interact. Ouyang, Tan, and Xu (2023) develop a sophisticated method to estimate hidden confounding from data in the setting of generalized linear models. For the same reasons as for the regression setting, it would be interesting to see whether one can apply ideas similar to the spectral deconfounding to classification objectives as well. There seems to be no literature in this direction yet.

Potential models we could consider $\eta \leftarrow f^0(X) + \delta^T H$ and then transform η to a binary response either with $Y \sim Ber(p_x)$ with $p_x = \frac{1}{1+e^\eta}$ or one could also consider $Y = \text{sign}(\eta + \nu)$. Interesting would be a model of the log-odds $\log(\text{odds}_y) = f(X) + \delta H$, which would give us an interpretable causal effect of X. It seems not immediately clear how to adjust the likelihood criteria to get an unbiased estimator for the causal effect in this setting.

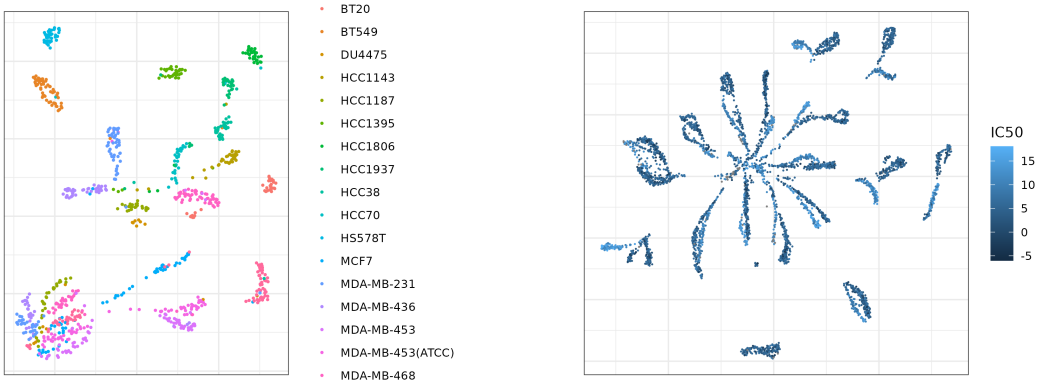
2 Proteomics

Joint work with:
Tiannan Guo: Guomics Laboratory for Proteome Complexity Science `https://guomics.com/`
Peter Bühlmann, Xinwei Shen, Roberto Desponds (Master student), Michael Vollenweider (potentially Master student), Marin Sola (potentially)

Single-drug treatments are extensively studied and tested to prove that they help fight breast cancer. In practice, however, a group of experts decides on a combination of drugs based on personalized information. This project aims to help in this decision by using additional information about the expression levels of the different proteins. Using the gained understanding of how the various drugs affect protein levels and what proteins play a role in fighting breast cancer, we might get closer to a foundation model in drug discovery for breast cancer.

2.1 Data

Guomics Laboratory provides proteomic data regarding breast cancer treatment (Gillet et al., 2012; Sun et al., 2023). We have data from 18 different protein plates defining the initial distribution of 5519 protein expressions at time point zero. The protein plates are treated with varying drug perturbations and measured after 6, 24, or 48 hours. See Figure 4 for a visualization. We have data for 63 individual drugs at a fixed concentration of 10 μmol. In addition to the individual drugs, we have data from experiments with 58 combinations of two drugs with varying concentrations. Figure 5 shows an



(a) Initial protein expression

(b) After 48 hours

Figure 4: Visualization of initial protein expression and the protein expression after 48 hours using UMAP. The initial protein expression in (a) is coloured using the different protein plates and the protein expressions after 48 hours in (b) are coloured according to the IC50 values.

example of the support for combining Lapatinib Ditosylate Hydrate and Olaparib. For most drug combinations, however, we only see marginal support in the data. For each treatment protein plate combination, we know the IC50 value. IC50 measures the drug concentration required to kill 50% of the cells. Therefore, a lower IC50 value corresponds to a better health outcome. Figure 6 shows the distribution of this response. Overall, around 60% of the protein expression values are missing in the data. These missing are not at random but indicate an expression that is too low so that the measurement techniques do not pick up on anything. Our current strategy is to impute the missing values by $0.8 \min(x_j)$ for each of the j proteins. The encoding of low-level proteins does not matter in the case of tree-based models, but we need to keep this in mind for other model classes. The overall process of how the different parts affect each other is shown in Figure 7. The protein plate initialises various protein expressions and evolves over time, perturbed by the drug treatment. The open question is how the proteins should behave at different times to provide optimal health outcomes. Not shown in the graph are different hidden confounders, such as other non-measured proteins.

2.2 Drug combination discovery

In the following, we denote the health outcome measured by IC50 by $Y \in \mathbb{R}$. $P_t \in \mathbb{R}^{5519}$ with $t = 0, 6, 24, 48$ encodes the protein expressions at time t and the sparse $D \in \mathbb{R}^{63}$ encodes the concentration of drugs used. We are interested

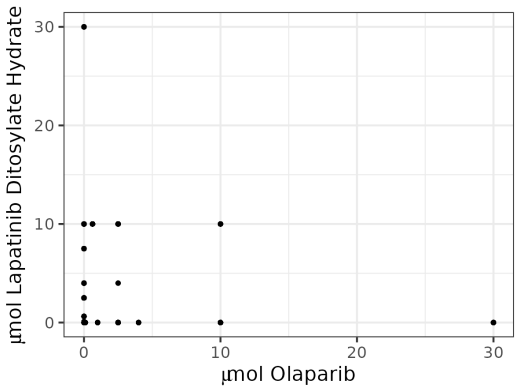


Figure 5: The Concentration of Olaparib and Lapatinib Ditosylate Hydrate in μmol used in the experiments.

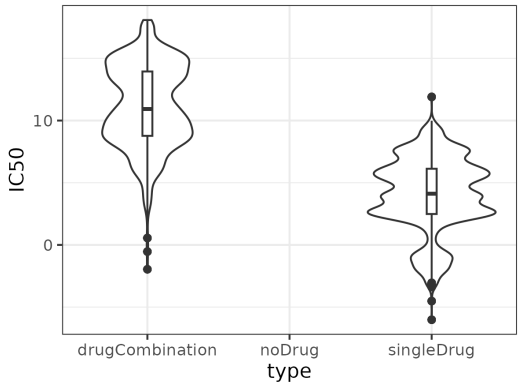


Figure 6: Distribution of IC50 values for single drug treatment and drug combinations.

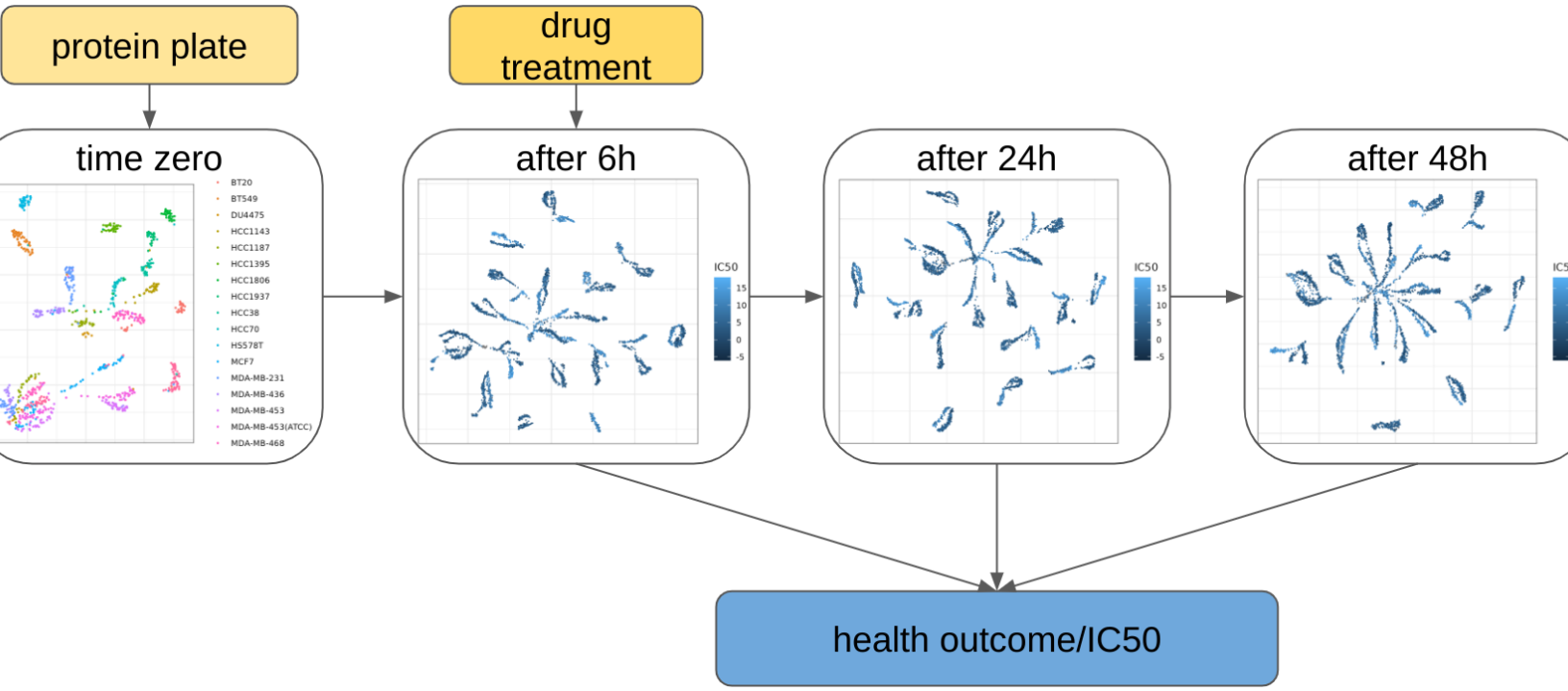


Figure 7: Overview of the evolution of the protein expressions over time with the main effects. The protein expressions at the different time points are visualized using UMAP.

in the interventional distribution $p^{do(D:=d)}(Y|P_0)$ of the health outcome Y if we change the treatment D given the initial protein expression in a patient P_0. Consider

$$p(Y|D,P_0) = \int p(Y|P_6,P_{24},P_{48},D,P_0)p(P_6,P_{24},P_{48}|D,P_0)dP_6dP_{24}dP_{48}$$

Since we assume that D is a root node, we also have

$$\begin{aligned} p^{do(D:=d)}(Y|P_0) &= \int p^{do(D:=d)}(Y|P_6,P_{24},P_{48},P_0)p^{do(D:=d)}(P_6,P_{24},P_{48}|P_0)dP_6dP_{24}dP_{48} \\ &= \int p(Y|P_6,P_{24},P_{48},P_0,D=d)p(P_6,P_{24},P_{48}|P_0,D=d)dP_6dP_{24}dP_{48}. \end{aligned}$$

Thus, also

$$\mathbb{E}^{do(D:=d)}[Y|P_0] = \int \mathbb{E}[Y|P_6,P_{24},P_{48},P_0,D=d]p(P_6,P_{24},P_{48}|P_0,D=d)dP_6dP_{24}dP_{48}.$$

We thus break up the problem in modelling:

- $\mathbb{E}[Y|P_6,P_{24},P_{48},P_0,D=d]$ being invariant w.r.t. D (Section 2.4)
- distributional model for $p(P_6,P_{24},P_{48}|P_0,D=d)$ (Section 2.5)

2.3 Interaction of drugs

According to Tiannan's group, the drugs do not have an additive effect on health outcomes; they interact with each other. For example, one could combine two drugs with a low IC50 and get a combination with a high IC50. We see in Figure 8 that this seems to be the case for most of the combinations tried. We can think of the two models $\mathbb{E}[P_t|P_0,D] = f(P_0,D)$ or $\mathbb{E}[P_t|P_0,D] = \sum_{j=1}^{63} f_j(P_0,D_j)$ and estimate both of these model using the data. The first model will fit the data better, but it would be informative to understand how important interactions of different drugs are in the model and whether some drug combinations benefit more from the interactions than others. (Xinwei has more results on that, and Roberto is also looking at significant interaction terms in the linear model.)

2.4 Prediction of IC50

We assume that the drugs affect the health outcome via the observed proteins. So what we want to estimate is $E[Y|P_6,P_{24},P_{48},P_0,D=d]$ being invariant w.r.t. D. We have, in total, around 15 thousand protein expression

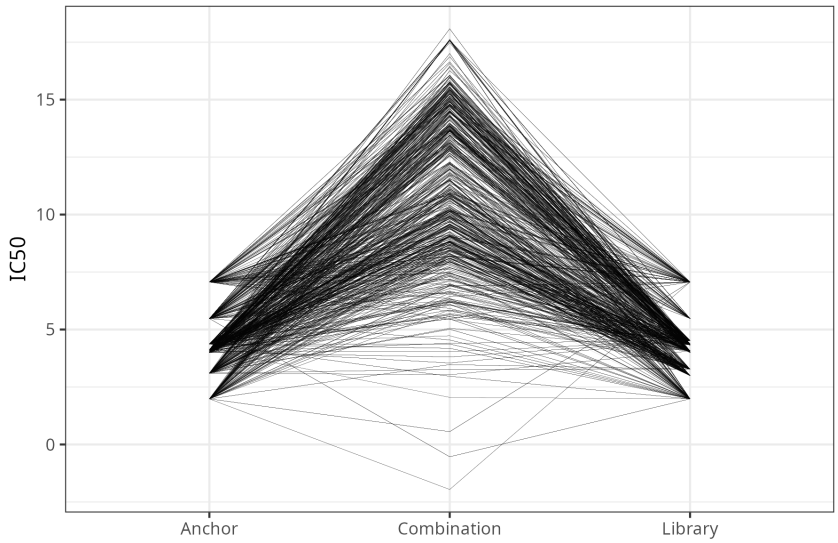


Figure 8: IC50 value for the different drug combination experiments. The Anchor IC50 values correspond to the mean over the IC50 values of single drug treatments using the Anchor part of the drug combination. Library corresponds to the mean over the IC50 values of single drug treatments using the Library part used as a single drug treatment.

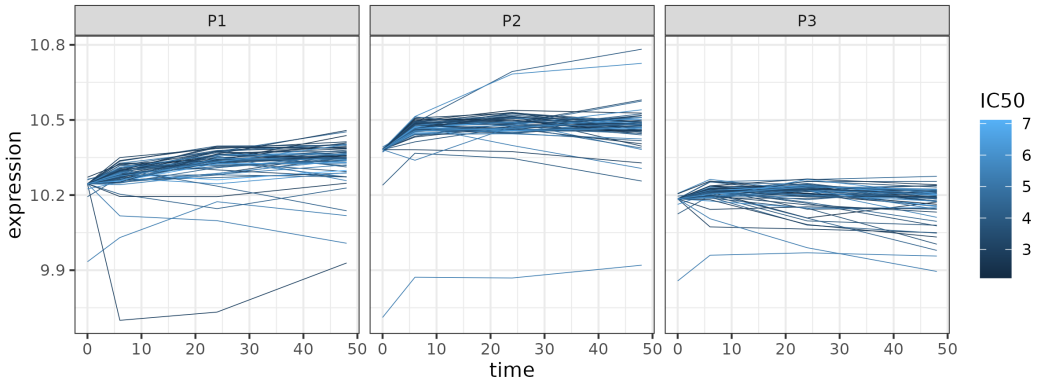
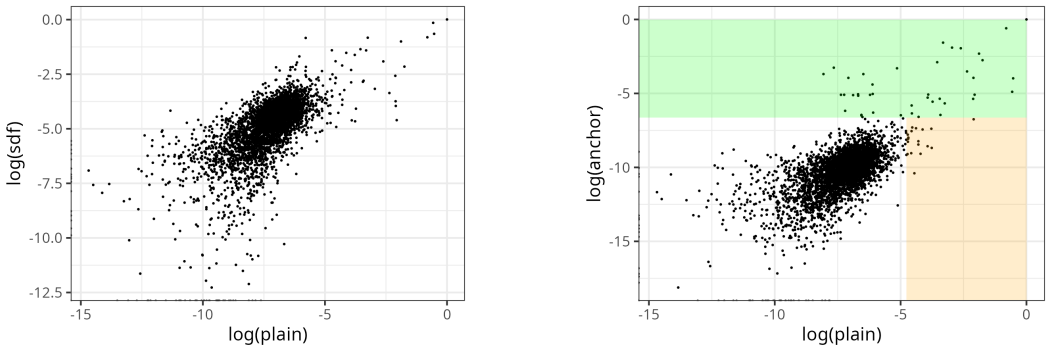


Figure 9: Temporal graph of three example proteins. The mean expression over different protein plates for the different drug perturbations is shown against the time after drug intervention. The temporal expressions are coloured according to the corresponding IC50 value.

experiments. Around 13 thousand are experiments with drug perturbations and measurement after one specific time point. We do not have cell death measurements for individual experiments, only theoretical values (from other experiments) for protein plate drug perturbation combinations. This means for a specific protein plate and drug treatment, there is no variance in Y. A caveat is that we do not have measurements of the same experiment after different time points. For a given protein plate drug perturbation combination, we have three individual experiments measured after 6 hours, 24 hours, and 48 hours. To get single observations with protein expressions at different time points, we aggregate them for each time point, protein plate, and perturbation using the median over the expression values. This results in 1600 usable observations, where 1100 correspond to single-drug treatments. Figure 9 shows three examples of protein expressions over time. If we would only need, for example, $p(Y|P_6, P_0, D = d)$, we would already have three times the observations to work with. (Even though the response for these three observations would be the same.) We work with the log-transformed values of the highly skewed protein expression. The change in protein expression instead of the absolute log expressions is more informative. Some protein expressions rise in the first 6 hours to some level and stay the same until the end of the experiments.

We use the differences between the different time points to model IC50. We estimate models for IC50 using different Random Forest versions. We chose Random Forest because it is easy to estimate a nonlinear relation with interactions. In addition, it has a nice way of dealing with the imputed non-observed



(a) Deconfounded Random Forest (b) Anchor Random Forest

Figure 10: Comparison of the log variable importance for the deconfounded random forest and the anchor random forest against the plain random forest. In (b), green is the area with importance above the 99% quantile for the invariant model. The area with high association but not high invariant importance is orange.

protein expressions. In the regression trees, the non-observed imputed values simply get their partition. We estimate the Random forest using the mean squared objective, spectral deconfounding, and anchor regression using (Ulmer & Scheidegger, 2024). Figure 10 compares the aggregated importance of the different proteins for the different models. In comparing anchor regression with the plain Random Forest in Figure 10b, we colour the area with proteins that have a high association with IC50 but not high importance in the invariant model in orange. The area with invariant proteins is coloured green. We expect the green proteins to be interesting for our project and denote them as P^I.

In Figure 11, we compare the out-of-bag error distribution for different new drug combination environments against the out-of-environment error distribution, where we left each environment out for the training (we observe all the other drug combinations in training). Even though we see interesting differences in the variable importance of the three different models, there does not seem to be a strong distribution shift when going to new drug combinations.

2.5 Prediction of protein expression

distributional model for $p(P_6^I, P_{24}^I, P_{48}^I | P_0, D = d)$

- linear model (Roberto)

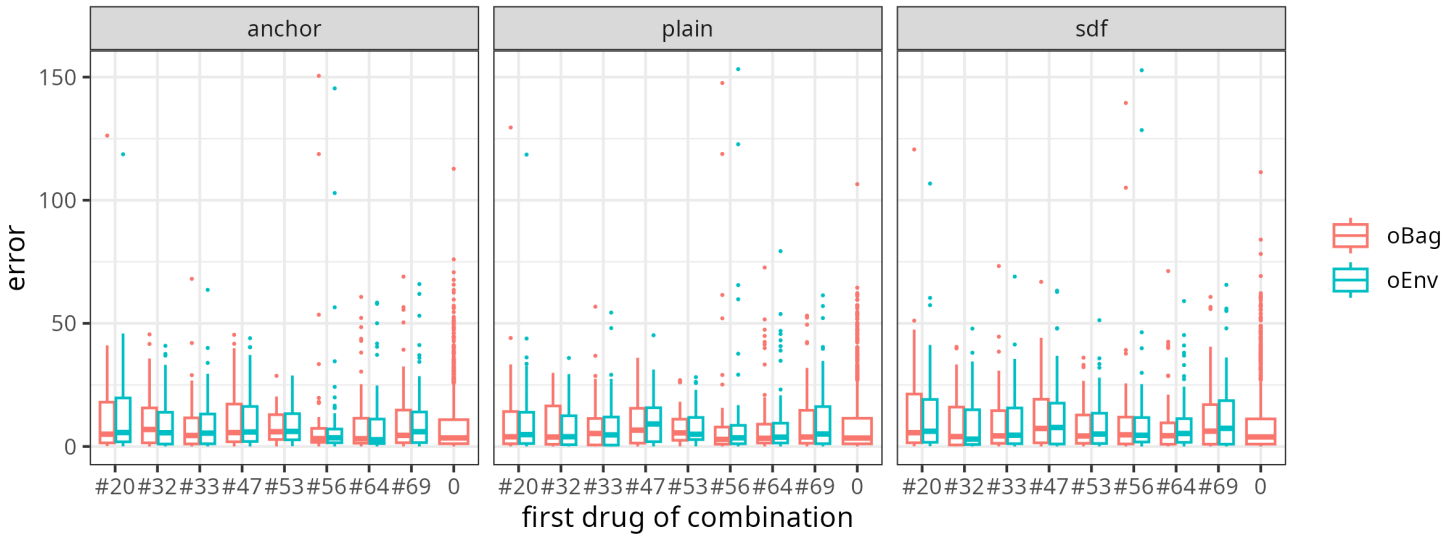


Figure 11: Comparison of Out-of-bag error with Out-of-environment error for plain Random Forest, Anchor Random Forest, and SDForest. On the y-axis are the different environments denoted by the first drug of a drug combination. Environment 0 corresponds to all single-drug treatments.

- how much does performance decrease when predicting new drug combinations
- applying extrapolation techniques

2.6 Estimation of causal subgraph

How does the sparse set of important protein P^I interact, and can we estimate the corresponding sub-DAG?

2.7 Foundation model, personalized medicine, ...

It's far away, but can we use the gained understanding of the dependencies between proteins and health outcomes for drug discovery in general? And can we use a patient's initial protein expression to find a personalized treatment?

3 Applied projects

Maybe with Cyrill
Potential application in machine health, forest ecosystems, or Aviation with