Perturbation-MMKPNN

Perturbation-MMKPNN: Interpretable Modeling of Single-Cell Perturbation Responses
Sally Yepes (2025)

---

1. Problem

Single-cell perturbation datasets are revolutionizing biology by revealing how cells respond to drugs and genetic interventions. However, current models such as scGen and CPA treat this task as black-box prediction, leaving fundamental questions unanswered: Which pathways are rewired by perturbations? Which subnetworks consistently mediate resistance across cell types? Explanations are often unstable across seeds or datasets, limiting biological trust and translational use.

---

2. Innovation

Perturbation-MMKPNN introduces **concept bottlenecks** into perturbation modeling. Expression features and perturbation metadata are forced to propagate through biologically curated modules — pathways, transcription factors, regulatory subnetworks — derived from Reactome, DoRothEA, and MSigDB. Predictions are therefore mediated by interpretable latent variables, transforming the model into a transparent mapping: *perturbation → regulatory program → transcriptional outcome*. This design stabilizes interpretability and grounds predictions in biological priors.

---

3. Validation

The framework will be validated across multiple datasets — scPerturb, Perturb-seq, L1000, DrugComb — to ensure robustness. Benchmarking against scGen, CPA, and linear baselines will test predictive accuracy and generalization. Novel metrics will include **attribution stability** (repeatability of bottleneck activations across runs) and **cross-dataset transfer** (train on one perturbation set, test on another). Biological case studies will identify conserved regulators of resistance, supported by counterfactual experiments silencing bottleneck nodes to simulate pathway inhibition.

---

4. Contribution

Perturbation-MMKPNN moves perturbation modeling beyond prediction to **mechanistic discovery**. By systematizing interpretability, reproducibility, and robustness, it provides a framework that can be adopted as a **benchmark standard** for single-cell perturbation analysis. Scientifically, it enables identification of pathways mediating drug resistance and

synthetic lethality. Methodologically, it demonstrates how concept-bottleneck models can stabilize explanations, setting a precedent for causal, reproducible AI in biology.