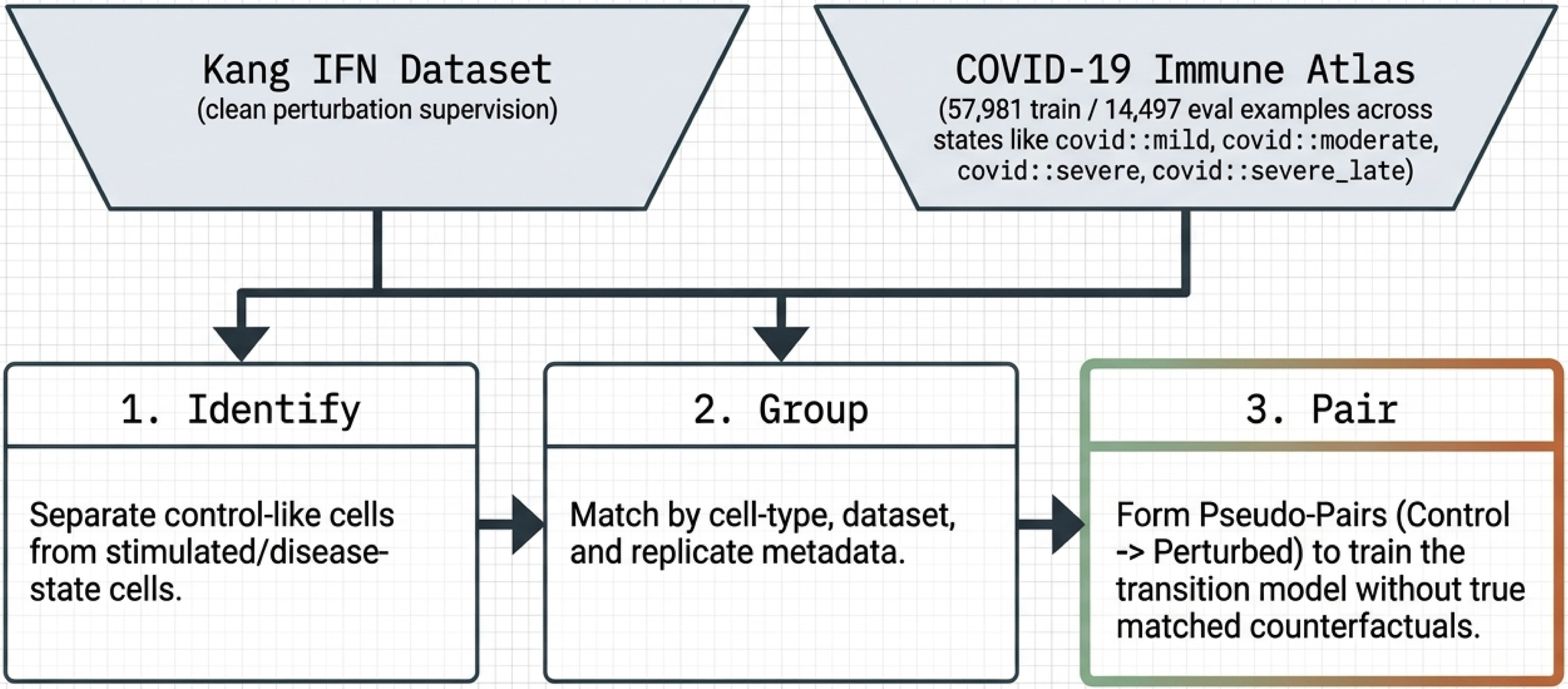
Given a cell in a baseline state, what would its expression profile look like under a specific perturbation?

Unobserved Counterfactual

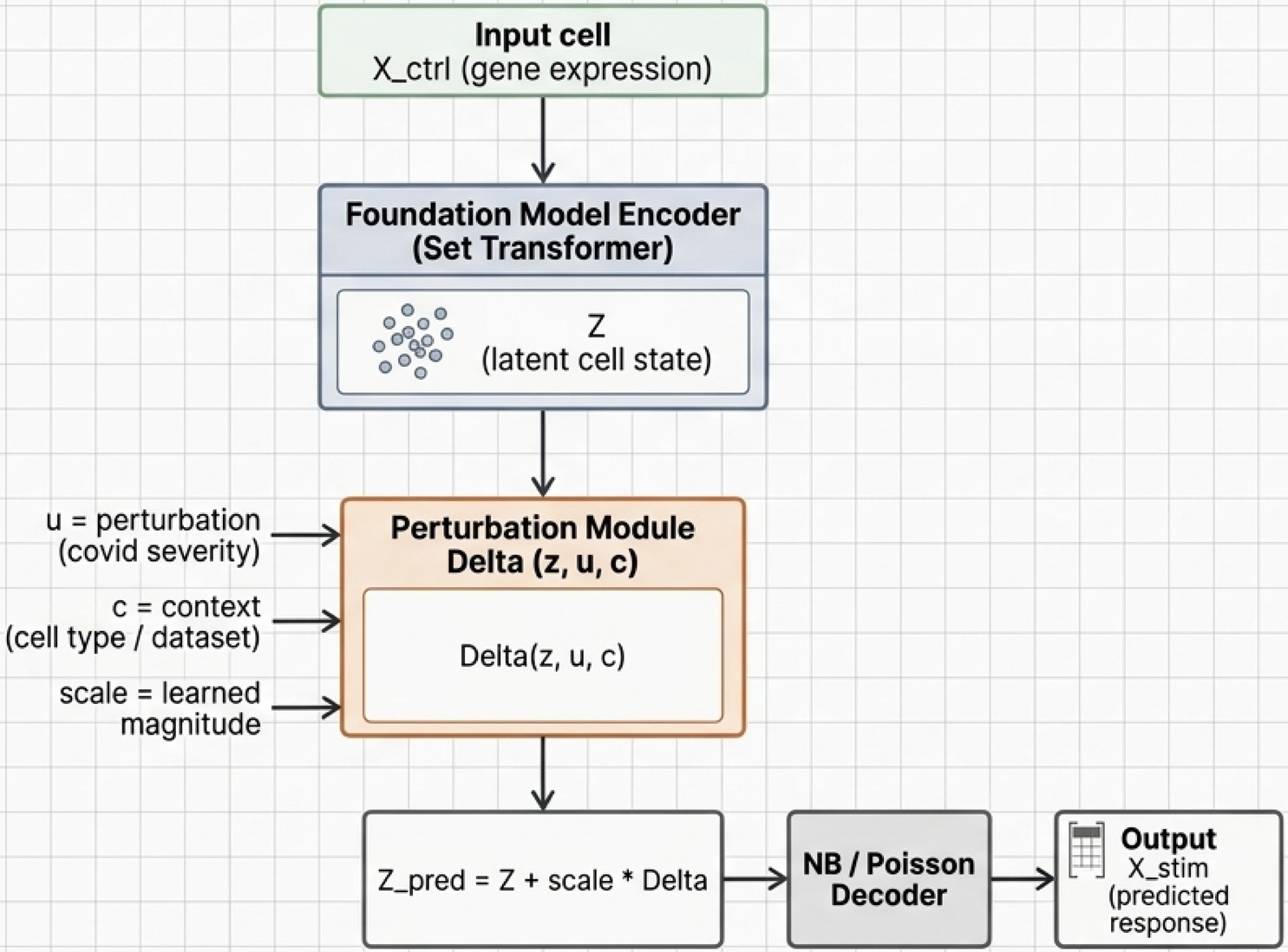


Traditional models rely on generic class labels. The Virtual IFN Cell shifts to structured, continuous latent shifts, capturing both the biological program and its intensity.

Data Engine: Constructing Pseudo-Pairs



Architecture: The Journey of a Cell



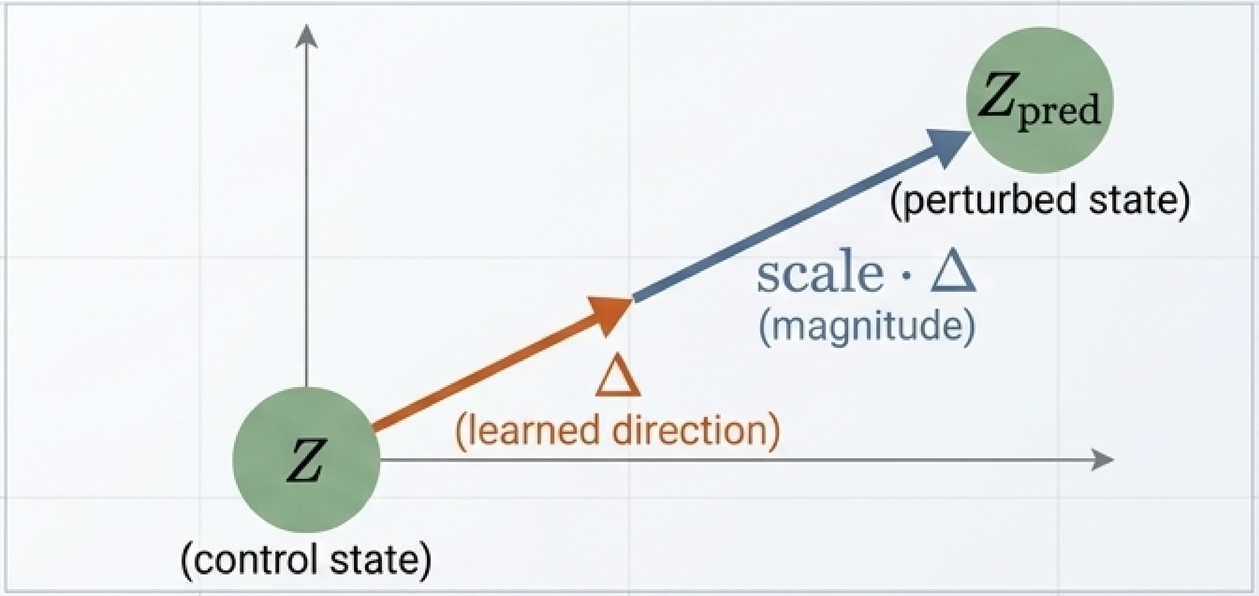
1. Input: Control Cell Gene Expression.

2. FM Encoder: A FiLM-conditioned Set Transformer. Tokenizes 5,000 top genes via K_top (512) and K_rand (512) embeddings into a latent state (Z).

3. Perturbation Module: A context-aware network applies the perturbation (u). Conditioned on dataset and cell type.

4. Decoder: Negative-Binomial decoder maps the predicted latent state (Z_pred) back to expected gene counts.

The Latent Perturbation Mechanism



Direction (Delta)
The biological program or response axis induced (e.g., IFN activation).
Represents WHERE to move.

Magnitude (Scale)
The strength of the response for a specific cell/context. Cell-aware (sigmoid-valued), preventing blind amplification.
Represents HOW FAR to move.

Key Idea: By separating direction and magnitude, the model accounts for varying biological intensities across different cell types and disease severities.

Disentangling Identity from State

Latent Vector (Z)

Z_{id} (64-dim)	Z_{state} (64-dim)
The stable cell identity. Remains untouched and frozen during the perturbation.	The dynamic biological state. This is is the only portion transformed by the perturbation network.

```
Z_pred = [Z_id, Z_state + scale * Delta]
```

Takeaway: This formulation **explicitly** prevents identity loss, ensuring the predicted cell remains the same cell type post-perturbation.

Evaluation I: Global Reconstruction Metrics

Tested on 14,497 held-out COVID examples.

Latent Cosine Similarity

0.8713

(High structural retention in latent space)

Mean Delta-Gene Correlation

0.6526

(Stable response direction learned)

Delta MSE

6.0724

IFN Gap Score

0.6103

The model successfully generalizes to unseen cells, proving that a stable latent response direction is learned and maintained outside the training set.

Evaluation II: Biological Pathway Validation

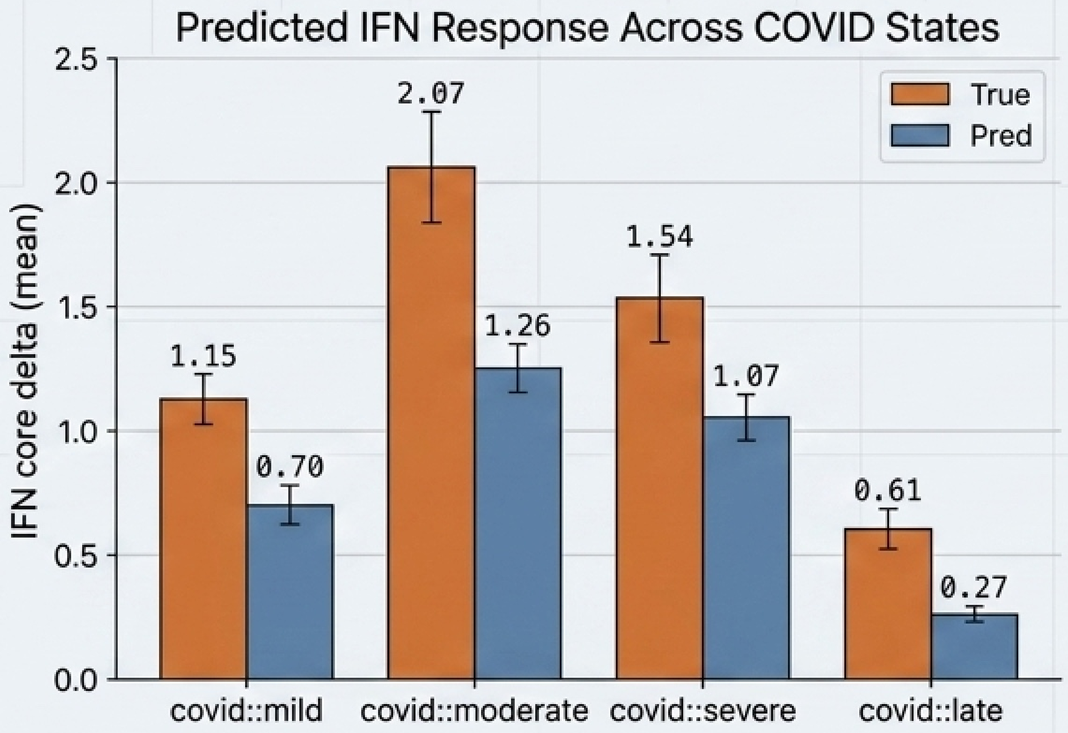
Diagnostic Comparison Matrix

Pathway	Direction Match	Gene Corr / Gap	Result
IFN Core	1.00 (Perfect)	Gene Corr: 0.98	Strong Success
IFN Extended	0.94	Gene Corr: 0.94	Strong Success
Myeloid Inflammation	0.37	Gap: 0.83	Weak/Failing
Cytotoxic	0.16	Gap: 0.75	Weak/Failing

Note: Top-50 response gene overlap is 0.54. The v1 model is a highly effective IFN specialist, perfectly capturing target biology, but struggles with unrelated pathway directions.

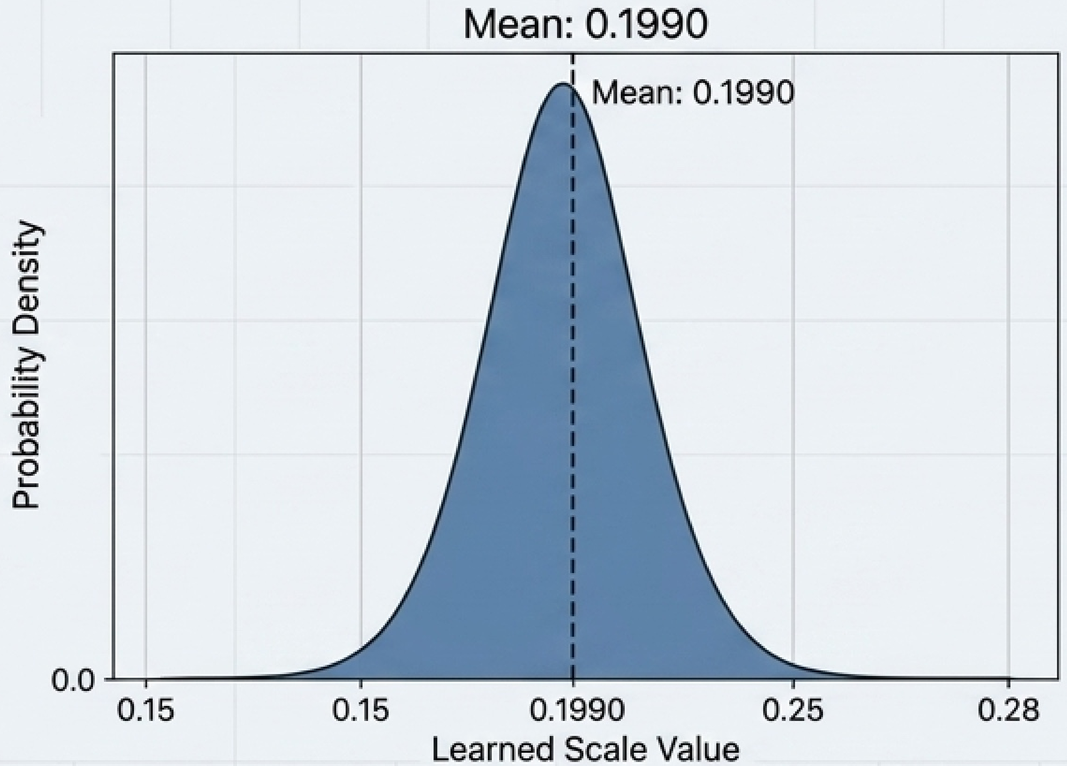
Diagnostics: Severity Progression & Learned Scale

Severity isn't Monotonic



Real-world clinical severity labels do not form a perfectly straight line of IFN activation. The model accurately tracks this non-linear biological reality.

Conservative Scale



Mean scale is 0.1990. The model resists aggressive, unrealistic amplification, heavily prioritizing biologically safe, conservative magnitude shifts.

Application: The Real-Cell Inference API

The model is not trapped in a training script. The lightweight VirtualIFNCell wrapper allows direct perturbation simulation on real control-cell vectors.

- **predict_response**: Apply a named perturbation (e.g., covid::severe_late) to a baseline control cell.
- **compare_perturbations**: Test multiple clinical severities on a single cell to simulate progression.
- **delta_summary**: Extract top up-regulated and down-regulated gene targets instantly.


```
{ "status": "success",
  "cell_id": "control_1A",
  "perturbation": "covid::severe_late",
  "shifts": {
    "gene_A": 0.85,
    "gene_B": -0.42,
    "gene_C": 1.20,
  },
  "delta_summary": {
    "top_up": ["gene_C", "gene_A"],
    "top_down": ["gene_B", "gene_D"]
  }
}
```


Structured Gene Shift Data (CSV Format)

Gene ID	Baseline (Control)	Perturbed (Severe)	Delta Shift
gene_C	0.20	1.40	↑ +1.20
gene_A	0.50	1.35	↑ +0.85
gene_B	0.80	0.38	↓ -0.42
gene_D	0.40	0.25	↓ -0.15

The Reality Check: v1 Limitations

Not a Clinical Model

Designed for research and prototype exploration, not for patient diagnostics or clinical severity prediction.

Not a Causal Simulator

Relies on observational pseudo-pairs. It is an associative latent transition model, not a true causal engine.

Underestimated Magnitude

While directionally accurate, the learned scale is systematically conservative, limiting the dynamic range of predicted expressions.

Exploratory STAT1 Knockout

Attempted causal gene-intervention (STAT1 KO) shows extremely weak suppression (<0.01). Causal behavior is currently inactive.

The v2 Roadmap: Toward a Causal Simulator

Explicit Data Integration

Move beyond pseudo-pairs by ingesting real perturb-seq and explicit IFN/JAK/STAT knockout datasets.

Pathway Guidance

Implement pathway-weighted losses to heavily anchor the model in known biological mechanics.

Magnitude Calibration

Develop stronger priors to safely expand the dynamic response range without hallucination.

Refined Latent States

Explicitly separate disease severity vectors from inflammatory state vectors for higher resolution control.

Open source. Built for iteration.
Available on GitHub.