Dose-response modeling in high-throughput cancer drug screenings: A case study with recommendations for practitioners

Wesley Tansey*[1,2], Kathy Li[3], Haoran Zhang[3], Scott W. Linderman[1,4], Raul Rabadan[2], David M. Blei[1,4,5], and Chris H. Wiggins[1,3]

[1]Data Science Institute, Columbia University, New York, NY, USA

[2]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA

[3]Department of Applied Mathematics and Applied Physics, Columbia University, New York, NY, USA

[4]Department of Statistics, Columbia University, New York, NY, USA

[5]Department of Computer Science, Columbia University, New York, NY, USA

Abstract

Personalized cancer treatments based on the molecular profile of a patient's tumor are becoming a standard of care in oncology. Experimentalists and pharmacologists rely on high-throughput, *in vitro* screenings of many compounds against many different cell lines to build models of drug response. These models help them discover new potential therapeutics that may apply to broad classes of tumors matching some molecular pattern. We propose a hierarchical Bayesian model of how cancer cell lines respond to drugs in these experiments and develop a method for fitting the model to real-world data. Through a case study, the model is shown both quantitatively and qualitatively to capture nontrivial associations between molecular features and drug response. Finally, we draw five conclusions and recommendations that may benefit experimentalists, analysts, and clinicians working in the field of personalized medicine for cancer therapeutics.

*wt2274@cumc.columbia.edu (corresponding author)

1 Introduction

1.1 High-throughput cancer drug screening

Genomic sequencing and high-throughput drug screening is becoming cheaper, enabling widespread adoption in both research institutions and hospitals (Muir et al., 2016). Many of these institutions have built datasets that enable scientists to explore potential connections between the molecular profile of a cell (i.e. its DNA and other related biological information) and its phenotypic response to treatment with a certain drug. Specifically in cancer therapeutics, large public datasets have become available with thousands of experiments testing different drugs on different types of cancer cell lines (e.g. Yang et al., 2012; Barretina et al., 2012; Haverty et al., 2016).

One goal in analyzing these datasets is to build a predictive model of drug response. The model takes in a set of molecular features of a cell line and predicts the expected outcome of using a candidate drug to treat it. The more accurate the predictor, the more it can be trusted to faithfully simulate wet lab results. If a good predictor can be constructed, it can accelerate the discovery of targeted therapies by refining experimentalists' hypotheses much faster than can be done in the lab. Thus, good predictors have high value to biologists.

We build a generative model of drug response in high-throughput cancer drug screenings. The model captures uncertainty inherent in cell line experiments, including measurement error, natural variation in cell growth, and drug response heterogeneity. We take an empirical Bayes approach to estimating the model parameters and also propose a method for detecting contaminated data. We use the model in a case study analyzing a dataset (Yang et al., 2012) containing hundreds of thousands of experiments. In our study, the model outperforms a state-of-the-art approach for estimating dose-response curves (Vis et al., 2016). Model predictions also recapitulate known biology involving nonlinear interactions between molecular features and drugs.

The experimental design used in our case study is similar to many pharmacogenomic profiling experiments. Consequently, we expect the proposed model and insights from our case study to be applicable to other high-throughput cancer drug screening studies. In the conclusions of the paper, we summarize these insights and propose several recommendations for practitioners analyzing other datasets, designing new experiments in the lab, or guiding data-gathering policy at hospitals.

1.2 Experimental setup and details

Our data come from the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2012; Garnett et al., 2012), a high-throughput screening (HTS) study on therapeutic response in cancer cell lines. The GDSC data comprise the results of testing 1072 cancer cell lines nearly-combinatorially against 265 cancer therapeutic drugs, *in vitro*. The experiments were conducted across two separate testing sites: the Wellcome Trust Sanger Institute ("Site 1") and Massachusetts General Hospital ("Site 2"). Experiments were carried out over the course of multiple years and used one of two assay types, depending on the cell type: suspended ("Assay S") or adherent ("Assay A"). For a given site and assay type, each experiment was carried out using a series of HTS microwell plates, each laid out as shown in Fig. 1.

A given plate contains hundreds of microwells, where each well is designated as either a negative control, positive control, or treatment. Negative control wells are left unpopulated and untreated, so as to calibrate the base level of machine output if all cells were to die. All other wells are populated with a constant volume of cells; due to natural variation in cell

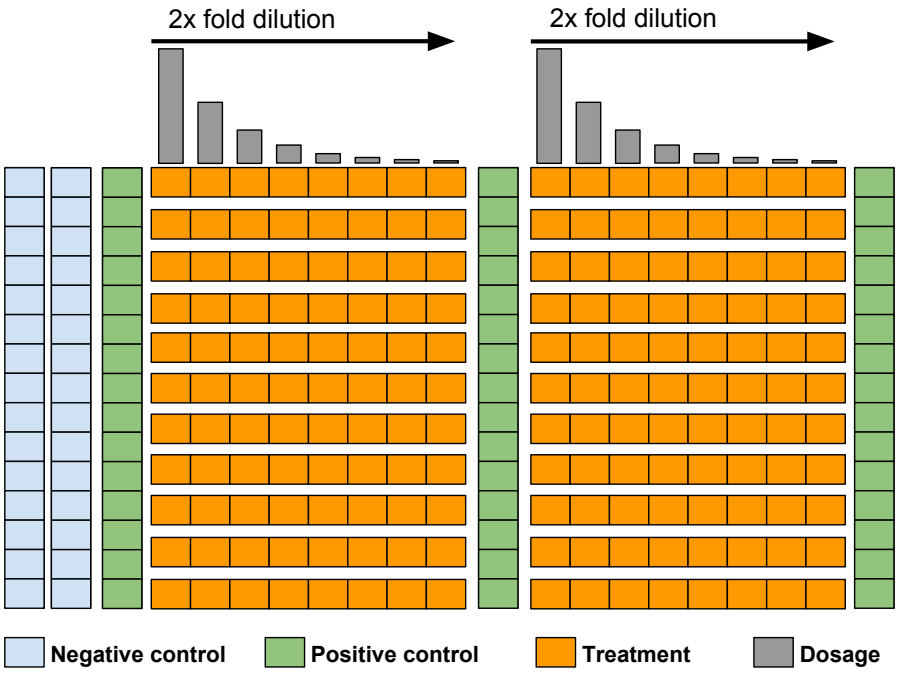


Figure 1: Layout of the high-throughput screening plates used in the GDSC experiment for a 9-level dosage schedule. Negative controls (blue) are unpopulated and untreated; positive controls (green) are populated but untreated; treatment wells (orange) are treated with one drug per row, with each well in a row treated at a different concentration.

volume, exactly how many cells occupy any given well is unknown. Positive control wells are left untreated so as to measure the base response for drugs that have no effect.[1] Finally, a series of drugs (one per orange row in Fig. 1) are applied to the remaining wells, with each well receiving a different concentration. A single plate is used to test only one specific cell line; all wells in a plate are filled with the specific assay fluid appropriate for the target cell line. Drug concentration ranges were derived based on previous experiments and were delivered at either 5 or 9 dosage levels, depending on the testing site. Both testing sites start at the same maximum dosage level, which is diluted at a 4-fold rate for the 5 dosage schedule and at a 2-fold rate for the 9 dosage schedule, resulting in the same final minimum dosage. Since the sites use the same minimum and maximum dosage, this dilution procedure yields missing data for odd-numbered dosage levels in the dataset. Once treated, cells are left to either grow or die for 72 hours.

Cell population size is approximated by a fluorescence assay. A fluorescent compound is added to the wells that is capable of penetrating cells and binding to a protein kept at relatively-constant levels in all living cells. When a cell dies, its structure breaks down and the target protein denatures, leaving no binding agent and resulting in no fluorescence. Robotically-controlled cameras photograph each well and the total luminescence of the image (i.e. pixel intensity count) is used as a relative measure of cell population size. Luminescence of positive and negative control wells is used to calibrate luminescence of treatment wells -

[1]The terms "negative" and "positive" control here are used differently from their common usage in biology. However, this is the terminology used by Garnett et al. (2012); we follow their usage.

that is, positive control wells provide an estimate for how much a well with all living cells will fluoresce while negative control wells provide an estimate for how a well with no living cells will fluoresce. We refer to this negative control measure as the baseline fluorescence bias, since it represents the fluorescence of an empty well. This baseline measure is also subject to machine or technical error from the specific equipment being used. Positive controls are subject to natural biological variation from the particular cell line being used. Population size after treatment, relative to the positive and negative controls, is the quantity of interest for each treatment microwell.

We aim to build a model that treats the molecular covariates as features that potentially convey predictive information about sensitivity and resistance to different therapies. For 963 of the cell lines, we have molecular information about gene mutations, copy number variations (CNVs), and gene expression. We preprocessed the mutations and CNVs, as described in Appendix A, to filter down to genes that are recurrently observed as altered in large-scale observational cancer studies. After preprocessing, we are left with 5822 binary gene mutations and 234 copy number counts; we keep all 17271 gene expression covariates. For a handful of cell lines (109), no molecular information is available; we treat these as missing data. Almost all drugs have been screened against all cell lines, yielding a dataset of 225385 (cell line, drug) experiments. We treat all missing experiments and all missing molecular features as missing at random.

2 A generative modeling of drug response

Consider N cell lines and M drugs. Each cell line $i = 1, \ldots, N$ is tested against each drug $j = 1, \ldots, M$, at dosage levels $t = 1, \ldots, D$. The study consists of plates $\ell = 1, \ldots, L$; we denote by $\ell(i, j)$ the plate on which a specific (i, j) pair was tested. The result of a plate experiment is fluorescence count data $(\mathbf{r}_\ell, \mathbf{q}_\ell, Y_\ell)$ where $\mathbf{r}_\ell = (r_{\ell 1}, \ldots, r_{\ell R})$ are the R negative control measurements used to estimate the baseline fluorescence bias, $\mathbf{q}_\ell = (q_{\ell 1}, \ldots, q_{\ell Q})$ are the Q positive control measurements used to estimate the fluorescence of a population of cells when no effective treatment has been applied, and Y_ℓ are the treatment well measurements for each of the drugs on plate ℓ. Since each (i, j) pair is tested only on a single plate, we index treatments by their cell line, drug, and dosage levels, respectively; thus, $\mathcal{Y} \in \mathbb{R}^{N \times M \times D}$ is a 3-tensor and y_{ijt} represents the result of treating cell line i with drug j at dosage level t. Each cell line i has an associated vector of mutation, copy number variation, and expression covariates X_i.

We propose a fully generative model of dose-response,

$$
\begin{aligned}
r_{\ell k} &\sim \text{Poisson}(c_\ell) \\
q_{\ell k} &\sim \text{Poisson}(\lambda_\ell + c_\ell) \\
y_{ijt} &\sim \text{Poisson}(\tau_{ijt}\lambda_{\ell(i,j)} + c_{\ell(i,j)}) \\
c_\ell &\sim \text{Gamma}(v_\ell, w_\ell) \\
\lambda_\ell &\sim \text{Gamma}(a_\ell, b_\ell) \\
\tau_{ijt} &= \frac{1}{1 + \exp(-\beta_{ijt})} \\
\boldsymbol{\beta}_{ij} &\sim \text{GP}^+(\boldsymbol{\mu}_{ij}, \Sigma) \\
\boldsymbol{\mu}_{ij} &= f^+(X_i, j; \theta)\,.
\end{aligned}
\tag{1}
$$

The plate-specific negative and positive control fluorescent counts, $r_{\ell k}$ and $q_{\ell k}$, are modeled as random variables for which we receive R and Q i.i.d. observations, respectively. The positive control fluorescence distribution is a function of the baseline fluorescence from the machine (c_ℓ) and the growth rate (λ_ℓ) of the cell line on plate ℓ. The outcome of experiment y_{ijt} is a sample from the positive control distribution, but with a treatment effect τ_{ijt} that reduces the growth rate of the cells.

The treatment effect τ_{ijt} has the direct interpretation as drug j killing $(1 - \tau_{ijt}) \times 100\%$ of cell line i on average when applied at dosage level t. The vector of responses $\boldsymbol{\tau}_{ij}$ is the dose-response curve – that is, it represents the percentage of cells that survive at different dosage levels; this is the primary quantity of interest in every experiment. Discovering effective drugs corresponds to finding curves that show sensitivity in some subset of cell lines, indicating the therapy is likely targeting some molecular property of the cells.

The dose-response curve is modeled nonparametrically through a latent constrained Gaussian process (GP). The logistic transform from the GP variable β_{ijt} to the effect τ_{ijt} constrains effects to be in the $[0, 1]$ interval, corresponding to expected cell survival percentage. The GP is constrained to the monotone-increasing half-space, encoding that the drug effect can only become stronger as the dosage increases. The mean response is a monotone function f of the molecular features for the target cell line and the ID of the drug to be applied.

The model for $\boldsymbol{\tau}$ encodes several scientific assumptions. We assume no drug has a positive effect on any cell – that is, no drug actually encourages growth. There are two biological motivations for this assumption. First, cancer cells are generally defined by uncontrolled proliferation. Cultivating non-cancerous cells *in vitro* is extremely difficult as most cells induce apoptosis (cell suicide) outside of their host environment. There is little room left biologically for a drug to encourage the cancer cell lines to grow even more. Second, the drugs chosen for the GDSC experiment have all been selected for their ability to stress and kill cells. A large portion of the drugs are established cancer therapeutics that are designed to kill cells by targeting pathways recurrently found altered in certain cancers. In addition to assuming all drugs do not encourage growth, we also assume that toxicity only increases with drug concentration. This assumption is again based on the notion of cancer drugs being highly toxic. We check these assumptions empirically in Appendix B.

3 Modeling details: contamination, batch effects, and empirical Bayes

3.1 Empirical Bayes estimation of model parameters

In practice, simultaneous estimation of all parameters in (1) is infeasible computationally. Furthermore, we discovered contamination in the GDSC data that must be handled carefully.

We instead take an empirical Bayes approach to estimation with the following model,

$$\begin{aligned} y_{ijt} \sim &\quad \text{Poisson}(\tau_{ijt}\lambda_{\ell(i,j)} + \hat{c}_{\ell(i,j)}) \\ \lambda_\ell \sim &\quad \text{Gamma}(\hat{a}_\ell, \hat{b}_\ell) \\ \tau_{ijt} = &\quad \frac{1}{1+\exp(-\beta_{ijt})} \\ \boldsymbol{\beta}_{ij} \sim &\quad \text{MVN}^+(\hat{\boldsymbol{\mu}}_{ij}, \widehat{\Sigma}) \\ \hat{\boldsymbol{\mu}}_{ij} = &\quad f(X_i, j; \hat{\theta}) \end{aligned} \tag{2}$$

The prior parameter estimates $(\hat{a}_\ell, \hat{b}_\ell, \hat{c}_\ell, \hat{\boldsymbol{\mu}}_{ij}, \hat{\theta}, \widehat{\Sigma})$ are obtained through a stepwise procedure. At a high level, there are four main steps:

1. Baseline fluorescence bias rate $\hat{c}_\ell$ is estimated from negative controls. These controls contain systematic biases due to technical error; we detail a denoising approach for negative controls in Sections 3.2 and 3.3.

2. Positive control priors $\hat{a}_\ell$ and $\hat{b}_\ell$ are estimated by maximum likelihood for each plate in Section 3.4.

3. The black box predictive model f is chosen to be a neural network; parameters $\hat{\theta}$ are estimated in Section 3.5 by maximizing the marginal log likelihood of the data with fixed control priors and an identity covariance matrix.

4. Dosage correlation structure $\widehat{\Sigma}$ is estimated in Section 3.6 through a rejection sampling procedure with fixed prior means $\hat{\boldsymbol{\mu}}$.

The final model enables full Bayesian posterior inference on $\boldsymbol{\tau}_{ij}$, the entire dose-response curve. The remainder of this section details the four estimation steps in our empirical Bayes estimation procedure.

3.2 Cross-contamination

An assumption in high-throughput cancer drug analyses is that each individual well is independent of the other wells. There is growing concern about this assumption among biologists, due to recent studies suggesting there is substantial spatial bias in microwell assays (Lachmann et al., 2016; Mazoure et al., 2017). We generally label this "cross-contamination" since microwell results are spilling over into neighboring wells. It is unknown whether this spatial bias is due to literal biological contamination by nearby cells or technical issues related to the fluorescence scanning process. These recent studies suggest that when information about the spatial layout of HTS plates is known, the spatial bias maps well to a simple weighted model and existing spatial denoising models can remove this bias (Mazoure et al., 2017).

While bias concerns have been well-investigated in other HTS assays, such concerns are not yet widespread in the cancer research community. Consequently, our high-throughput cancer screening data contains no spatial information about the microwell layout, making it difficult to leverage existing methods of removing spatial bias. Instead, the information available is a unique ID for each positive and negative control well (across plates in the same testing site and assay type); no location information about the treatment wells is available outside the generic plate design. Nonetheless, uncovering systematic spatial bias is still

crucial in the GDSC data, as the negative and positive control wells are positioned adjacent one another. Therefore, any bleeding over of the positive signal could drastically alter the estimate of the baseline fluorescence.

Fig. 2 shows an example of such positive-to-negative control well contamination. In this example, the positive controls are all roughly around $900K$, as indicated by the dashed red mean line at the top. The negative control wells mostly hover around $30K$ but some wells have obviously been contaminated by positive control wells and are wildly different. Estimating the negative control rate as the mean of the wells (dashed gray line) would put the estimate of the bias at approximately 14.5% of the positive control mean. The empirical Bayes procedure we describe below estimates the bias with the contamination effects removed at approximately 3.5%. Any treatments estimated from the mean method would therefore bias the drug effects upward, estimating treatment efficacy to be stronger than was likely true.

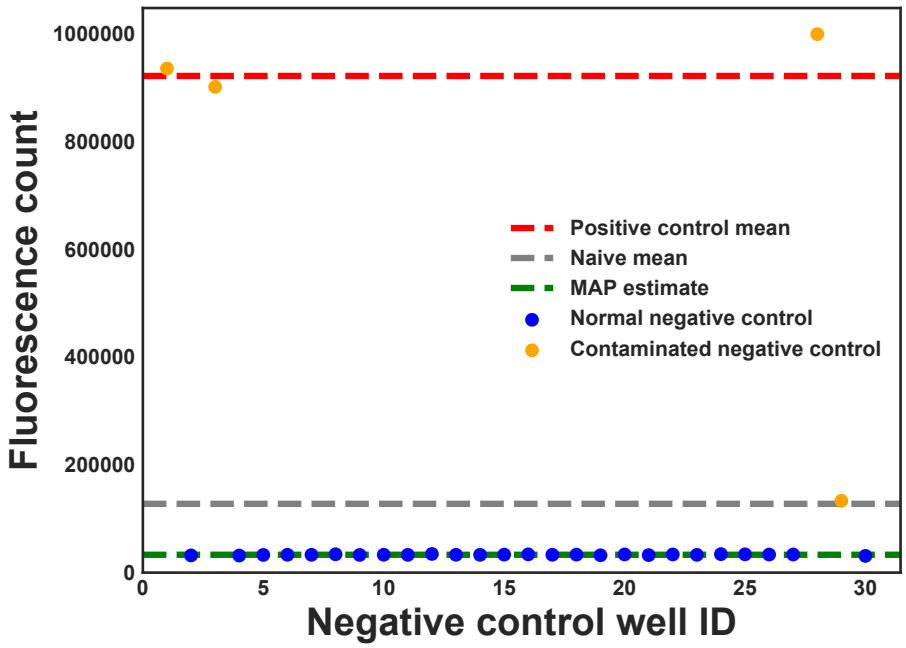


Figure 2: Example microwell contamination for one plate in the GDSC data. Each dot is a different negative control well. The dashed red line at the top is the mean of the positive control wells; the middle gray line is the mean of the negative controls; the bottom line is the negative control estimate after our debiasing steps. The improved estimate accounts for an additional 11% of treatment effect for any cells on this plate.

Without exact layout knowledge, removing the spatial effects is challenging. Fortunately, we have a large number of both negative (≈ 30) and positive (≈ 40) controls, along with well-specific IDs; these IDs consistently map to the same microwell location across all plates for the same assay type and at the same testing site. This enables us to determine which negative control wells in each site and assay stratification are being systematically contaminated. Fig. 3 shows the cross-correlation between all of the control wells in each of the four stratifications; different stratifications have different numbers of control wells. In each subplot, the upper left block represents the correlation structure of the negative controls, the lower right represents the positive controls, and the off-diagonal blocks represent

the correlation between negative and positive wells.

Every site shows evidence of at least some contamination, with non-zero entries for many of the off-diagonal blocks. In particular Site 1, Assay S (1S) shows evidence of a large degree of contamination, with the majority of the negative wells being correlated with the positive wells. As a conservative measure, we remove from the dataset any negative wells whose maximum positive well correlation is greater than 0.15 in magnitude. For three of the four testing sites, less than 10 of the negative wells are dropped; for 1S, only 7 negative controls remain.

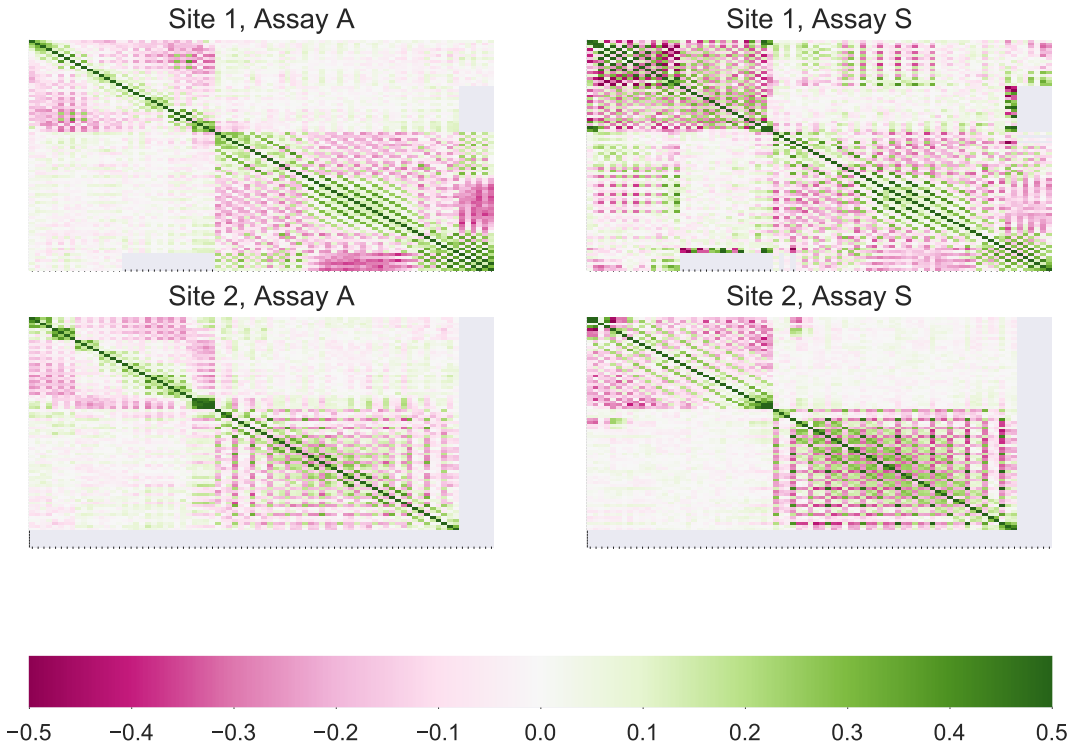


Figure 3: Evidence of potential cross-contamination or other systematic biases in the GDSC data. The four figures show control microwell correlation across all plate experiments, stratified by testing site and assay type. The top left corner in each subplot contains the negative control wells; the bottom right contains the positive control wells. Site 1, Assay S in particular shows clear signs of correlation between the two well types, as denoted by the large off-diagonal correlations.

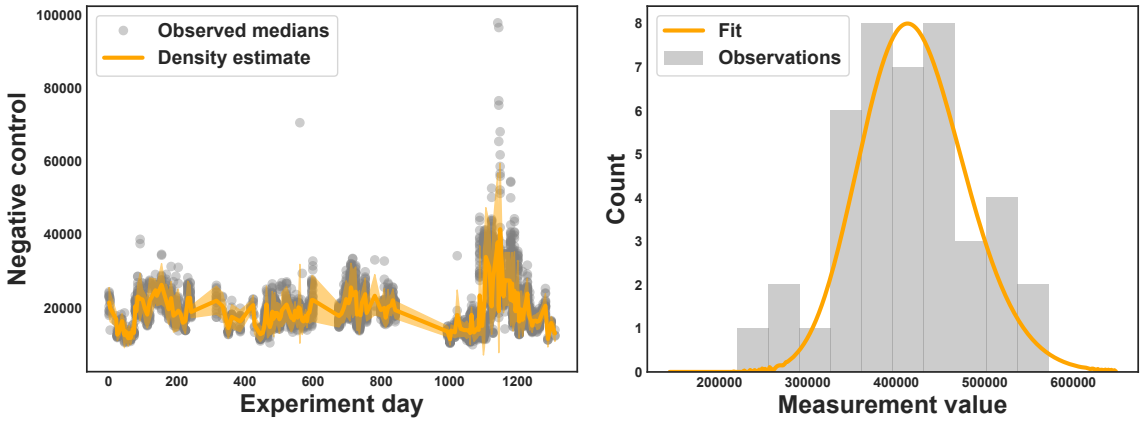
3.3 Temporal batch effects

Separate from the independent-well assumption is the assumption of independent plates. The assumption is that each plate of observations is independent, regardless of when or where it was tested; as we show, this assumption is also violated for the GDSC data. Unlike the cross-contamination issue, however, this temporal dependency is well-established in the literature (Johnson et al., 2007; Leek et al., 2010) and falls broadly under the term "batch effects." These are technical artifacts that cause otherwise-independent experiments to yield dependent results. Myriad causes can introduce temporal batch effects: using the same

test site, the same equipment, preparation by the same technician, or even conducting the experiment at the same ambient temperature. Generally batch effects are seen as a nuisance that reduces the signal in experimental data and must be removed as a preprocessing step, if possible.

In the case of negative control estimation, we use the temporal batch effects to our advantage. Dropping most of the wells in 1S reduces power to estimate the negative control mean; even in stratifications where most wells were kept, the sample size is still fairly low (≈ 20). If temporal dependencies exist between negative controls, we can regain some power by sharing statistical strength between plates run at similar times.

Fig. 4 (left) shows one of the four (site, assay) stratifications, with negative control medians in gray.[2] The x-axis in each subplot is the date ID of the plate (the date when the plate was screened); all dates are relative to the first day of the study. There is clear visual evidence of temporal dependence, with trends followed by sharp discontinuities.



(a) Negative control uncertainty

(b) Positive control uncertainty

Figure 4: Left panel: Time-evolving estimate of the negative control density for an example testing site and assay type. The x-axis is the date a plate was screened, relative to the start of the study; y-axis is the median control well measurements after removing contaminated wells. The trend filtering density regression fit is in orange (line: mean, bands: 90% regions). Right panel: Maximum likelihood estimate of an example positive control density, after fitting the negative control MAP estimate.

We leverage this dependence through an empirical Bayes procedure that shrinks the differences between median estimates of negative control wells for plates screened on similar days. We use the relevant portion of the generative model for a single negative control well,

$$\begin{aligned} r_{\ell k} &\sim \quad \text{Poisson}(c_\ell) \\ c_\ell &\sim \quad \text{Gamma}(v_{d(\ell)}, w_{d(\ell)}) \,, \end{aligned} \tag{3}$$

where $d(\ell)$ denotes the specific day d that plate ℓ was screened. We use the median observation $\widetilde{c}_\ell$ as a noisy approximation to the true rate. We then fit a time-evolving density

[2]We use medians rather than means to avoid spurious contamination not removed in Section 3.2.

to model the prior distribution of the medians on each day,

$$\underset{\mathbf{v},\mathbf{w}\in\mathbb{R}^+}{\text{minimize}} \quad -\sum_d \sum_{\ell\in\{\ell:\ d=d(\ell)\}} \log(\text{Gamma}(\widetilde{c}_\ell; v_d, w_d)) + \rho_1 \left|\left|\Delta^{(1)}\frac{\mathbf{v}}{\mathbf{w}}\right|\right|_1 + \rho_2 \left|\left|\Delta^{(0)}\frac{\mathbf{v}}{\mathbf{w}^2}\right|\right|_1 , \tag{4}$$

where $\Delta^{(k)}$ is the k^{th}-order trend filtering matrix (Tibshirani, 2014) using the falling factorial basis (Wang et al., 2014) to handle the irregular grid of days. The solution to (4) finds densities that are piecewise-linear in their mean and piecewise constant in their variance.

The regularization parameters ρ_1 and ρ_2 are chosen via 5-fold cross-validation and the model is fit with stochastic gradient descent. Since the counts are large, we use a normal approximation to the gamma, parameterized in natural parameter space for computational convenience; shape and rate parameters are reconstructed from the mean and variance of the normal. Since (4) is non-convex, we only find a local optimum, but empirically we observe good fits across a wide array of simulated data. The orange line and bands in the left panel of Fig. 4 show the results for one stratification, where the outliers have been shrunk substantially.

Given the learned prior from (4), we calculate the *maximum a posteriori* (MAP) estimate for the negative control mean,

$$\hat{c}_\ell = \frac{\hat{v}_{d(\ell)} + \sum_k r_{\ell k}}{\hat{w}_{d(\ell)} + \sum_k 1}\,. \tag{5}$$

While there is technical variation in the bias rate, it is relatively small after corrections and thus as a practical matter we simply use the MAP estimate for the Poisson rate of the negative controls.

3.4 Natural variation in cell line growth

Even under ideal conditions without any batch effects or spatial plate bias, cell population growth and response exhibits a large degree of natural variation. For instance, Fig. 4 (right) shows the distribution of the positive control wells for one example plate. The variance in this cell line is so large that a decrease in fluorescence of even 50% compared to the mean would not be highly unlikely. Furthermore, population growth variance varies substantially between cell lines, testing sites, and assay types.

Since we have a reasonably large number of positive control wells on each plate, we estimate the population directly,

$$\underset{a,b\in\mathbb{R}^+}{\text{maximize}} \prod_q \int \text{Poisson}(q; \lambda + \hat{c}_\ell)\text{Gamma}(\lambda; a, b)d\lambda\,. \tag{6}$$

The integral in (6) can be resolved analytically to be an incomplete gamma; however, we found a finite grid approximation to be more numerically stable. The maximum likelihood problem is also nonconvex, so we again rely on a local optima approximation found via a sequential least squares solver. We found the fits in simulation to be close to the ground truth when using the same number of control replicates as in the GDSC data. The orange line in the right panel of Fig. 4 shows the resulting fit from optimizing (6) on the example plate.

3.5 Fitting the black box response prior

We use a deep neural network for f, parameterized by weights θ. The architecture uses a 300-dimensional linear projection of X_i and a 100-dimensional embedding of each drug. The concatenated 400-dimensional vector is then passed through a $200 \times 200 \times 9$ ReLU network. For cell lines with missing molecular features, we learn an embedding from the cell line ID to the same 300-dimensional space. The 9 outputs $\boldsymbol{\phi}_{ij}$ are then constrained to be monotone,

$$\mu_{ijt} = \phi_{ij9} + \sum_{9>k\geq t} \log(1 + \exp(\phi_{ijk}))\,, \tag{7}$$

where the right-hand side of (7) is the cumulative sum of the softplus operator applied to the raw outputs from dosage t upward, allowing the maximum dosage level to set the offset. We fix $\widehat{\Sigma}$ to the identity matrix; we found the results did not improve by using off-diagonal terms.

We optimize θ by maximizing the log-likelihood of the data,

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{i=1}^{N}\sum_{j=1}^{M} \log p\left(y_{ij} \mid f(x_{ij};\theta), \widehat{\Sigma}, \hat{a}, \hat{b}, \hat{c}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} \log \int \left[\prod_t p\left(y_{ijt} \mid \beta_{ijt}, \hat{a}, \hat{b}, \hat{c}\right)\right] p(\beta_{ij} \mid f(x_{ij},\theta), \widehat{\Sigma})\,\mathrm{d}\beta_{ij}\,,\end{aligned} \tag{8}$$

where

$$p\left(y_{ijt} \mid \beta_{ijt}, \hat{a}, \hat{b}, \hat{c}\right) = \int \mathrm{Po}(y_{ijt} \mid \sigma(\beta_{ijt})\lambda_{ijt} + \hat{c})\,\mathrm{Ga}(\lambda_{ijt} \mid \hat{a}, \hat{b})\,\mathrm{d}\lambda_{ijt}\,.$$

We approximate the inner integral in (8) with a numeric grid over the values of λ_{ijt}.

We split the data into 10 cross-validation (CV) folds. For each fold, we use 90% of the other folds as training and 10% as validation in early-stopping. We optimize (8) using RMSprop (Tieleman and Hinton, 2012) for 50 epochs with 100 samples per mini-batch. We check empirical risk on the validation set after every epoch and keep the best model over the entire run. The final model predictions on the held out test fold are then used for evaluation of the prior in Section 4.

3.6 Estimating dosage covariance

After fitting $\hat{\theta}$, we reuse the validation set to estimate the covariance matrix $\widehat{\Sigma}$. We randomly sample 1000 experiments from the validation set and estimate an approximate marginal distribution for $\widehat{\Sigma}$ via MCMC with a weakly informative inverse Wishart prior. We evaluated two MCMC methods: 1) a fully-conjugate Gibbs sampler implemented via Polya-Gamma augmentation (Polson et al., 2013) where the sampled posterior MVN logits are projected to be monotone using the pool adjacent violators algorithm as in (Lin and Dunson, 2014), and 2) rejection sampling with an elliptical slice sampling (Murray et al., 2010) proposal.

We found the Gibbs sampler to have high sample complexity due to the need to sample both λ_{ijt} and τ_{ijt}, where the value of one tightly constrains the distribution over the other. The elliptical slice sampler, by contrast, approximated the posterior better with fewer samples and only required a single MCMC chain. We ran the elliptical slice sampler for 2000

iterations with the first 1000 iterations discarded as burn-in samples. The average covariance matrix over the remaining 1000 samples is then used for evaluation of the posterior in Section 4.

4 Model comparison and evaluation

4.1 Baseline approach

A typical approach to modeling fluorescent count experiments follows a pipeline approach. At each step in the pipeline, data is processed into a more refined stage that simplifies downstream analyses. Here we describe the current state of the art used for the GDSC dataset. The pipeline is similar to those used in other high-throughput cancer drug screening analyses (e.g., Barretina et al., 2012).

First, negative and positive controls are averaged to obtain $\bar{r}_\ell$ and $\bar{q}_\ell$. These are point estimates of baseline fluorescence when all cells die or survive, respectively, on plate ℓ. The control point estimates are then used to calculate the expected percentage of cells that survived each treatment experiment,

$$\tilde{\tau}_{ijt} = \max\left(0,\ \min\left(1, \frac{\bar{q}_{\ell(i,j)} - y_{ijt}}{\bar{q}_{\ell(i,j)} - \bar{r}_{\ell(i,j)}}\right)\right). \tag{9}$$

The $\tilde{\tau}$ estimates are then treated as observations of the percentage of cells surviving. A logistic curve is fit for every (i, j) pair using a multilevel mixed effects model (Vis et al., 2016),

$$\begin{aligned} \hat{\tau}_{ijt} = &\quad \frac{1}{1 + e^{-\frac{t - \beta_1 + b_{1i} + b_{1ij}}{\beta_2 + b_{2i}}}} \\ [b_{1i}, b_{2i}] \sim &\quad \mathcal{N}(0, \Psi) \\ b_{1ij} \sim &\quad \mathcal{N}(0, \sigma^2), \end{aligned} \tag{10}$$

where b_{1i} and b_{2i} enable the model to share statistical strength between drugs tested on the same cell line; the model is fit by maximum likelihood estimation. Curves are summarized by integrating out the dosage parameter t to obtain a summary statistic, such as the estimated concentration required to kill 50% of cells (IC50). Log-IC50 values are used as targets for predictive modeling of molecular features. For each drug, an elastic net (Zou and Hastie, 2005) model is fit with hyperparameters chosen through cross-validation.

4.2 Performance comparison

We compare the Bayesian model in Eq. (1) to the above pipeline approach. To compare the two methods, we consider the task of imputing a missing dosage given observations of the other dosage levels in the experiment. This checks the ability of the model to capture the shape of the dose-response curve. For each (cell line, drug) experiment, we hold out one dosage level at random and treat it as missing data that must be imputed. Error is measured in terms of variance-adjusted raw count values on the held out data,

$$e_{ijt} = \left(\frac{y_{ijt} - \hat{y}_{ijt}}{\sigma(\mathbf{q}_{\ell(i,j)})}\right)^2, \tag{11}$$

	(min)	Dosage level							(max)	
Model	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**All**
Pipeline	2.7	2.8	3.4	3.56	4.19	4.7	4.66	6.76	13.84	5.26
Hybrid	2.82	2.8	3.54	3.47	3.96	4.27	4.19	**6.31**	14.22	5.16
Bayesian model	**2.49**	**2.40**	**2.43**	**2.44**	**2.8**	**3.7**	**3.57**	6.48	**11.84**	**4.29**

Table 1: Mean squared error results on the single-dosage imputation benchmark. The pipeline model from Section 4.1 is slightly improved by using corrected controls (Hybrid), but overall is not flexible enough to fully model the observed dose-response curves; the Bayesian model outperforms both pipelined approaches.

where $\hat{y}_{ijt}$ is the model prediction and $\sigma(\mathbf{q}_{\ell(i,j)})$ is the standard deviation of the raw positive controls. For the pipeline, raw predictions are backed out from (9) and (10),

$$\hat{y}_{ijt}^{(\text{pipeline})} = \hat{\tau}_{ijt}^{(\text{pipeline})}(\bar{q}_{\ell(i,j)} - \bar{r}_{\ell(i,j)}) + \bar{r}_{\ell(i,j)}\,. \tag{12}$$

For the Bayesian model, we use a MAP estimate of the raw count,

$$\hat{y}_{ijt}^{(\text{Bayes})} = \hat{\tau}_{ijt}^{(\text{Bayes})} \times \hat{a}_{\ell(i,j)} \times \hat{b}_{\ell(i,j)} + \hat{c}_{\ell(i,j)}\,, \tag{13}$$

where $\hat{\tau}_{ijt}^{(\text{Bayes})}$ is the posterior mean estimate of τ_{ijt} in (2). We also consider a hybrid version of the pipeline that uses only the control correction technique. For this method, we replace the control means with the empirical Bayes estimates,

$$\hat{y}_{ijt}^{(\text{hybrid})} = \hat{\tau}_{ijt}^{(\text{pipeline})} \times \hat{a}_{\ell(i,j)} \times \hat{b}_{\ell(i,j)} + \hat{c}_{\ell(i,j)}\,. \tag{14}$$

Table 1 presents the results for this benchmark. Using the corrected controls in the hybrid model improves the predictions of the pipeline model slightly, suggesting the correction procedures from Sections 3.2 to 3.4 are useful independent of the Bayesian model. However, the Bayesian model outperforms the predictions from the corrected pipeline model by $\approx 20\%$. We also investigated a featureless version of the Bayesian model and found similar improvements over the pipeline, suggesting the improvements are due to the flexibility of the nonparametric Gaussian process prior. The log-linear constraint of the pipeline model imposes a strong assumption that the true dose response curve has a sigmoidal shape. By contrast, the Bayesian model only assumes monotonicity and learns the shape priors from the data.

4.3 Assessing feature importance

A current debate in precision oncology is whether all molecular feature types contain predictive power. Gene panels used in hospitals, such as MSK-IMPACT (Cheng et al., 2015), only consider mutations and copy number, while others argue that gene expression is sufficient to predict response (Piovan et al., 2013; Rodriguez-Barrueco et al., 2015). We investigate this here by fitting separate models for every possible subset of features (mutations, copy number variations, and expression). If a feature set does not contain worthwhile information, it will effectively introduce noise into the model and either not improve performance or lower it (e.g. due to finite samples and a nonconvex optimization procedure).

To measure predictive power of a model, we consider the task of predicting entire out-

	(min)	Dosage level							(max)	
Model	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**All**
Mutations	3.90	4.70	5.81	7.71	9.10	12.81	13.83	18.55	20.71	10.43
CNV	3.84	4.65	5.78	7.89	9.48	13.76	15.00	20.51	23.08	11.14
Expression	3.72	4.49	5.55	7.50	8.92	12.77	13.72	18.29	20.90	10.29
Mut+CNV	3.85	4.56	5.73	7.64	9.07	12.80	13.75	18.57	20.57	10.37
Mut+Exp	3.67	4.40	5.45	7.36	8.76	12.51	13.54	**18.07**	**20.41**	**10.11**
CNV+Exp	**3.63**	4.39	**5.42**	7.37	8.77	12.56	13.63	18.27	20.75	10.17
All	3.68	**4.37**	**5.42**	**7.28**	**8.71**	**12.45**	**13.52**	18.21	20.56	**10.11**

Table 2: Mean error results on the curve prediction benchmark. Overall performance generally increases as more features are added, suggesting each subset conveys valuable predictive information not captured by the other two.

of-sample experiments at all dosage levels. We again measure error on variance-adjusted raw count predictions. However, the curve prediction task is a prior predictive check of the model, rather than a posterior one. This is a more challenging task, as the model does not see any outcomes from the specific experiment when making predictions. We use the logistic-transformed mean as the predicted drug effect,

$$\hat{y}_{ijt}^{(\text{prior})} = \frac{\hat{a}_{\ell(i,j)} \times \hat{b}_{\ell(i,j)}}{1 + e^{-f_t^+(X_i, j; \hat{\theta})}} + \hat{c}_{\ell(i,j)} \,. \tag{15}$$

The pipeline approach does not provide a way to make raw predictions from feature subsets; we therefore only consider the Bayesian model for this task.

We evaluate the predicted priors from the held out cross-validation folds, across all folds. We measure mean error by first taking the average error on the entire curve, then averaging across all curves. Table 2 shows the curve prediction results. The model generally has lower error as more feature subsets are included, suggesting that all three feature subsets add predictive value.

4.4 Qualitative evaluation

In addition to the quantitative results above, we also evaluate whether the model recapitulates known biology in its learned prior. This qualitative check provides reassurance to biologists that the patterns discovered by the model are reliable and not likely just functions of undetected experimental artifacts. We generated marginal prior predictive curves with different subsets of drugs and cell lines. The subsets of cells are chosen based on features that are known to be targeted by certain drugs. Drugs that target one subset should produce more sensitive marginal predicted response curves.

BRAF inhibitors 6 of the drugs in the dataset belong to a family of therapeutics known as *BRAF inhibitors*. These drugs all target a well-studied, commonly-mutated oncogene called BRAF. Cells that express a mutated form of BRAF are prone to unproliferated growth as natural inhibitors of BRAF do not bind to the mutated variant. BRAF inhibitors correct for this imbalance by binding to the mutated BRAF protein or otherwise down-regulating BRAF pathway. As shown in the top left panel of Fig. 5, the model successfully captures the differential efficacy of BRAF inhibitors between cell lines possessing mutant and wild type variants.

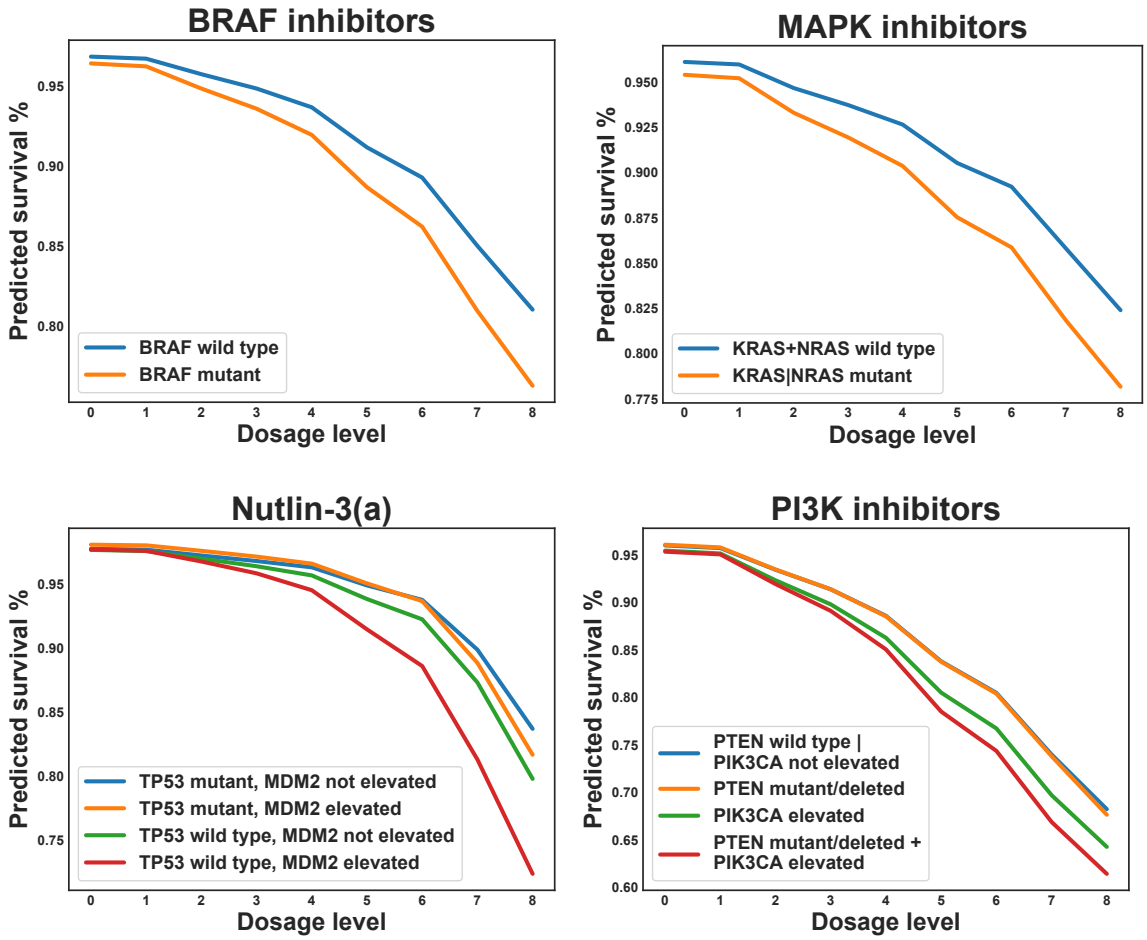


Figure 5: Prior predictive checks recapitulating known biology. Top left: cell lines with and without a mutated BRAF gene, when treated with any of the six drugs designed to target such cells. Top right: MAPK inhibitors that shut down the oncogenic pathway caused by mutations in RAS-type genes. Bottom left: Nutlin-3(a) (an MDM2 inhibitor) is effective only when both MDM2 levels are high and TP53 is not mutated. Bottom right: PI3K inhibitors target cells exhibiting PTEN loss and PIK3CA elevation.

MAPK inhibitors 18 of the drugs in the dataset belong to a family of therapeutics known as *MAPK inhibitors*. These drugs target a signaling pathway involved in cell growth and proliferation. Mutations in the oncogenes KRAS and NRAS are known to lead to unproliferated growth through activation of the MAPK pathway. The top right panel of Fig. 5 shows that the model successfully predicts more sensitivity to MAPK inhibitors when a cell line has one of these mutations.

Nutlin-3(a) The drug Nutlin-3(a) binds to and suppresses MDM2. However, the biological mechanism of action behind Nutlin-3(a) is more complicated than in the BRAF inhibitors. In this case, a mediator gene known as TP53 is necessary. TP53 is the most commonly mutated gene in cancer and is known as the "guardian of the genome." One of its primary roles is to check for DNA damage such as excessive or dangerous mutations. When damage is found, TP53 will attempt to repair the damage or, if the damage is too severe, will initiate

apoptosis and force the cell to die. When TP53 is mutated or suppressed, these safety mechanisms are disabled and they allow downstream damage to occur unchecked, leading to oncogenic growth. MDM2 is an inhibitor of TP53 and when elevated can suppress TP53 entirely. Thus, in order for Nutlin-3(a) to be effective, we need two conditions to be true: 1) MDM2 needs to be elevated above normal levels such that it inactivates TP53, and 2) TP53 must not be mutated, such that if MDM2 is suppressed then TP53 will be functional and capable of initiating apoptosis. In this case, we consider a cell line to have an elevated level of MDM2 if it is expressed at least one standard deviation above the mean expression level in the dataset. The bottom left panel of Fig. 5 confirms that the model predicts this nonlinear efficacy pattern.

PI3K inhibitors There are 24 drugs in the dataset that are *PI3K inhibitors*. These drugs broadly target the PI3K-AKT-mTOR pathway, which plays a prominent role in cell growth and division. Cells that exhibit oncogenic malfunctions in this pathway are characterized by two common hallmarks: 1) a loss or mutation of the tumor suppressor gene PTEN, and 2) elevated levels of the oncogene PIK3CA. As in the Nutlin-3(a) scenario, these two show a nonlinear interaction that is replicated by the learned prior, as seen in the bottom right panel of Fig. 5.

The qualitative checks above provide a useful reasurrance, but we acknowledge that they are not foolproof. It is possible that latent confounders, such as the type of cancer or DNA methylation, are correlated with the features that we explored. Even among the observed features there are correlations which may lead us to believe the model is capturing biological knowledge when it is really learning information about other features that have strong dependencies with the chosen examples. Making strong causal inference statements about drug sensitivity would require follow-up wet lab experiments. The predictive model does not replace the need for the validation experiments. Instead, it enables biologists to guide their experimental planning by suggesting potential drivers of sensitivity and resistance.

5 Discussion

5.1 The benefits of modeling uncertainty

Predictive models for cancer cell line drug response enable science to move at a faster pace. If a predictor can faithfully replicate the outcome of a wet lab experiment, scientists can screen drugs quickly in simulation to find potential therapies worth investigating. The usefulness of good predictors has led biologists to organize predictive modeling competitions to crowdsource better predictors (Costello et al., 2014) and to build bespoke machine learning models to predict drug response (Menden et al., 2013; Ammad-ud din et al., 2016; Rampasek et al., 2017). While these efforts have led to models with improved predictive performance, the target of prediction is a summary statistic derived from preprocessing pipelines such as the approach in Section 4.1.

Compressing each experiment down to a single point estimate of a summary statistic is problematic. At each step in the pipeline, simplifying assumptions remove structure from the model and obfuscate the inherent uncertainty in the measurements, effects, and outcomes. Specifically, (i) averaging the negative controls ignores measurement noise and technical error; (ii) averaging positive controls fails to capture natural variability in cell growth; (iii) a log-linear dosage model makes strong assumptions about the effect of different

drug concentrations; and (iv) the summary statistic reported contains no information about the uncertainty in effect size of a given drug at any specific dosage.

With these considerations in mind, we proposed a Bayesian approach to modeling dose-response in high-throughput cancer cell line experiments. The Bayesian model addresses the issues with the typical pipeline approach by directly modeling uncertainty at every step: (i-ii) both positive and negative controls are treated as random quantities, with plate-level uncertainty quantification of machine bias and cell growth; (iii) the dosage effect is constrained to be monotonic, but is otherwise fully flexible and not limited to the log-linear regime; and (iv) the model enables full Bayesian posterior inference over the entire dose-response curve for every experiment.

The Bayesian model outperforms the pipelined approach in benchmarks on the GDSC dataset. However, there is still ample room for improvement. The neural network architecture and training method we used was not explored extensively. Other models or architectures such as those from the existing literature in computational biology (e.g. Rampasek et al., 2017) may yield better performance if combined with a Bayesian model of high-throughput screening experiments. The model could also be improved to better match the data. For instance, some drugs induce total cell death at high dosages, and some have no effect at all below a certain concentration. The logistic Gaussian process in our model may be a poor fit in these cases since they push the logits to extreme values and consequently dominate the likelihood. Finally, we followed a typical processing pipeline to determine mutations, CNVs, and expression levels. These feature pipelines remove uncertainty in the sequencing process that, if modeled directly, may also lead to a better predictive model. We plan to investigate these extensions in future work.

5.2 Conclusions and recommendations

Beyond the specific model and metrics, our experience led to several observations about good practice in high-throughput cancer drug screening. We draw five conclusions, each with a corresponding recommendation for practitioners.

1. **Spatial and temporal batch effects exist in high-throughput data.** We showed evidence in Sections 3.2 and 3.3 that experimental controls are systematically contaminated, creating dependency between observations. We detailed a correction approach for legacy data that first detects and discards contaminated observations, then leverages temporal dependencies to compensate for the loss of data.

   Recommendation: Experimentalists abandon the *de facto* standard of sequential plate layouts and instead randomize their plate layouts as much as possible to facilitate denoising.

2. **The one-trial-per-dose paradigm is problematic.** Natural variation among cell lines was observed to be substantial in Section 3.4, with swings as high as $+/-50\%$ of the median control response being common in many experiments. This complicates inference by creating high degrees of uncertainty about any individual experiment.

   Recommendation: Experimentalists conduct multiple replicates of every experiment; computational biologists rely on heteroskedastic regression or Bayesian models when fitting dose-response curves.

3. **A nonparametric predictive dose-response model outperforms the state of the art.** The Bayesian model was shown in Section 4.2 to outperform a state-of-the-art

technique for dose-response modeling developed for the GDSC data. The Gaussian process posterior from the model has $\approx 20\%$ lower mean squared error when predicting a held out dosage level.

Recommendation: Computational biologists discard parametric dose-response curve assumptions, such as sigmoidal shape, in favor of a more flexible, nonparametric model such as a monotonic Gaussian process.

4. **All molecular feature types contain relevant information.** Through an ablation study in Section 4.3, we showed that the model predictions are improved by each of the three subsets of features (mutations, copy number, and expression).

   Recommendation: Clinicians and hospitals gather all three sets of information; this will likely lead to more accurate recommendation models for personalized therapies.

5. **A deep, probabilistic model can generate biologically-meaningful, nonlinear hypotheses.** A series of examples in Section 4.4 showed that the model recapitulates known biology. In some cases, this involved nonlinear combinations of features such as needing a high level of expression in one gene and a corresponding wild type of another gene for a specific drug to be effective. This suggests the approach has the potential to be used in exploratory drug discovery experiments where candidate drugs are often tried without a known mechanism of action.

   Recommendation: Computational biologists avoid linear models when dealing with cell line data. The richness of the interactions, combined with their high dimensional nature, make black box models the preferred approach.

Acknowledgements. The authors thank Victor Veitch, Mykola Bordyuh, Antonio Iavarone and Anna Lasorella for many helpful conversations. WT is supported by the a seed grant from the Data Science Institute of Columbia University and the NIH (U54-CA193313). SWL is supported by the Simons Foundation (SCGB-418011). CHW is supported by the NSF (1305023, 1344668) and NIH (U54-CA193313). DMB is supported by ONR (N00014-17-1-2131, N00014-15-1-2209), NIH (1U01MH115727-01), and DARPA (SD2 FA8750-18-C-0130).

References

M. Ammad-ud din, S. A. Khan, D. Malani, A. Murumägi, O. Kallioniemi, T. Aittokallio, and S. Kaski. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, 32(17):i455–i463, 2016.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of molecular diagnostics*, 17(3): 251–264, 2015.

J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, P. Hintsanen, S. A. Khan, J.-P. Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(96–104), 2004.

B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008.

M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

P. M. Haverty, E. Lin, J. Tan, Y. Yu, B. Lam, S. Lianoglou, R. M. Neve, S. Martin, J. Settleman, R. L. Yauch, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, 533 (7603):333–337, 2016.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

A. Lachmann, F. M. Giorgi, M. J. Alvarez, and A. Califano. Detection and removal of spatial bias in multiwell assays. *Bioinformatics*, 32(13):1959–1965, 2016.

J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

L. Lin and D. B. Dunson. Bayesian monotone regression using gaussian process projection. *Biometrika*, 101(2):303–317, 2014.

B. Mazoure, R. Nadon, and V. Makarenkov. Identification and correction of spatial bias are essential for obtaining quality data in high-throughput screening technologies. *Scientific reports*, 7(1):11921, 2017.

M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4):e61318, 2013.

P. Muir, S. Li, S. Lou, D. Wang, D. J. Spakowicz, L. Salichos, J. Zhang, G. M. Weinstock, F. Isaacs, J. Rozowsky, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17(1):53, 2016.

I. Murray, R. Adams, and D. MacKay. Elliptical slice sampling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 541–548, 2010.

E. Piovan, J. Yu, V. Tosello, D. Herranz, A. Ambesi-Impiombato, A. C. Da Silva, M. Sanchez-Martin, A. Perez-Garcia, I. Rigo, M. Castillo, et al. Direct reversal of glucocorticoid resistance by akt inhibition in acute lymphoblastic leukemia. *Cancer cell*, 24(6):766–776, 2013.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

L. Rampasek, D. Hidru, P. Smirnov, B. Haibe-Kains, and A. Goldenberg. Dr. vae: Drug response variational autoencoder. *arXiv preprint arXiv:1706.08203*, 2017.

R. Rodriguez-Barrueco, J. Yu, L. P. Saucedo-Cuevas, M. Olivan, D. Llobet-Navas, P. Putcha, V. Castro, E. M. Murga-Penas, A. Collazo-Lorduy, M. Castillo-Martin, et al. Inhibition of the autocrine il-6–jak2–stat3–calprotectin axis as targeted therapy for hr-/her2+ breast cancers. *Genes & development*, 2015.

R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

D. J. Vis, L. Bombardelli, H. Lightfoot, F. Iorio, M. J. Garnett, and L. F. Wessels. Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics*, 17(7):691–700, 2016.

Y.-X. Wang, A. Smola, and R. Tibshirani. The falling factorial basis and its statistical applications. In *International Conference on Machine Learning*, pages 730–738, 2014.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, et al. Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, 45(10):1134, 2013.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

A Feature preprocessing

A.1 Mutations

We consider a gene *mutated* based on the following SNP-level filter:

- Silent mutations are ignored.
- Insertion, deletion, nonsense, and nonstop mutations are automatically considered mutated.
- Missense SNPs are considered mutated if they are present in at least 3 TCGA (Tomczak et al., 2015) samples.
- Splice site mutations are considered mutated if they have an "IMPACT" score of high.

We also include the two gene fusions that were detected in the GDSC data, EWSR1-X and BCR-ABL. This generates a total of 5822 mutation features.

A.2 Copy Number Variations

Copy number alterations occur when a swath of the genome is lost or amplified. In most cases, this occurs across multiple gene coding regions at once. The result is a highly correlated set of features. We filter this list down to known and potential drivers as follows:

- We keep the list of 193 known driver genes identified in the original GDSC analysis.
- We add an additional list of 87 potential driver genes identified in a follow up study (Zack et al., 2013).
- We remove genes from the second list if their copy number is constant across all samples or has correlation coefficient ≥ 0.95 with any gene from the first list.

For all cell lines, we consider their median copy number to be the baseline level. Despite the GDSC data coming from human tumor cultured cell lines, the median copy number is actually 3. Since copy number mostly impacts differential protein expression, relative abundance in a cell is the most important feature here and we thus consider a cell's median copy number to be its baseline, even though in a normal human tissue a copy number of 3 would be a gain.

A.3 Expression

We retain all 17271 protein-coding genes sequenced in the GDSC dataset. We use the same preprocessed dataset as in the original GDSC analysis.

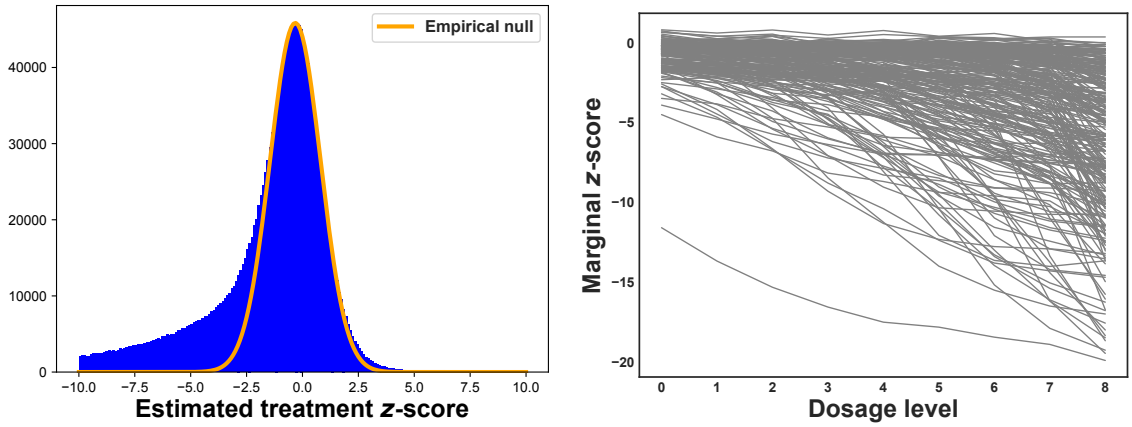
B Checking drug effect assumptions

As a pragmatic check on our assumption that drugs can only hinder growth, we use the estimated control parameters from Sections 3.3 and 3.4 to calculate approximate z-scores for the drug effects,

$$\widetilde{z}_{ijt} = \frac{y_{ijt} - \hat{a}_{\ell(i,j)} \times \hat{b}_{\ell(i,j)} - \hat{c}_{\ell(i,j)}}{\hat{a}_{\ell(i,j)} \times \hat{b}^2_{\ell(i,j)}} \,. \tag{16}$$

Figure 6a shows the marginal distribution of $\widetilde{\mathbf{z}}$ across all experiments and all dosage levels in the dataset, omitting z-scores less than -10. The z-scores are likely slightly biased due to experimental, technical, and preprocessing error, making the theoretical $\mathcal{N}(0,1)$ null distribution poorly specified; this is a common phenomenon in HTS experiments (Efron, 2008). We calculate an empirical null using a polynomial approximation to Efron's method (Efron, 2004). The empirical null density suggests that the alternative (i.e. non-null treatment effect) distribution has effectively no mass in the positive (i.e. growth-encouraging) region of z-scores, supporting our death-only assumption on drug effects.

Figure 6b presents the same marginal z-score check from (16), broken down by individual drugs as a function of dosage level. Every drug appears to have a clear monotone marginal distribution, supporting our montonicity assumption on drug dosage.



(a) All experiments

(b) Individual drugs

Figure 6: The marginal distribution of approximate z-scores from (16) across (a) all experiments in the dataset, and (b) individual drugs. In (a), we overlay an empirical estimate of the null distribution; most drugs contain no effect and there appears to be no evidence that any drugs have a positive effect on growth. Similarly, in (b) we observe that all drugs seem to have stronger marginal effects as the dosage level increases, suggesting that the monotone toxicity assumption is also reasonable.