Master Thesis

michavol

March 2025

1 Data

1.1 Data-Generating Mechanism

Consider the observational (or unperturbed) data distribution P_X, where $X = \{X_1, ..., X_d\}$ is a $d-$dimensional random vector. Further, let $\mathcal{S} = \{s_1, ..., s_m\}$ be the set of possible atomic interventions (or perturbations) to the system. Then, we can denote n samples drawn from the interventional distribution $P_X^{do(\{s_i\})}$ as $X^{s_i} \in \mathbb{R}^{n,d}$ and from $P_X^{do(\{s_i,s_j\})}$ as $X^{s_i,s_j} \in \mathbb{R}^{n,d}$. The latter describes samples from a system that has been intervened on with s_i and s_j simultaneously. Our goal is to find an estimate $\hat{X}^{s_i,s_j}$, relying only on X^{s_i} and X^{s_j}. To train a suitable estimator, we assume observations from some, but not necessarily all possible combinations of interventions. Note that the number of possible intervention pairs $\{s_i, s_j\} \in \mathcal{S} \times \mathcal{S}$ grows quadratically with m. In the following, we describe three datasets which follow such a data-generating mechanism.

1.2 Transcriptomics & CRISPR-Mediated Activation

This dataset captures a large-scale exploration of gain-of-function genetic interactions in K562 cells, employing a CRISPR activation (CRISPRa) approach for single and pairwise gene perturbations. Specifically, **?** considered 112 genes whose overexpression significantly impact the growth of K562 cells, a lymphoblast cell line derived from the bone marrow of a 53-year-old chronic myelogenous leukemia patient. Using CRISPR interference (CRISPRi) they systematically screened for gain-of-function gene interactions (GIs) between any two of the 112 genes (6216 possible pairs). To study promising GIs in greater detail, they performed a Perturb-seq analysis on 132 selected pairs of genes using CRISPR activation (CRISPRa) systems. They also overexpressed all genes individually to allow for direct comparison of individual and pairwise perturbations. This resulted in transcriptional readouts for 287 perturbations measured across approximately 110,000 single cells with a median of 273 cells per condition. We preprocess the transcriptomics data according to the steps established by **?**. This includes normalizing and log-transforming the data using SCANPY and selecting 5,000 highly variable genes (HVGs).

In terms of the data-generating mechanism from above, $X^{s_i} \in \mathbb{R}^{n,d}$ represents transcriptional reads for $d = 5000$ HVGs across all n cells following overexpression of gene s_i, while X^{s_i,s_j} captures the joint effect of simultaneously activating genes s_i and s_j.

1.3 Proteomics & Drug Perturbations

We derived this dataset from a large-scale screen in which **?** exposed 16 triple-negative breast cancer (TNBC) and 2 non-TNBC cell lines to a set of FDA-approved small-molecule inhibitors, both individually and in pairs. They seeded cells in 96-well plates, treated them with varying concentrations of each drug (0.5, 5, 50, 200 μM), and measured cell viability using a CCK8 proliferation assay. From these viability profiles, they determined drug efficacies and approximate IC50 values, which enabled them to select effective drugs for further proteomic analysis. For each single-drug perturbation, they used a final concentration of 10 μM to lyse cells and extract proteins, following established protocols (**??**). For combination treatments, they selected concentrations according to a comprehensive assay of 2025 drug combinations (**?**). Next, they analyzed the resulting peptides via data-independent acquisition (DIA) on a TripleTOF mass spectrometer, obtaining quantitative readouts of thousands of proteins. In total, the authors collected several hundred proteomic profiles covering both single and pairwise drug perturbations, as well as control (untreated) conditions. They applied a stringent quality control pipeline: discarding low-coverage samples, imputing missing values at a fraction of the minimum detected intensity, and assessing reproducibility using Pearson correlation and the coefficient of variation. After some more post-processing, they obtained expression levels of 5585 proteins under each perturbation condition (63 single drugs, 58 drug combinations, controls) across multiple cell lines and time points (0h, 6h, 24h, and 48h).

Each drug perturbation s_i can be regarded as an atomic intervention on the system, and drug combinations s_i, s_j jointly alter the proteomic landscape. Accordingly, $X^{s_i} \in \mathbb{R}^{n,d}$ denotes the measured proteomic features for $d = 5585$ proteins across n replicates following a single-drug treatment, whereas $X^{s_i,s_j} \in \mathbb{R}^{n,d}$ represents samples under the combined effects of drugs s_i and s_j. As with the previous dataset, only a fraction of all possible pairwise interventions is practically feasible to measure, underscoring the need for predictive models that can generalize to untested perturbation combinations.

1.4 Simulated

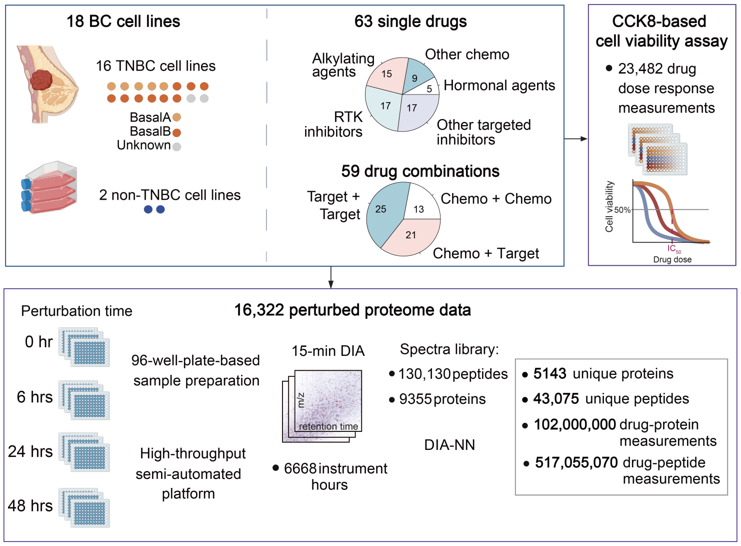

Figure 1: Details for Proteomics Data