From Prediction to Causation: Active Experiment Design for Robust Discovery

Discovering reliable cause–effect relations in biology requires not only large-scale data but also principled experimental design. Existing active learning approaches typically prioritize predictive accuracy when selecting new perturbations, which can result in redundant measurements and limited causal insight. Prior work has shown that carefully chosen interventions can drastically reduce the number of experiments needed to identify causal relations, and that active strategies can improve causal structure discovery. Building on this foundation, we propose a framework for **active causal experiment design** that explicitly prioritizes new experiments by their ability to reduce **causal uncertainty**.

Our methodology introduces **causal acquisition functions** based on the variance of average treatment effect (ATE) estimates. At each iteration of an active learning loop, candidate drug–cell perturbations are ranked according to expected reduction in ATE variance, and the top-ranked perturbation is selected. This approach contrasts with prediction-driven strategies, which rely on expected error reduction, and directly ties experiment selection to improved causal identification. The design is inspired by recent work in causal bandits but tailored to high-dimensional multimodal biological data.

To evaluate the framework, we simulate active design using public datasets that integrate **chemical structures (SMILES), morphological profiles (Cell Painting), and transcriptomic signatures (LINCS L1000)**. Ground-truth causal effects are approximated from held-out perturbation experiments. We compare causal acquisition strategies against standard active learning baselines on metrics including speed of mediator recovery, accuracy of effect estimation, and robustness under constrained experimental budgets.

Preliminary results indicate that causal acquisition outperforms predictive criteria in identifying true mediators and stabilizing effect estimates across modalities. Beyond drug discovery, this framework generalizes to domains such as **synthetic biology, personalized healthcare, and climate science**, where experimental resources are limited and causal inference is essential. By aligning experiment selection with causal objectives, our work provides a path toward more efficient, autonomous experimental design.