PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Sparse discriminative latent characteristics for predicting cancer drug sensitivity from genomic features

David A. Knowles[1,2], **Gina Bouchard**[1], **Sylvia Plevritis**[1,3]*

1 Department of Radiology, Stanford University School of Medicine, Stanford, California, USA, **2** Department of Genetics, Stanford University School of Medicine, Stanford, California, USA, **3** Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA

* plevriti@stanford.edu

OPEN ACCESS

Citation: Knowles DA, Bouchard G, Plevritis S (2019) Sparse discriminative latent characteristics for predicting cancer drug sensitivity from genomic features. PLoS Comput Biol 15(5): e1006743. https://doi.org/10.1371/journal.pcbi.1006743

Editor: Florian Markowetz, University of Cambridge, UNITED KINGDOM

Received: September 7, 2017

Accepted: December 21, 2018

Published: May 28, 2019

Data Availability Statement: All drug sensitivity and cell line characterization is available through https://pharmacodb.pmgenomics.ca/. Both data types were accessed programmatically using the R/Bioconductor package PharmacoGx (version 1.10.3) using the function downloadPSet followed by summarizeSensitivityProfiles and summarizeMolecularProfiles for the drug and molecular data respectively.

Funding: This work was supported by NIH grants U54-CA209971, U54-CA149145 and R25-CA180993, and Le Fonds de recherche du

Abstract

Drug screening studies typically involve assaying the sensitivity of a range of cancer cell lines across an array of anti-cancer therapeutics. Alongside these sensitivity measurements high dimensional molecular characterizations of the cell lines are typically available, including gene expression, copy number variation and genomic mutations. We propose a sparse multitask regression model which learns discriminative latent characteristics that predict drug sensitivity and are associated with specific molecular features. We use ideas from Bayesian nonparametrics to automatically infer the appropriate number of these latent characteristics. The resulting analysis couples high predictive performance with interpretability since each latent characteristic involves a typically small set of drugs, cell lines and genomic features. Our model uncovers a number of drug-gene sensitivity associations missed by single gene analyses. We functionally validate one such novel association: that increased expression of the cell-cycle regulator C/EBPδ decreases sensitivity to the histone deacetylase (HDAC) inhibitor panobinostat.

Author summary

A core tenant of precision medicine is that treatment should be tailored to the patient. In the context of cancer, large-scale screens, assaying the sensitivity of many cell-lines to panels of drugs, have the potential to enable discovery of biomarkers of sensitivity to specific therapeutics. However, existing computational approaches have not taken full advantage of these data. We develop a novel multi-task regression model, `Lacrosse`, which uses a Bayesian non-parametric prior to model "latent characteristics" of cell-lines that confer sensitivity to specific drugs and are predictable from genomic features. The resulting algorithm improves upon existing work by: a) jointly modeling multiple drugs to share statistical signal b) incorporating prior knowledge in terms of known inhibition targets c) using a sparse latent variable regression approach giving interpretable summaries of detected gene-drug associations. In particular, our analysis uncovers groups of drugs whose efficacy depends on genomic features in a similar way. We find new potential biomarkers of

Québec - Santé FRQ-S 35603. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

drug sensitivity, one of which we validate experimentally: that panobinostat is less effective when C/EBPδ is over-expressed.

Introduction

Several drug screening studies have assayed the sensitivity of a library of cancer cell lines to an array of anti-cancer compounds. Notable examples are the Cancer Cell Line Encyclopedia [1, CCLE], Genomics of Drug Sensitivity in Cancer [2, 3] the 2012 DREAM challenge [4, 5] and the Cancer Therapeutics Response Portal v2 [6, CTRPv2]. Along with viability response curves, these studies provide high-dimensional molecular profiling of the assayed cell lines. For example, CCLE includes gene expression microarrays, copy number variation (CNV), and oncogene mutation status assays.

These data have the potential to both help understand the key differences between cancers and cancer subtypes that drive resistance to specific drugs, and to one day help choose the appropriate drug (or combination of drugs) for an individual patient, the core idea of precision medicine. As a result there is a need for analyses that both identify what differences between cancers cause the observed sensitivity patterns, and which differences can accurately predict what drugs will be efficacious for a tumor based on its genomic profile. The existing analyses of these datasets involve simple per drug regressions, such as elastic net [7]. While these methods are able to pick out the strongest signals in the data, they suffer from not taking advantage of known relationships between drugs and between genomic features. For example, we know which drugs have the same molecular target, and which features are related to the same gene, e.g. gene expression, CNV and mutation status will all typically be assayed for a given gene.

Cancer is highly heterogeneous in terms of its genomic features but many cancers share common phenotypic characteristics (Fig 1a), the "hallmarks of cancer" [8]: broken apoptosis or cell cycle regulation [9], disrupted DNA repair mechanisms [10], or "addiction" to specific oncogene pathways [11]. Because these phenotypic features can not be directly observed from genomic data, we regard them as unobserved, latent characteristics in this work. We expect these unobserved, latent characteristics to be associated with genomic cell line features such as gene expression. Moreover, we expect that the presence or absence of these latent characteristics confers sensitivity or resistance to specific therapeutic compounds.

To predict cancer drug sensitivity based on latent characteristics derived from genomic-drug screening data, we propose a novel approach, LAtent ChaRacteristics Of Small-molecule SEnsitivity (`Lacrosse`). The statistical model underlying `Lacrosse` is a discriminative, Bayesian non-parametric, sparse factor analysis, which falls under the general class of multi-task regression models [12]. A handful of related statistical methods have been used for drug sensitivity prediction. Menden et al. [13] used a single hidden layer feed-forward neural network [14]. While multitask neural networks are easily constructed by allowing multiple outputs to share the same hidden layers, Menden et al. took the alternative approach of featurizing the drugs using the chemoinformatic program PaDEL [15] and learning a single predictive model learned using both the drug and cell lines features. In contrast, `Lacrosse` does not featurize the drugs.

`Lacrosse` is closely related to kernelized Bayesian multitask learning [16, KBMTL]. While their approach gives excellent predictive performance, interpretability is somewhat lacking since the kernels (similarity between cell lines) are calculated on entire "views" (e.g. continuous gene expression) so that the influence of specific genes or pathways is not elucidated. We

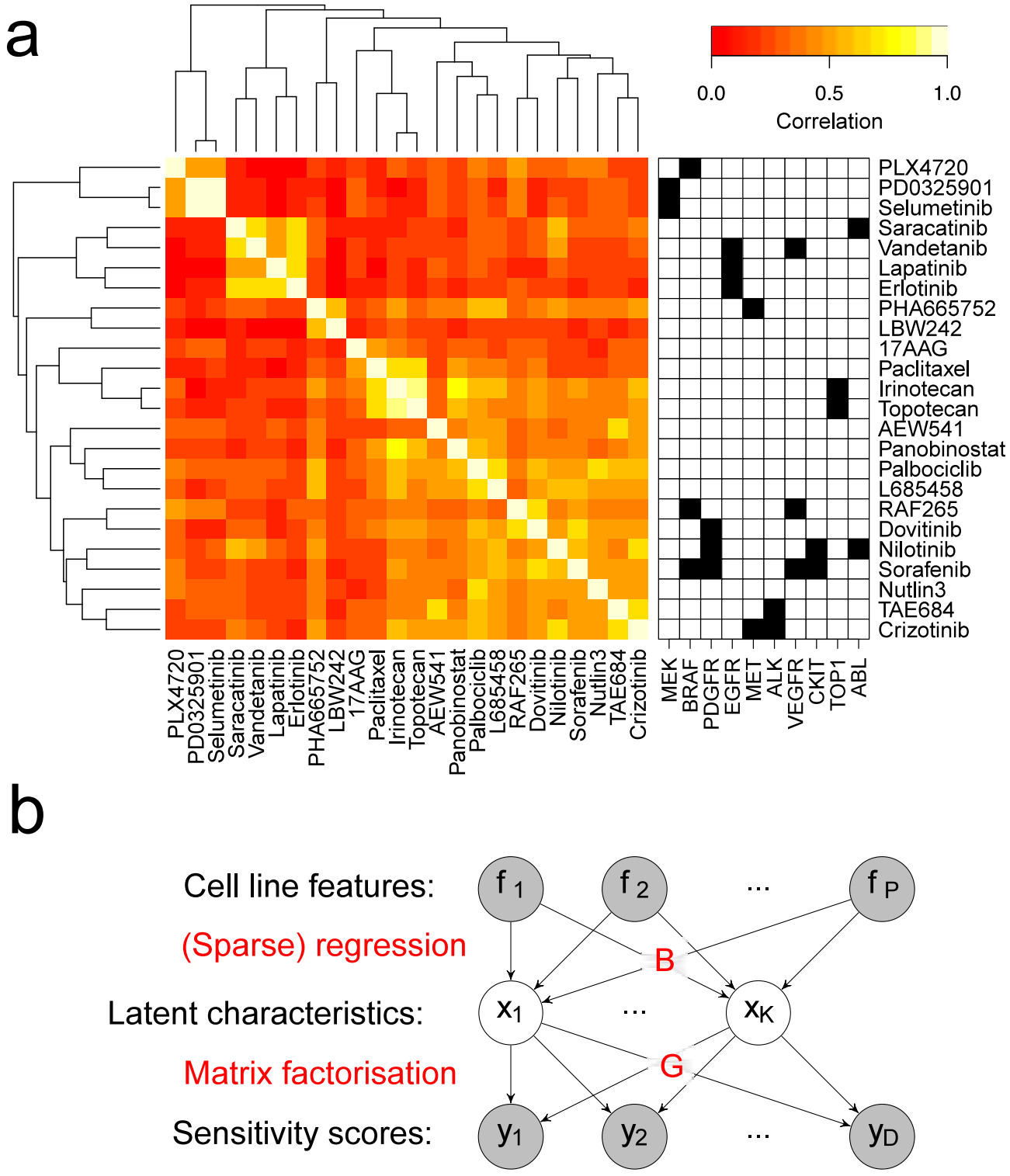

Fig 1. Joint analysis of cell line sensitivity across multiple drugs has the potential to improve predictive accuracy as well as model interpretability. a. While cancers are genetically heterogeneous there are phenotypic characteristics shared by many cancers subtypes, some of which are illustrated here. We refer to these as *latent* characteristics (LCs) because they are not directly observable from genomic data, but we hypothesize that each characteristic will have a detectable genetic signature and a defined influence on drug sensitivity. **b.** Pearson correlation of drug sensitivity profiles *(active area scores)* across CCLE, annotated by known inhibition targets. **c** The `Lacrosse` model consists of two components, shown here graphically. The first is a sparse linear regression from cell line features, **F** (gene expression, copy number variation and genomic mutations) to continuous valued latent characteristics, **X**. The second is a sparse factor analysis (matrix factorization) where the latent characteristics (the factors) explain the observed sensitivity scores through a sparse loading matrix **G**.

https://doi.org/10.1371/journal.pcbi.1006743.g001

compare to KBMTL here and show comparable, or even slightly improved predictive performance, with the added benefit of increased interpretability.

A Bayesian multitask multiview linear regression (MVLR) was recently developed [17]. Sparse Cauchy priors are used to select features and a Dirichlet prior is used over a parameter vector that selects predictive views. While the approach is multitask the relationship between a feature and drug response is the same for all drugs included in the model, up to a positive multiplicative weight. This is analogous to `Lacrosse` restricted to only one latent characteristic.

A distinct approach is taken by Knijnenburg et al. [18] who develop 'Logic Optimization for Binary Input to Continuous Output' (LOBICO) which finds small, interpretable logic networks predictive of drug response for individual drugs. LOBICO is limited to binary cell-line features.

Two other recent studies have developed analyses that share ideas with `Lacrosse` but which provide exploratory analyses rather than predictive models of drug response. This makes it infeasible to compare to our approach in terms of predictive performance. First, Seashore-Ludlow et al. [6] developed Annotated Cluster Multidimensional Enrichment (ACME), which tests whether drug/cell line sensitivity biclusters have coherent biological signal in terms of enriched protein targets (for the drugs) and mutations/lineage (for the cell lines). Second, El-Hachem et al. [19] applied Similarity Network Fusion [20] to drug-drug networks derived similarity in terms of chemical structure, drug sensitivity, and drug perturbation response.

`Lacrosse` also has similarities to nuclear norm regression [21], an extension of L1-regularized regression [7] to the multitask setting by penalizing the trace/nuclear-norm (the sum of singular values) of the coefficient matrix. Compared to their optimization based approach, DNFSA, as a Bayesian probabilistic approach has the advantage of allowing incorporation of prior knowledge in the form of known drug-drug relationships.

An additional unique aspect of `Lacrosse` is that it allows a graph over the drugs to be specified encoding prior knowledge about which drugs are likely to have similar properties. This graph is used to specify a Markov-random field which explicitly encourages drugs which share edges to have more similar coefficients in the model.

Results

Summarizing dose-response curves using active area

Large-scale viability screens such as CCLE have typically summarized dose-response curves in terms of the drug concentration required for 50% inhibition of growth, "IC50". IC50 however has several weaknesses: a) it is undefined if 50% inhibition is never reached, b) it is noisy due to being overly reliant on viability measurements close to the IC50 value, and c) it ignores differences in effectiveness for doses above the IC50 point. An simple alternative summary is "active area", the integrated area above the dose-response curve (S1 Fig). We have found active area to be more predictable from molecular profiles than IC50: 10-fold cross-validation on CCLE using group LASSO explains 27.5% of heldout variance in active area scores across drugs, compared to only 14.4% for IC50. We therefore choose to use active area as the drug sensitivity metric throughout this work.

Known drug targets only partially explain sensitivity profiles

For the 24 drugs in CCLE we analyzed the between-drug correlation of sensitivity profiles (active area scores) across 432 cell lines (Fig 1b). In some cases the high correlation between profiles is explained in terms of shared inhibition target: e.g. the MEK inhibitors PD0325901 and selumetinib have a highly similar sensitivity profile across cell lines. Similar statements can be made for inhibitors of EGFR (erlotinib, lapatinib, vandetanib), TOP1 (topotecan, irinotecan) and ALK (crizotinib, TAE684). However, in other cases are less easily explained:

PHA665752, a c-MET inhibitor, and LBW242, a SMAC mimic, have significantly correlated profiles but no known shared mechanism of action. These observations motivate using known inhibition targets as soft prior knowledge rather than hard constraints in our methodology.

`Lacrosse` overview

`Lacrosse` is a Bayesian-nonparametric, sparse, multitask regression that jointly predicts viability for multiple drugs using cell line genomic features. `Lacrosse` posits that subsets of cancer cell lines possess latent characteristics (LCs) which are predictive of their sensitivity profile across different drugs. Whether a cell line possesses a particular LC is identified by the presence of genomic features specific to the LC, that is, the LCs are generated by a sparse regression on the cell line features.

Explicitly, given the matrix **F** of genomic features (# features × # cell lines), `Lacrosse` models the matrix **Y** of viability scores (# drugs × # cell lines) as

$$\mathbf{Y} \approx \mathbf{GX}, \qquad \mathbf{X} \approx \mathbf{BF}, \tag{1}$$

where **X** are the LCs, and **G** and **B** are matrices of (sparse) regression coefficients to be learned (Fig 1c). We use 40,492 genomic features from CCLE spanning gene expression, CNV and mutations. These LCs represent a low dimensional embedding of the cell lines which preserves information salient to their drug sensitivity profiles. Using this graphical model the LCs are primarily focused on modeling the drug sensitivity patterns, but are also constrained to be predictable from cell line genomic features.

We additionally extend `Lacrosse` to allow prior knowledge about both drugs and genomic features to be incorporated in the form of a graph where related drugs (or genomic features) share edges. Connected nodes are encouraged to have the same regression coefficient sparsity pattern using a Markov-random field (MRF) approach. In practice we use this capability to inform the model about which drugs share molecular targets (hand-curated, Fig 2a) and which cell line features correspond to the same gene.

Predictive performance

To assess the predictive performance of `Lacrosse` compared to existing state-of-the-art methods we performed 10-fold cross-validation holding out 10% of cell lines for each fold and calculated the proportion of variance explained (PVE) for each fold. We compared `Lacrosse` to

- FA: sparse, nonparametric factor analysis jointly over the molecular characteristics and sensitivity [22]
- REG: spike and slab Bayesian linear regression [23]
- L1: LASSO L1-regularized linear regression [24] using the `glmnet` R package
- L1 multi: multiresponse regression using group LASSO [25] resulting in a shared sparsity pattern across all drugs, again using `glmnet`
- KBMTL: kernelized Bayesian multitask learning [16].
- Ridge: ridge regression.
- Elastic Net: with $\alpha = \frac{1}{2}$
- Bayesian multitask multiview linear regression (MVLR) [17]

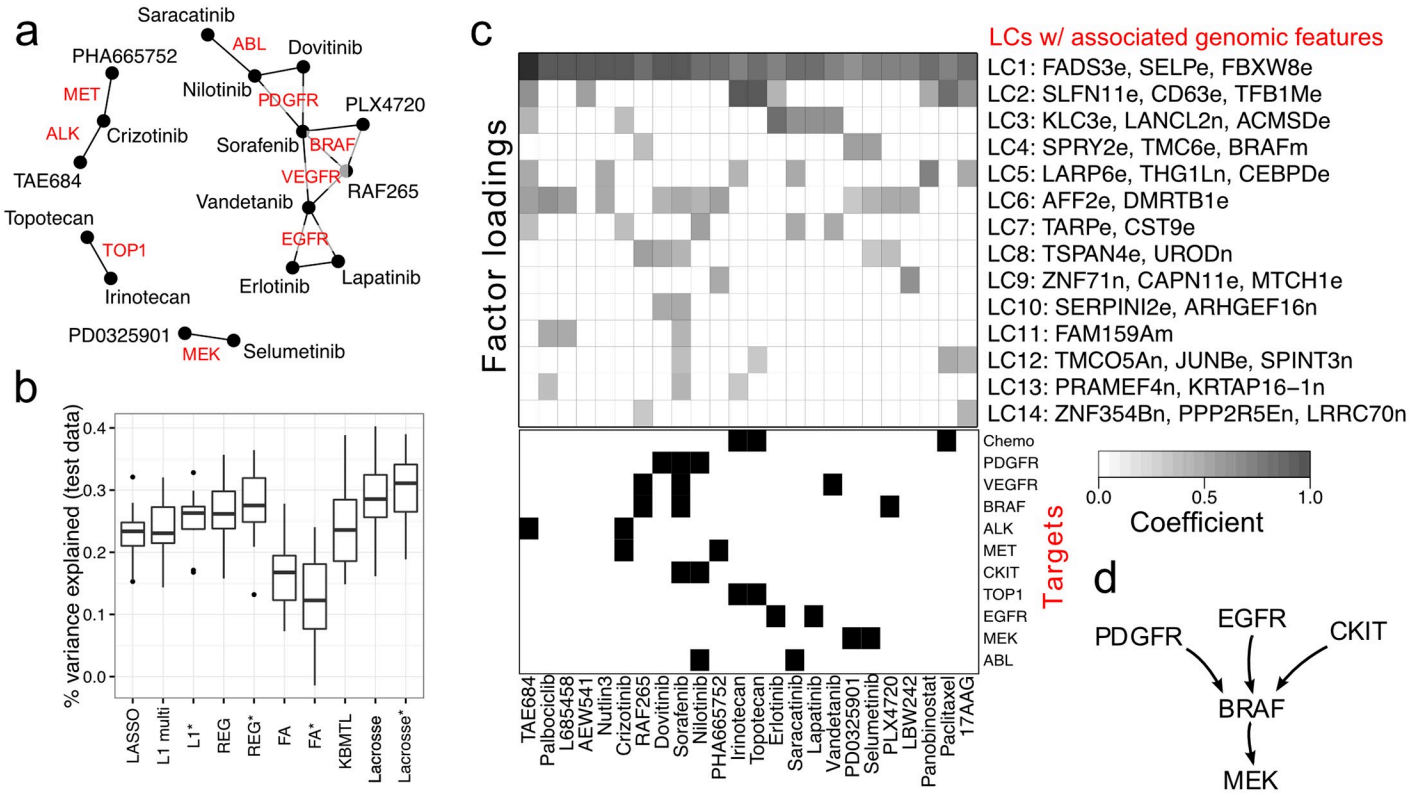

Fig 2. Lacrosse combines prior knowledge (shared inhibition targets) and observed similarity between drugs to improve predictive performance and define overlapping clusters of drugs with related dependence on genomic features. a. A Markov random field (MRF) is used to encode prior knowledge of which drugs share inhibition targets. `Lacrosse` does not use the inhibition targets themselves, only that if two drugs share an inhibition target then they are more likely to share sensitivity to a particular latent characteristic. **b**. Predictive performance results on CCLE. Here LASSO = L1 regression, REG = spike & slab regression. `Lacrosse` = discriminative factor analysis. In general incorporating prior knowledge about between drug relationships (the methods denoted with *) improves predictive performance. The Bayesian spike and slab regression based methods (REG and `Lacrosse`) also perform somewhat better than the L1 optimization method (LASSO, L1 multi, and L1*), although this comes at considerable computational expense. The factor analysis models (FA and FA*) have quite poor performance, likely to due to not being discriminative. The performance of KBMTL is similar to that of LASSO. **c**. `Lacrosse`* factor loadings: 14 latent characteristics (LCs) × 24 drugs, known targets for each drug, and associated gene features (i.e. non-zero coefficients in **B**) for each LC. Encoding of gene features: e = expression, n = copy number variation, m = mutation. **d**. A highly simplified layout of the MAPK pathway associated with latent characteristic 4.

https://doi.org/10.1371/journal.pcbi.1006743.g002

For each of these approaches we additionally consider an MRF extension (see Methods) which encourages drugs with shared targets to have similar sets of coefficients with non-zero weights. For LASSO regression we analogously enforce that any drugs sharing an inhibition target have the same sparsity pattern. These extensions are denoted with a superscript asterisk (*) on the method acronym.

`Lacrosse` outperforms these baselines in terms of its ability to predict sensitivity across drugs for heldout cell lines in CCLE (see Fig 2b). The sparse factor analysis (FA) approaches perform poorly, presumably because so much modeling capacity is effectively wasted in modeling the thousands of genomic features that are not be associated with sensitivity. Of the regression based approaches LASSO, group LASSO and KBMTL perform comparably with PVE around 24%. Bayesian spike-and-slab sparse regression (REG) somewhat outperforms these methods, with PVE = 26%. Finally `Lacrosse`, being able to share statistical signal across drugs, performs best, with PVE = 28.5%. In all cases apart from FA, we additionally find that explicitly incorporating known relationships between drugs (in terms of shared inhibition targets) and genomic features (corresponding to the same gene) improves predictive performance. Performance for FA may have dropped using known relationships since the model

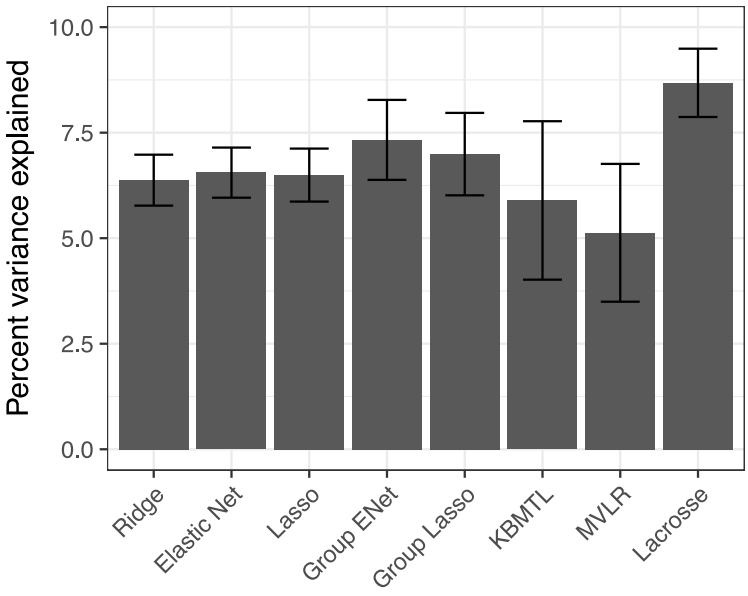

Fig 3. Cross-validated predictive performance on the large CTRPv2 sensitivity screen. `Lacrosse` significantly outperforms Group Lasso and Elastic Net, which themselves surpass the single task methods as well as KBMTL and MVLR.

https://doi.org/10.1371/journal.pcbi.1006743.g003

capacity is already insufficient to accurately associate genomic features with sensitivity, and added constraints further reduce the effective model capacity. `Lacrosse` including the MRF-extension achieves a median PVE across folds of 31.5%.

We additionally assessed predictive performance in terms of PVE on the 545 drugs, 783 cell lines CTRPv2 dataset (Fig 3), again using 10-fold cross-validation. The three tested single task methods—ridge regression, Elastic Net and LASSO—all performed similarly. Group LASSO outperforms LASSO ($p = 0.032$, paired t-test), and is itself outperformed by Group Elastic Net ($p = 0.0028$, paired t-test). In our hands KBMTL underperforms Group LASSO ($p = 0.036$) and Group Elastic Net ($p = 0.005$) and has comparable performance to the single task methods. MVLR significantly underperforms the single task methods ($p = 0.007$, 0.005, 0.006 for ridge, LASSO and Elastic Net respectively) although we emphasize that MVLR is only designed for joint modeling of closely related drugs (e.g. those with shared inhibition targets) so it is unsurprisingly it performs poorly at simultaneous modeling of over 500 drugs with many different mechanisms-of-action. Finally `Lacrosse` outperforms Group Elastic Net by a significant margin ($p = 0.0008$, paired t-test). The PVE across folds was not found to be significantly non-normal by a Shapiro-Wilk test for any method, but we none-the-less confirmed all comparisons found significant by t-test were also significant by Wilcoxon signed rank test, apart from KBMTL outperforming Group LASSO. We confirmed the qualitative ordering of the methods held up when assessing performance using the concordance index (S2 Fig). While `Lacrosse` obtains a higher mean c-index than Group Elastic Net (0.613 vs 0.608) this difference is not statistically significant.

We varied `Lacrosse`'s MRF strength parameter and found performance to be robust within a wide range of values (S3 Fig). The 29 LCs discovered running on the full CTRPv2 dataset are available here: https://www.dropbox.com/s/cucayg3nulkikjb/CTRPv2_LCs.zip?dl=0.

Comparison with LOBICO [18] is complicated by the fact that it operates on binary features (typically mutations) and response (binarized AUC or IC50) (the continuous response is used

to weight samples however). Considering this we compared LOBICO to single-task LASSO either using continuous AUC as response, or L1-regularized logistic regression using the same binarization as for LOBICO. For all three methods we used mutation data only (for 1600 genes) to accommodate LOBICO's requirement for binary features. Across 11 randomly chosen drugs from CTRPv2 LOBICO is substantially and consistently outperformed by both L1-regression based models (S4 Fig). We hypothesize that while LOBICO can explore a small hypothesis space of genes (<100) known to be important in conferring sensitivity or resistance, it is unable to effectively screen a larger number of potentially relevant features. While Iorio et al. [3] were able to build significantly predictive and interpretable LOBICO models with > 600 features, they did not compare predictive performance to sparse regression methods.

We explored whether the predictive model learnt on CTRPv2 would generalize to independent data. We first assessed generalization performance in CCLE (S5 Fig). For all 14 drugs common to both datasets we see statistically significant prediction (Benjamini-Hochberg adjusted Spearman correlation between predicted and observed AUC $p < 0.004$), and qualitatively the concordance indices are comparable to in-sample accuracy assessed using pre-validation in CTRPv2, with the exceptions of Nutlin-3 and RAF265. Since CTRPv2 includes all cell lines assayed in CCLE we next assayed generalization in The Genentech Cell Line Screening Initiative (gCSI) [26]. 37 cell lines were profiled in gCSI that were not in CTRPv2. We show results both for all cell lines in gCSI and the 37 non-overlapping cell lines alone in S6 Fig. Of the 10 drugs in common between CTRPv2 and gCSI, 7 are statistically significantly predicted on the full gCSI cell line collection (Benjamini-Hochberg adjusted Spearman correlation $p < 0.05$), but MS-275, Bortezomib and Crizotinib are not. On the 37 shared cell lines only Vorinostat shows significant prediction, but this maybe due to limited power with only $n = 37$ test points. Finally we applied the CTRPv2 model in GDSC1000 [3] for 69 shared drugs. Across all $n = 1110$ cell lines in GDSC1000, 59 drugs showed significant prediction, compared to 50/69 when using the $n = 501$ non-overlapping cell lines (Benjamini-Hochberg adjusted Spearman correlation $p < 0.05$, S7 Fig).

Drug-gene associations

Analysis of the latent characteristics can provide insights into signaling and regulatory pathways predictive of drug response. We tested pairwise relationships between the drugs and genomic features in each LC from CCLE using Spearman correlation, with selected LCs of interest shown in Fig 4. Remaining LCs are shown in S8 Fig. While many strong associations are seen, other associations are weaker, suggesting that the relationships uncovered by `Lacrosse` rely on sharing statistical power across drugs within an LC. LC 1 has positive loading for all drugs, suggesting there is some shared behavior across all the drugs in this dataset, a phenomenon previously described as "general levels of drug sensitivity" (GLDS) [27]. Indeed, LC 1 is highly correlated with the first GLDS ($R^2 = 0.75$, $p < 2 \times 10^{-16}$, S9 Fig). The expression level of FBXW8 is a genomic feature of LC 1, which is interesting since this gene is known to play an essential role in cancer cell proliferation [28]. This association was not picked up by the per drug regression, presumably because `Lacrosse` gains statistical power by analyzing all the drugs simultaneously.

Some LCs include known related drugs, e.g. LC 2 involves the DNA-damaging agents irinotecan and topotecan, as well as another chemotherapeutic agent, paclitaxel. It is reassuring to see SLFN11 expression modulating sensitivity to irinotecan and topotecan in this LC, since this is a known and experimentally validated relationship [29]. Increased CD63 expression decreases sensitivity to the drugs in LC 2, particularly paclitaxel, irinotecan, topotecan and

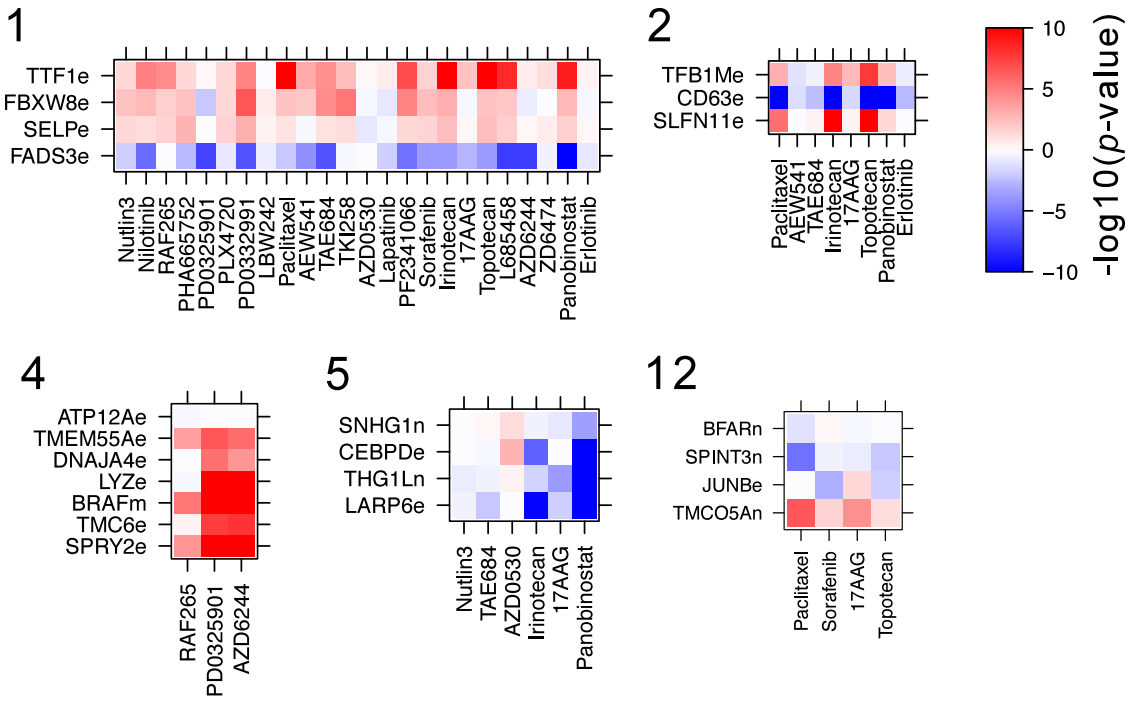

Fig 4. A subset of the learned latent characteristics represented as clusters of genomic features (rows) and drugs (columns). Numbering corresponds to the internal LC identifier and is arbitrary. Colors represent $-\log_{10} p$ for the relationship between the feature and drug, with the sign representing positive or negative associations (using Spearman correlation). Genomic feature suffixes e = expression, n = copy number variation, m = mutation. The remaining LCs are shown in S8 Fig.

https://doi.org/10.1371/journal.pcbi.1006743.g004

panobinostat. While this association is novel to the best of our knowledge, CD63 is known to be an exosomal marker that correlates with invasiveness in ovarian cancer cell lines [30].

LC 4 includes mutation of B-RAF as a genomic feature and drives sensitivity to the B-RAF inhibitor RAF265 and MEK inhibitors (PD0325901 and selumetinib/AZD6244). As a result, we conclude LC 4 corresponds to the MAPK pathway, a signaling cascade involving B-RAF, MEK and ERK (see Fig 2d). Since `Lacrosse` builds a sparse predictive model so may exclude some meaningful associations. Indeed we find LC4 level is correlated with B-RAF ($p = 6 \times 10^{-20}$) and KIT ($p = 0.016$) mutations, PDGFRB expression ($p = 1 \times 10^{-4}$) and MAP2K1 copy number ($p = 0.03$). 86/253 (34%) genes in the KEGG MAPK pathway are significantly correlated with LC4 level across cell lines, a 1.5× enrichment compared to background (Fisher's exact $p = 0.002$). Other interesting associations in LC 4 are those with TMC6 and SPRY2. SPRY2 is downstream of the MAPK pathway, but SPRY2 inhibition also activates MAPK and can lead to tumorigenesis [31], suggesting an as yet poorly characterized feedback loop. While TMC6 is not known to be involved in MAPK signaling, there are suggestive hints: TMC6/8 forms a complex with ZnT-1 [32], a zinc-finger protein which itself activates the MAPK pathway [33]. Potentially also of interest is that nonsense mutations in TMC6 cause a hereditary condition called Epidermodysplasia verruciformis involving susceptibility to human papillomavirus and resulting in cutaneous squamous cell carcinomas [32].

The influence of JunB on the VEGF inhibitor sorafenib and heat shock protein 90 (HSP90) inhibitor tanespimycin in LC 12 is likely due to regulation of the VEGF pathway by JunB [34], and the fact that HSP90 is required for VEGF induction. Finally we note that hypermethylation of TMCO5A is associated with worse prognosis in ovarian cancer [35].

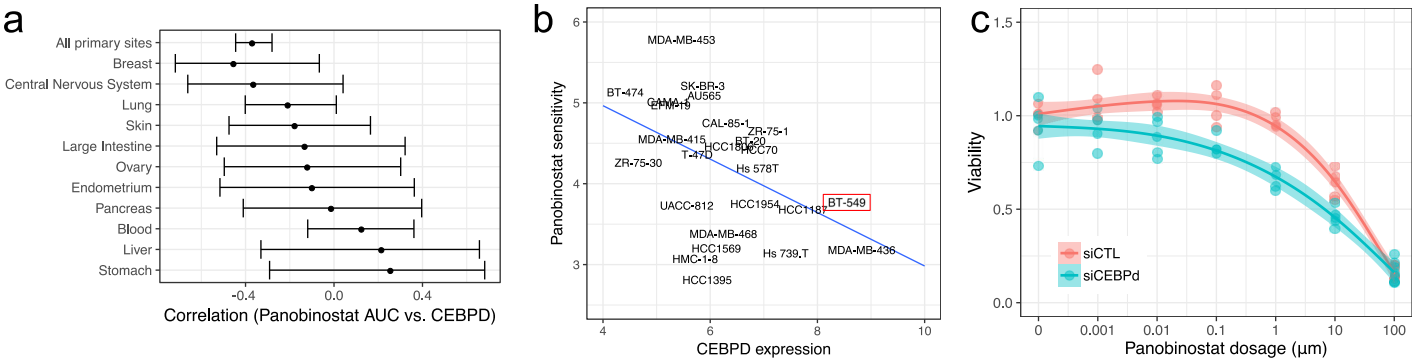

Fig 5. Knock-down of the cell-cycle regulator C/EBPδ increases sensitivity to the HDAC-inhibitor panobinostat *in vitro*. a. The negative association between C/EBPδ expression and sensitivity to panobinostat is seen across a range of cancers, but is most pronounced in breast cancer cell lines. Whiskers denote 95% confidence intervals. **b**. Within breast cancer cell-lines we chose to use BT-549 since it has relatively high C/EBPδ expression and low panobinostat sensitivity. **c**. siRNA mediated knock-down of C/EBPδ increases the sensitivity of BT-549 cells to panobinostat.

https://doi.org/10.1371/journal.pcbi.1006743.g005

Functional validation

In LC 5 we observe that sensitivity to the HDAC-inhibitor panobinostat is associated with C/EBPδ, a transcription factor that regulates cell cycle progression and apoptotis, and is therefore a putative tumor suppressor [36]. While C/EBPδ has not previously been associated with drug sensitivity, high expression is known be associated with poor prognosis across glioblastomas [37] and the closely related C/EBPβ is recognized as a synergistic master regulator with STAT3 of the mesenchymal phenotype in aggressive glioma [38]. Since this association was not reported using single drug analyses of the CCLE dataset and because of the significant current interest in HDAC-inhibitors, we aimed to functionally validate this finding. The association between C/EBPδ mRNA expression and sensitivity to panobinostat is replicated in LC 2 from CTRPv2, with the pairwise relationship also being highly significant (Spearman $\rho = -0.30$, $p < 1 \times 10^{-15}$).

The relationship between C/EBPδ expression and sensitivity to panobinostat is strongest in breast cancer cell lines (Fig 5a). Of these, we chose the invasive ductal carcinoma cell line BT-549 as our model system, since BT-549 expresses relatively high levels of C/EBPδ and has low sensitivity to panobinostat (Fig 5b), allowing us to effectively test the hypothesis that knocking down C/EBPδ will increase panobinostat sensitivity. We treated BT-549 cells overnight with either a short interfering (si) RNA targeting C/EBPδ or a control non-targeting siRNA. We confirmed cells remained viable after C/EBPδ knock-down (S10a Fig) and that the knock-down was successful by Western blot (S10b Fig). Cells were then incubated for 24h at different panobinostat concentrations. The C/EBPδ knock-down (KD) increases sensitivity to panobinostat: a reduction in viability to 66% is detectable for the KD condition at a drug concentration of 1μM, whereas for the control cells viability remains at 97% (Fig 5c). At 10μM viability is reduced to 46% and 63% for the KD and control respectively. Using 4-parameter log-logistic growth curves estimated using the `drc` R package [39] IC50 values are 5.1μM (KD) and 17.0μM (control). Based on a ANOVA comparing one vs. two cubic spline fits, the difference in these response curves is statistically significant ($p = 0.03$).

Discussion

We introduced `Lacrosse`, a model capable of learning latent characteristics which explain observed sensitivity of cancer cell lines to groups of related drugs, and which are predictable from genomic features. Using a Bayesian nonparametric approach `Lacrosse` is able to

adaptively learn an appropriate number of latent characteristics from data. Compared to kernel methods such as KBMTL, `Lacrosse`'s underlying sparse regression allows straightforward interpretation via visual inspection of the small number of LCs. `Lacrosse` allows straightforward incorporation of "soft" prior knowledge in the form of a graph over drugs representing known drug-drug relationships (shared inhibition targets in our analysis). Including such prior information in a non-Bayesian method such as nuclear norm regression would be challenging. `Lacrosse` uncovers both known and novel associations, one of which we were able to validate experimentally. Our finding that reducing C/EBPδ expression increases sensitivity to panobinostat suggests a therapeutic avenue if C/EBPδ expression could be reduced *in vivo* using a similar siRNA approach or alternative targeted methods such as anti-sense oligonucleotides [40].

Our positive functional validation result is encouraging, but there are limitations of our analysis. Genomic features, especially gene expression and CNV, have high levels of correlation so it is difficult to say which out of set of correlated features is most likely to be causally related to sensitivity. However, correlation is less problematic if we are primarily interested in the model's ability to predict sensitivity, a valuable task in its own right due to its potential application in precision medicine. `Lacrosse` is relatively computational intensive, taking around 10h to complete 10,000 Gibbs sampling iterations on a modern workstation, compared to only around 20min for group Lasso (including cross-validation) on the same data. However, any future clinical application would not require retraining the model, only performing prediction, which is extremely fast once the model is trained. We experimented with running `Lacrosse` on subsets of cell-lines corresponding to specific cancer types, but found the reduced sample size resulted in substantially reduced predictive performance and a smaller number of LCs. One interesting future direction would be to extend `Lacrosse` to be multi-task over not only drugs but also cancer types, allowing sharing of statistical strength across all cell lines but allowing tissue-of-origin specific associations when supported by the data.

We focused here on the CCLE and CTRPv2 dataset due to their dense coverage of drug-cell line pairs. In principle it would be beneficial to integrate additional large scale drug viability assays such as GDSC [3]. However, the agreement between these screens is modest due to the use of different growth assays and other technical variation [41] which would make such an analysis extremely daunting. CCLE does not include any epigenomic features such as methylation so we were not able to incorporate this into the model. However, it would be straightforward to include such data if it were available for a different dataset.

An intriguing direction is to utilize more features of the drugs than just known inhibition targets. For example, Menden et al. [13] showed promising performance summarizing the chemical structure of each drug as a feature vector using PaDEL-Descriptor [15], a popular chemo-informatics tool. An alternative would be to leverage a graph kernel like Weisfeiler-Lehman [42]. This avenue opens the exciting possibility of generalizing not only to new cell lines but also to new drugs.

Methods

Data acquisition and pre-processing

Cell line molecular profiles and drug sensitivity scores were obtained from the CCLE data portal. For mRNA expression gene-centric RMA-normalized values were log and then z-transformed (analysis date 2012-09-29). Mutations in 1651 genes determined by hybrid capture sequencing were summarized to a binary variable for each assayed gene denoting 1 for any nonsynonymous change and 0 otherwise (analysis date 2012-05-07). We used the CCLE recommended data where variants that are a) common polymorphisms, b) have allelic fraction $<10\%$, c) are putative neutral variants or d) are located outside of the CDS for all transcripts

are removed. Per gene DNA copy number as determined using Affymetrix SNP6.0 arrays were obtained (analysis date 2013-12-03). Pharmacologic profiles were taken from *CCLE_NP24.2009_Drug_data_2015.02.24.csv*, specifically columns "IC50" and "ActArea".

For CTRPv2 data was obtained using PharmacoGx [43]. The mRNA expression, copy number and mutation data is the same as for CCLE and were pre-processed analogously. PharmacoGx recomputed active area scores were used as the response.

The Indian buffet process

The Indian buffet process [44, IBP] defines a distribution over infinite binary matrices, which can be used to construct latent feature models where the number of features is unbounded a priori. Models constructed using the IBP are sparse in that only a small number of features are typically active for each entity. The IBP is the infinite limit of the finite K model,

$$v_k|\alpha \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right) \tag{2}$$

$$Z_{dk}|v_k \sim \text{Bernoulli}(v_k) \tag{3}$$

Taking the limit $K \to \infty$, and rearranging columns of $\mathbf{Z}$ carefully, we obtain a stochastic process most easily described by a culinary metaphor. Consider a buffet with a seemingly infinite number of dishes (latent factors/LCs) arranged in a line. The first customer (drugs) starts at the left and samples Poisson(α) dishes. The dth customer (drug) moves from left to right sampling dishes with probability $\frac{m_k}{d}$ where m_k is the number of customers to have previously sampled dish k. Having reached the end of the previously sampled dishes, he tries Poisson$\left(\frac{\alpha}{d}\right)$ new dishes. Element Z_{dk} of the $D \times K$ binary feature allocation matrix $\mathbf{Z}$ is 1 if and only if customer d tried dish k. In `Lacrosse` element Z_{dk} corresponds to whether drug d is influenced by LC k.

Statistical model

In the `Lacrosse` generative process the Indian buffet process matrix, $\mathbf{Z}$, determines which elements of a factor loadings matrix, $\mathbf{G}$ are non-zero. $\mathbf{G}$ is then used in the model,

$$\overset{[D\times N]}{\mathbf{Y}} = \overset{[D\times K]}{\mathbf{G}}\ \overset{[K\times N]}{\mathbf{X}} + \epsilon, \qquad \overset{[K\times N]}{\mathbf{X}} = \overset{[K\times P]}{\mathbf{B}}\ \overset{[P\times N]}{\mathbf{F}} + w,$$

where ϵ, w represent noise, $\mathbf{Y}$ are the sensitivity measurements, $\mathbf{X}$ are the latent factors, $\mathbf{B}$ are regression coefficients, and $\mathbf{F}$ are the genomic features. For our application D is the number of drugs, N is the number of cell lines and P is the number of molecular features, including gene expression, CNV and mutations. Here K is the number of latent factors. The model is closely related to reduced rank regression, with the addition of the noise w on $\mathbf{X}$.

We use a spike and slab prior on elements of $\mathbf{G}$:

$$G_{dk}|Z_{dk} \sim Z_{dk}\mathcal{N}(0,1) + (1 - Z_{dk})\delta_0 \tag{4}$$

where $Z_{dk} \in \{0, 1\}$ indicates whether G_{dk} is non-zero, $\mathcal{N}$ is the Gaussian distribution, and δ_0 is a delta point mass at 0. By using the Indian buffet process as a prior on $\mathbf{Z}$ we allow the model an unbounded number of latent characteristics K.

We also use a spike and slab prior on $\mathbf{B}$,

$$\begin{aligned} B_{kp}|V_{kp} &\sim V_{kp}\mathcal{N}(0,1) + (1 - V_{kp})\delta_0, \\ V_{kp} &\sim \text{Bernoulli}(\pi_p) \qquad \pi_p \sim \text{Beta}(\beta/P, 1) \end{aligned} \tag{5}$$

We know that the sensitivity profile for some drugs is more easily predicted than others, so we use diagonal rather than spherical (isotropic) noise on **Y**, specifically the hierarchical prior

$$\begin{aligned} \epsilon_{dn}|\lambda^{e} &\sim N(0, 1/\lambda_d^e), \\ \lambda_d^e|b &\sim G(1, 1/b), \\ b &\sim G(1, 1) \end{aligned} \tag{6}$$

where G is the Gamma distribution. We use an analogous prior on the noise for **X**, denoted w, since some latent characteristics may be better predicted by the molecular characteristics than others.

Inference is performed using 10,000 iterations of standard Gibbs sampling [45] implemented in C++ using `Eigen` (eigen.tuxfamily.org) and interfaced to R using `RCpp/RCppEigen` [46].

Markov random field extension

The relationships between drugs and between features are easily represented by a Markov random field (MRF). For example, two drugs sharing a common target molecule are linked in the MRF, and two any features such CNV and gene expression for the *same* gene will have an edge between them. Following [47] we modify the IBP probability (see Eq 2) of a column $Z_{:k}$ to be

$$P(Z_{:k}|\tau_k) \propto \underbrace{\exp\left(\sum_{d'<d} w_{d'd} Z_{d'k} Z_{dk}\right)}_{\text{MRF term}} \underbrace{\prod_d \nu_k^{Z_{dk}} (1-\nu_k)^{1-Z_{dk}}}_{\text{usual IBP term}} \tag{7}$$

where $w_{d'd}$ is the edge weight between drugs d' and d in the MRF. We use an analogous prior on V_{kp} to couple the probability of having non-zero coefficients in **B** for different features associated with the same gene.

Cell culture and siRNA transfections

Human breast carcinoma cell line BT-549 was obtained from American Type Culture Collection (Manassas, VA) and cultured in RPMI-1640 medium, 2mM L-Glutamine (Gibco, ThermoFisher Scientific, Waltham, MA), supplemented with 10% FBS (Gibco, ThermoFisher Scientific, Waltham, MA), and 0.023 U/ml (BT-549) in a humidified atmosphere with 5% CO_2 at 37˚C.

For CCAAT/enhancer binding protein delta (C/EBPδ) silencing, cells were seeded (150 000 into 6-well plate or 10 000 cells into 96-well plate) and grown overnight prior to transfection. Cells were transfected using a non-targeting *Silencer* Select Negative Control siRNA (4390843, ThermoFisher Scientific, Waltham, MA) or siRNA targeting C/EBPδ (s2895, 4392420, ThermoFisher Scientific, Waltham, MA) using Lipofectamine 2000 (Invitrogen, ThermoFisher Scientific, Waltham, MA) according to the manufacturer protocol. Reagents were diluted in Opti-MEM reduced serum medium (Gibco, ThermoFisher Scientific, Waltham, MA) and transfection complexes were added to the cells at a final concentration of 20 nM. Transfection media was replaced with 10% FBS antibiotic-free RPMI with panobinostat (range of concentration from 0.01μM to 10μM) after overnight incubation with siRNAs and incubated for 24 h. Cells were harvested for assessment of knock-down efficiency using Western blot analysis or viability using RealTime-Glo MT Cell Viability Assay (Promega, Madison, WI) according to the manufacturer protocol.

Western blot

Cell protein extracts were prepared with lysis buffer containing 50 mM Tris pH 8, 2% sodium dodecyl sulphate (SDS), 5 mM Ethylenediaminetetraacetic acid (EDTA), 5 mM Ethylene glycol-bis (2-aminoethylether)-N,N,N',N'-tetraacetic acid (EGTA), 25 mM sodium fluoride (NaF) and 1 mM sodium orthovanadate (Na_3VO_4) supplemented with the protein inhibitor cocktail Complete Mini, EDTA-free (Roche, Indianapolis, IN). Lysates were briefly sonicated, vortexed, incubated 5 min at 4˚C and vortexed again. Cellular debris were cleared by centrifugation at 12 000 rpm during 10 min. Supernatants were aliquoted and stored at -80˚C for further use. Protein quantification assay was performed using a BCA Protein Assay kit (Pierce, ThermoFisher Scientific, Waltham, MA). The protein extracts (15 μg) were applied on a 12% polyacrylamide-SDS gel electrophoresed at 200V during 45 min and transferred to a Immobilon transfer membrane (EMD Millipore, Billerica, MA) using the Mini Trans-Blot Cell (Bio-Rad, Hercules, CA) settled at 160V for 1 h. The membrane was blocked with 5% reconstituted skim milk powder in TBST solution (10 mM Tris–HCl pH 7.4 containing 150 mM NaCl and 0.05% Tween 20). The blots were incubated with CEBPD antibody (1:500, ab65081, Abcam, Burlingame, CA) in TBST overnight at 4˚C. After washing with TBST, horseradish peroxidase-conjugated secondary antibodies (1:10 000, ab97051, Abcam, Burlingame, CA) were applied and the blots developed by the Enhanced Chemiluminescence Detection System (Pierce, ThermoFisher Scientific, Waltham, MA). Levels of beta-tubulin were used as an internal standard for equal loading.

Supporting information

S1 Fig. Active area is an alternative summary metric for dose-response curves. It has advantages relative to the more popular IC50: it is well-defined even if no growth inhibition is observed at any tested drug concentration, it is less sensitive to measurements around the IC50 point, and better represents whether there is a tail of resistance cells.
(EPS)

S2 Fig. Predictive performance on CTRPv2 assessed using mean concordance index across drugs (10-fold cross-validation).
(EPS)

S3 Fig. Lacrosse's performance on CTRPv2 is robust to the exact choice of the MRF strength parameter (edge weight).
(EPS)

S4 Fig. LOBICO [18] attempts to find small logical networks that combine mutation status to predict binarized drug response. We tested LOBICO using only mutation data from CTRPv2 and found it was consistently and severely outperformed by LASSO in terms of predictive performance. LASSO here refers to linear regression on the AUCs, whereas GLM is L1-regularized logistic regression on the same binarized response as for LOBICO. 10-fold cross-validation was used for all methods: to choose λ for the regression approaches and the model complexity for LOBICO (using the same 8 complexity settings as used in the LOBICO paper). It is possible that LOBICO would be more competitive if we used a smaller set of known important mutations. Since we are interested in discovering such relationships *de novo* we did not explore this approach further.
(EPS)

S5 Fig. Lacrosse's generalization performance in CCLE having been trained on CTRPv2.
(EPS)

S6 Fig. Lacrosse's generalization performance in gCSI having been trained on CTRPv2.
(EPS)

S7 Fig. Lacrosse's generalization performance in GDSC1000 having been trained on CTRPv2.
(EPS)

S8 Fig. Other latent characteristics discovered by `Lacrosse` using CCLE, showing the drugs in the LC and predictive genomic features. *p*-values are from a Spearman correlation test. LCs noted in the text are shown in Fig 3.
(EPS)

S9 Fig. Comparing the global latent characteristic discovered by `Lacrosse` to the general level of drug sensitivity (GLDS) described by Geeleher et al. [27].
(EPS)

S10 Fig. Successful knock-down of C/EBPδ using siRNA in the BT-549 breast cancer cell-line. a. By bright-field microscopy cells appear healthy/viable after knock-down. **b**. Western blot analysis confirms that C/EBPδ protein levels are substantially reduced following overnight (O/N) treatment with the targeting siRNA, and that this knock-down remains substantial after 24 hours.
(TIFF)

Author Contributions

Conceptualization: David A. Knowles, Sylvia Plevritis.

Formal analysis: David A. Knowles.

Funding acquisition: Sylvia Plevritis.

Investigation: David A. Knowles, Gina Bouchard, Sylvia Plevritis.

Methodology: David A. Knowles, Sylvia Plevritis.

Project administration: Sylvia Plevritis.

Software: David A. Knowles.

Supervision: Sylvia Plevritis.

Validation: Gina Bouchard.

Visualization: David A. Knowles.

Writing – original draft: David A. Knowles, Gina Bouchard, Sylvia Plevritis.

Writing – review & editing: David A. Knowles, Sylvia Plevritis.

References

1. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483 (7391):603–607. https://doi.org/10.1038/nature11003 PMID: 22460905
2. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012; 483(7391):570–575. https://doi.org/10.1038/nature11005 PMID: 22460902
3. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016; 166(3):740–754. https://doi.org/10.1016/j.cell.2016.06.017 PMID: 27397505

4. Heiser LM, Sadanandam A, Kuo WL, Benz SC, Goldstein TC, Ng S, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proceedings of the National Academy of Sciences. 2012; 109(8):2724–2729. https://doi.org/10.1073/pnas.1018854108
5. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology. 2014; 32(12):1202–1212. https://doi.org/10.1038/nbt.2877 PMID: 24880487
6. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. 2015; 5(11):1210–1223. https://doi.org/10.1158/2159-8290.CD-15-0235 PMID: 26482930
7. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
8. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. cell. 2011; 144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230
9. Evan GI, Vousden KH. Proliferation, cell cycle and apoptosis in cancer. Nature. 2001; 411(6835):342–348. https://doi.org/10.1038/35077213 PMID: 11357141
10. Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. Cancer Epidemiology Biomarkers & Prevention. 2002; 11(12):1513–1530.
11. Weinstein IB, Joe AK. Mechanisms of disease: oncogene addiction, a rationale for molecular targeting in cancer therapy. Nature Clinical Practice Oncology. 2006; 3(8):448–457. https://doi.org/10.1038/ncponc0558 PMID: 16894390
12. Caruana R. Multitask learning. Machine learning. 1997; 28(1):41–75. https://doi.org/10.1023/A:1007379606734
13. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS ONE. 2013; 8(4):e61318. https://doi.org/10.1371/journal.pone.0061318 PMID: 23646105
14. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958; 65(6):386. PMID: 13602029
15. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 2011; 32(7):1466–1474. https://doi.org/10.1002/jcc.21707 PMID: 21425294
16. Gönen M, Margolin AA. Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. Bioinformatics. 2014; 30(17):i556–i563. https://doi.org/10.1093/bioinformatics/btu464 PMID: 25161247
17. Ammad-ud din M, Khan SA, Wennerberg K, Aittokallio T. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics. 2017; 33(14):i359–i368. https://doi.org/10.1093/bioinformatics/btx266 PMID: 28881998
18. Knijnenburg TA, Klau GW, Iorio F, Garnett MJ, McDermott U, Shmulevich I, et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. Sci Rep. 2016; 6:36812. https://doi.org/10.1038/srep36812 PMID: 27876821
19. El-Hachem N, Gendoo DMA, Ghoraie LS, Safikhani Z, Smirnov P, Chung C, et al. Integrative Cancer Pharmacogenomics to Infer Large-Scale Drug Taxonomy. Cancer Res. 2017; 77(11):3057–3069. https://doi.org/10.1158/0008-5472.CAN-17-0096 PMID: 28314784
20. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11(3):333–337. https://doi.org/10.1038/nmeth.2810 PMID: 24464287
21. Evgeniou A, Pontil M. Multi-task feature learning. Advances in neural information processing systems. 2007; 19:41.
22. Knowles DA, Ghahramani Z. Nonparametric Bayesian sparse factor models with application to gene expression modeling. The Annals of Applied Statistics. 2011; 5(2B):1534–1552. https://doi.org/10.1214/10-AOAS435
23. Hernández-Lobato JM, Hernández-Lobato D, Suárez A. Expectation propagation in linear regression models with spike-and-slab priors. Machine Learning. 2015; 99(3):437–487. https://doi.org/10.1007/s10994-014-5475-7
24. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; 58:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
25. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

26. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol. 2015; 33(3):306–312. https://doi.org/10.1038/nbt.3080 PMID: 25485619

27. Geeleher P, Cox NJ, Huang RS. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. Genome Biol. 2016; 17(1):190. https://doi.org/10.1186/s13059-016-1050-9 PMID: 27654937

28. Okabe H, Lee SH, Phuchareon J, Albertson DG, McCormick F, Tetsu O. A critical role for FBXW8 and MAPK in cyclin D1 degradation and cancer cell proliferation. PloS one. 2006; 1(1):e128. https://doi.org/10.1371/journal.pone.0000128 PMID: 17205132

29. Zoppoli G, Regairaz M, Leo E, Reinhold WC, Varma S, Ballestrero A, et al. Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. Proceedings of the National Academy of Sciences. 2012; 109(37):15030–15035. https://doi.org/10.1073/pnas.1205943109

30. Kobayashi M, Salomon C, Tapia J, Illanes SE, Mitchell MD, Rice GE. Ovarian cancer cell invasiveness is associated with discordant exosomal sequestration of Let-7 miRNA and miR-200. J Transl Med. 2014; 12(4).

31. Sanchez A, Setien F, Martinez N, Oliva J, Herranz M, Fraga M, et al. Epigenetic inactivation of the ERK inhibitor Spry2 in B-cell diffuse lymphomas. Oncogene. 2008; 27(36):4969–4972. https://doi.org/10.1038/onc.2008.129 PMID: 18427547

32. Lazarczyk M, Pons C, Mendoza JA, Cassonnet P, Jacob Y, Favre M. Regulation of cellular zinc balance as a potential mechanism of EVER-mediated protection against pathogenesis by cutaneous oncogenic human papillomaviruses. Journal of Experimental Medicine. 2008; 205(1):35–42. https://doi.org/10.1084/jem.20071311 PMID: 18158319

33. Mor M, Beharier O, Levy S, Kahn J, Dror S, Blumenthal D, et al. ZnT-1 enhances the activity and surface expression of T-type calcium channels through activation of Ras-ERK signaling. American Journal of Physiology-Cell Physiology. 2012; 303(2):C192–C203. https://doi.org/10.1152/ajpcell.00427.2011 PMID: 22572848

34. Schmidt D, Textor B, Pein OT, Licht AH, Andrecht S, Sator-Schmitt M, et al. Critical role for NF-κB-induced JunB in VEGF regulation and tumor angiogenesis. The EMBO journal. 2007; 26(3):710–719. https://doi.org/10.1038/sj.emboj.7601539 PMID: 17255940

35. Bauerschlag DO, Ammerpohl O, Bräutigam K, Schem C, Lin Q, Weigel MT, et al. Progression-free survival in ovarian cancer is reflected in epigenetic DNA methylation profiles. Oncology. 2011; 80(1-2):12–20. https://doi.org/10.1159/000327746 PMID: 21577013

36. Gery S, Tanosaki S, Hofmann WK, Koppel A, Koeffler HP. C/EBPδ expression in a BCR-ABL-positive cell line induces growth arrest and myeloid differentiation. Oncogene. 2005; 24(9):1589–1597. https://doi.org/10.1038/sj.onc.1208393 PMID: 15674331

37. Cooper LAD, Gutman DA, Chisolm C, Appin C, Kong J, Rong Y, et al. The tumor microenvironment strongly impacts master transcriptional regulators and gene expression class of glioblastoma. American Journal of Pathology. 2012; 180(5):2108–2119. https://doi.org/10.1016/j.ajpath.2012.01.040 PMID: 22440258

38. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010; 463(7279):318–325. https://doi.org/10.1038/nature08712 PMID: 20032975

39. Ritz C, Baty F, Streibig JC, Gerhard D. Dose-Response Analysis Using R. PLOS ONE. 2015; 10 (e0146021).

40. Chan JH, Lim S, Wong W. Antisense oligonucleotides: from design to therapeutic application. Clinical and Experimental Pharmacology and Physiology. 2006; 33(5-6):533–540. https://doi.org/10.1111/j.1440-1681.2006.04403.x PMID: 16700890

41. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, et al. Inconsistency in large pharmacogenomic studies. Nature. 2013; 504(7480):389–393. https://doi.org/10.1038/nature12831 PMID: 24284626

42. Shervashidze N, Schweitzer P, Leeuwen EJv, Mehlhorn K, Borgwardt KM. Weisfeiler-lehman graph kernels. Journal of Machine Learning Research. 2011; 12(Sep):2539–2561.

43. Smirnov P, Safikhani Z, El-Hachem N, Wang D, She A, Olsen C, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. Bioinformatics. 2016; 32(8):1244–1246. https://doi.org/10.1093/bioinformatics/btv723 PMID: 26656004

44. Griffiths TL, Ghahramani Z. The Indian buffet process: An introduction and review. Journal of Machine Learning Research. 2011; 12:1185–1224.

45. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984; 6:721–741. https://doi.org/10.1109/TPAMI.1984.4767596 PMID: 22499653

46. Eddelbuettel D, François R. Rcpp: Seamless R and C++ Integration. Journal of Statistical Software. 2011; 40(8):1–18. https://doi.org/10.18637/jss.v040.i08

47. Hai-son PL, Bar-Joseph Z. Inferring interaction networks using the IBP applied to microRNA target prediction. In: Advances in Neural Information Processing Systems. Curran Associates; 2011. p. 235–243.