*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

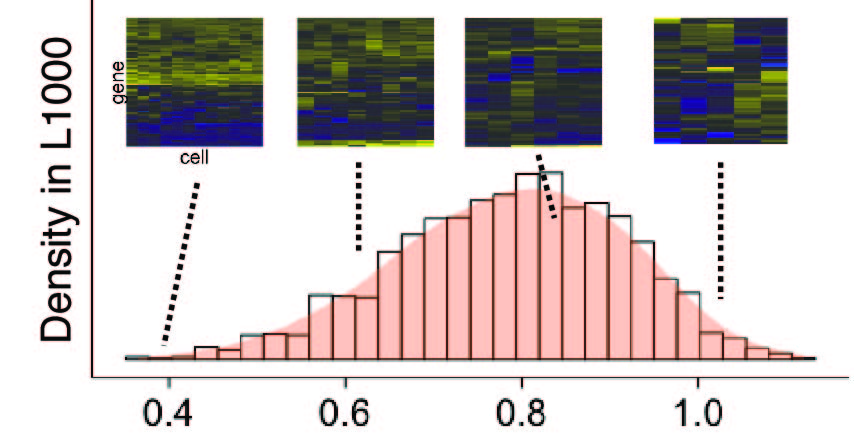
Original Paper

|  |
| --- |
| Gene expression  **Cell-specific predictionof drug-induced gene expression profiles**  Rachel Hodos1,2, Hao Chih Lee1, Ping Zhang3, Qiaonan Duan1, Zichen Wang1, Neil R. Clark1, Avi Ma’ayan1, Fei Wang3,4, Brian Kidd1, David Sontag2, Jianying Hu3,Joel Dudley1\*  1Icahn School of Medicine at Mt. Sinai, New York, NY, 10029, 2Courant Institute of Mathematical Sciences, New York University, New York, NY, 10012, 3IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598, 4Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York, NY, 10065  \*To whom correspondence should be addressed.  Associate Editor: XXXXXXX  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Abstract  **Motivation:** Gene expression profiling of *in vitro* drug perturbations is useful for many biomedical discovery applications including drug repurposing and elucidation of drug mechanisms. However, limited data availability across cell types has hindered our capacity to leverage or explore the cell-specificity of these perturbations. While recent efforts have generated a large number of drug perturbation profiles across a variety of cell types*,* many gaps remain in this combinatorial drug/cell space. Hence, we asked whether it is possible to fill these gaps by predicting cell-specific drug perturbation profiles using available expression data from related conditions--i.e. from other drugs and cell types.  **Results:** We developed a computational framework that first arranges existing profiles into a three-dimensional array (or tensor) indexed by drugs, genes, and cell types, and then uses either local (nearest neighbors) or global (tensor completion) information to predict unmeasured profiles. We evaluated the results using cross-validation and drug repurposing metrics, as well as comparison to external drug profiles. The two methods had generally comparable performance, with some distinctions depending on the input data dimensions. Predictions achieved correlations of 0.54 between true and predicted expression values, and in many cases had stronger correspondence with independently measured drug profiles than their measured counterparts. Further, the predicted profiles maintain biologically and therapeutically relevant properties such as accurate differentially expressed genes (AUC 0.81), gene correlation structure, and connection to disease signatures for drug repurposing.  **Contact:** joel.dudley@mssm.edu  **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# Introduction

Genome-wide expression profiling of *in vitro* drug perturbations has proven to be useful for many aspects of drug discovery and development (Qu and Rajpal, 2012). Applications include elucidation of drug mechanisms (Wei, et al., 2006), lead identification (Hassane, et al., 2008), and drug repurposing (Dudley, et al., 2011; Hu and Agarwal, 2009). Despite this success, the capacity to leverage cell-specific responses has been hindered by limited data availability across cell types (Khan, et al., 2014; Parkkinen and Kaski, 2014). To address this limitation, the Library of Integrated Cellular Signatures (LINCS) program (Duan, et al., 2014; NIH, 2014) has greatly expanded the publicly available data to nearly one million profiles characterizing thousands of drugs exposed to dozens of cell types. However, this combinatorial space of drugs and cell types is vast, and many gaps remain in this space (see Figure 7A). These gaps present difficulties both for large-scale analysis as well as for making cell-resolved comparisons, e.g. between two drugs or between drug and disease. Therefore, we asked whether it is possible to leverage existing expression profiles to predict the remaining unmeasured profiles.

Expression responses to drug exposure are often highly cell-specific, e.g. due to differences in expression of drug targets. Indeed, we observe a high degree of cell-specificity for many drugs in the LINCS data (see Figure 1). The utility of such cell-specific gene expression has previously been demonstrated for a variety of applications. For example, (Iwata, et al., 2017) showed that using transcriptional similarity to predict drug-target interactions works better when comparing drug profiles in the same cell line. Another study (Khan, et al., 2014) identified cell-specific expression responses with potential therapeutic relevance, e.g. anti-estrogen drugs inducing a common response that was specific to breast cancer cells.



Prior studies have described methods to predict expression profiles using outside information. For example, (Gamazon, et al., 2015) predict tissue-specific expression profiles from genetic variants, but are limited to basal-state expression. Conversely, (Lagunin, et al., 2013) predict drug-induced expression responses from a drug’s chemical structure, but are agnostic to cell type. There are also many techniques to impute missing entries of a gene expression matrix, generally using either local (e.g. nearest neighbors) or global (e.g. low-rank matrix approximation) information (Brock, et al., 2008; Troyanskaya, et al., 2001). However, most of these methods are not directly applicable to our setting, as they rely on having at least some measurements available in the target experimental setting.

Fig. 1: Distribution of cell-specificity of 2,130 drugs in the L1000 dataset. The cell-specificity of each drug is quantified as the mean pairwise cosine distance between its expression profiles in all available cell lines, with a range of 0 (all cells identical) to 2 (perfect anti-correlation). Four representative examples are shown (L to R: homoharringtonine, terfenadine, dexamethasone, and JNJ-38877605). While some drugs such as homoharringtonine induce very similar expression across cell types, the majority of L1000 drugs have higher cell-specificity corresponding to distinctive patterns in different cell types.

Here, we draw inspiration from this prior work to solve a new problem: predicting *entire* expression profiles for cell-specific drug perturbations. Our two approaches are complementary in their use of local vs. global information. The “local” algorithm, *Drug Neighbor Profile Prediction* (DNPP) is inspired by K-nearest neighbors but adapted to this *de novo* prediction setting. The “global” algorithm, the *Fast, Low-Rank Tensor Completion* (FaLRTC) algorithm (Liu, et al., 2013) fills in the missing entries of a tensor (see Figure 2A) using the observed entries. The underlying assumption behind this and many other tensor completion algorithms is that the data are low-rank, i.e. some small set of underlying factors (e.g. drug targets) explain most of the variation in the data.

We evaluate our methods, along with two simple baselines, using several approaches. We use cross-validation (CV) to measure performance readouts including correlation of predicted and true expression values, accuracy of differentially expressed genes (DEGs), preservation of gene correlation structure, and preservation of cell-specificity. We analyze accuracy per drug, per gene and per cell type. We also study the dependence of accuracy on the amount of input data. In addition to these internal, CV-based evaluations, we perform external evaluation against an independently measured drug profile database. Finally, we perform an application-oriented evaluation of method performance by assessing preservation of connection to disease signatures in a drug repurposing framework.

# Methods

## Notation and terminology

*T* refers to a data tensor, where *Td,g,c*is the effect of drug *d* on gene *g* in cell *c*. A colon replacing any of these indices refers to all elements along that dimension. *Cd* and *Dc* respectively refer to the set of cell lines measured for drug *d*,and the set of drugs measured in cell *c.* All error bars in figures and text refer to ± one standard deviation. All correlations are Pearson correlations, denoted in equations as or by the letter *r*. We use the term ‘drug’ loosely to refer to compounds represented in the data, which includes approved drugs, drug-like compounds, and tool compounds.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cell line | MCF7 | VCAP | PC3 | A375 | A549 | HA1E | HT29 | HCC515 | HEPG2 | NPC |
| tissue | breast | prostate | prostate | skin | lung | kidney | colon | lung | liver | brain |
| # profiles | 1505 | 1368 | 1340 | 1168 | 1139 | 1127 | 1022 | 934 | 798 | 441 |

Table 1. The top ten cell types represented in the (largest) data tensor, along with the tissue of origin and number of drug profiles available.

## Data processing

**The LINCS drug expression data** (herein, the “L1000 data”) is measured on a targeted expression profiling platform called L1000. The platform measures the expression of 978 “landmark” genes (roughly 1000, hence the name), selected to be maximally predictive of the other genes while being widely expressed across many cell and tissue types.

**Differential expression** computed from the level 3 L1000 data were downloaded from <amp.pharm.mssm.edu/public/L1000CDS_download>. The dataset was generated using the Characteristic Direction (CD) method (Clark, et al., 2014) and is validated and described more fully in (Duan, et al., 2016). Briefly, a CD was calculated for each replicate using linear discriminate analysis, to find the direction in gene space that best separates cases and controls. Replicates were averaged and normalized to unit length. Average cosine distance (ACD), i.e. the mean pairwise cosine distances between an experiment’s CD replicates, was used to estimate significance. The null distribution of the ACDs was calculated per batch using random sampling (*n* = 10,000) of replicates in the same batch. A *p*-value for each profile (ACD *p*-value) was computed by comparing its ACD to the null.

**Tensor construction:** The 201,484 CD profiles (20,413 drugs, 72 cell types) were filtered to 34,716 profiles (6928 drugs, 72 cell types) with ACD *p* ≤ 0.1 in order to remove the most unreliable data. Drugs and cell types with < 3 remaining experiments were removed, as well as duplicate drug id’s corresponding to the same drug, for a final count of 25,672 profiles (2130 drugs, 71 cells, 12.7% of all CDs). Profiles were averaged across all available concentration and time points, renormalized, and then arranged into a tensor (see Figure 2A). Of the 151,230 possible drug-cell pairs, the tensor contained 15,855, corresponding to 10.5% observation density. Herein we refer to this as *Tlarge* or ‘the large tensor.’ Three subtensors were also defined (see Figure 7A) by varying the drug and cell-type dimensions while keeping all genes, to define: *Tmany‑drugs* (2000 drugs x 15 cell types, 39.5% observation density); *Tmany‑cells* (300 x 50, 38.5%) and *Tsmall* (300 x 15, 71.4%), also called ‘the small tensor.’ Note that *Tsmall* is the intersection of *Tmany‑drugs* and *Tmany‑cells*. The tensor element *Td,g,c* is the *g*th coordinate of the CD vector for drug *d* in cell *c.* All values lie in the range [‑1,1] after normalization, where a positive [negative] value corresponds to up- [down-] regulation. The 10 cell lines with the most data in *Tlarge* are listed in Table 1, along with the corresponding tissue of origin and number of profiles (i.e. drugs) present. Most of the 71 cell lines are cancer cell lines, and represent a range of human tissues including skin, lung, brain, kidney, and prostate. A list of all drugs, genes and cell lines in this tensor, along with some additional information (including prediction accuracy for each entity, see Section 3.6) is provided in Table S1.

**Connectivity Map data:** Expression profiles from the Connectivity Map (CMap) database (Lamb, et al., 2006) were downloaded from <http://research.cs.aalto.fi/pml/software/ProbCMap/>. The data had been processed as described in (Parkkinen and Kaski, 2014). Briefly, 718 drugs were selected out of the original 1,309 in CMap, due to their data availability across all three cell lines (MCF7, PC3, and HL60). The samples were RMA normalized, and then differential expression was computed by comparing each sample to the within-batch mean. Genes were intersected with the L1000 genes to give 930 genes in the overlap. Profiles were matched on both drug and cell type to those in *Tlarge*, yielding 210 profiles representing 135 unique drugs present in both CMap and the L1000 data tensor.

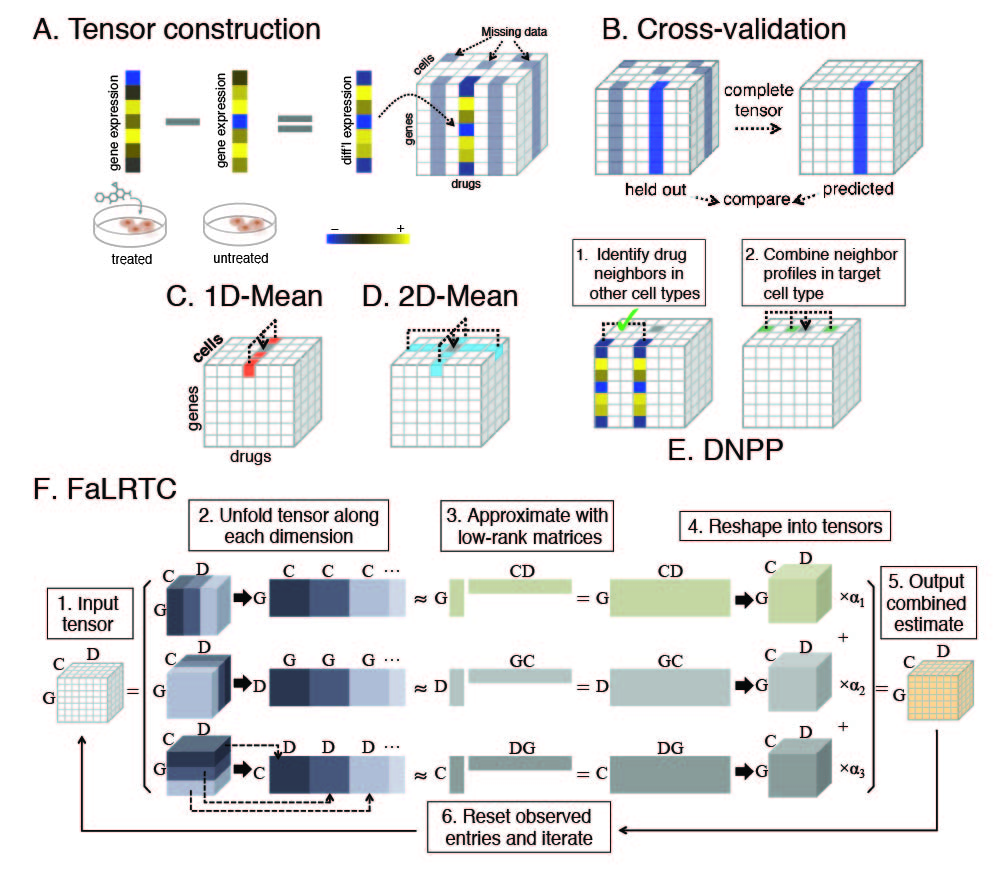


Fig. 2: Schematic overview. A. Differential expression profiles are compiled into a tensor of 978 genes x 2,130 drugs x 71 cell types. Profiles are either completely missing, in grey, or available for all genes, denoted by both the white and multicolor columns.B. CV setup, where entire expression profiles are held out. C-D Averaging baselines; the target expression value is shown in grey and the data used for averaging are colored. E. DNPP algorithm. The target value is in grey. First, drug neighbors are identified by comparing profiles in other cell lines. Then neighbor profiles in the target cell line are combined to form the prediction. F. FaLRTC algorithm. The data tensor is input on the left, and then unfolded in step 2 to form three matrices with dimensions G x CD (top), D x GC (middle), and C x DG (bottom). Each matrix is approximated using a spectral method and then reshaped into a tensor. The three tensors are then combined into one. Observed entries are reset to their initial values, and the process is iterated to minimize the matrix trace norms.

**Disease signatures:** 878 disease signatures were previously curated (see <http://amp.pharm.mssm.edu/CREEDS/downloads/Disease_signatures.json>), by the Ma’ayan Lab, using the Gene Expression Omnibus (GEO) (Edgar, et al., 2002). Differential expression between disease and normal tissues was computed using CD (Clark, et al., 2014), and the top 600 DEGs were identified for each disease signature, along with the corresponding CD values. For each disease signature, the combined ‘up’ and ‘down’ gene sets were intersected with the L1000 genes, resulting in 74 disease signatures (corresponding to 66 unique diseases, see Table S2) with at least 40 DEGs in common with the L1000 genes.

### Cross-validation setup

10-fold CV experiments were performed, where *entire expression profiles* are held out and then predicted (see Figure 2B), randomly selecting 10% of the profiles per fold. All of these predictions are compiled into a tensor, , with the same dimensions and pattern of missing entries as the original tensor. ­Accuracy was measured as the Pearson correlation with truth (PCT, also called gtPCC in (Donner, et al., 2012)). This is defined as where Ω corresponds to some subset of the tensor, the correlation is taken element-wise, and unobserved entries are ignored. For example, Ω might correspond to a particular gene, drug, cell (see Figure 6A, e.g. ), a CV fold, or the entire tensor.

* 1. **Baseline averaging schemes**

While many methods exist to impute randomly missing entries in a gene expression matrix, we are not aware of prior work predicting entire expression profiles without additional data inputs. Thus we use two simple baselines that make predictions by averaging relevant subsets of data.

“1D-Mean” (Figure 2C) predicts missing expression profiles for each drug by averaging all profiles available for that drug in the tensor (i.e. across cell lines). The prediction for drug *d*, cell *c* and gene *g* is given by:

“2D-Mean” (Figure 2D), combines the 1D-Mean average across cell lines with a similar average in the other dimension across drugs, i.e.

We use *λ* = ½ based on CV experiments (results not shown).

* 1. **The Drug Neighbor Profile Prediction algorithm**

The DNPP algorithm (Figure 2E) is an adaptation of K-nearest neighbors (KNN) to the *de novo* prediction setting. In other words, KNN normally requires at least some data present in the target condition in order to identify neighbors. To overcome this limitation, DNPP defines similarity between drugs instead of profiles. The similarity (*S*)between two drugs *d* and *d’* is defined based on average correlation between the two drugs’ profiles as measured in other cell types:

DNPP then estimates the profile for drug *d* and cell *c* as a weighted average of (up to) *K* profiles from cell type *c* corresponding to neighboring drugs. To generate a prediction for (*d, c*), drug neighbors of *d* are chosen only amongst drugs that have data in cell *c*, and hence neighbors can differ per cell type. Finally, the weights on the *K* profiles are chosen proportional to , normalized to sum to 1. We use *K =* 10 (CV results not shown).

### Benchmarking tensor completion algorithms

Six tensor completion algorithms were benchmarked for both speed and accuracy. The algorithms are: Tmac (Xu, et al., 2013); SiLRTC, HaLRTC, and FaLRTC (Liu, et al., 2013) and ‘Constrained’ and ‘AsMatrix’ (Tomioka, et al., 2010). MATLAB code for all algorithms was downloaded and used without any algorithmic changes. Some hyperparameters were hand-tuned while others were kept at default values (see Table S3). 10-fold CV experiments were run on the small tensor. Results and runtimes are shown in Figure S4, along with 1D-Mean, 2D-Mean and DNPP. FaLRTC was selected for further study due to its superior accuracy and efficiency.

### The Fast, Low-Rank Tensor Completion algorithm

The FaLRTC algorithm (Liu, et al., 2013) is sometimes referred to herein as simply ‘Tensor’ or ‘the tensor approach.’ We briefly describe the algorithm here, in a simplified form (see Figure 2F). Like most tensor completion algorithms, FaLRTC assumes that the data has some low-rank structure. While there is a notion of rank for a tensor (Kolda and Bader, 2009), this is in general hard to compute. Hence, FaLRTC resorts to low-rank matrix approximations instead. A three-dimensional tensor can be reshaped or ‘unfolded’ into matrices in three mathematically distinct ways (Kolda and Bader, 2009), i.e. a tensor can be unfolded into a , a , and a matrix. The algorithm forms all three such matrices, and then performs low-rank matrix approximation via a spectral method. The prediction of missing values is based on a weighted combination of the three matrix-derived estimates, where these weights (are user-defined parameters, constrained to be positive and sum to one. Observed elements are reset to their true values, and then this process is iterated using gradient descent to minimize (an upper bound on) the matrix ranks. Due to the column-structured pattern of missing entries in our tensors, gene correlation structure is less useful for predictions than correlations in the other two dimensions, and hence estimates from the matrix (that most strongly leverage gene correlations, were down-weighted by a factor of 100 relative to the other two (i.e. ). This can be seen as an adaptation of the algorithm to the present setting defined by the column-structured pattern of missing entries.

### Calling differentially expressed genes

Since the expression profiles in the tensors actually represent *differential* expression, genes can be labeled as DEGs simply based on the magnitude of their values in the tensor. For each drug-cell profile, genes are ranked by the magnitude of expression, and then genes at (or above) the *p*thpercentile of genes are called as differentially expressed. Note that, based on this approach, a different number of genes could be identified as up-regulated versus down-regulated. Alternative methods for calling DEGs were evaluated (see Figure S5) and found to have only minor effects on the results.

# Results

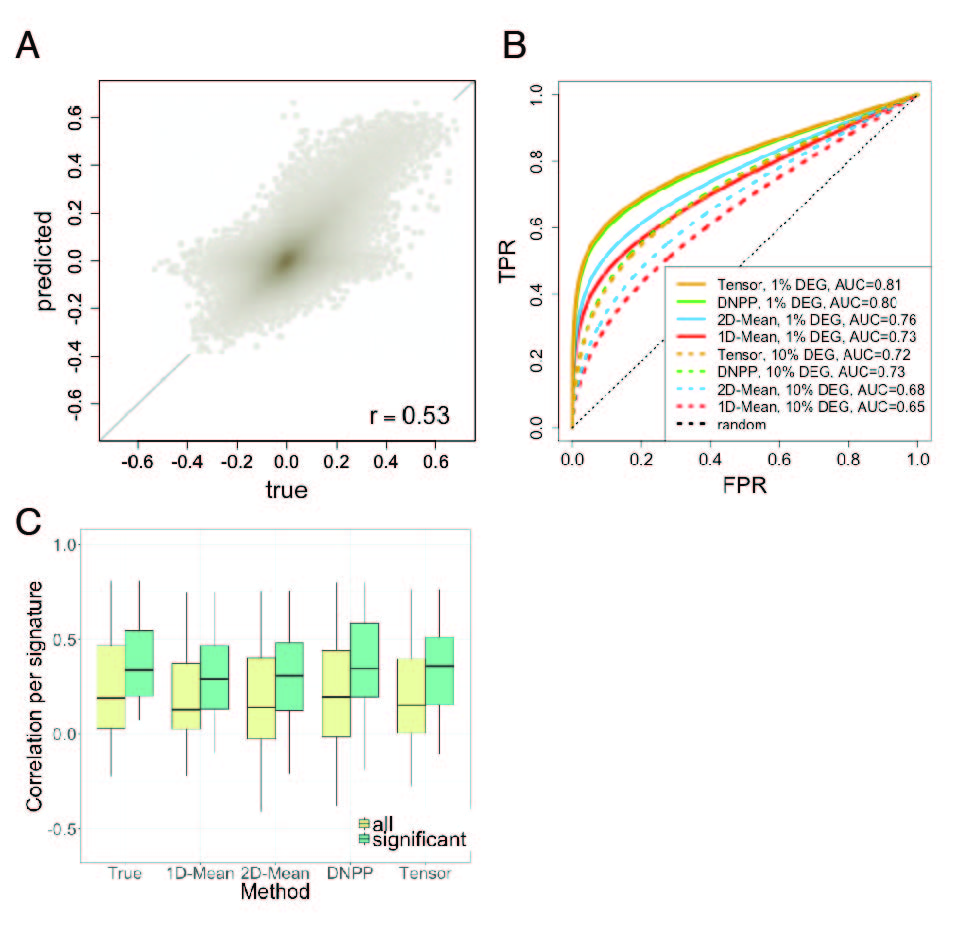
## Unless otherwise noted, experiments were performed on the large tensor, and ordered results correspond to: (Tensor; DNPP; 2D-Mean; 1D-Mean).

## Overall accuracy

First we evaluate the overall correlation between true and predicted values. Figure 3A shows a smoothed scatterplot of all Tensor (FaLRTC) predictions versus true values, where each point corresponds to a single, numeric entry in the tensor. The Pearson correlation with truth (PCT, see Methods) for each method was 0.53, 0.54, 0.46, and 0.40 (see Figure S6).

## Accuracy of DEGs

Next, we evaluated the accuracy of calling DEGs. For each profile, a gene was labeled as a DEG if its absolute expression value was at or above the *p*th percentile relative to all genes in the profile, where *p* was either 1% or 10% (see 2.8). ROC curves shown in Figure 3B were generated by varying an analogous percentile threshold across the range 0-100% for the values in the CV tensors, which defines a set of predicted DEGs for each profile at each possible threshold value. Each ROC curve represents aggregate results across all profiles in the tensor. The methods achieved area under the ROC curve (AUC) values of 0.81, 0.80, 0.76, and 0.73 at *p* = 1%. At *p* = 10%, a similar relationship between methods was observed (0.72, 0.73, 0.68, 0.65). For all four methods, AUC’s were higher at the 1% threshold relative to the 10% threshold, and this pattern is observed more generally (see Figure S5), where smaller values of *p* correspond with higher accuracy. This makes sense in that smaller percentile thresholds correspond to genes with stronger differential expression signals.



**Fig. 3. Overview of accuracy.** A. Scatterplot of Tensor predicted vs. true values. B. ROC curves assessing prediction of DEGs. C. Comparison of either measured (left) or predicted (all others) L1000 drug profiles to a corresponding set of profiles from CMap. Yellow corresponds to all 210 comparisons, while green shows the results on 138 drug profiles for which the correlations were significant and positive.

### Tradeoffs in accuracy with varying input data

As a first step in exploring the applicability of our prediction framework to other datasets or other subsets of the L1000 drug-cell space, we studied the dependence of accuracy on the amount of input data, varying both the tensor dimensions as well as the percent of observed profiles in the tensor.

Tensor dimensionswere varied along the drug and cell-type axes to define four tensors of different sizes, namely *Tlarge*, *Tsmall*, *Tmany‑drugs*, and *Tmany‑cells* (see 2.2 and Figure 7A). 10-fold CV results for all four tensors are shown in Figure 7B. All methods have the best performance on the small tensor (all *p* < 0.01, *t*-test), likely due to its higher observation density, with results on *Tmany‑cells* following close behind. While *Tmany‑cells*and *Tmany‑drugs* have comparable observation density around 39%, all methods have considerably better performanceon *Tmany‑cells*, perhaps indicating that shared information across cell types is more useful than shared information across drugs. These two tensors also help to further highlight a difference between DNPP and Tensor, where DNPP is the top-performer in the case with many drugs, while Tensor dominates with many cell types.

Independently, observation density was varied by subsampling profiles in the small tensorin 10% intervals from 10-60%, evaluating on a held-out set covering another 10% of the tensor. This sampling process was repeated 25 times generating the error bars in Figure 7C. At or above an observation density of 30%, Tensor has superior performance, while at lower densities, 2D-Mean is the top performer. Also, improvement in performance as a function of observation density is most dramatic for the tensor approach.

## Prediction accuracy for each drug, gene and cell type

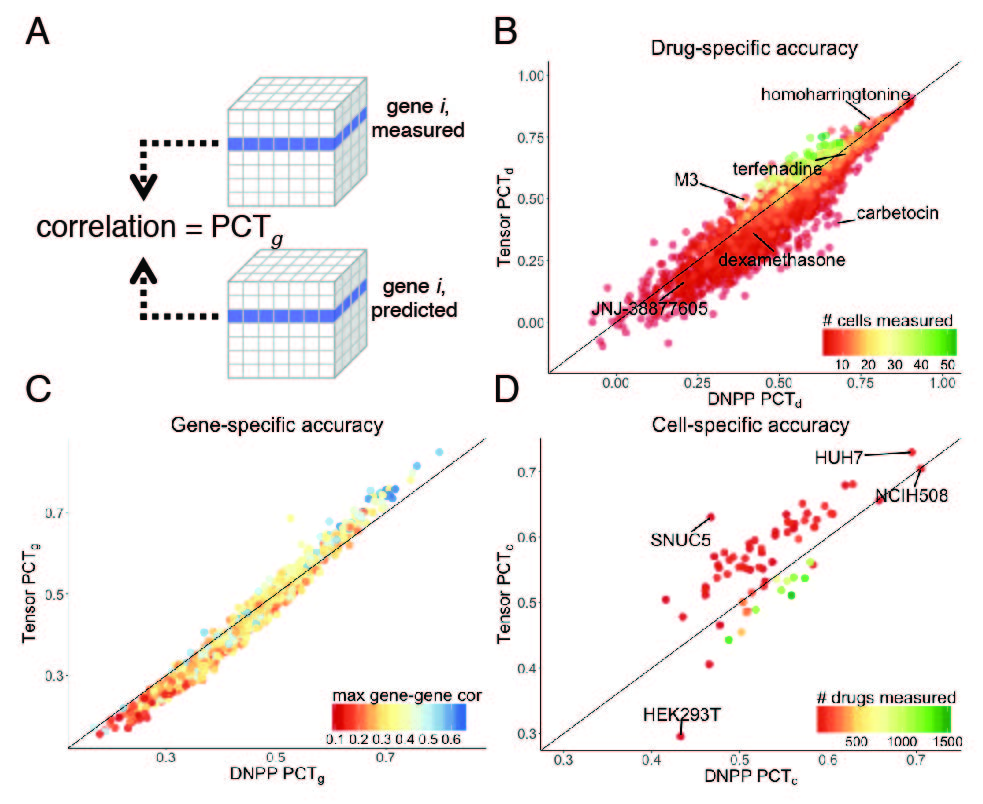


Fig. 6. Entity-specific accuracy. A. Accuracy was evaluated individually for each drug, gene, and cell type, by computing the correlation between true and predicted values for one slice of the tensor corresponding to the chosen entity. The figure illustrates the computation of PCTg. B-D. Drug-, gene-, and cell-specific accuracy, with Tensor results on the y-axis and DNPP results on the x-axis. Each scatterplot is colored by a factor found to be associated with the results (see text for details), with some examples and outliers labeled for reference.

While results may look promising when evaluating across the entire tensor, the question remains as to how accuracy varies in specific cases. We address this here by looking at accuracy for each entity represented in the tensor (see Figure 6A), starting with drugs, and then continuing with genes and cell types. We focus here on the top performers, Tensor and DNPP, but results for all methods are included in Table S1.

**Drug-specific accuracy** is quantified using PCT*d* (see 2.3). Figure 6B shows a scatterplot with points corresponding to each of the 2,130 drugs, with DNPP accuracy on the *x*-axis and Tensor accuracy on the *y*-axis. While DNPP had better performance for the majority (71.9%) of drugs, Tensor outperformed DNPP for drugs that were measured in a large number of cell types (the green region). For both methods, accuracy varies widely across drugs, but is highly correlated (*r* = 0.93) between the two methods. Table S11 lists the top and bottom 40 drugs according to Tensor accuracy. Many of the drugs with high predictive performance are HDAC inhibitors. No strong patterns were observed among the drugs with the lowest performance, but include a wide variety of substances such as CYP inhibitors, lipid-regulating agents, and β-lactam antibiotics.

**Gene-specific accuracy** is similarly presented in Figure 6C. There is once again a high degree of correlation between Tensor and DNPP accuracy (*r* = 0.99). However, a pattern emerges where Tensor outperforms DNPP for genes with higher predictive performance, whereas DNPP performs better with the “more difficult” genes with low predictive performance. One plausible explanation for this difference is the local vs. global nature of the two algorithms; genes that have very context-dependent behaviors would generally be expected to have lower prediction performance and yet be better approximated locally, whereas genes with highly consistent response patterns would likely have better performance with a global algorithm.

There is also a wide range of accuracy across genes. In an attempt to explain this variation, we found that the strongest identified association with PCT*g* for all four methods was the maximum absolute correlation between the gene in question and any other measured gene (*r* = 0.48, 0.45, 0.52, 0.52). Since neither of the averaging baselines shares information across genes, such gene correlations cannot possibly be directly leveraged here. Gene set enrichment analysis (see File S7) revealed that genes with higher predictive performance were enriched for several terms (see File S7) including “cell cycle”, “positive regulation of apoptotic process”, and “mitotic G2 DNA damage checkpoint,” perhaps reflecting that processes such as mitosis and apoptosis are well-conserved across cell types.

**Cell-specific accuracy** is shown in Figure 6D. Tensor and DNPP results are again correlated (*r* = 0.74), although less so than for genes or drugs. A factor that clearly distinguishes performance between the two methods is the number of drugs profiled in each cell type, ranging from 1,505 for MCF7 cells to only 3 for HS27A cells (see Figure 6D). We observe that Tensor generally outperforms DNPP in cell types with fewer drug profiles, while DNPP has better performance in cases with many drugs represented.

**Analysis of cell-specificity**

While some L1000 drugs show very similar responses across cell types, others induce highly cell-specific responses. One such example is M‑3M3FBS (herein “M3”, see Figure 4A), a PLC agonist that induces a variety of effects ranging from modulation of neutrophil function to apoptosis. The large tensor contains M3 profiles in 15 different cell lines, shown on the left of Figure 4A (“True”). Responses cluster into two primary groups, with one group (on the left) enriched for down-regulation of both spindle pole genes as well as valine, leucine, and isoleucine degradation, perhaps indicating a pre-apoptotic response (see File S7 for details of all enrichment analyses). The mean profile of the second group (A549, AGS, RKO, and MCF7 cells) is enriched for very different types of processes including up-regulation of Akt signaling, insulin signaling, and salivary secretion, all of which have established connections to PLC (Eichhorn, et al., 2002; Parrales, et al., 2011). Figure 4A shows that the tensor approach was able to accurately recapitulate these two classes of responses. DNPP, on the other hand, seems to “misclassify” some of the cell types into the wrong group, while 1D-Mean and 2D-Mean predictions are nearly identical across cell types.

Another example (Figure 4B) with highly cell-specific expression patterns is Carbetocin, an oxytocin analog. In contrast to the previous example, here DNPP outperforms the tensor approach. One explanation for DNPP’s success with Carbetocin is that all three measured cell lines (MCF7, A549 and VCAP) are among the top five most-sampled cell lines in the tensor, and therefore have many drug neighbors from which to choose (see 3.6). On the other hand, M3 has data in many cell types, which is associated with better Tensor predictions (again, see 3.6). In addition to M3 and Carbetocin, two more cell-specific examples are shown in Supplementary Figure S8, one (ABT-751) in which both methods do similarly well, and a second (GNF-2), where both have similarly poor performance.

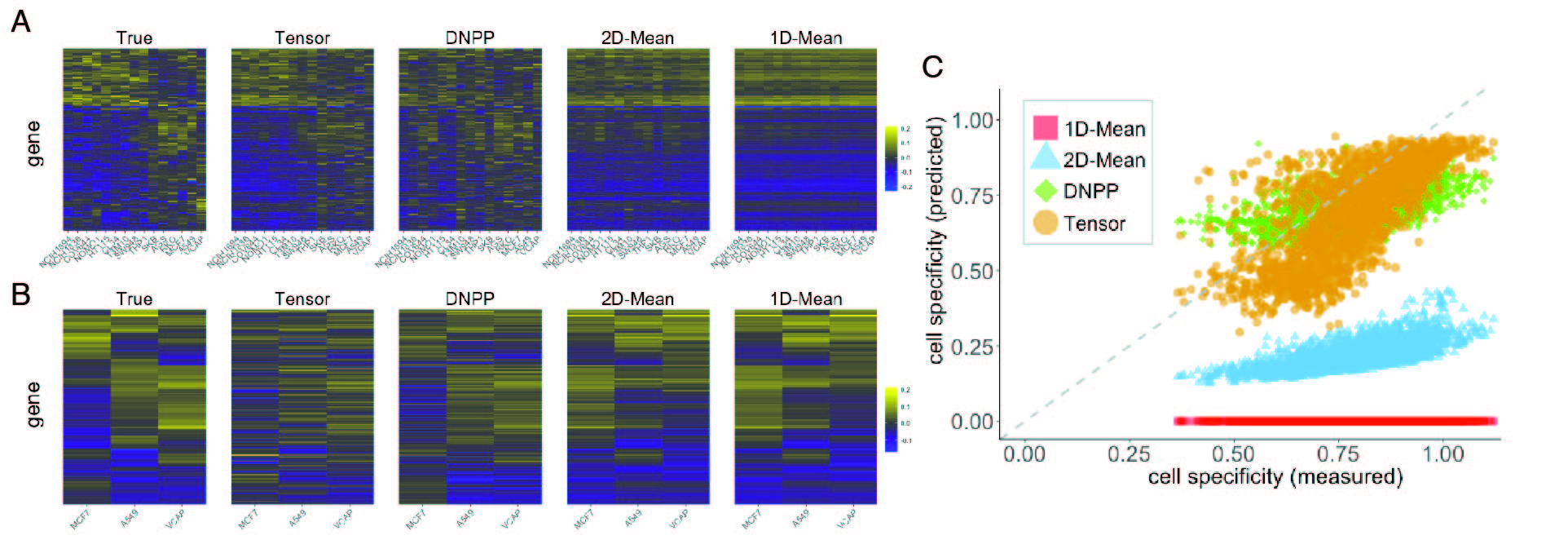


Fig. 4: Cell-specificity of predictions. A. All available expression profiles (measured and predicted) for the compound M-3M3FBS (see text for details). Rows correspond to genes, and columns to cells. B. Analogous plots for Carbetocin, in the three available cell lines, MCF7, A549, and VCAP. See text for details. C. A systematic evaluation of cell-specificity of predictions, comparing the cell-specificity per drug in predicted profiles (i.e. among cell lines not available in the data tensor) compared with cell-specificity of the same drug in measured profiles. Consistent with the examples in panels A and B, Tensor and DNPP results lie roughly near the diagonal, while 1D-Mean and 2D-Mean predictions have lower cell-specificity than measured. See Figure 1 for more on cell-specificity.

In order to further probe the cell-specificity of the results on a systematic level, we posited that if a drug induces highly specific responses among the measured subset of cell types, a similar degree of specificity would generally be expected among predictions in the remaining cell types. Hence for each drug we compared the cell-specificity (see Figure 1) on measured vs. predicted cell lines, treating the former quantity as a silver-standard truth. Results are shown in Figure 4C. Both Tensor and DNPP maintain similar cell-specificity among predicted profiles (both with mean-squared error, MSE, of 0.02), whereas 2D-Mean has a much lower degree of cell-specificity (MSE 0.33) and 1D-Mean predictions have no specificity by design (MSE 0.63). Similar results are observed when, instead of comparing to the cell lines that were not measured for a given drug, we consider either the CV predictions, or the union of both of these (see Figure S9).

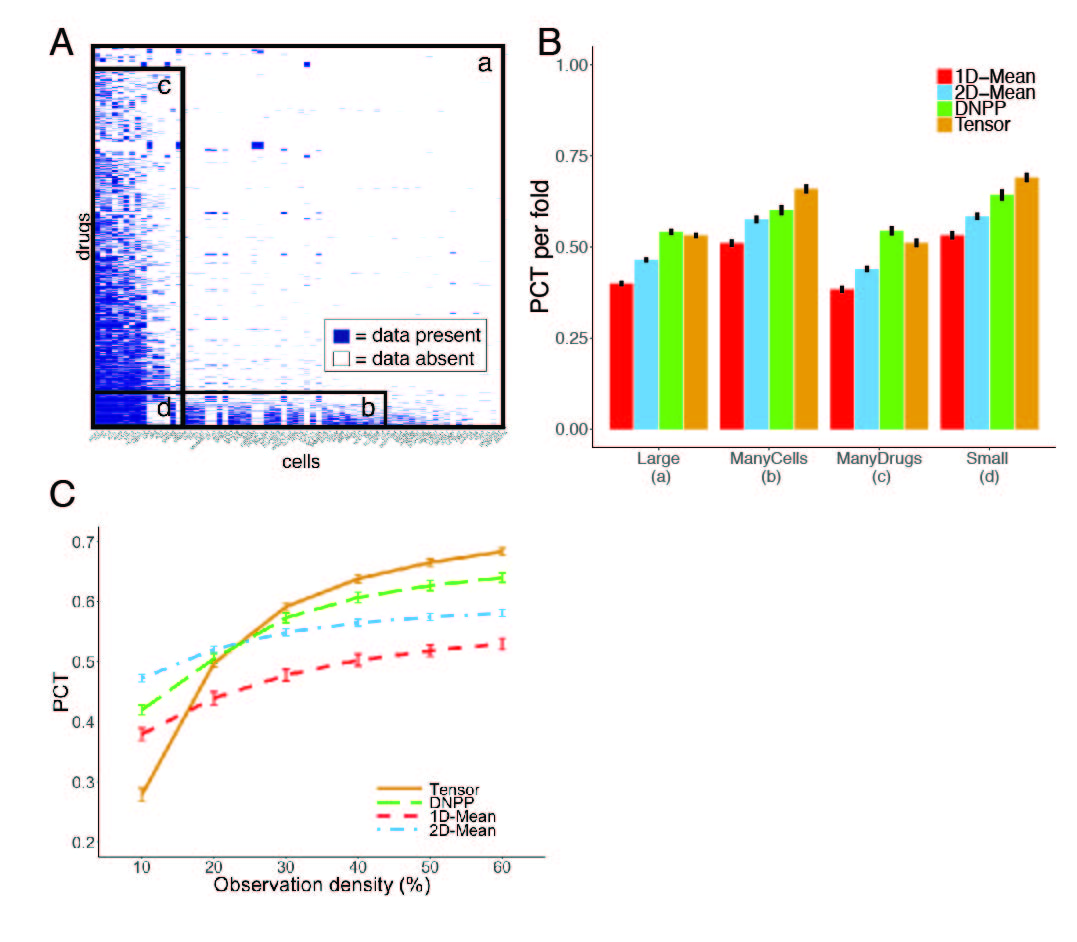


Fig. 7. Dependence of accuracy on amount of input data. A. Definition of four tensors where the full drug-cell matrix shown (a) corresponds to the ‘top’ view of the large data tensor used throughout the present work. Three subtensors are defined, which we refer to as (b) *Tmany-cells*, (c) *Tmany-drugs*, and (d) *Tsmall*. B. 10-fold CV results on these four tensors. C. Dependence of accuracy on observation density in *Tsmall*. See text for details.

# Discussion

Expression profiles characterizing *in vitro* drug perturbations are useful for a variety of applications in drug discovery. While many thousands of such expression profiles have been measured, large gaps remain in the combinatorial space across drugs and cell types. Hence, we asked whether it is possible to leverage existing data from other drug-cell combinations to predict unmeasured profiles. We tested both a local and a global approach, finding that predictions are not only accurate in an overall sense but preserve signal that is biologically and therapeutically relevant, e.g. maintaining gene correlation structure and connections to disease signatures.

Several results are consistent with our prior expectations for predictive performance. First, we observed that more extremely differentially expressed genes (e.g. the top 1%) were easier to recover than more moderately differentially expressed genes (e.g. the top 10%), perhaps due to a more reliable signal in the underlying measurements. Also, results were generally stronger across all methods with increasing observation density, suggesting that the underlying assumptions made by each method are reasonable. Finally, drug responses that were highly individual across cell lines were generally more challenging to predict, but the proposed approaches were more robust to this (i.e. less correlated) than the two baselines.

Both Tensor and DNPP almost uniformly outperformed the averaging baselines, with generally similar performance between the two methods. They particularly stood out in their ability to preserve both gene correlation structure, as well as cell-specificity. Despite their overall similarity, several distinctions may help guide users in deciding which approach to use. First, we found that Tensor outperformed DNPP when a lot of information was available per drug (not surprising due to its global usage of data), whereas DNPP had better performance when many drugs are represented (also not surprising, due to the increased chance of finding similar drug neighbors). We also found that Tensor outperformed DNPP at high observation densities. On the other hand, DNPP is conceptually simpler, uses only a single parameter, and requires less computation time.

Our framework produces testable and usable predictions at the L1000 profile level. More specifically, each value corresponds to the differential expression (CD) value of one gene in one cell line perturbed by one drug. However, the CD values do not map directly to measurable gene-level quantities such as fold change. Therefore, we advise that, unless one compares predictions to the result of a CD analysis, predictions should either be treated at the level of a ranked list of genes, or thresholded to define DEGs.

We noticed while processing the L1000 data that roughly 2/3 of the >20K drugs did not have any experiments with reliable (i.e. nominal ACD *p* < 0.1) measurements between replicates. While replicate consistency may improve with advances in data processing, it is likely that many of the drugs simply do not induce a strong enough expression response above biological or measurement noise levels. We believe that this should be considered carefully for any project working with L1000 drug profiles.

One limitation of our approach is the lack of established baselines. The baselines used in this study were relatively basic, but help to demonstrate that our predictions generally outperform alternatives that might be considered safe and intuitive. While few methods currently exist for systematic prediction of cell-type specific drug expression profiles, we expect that the methods and results presented in this study would serve as useful baselines for future work on improved methods.

Several factors may have introduced bias into the results. First, almost all of the cell lines are cancerous, which may result in more homogeneous expression compared with non-cancerous and/or patient-derived cells. Second, the selection of landmark genes may have biased the results. One line of thinking is that, due to the way these genes were selected, one would expect them to be relatively independent and therefore more difficult to predict than a random set of genes. If true, this would bias the results in a more conservative direction. Third, the presence of chemically similar drugs in the tensor could potentially make the prediction problem easier than otherwise. However, (Chen, et al., 2015) recently showed that there is a limited correspondence between chemical similarity and expression, and furthermore, the quality of the results are quite consistent for tensors defined by random drug subsets (results not shown). Fourth, our CV experiments reduced observation density by 10%, and hence results would likely be further improved by using all available data. Finally, the L1000 data has highly imbalanced sampling across the drug-cell space (see Figure 7A), and this is likely a source of positive bias for the large tensor results. Predictions made in the less-dense regions of the drug-cell space should therefore be used with caution and would likely benefit the most from methodological improvements.

Some methodological directions that we believe are worth exploring are: 1) incorporating nonlinear modeling; 2) incorporating auxiliary similarity information; 3) leveraging/modeling measurement reliability; and 4) adopting a probabilistic framework. This last direction, in addition to improving accuracy, could provide a measure of confidence around each prediction.

Our approach could readily be applied and evaluated on many other biological datasets where data span at least three categorical axes. Such datasets include CMap, with dimensions of drugs, genes, and cell types, and the Genotype-Tissue Expression (GTEx) (Melé, et al., 2015) and Braineac datasets (Glass, et al., 2013), each spanning individuals, genes, and tissues.

One exciting extension of this work would be to develop an algorithm to choose experiments among the unmeasured drug-cell combinations to assay on the L1000 platform in order to most efficiently map out the space and improve subsequent predictions. Toward this end, we envision an active learning framework similar to (Kangas, et al., 2014). Another extension of the work is to make possible ‘out-of-sample’ predictions (Wang, et al., 2016), which would be particularly useful for situations in which measurements are difficult to obtain (e.g. for human *in vivo* brain tissue expression) but where related measurements in more accessible tissues (e.g., neuronal cell types from induced pluripotent stem cells) could be obtained. This would likely require an integrative approach leveraging additional data sets and metrics (e.g., cell line genetic similarity as auxiliary data for tensor completion).

To the best of our knowledge, this work is the first attempt at prediction of expression profiles using only expression from related experimental conditions. Hence, we consider this work to be a compelling proof-of-concept, demonstrating the feasibility of this prediction task and paving the way for applications to new types of data in the future.

Funding

This work was supported by the following grants from the NIH: Illuminating the Druggable Genome (IDG) sponsored by NIH Common Fund and the NCI [U54CA189201]; NIDDK [R01DK098242] and the National Center for Advancing Translational Sciences [UL1TR000067] Clinical and Translational Science Award, to RH, BK, HCL, and JD. Additional grants from the NIH [grant numbers R01GM098316, U54HG008230 and U54CA189201] supported QD, ZW, AM, and NC.This work was also supported by an NSF Career Award [1350965] to DS. *Conflicts of Interest:* JD owns equity in NuMedii Inc., Ayasdi Inc., and LAM Therapeutics.

References

Brock, G.N.*, et al.* Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC bioinformatics* 2008;9(1):1.

Chen, B.*, et al.* Relating Chemical Structure to Cellular Response: An Integrative Analysis of Gene Expression, Bioactivity, and Structural Data Across 11,000 Compounds. *CPT: pharmacometrics & systems pharmacology* 2015;4(10):576-584.

Clark, N.R.*, et al.* The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC bioinformatics* 2014;15(1):79.

Donner, Y.*, et al.* Imputing gene expression from selectively reduced probe sets. *Nature methods* 2012;9(11):1120-1125.

Duan, Q.*, et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic acids research* 2014:gku476.

Duan, Q.*, et al.* L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *NPJ Systems Biology and Applications* 2016;2:16015.

Dudley, J.T., Deshpande, T. and Butte, A.J. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics* 2011:bbr013.

Edgar, R., Domrachev, M. and Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 2002;30(1):207-210.

Eichhorn, J.*, et al.* PLC-γ1 enzyme activity is required for insulin-induced DNA synthesis. *Endocrinology* 2002;143(2):655-664.

Gamazon, E.R.*, et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* 2015;47(9):1091-1098.

Glass, D.*, et al.* Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol* 2013;14(7):R75.

Hassane, D.C.*, et al.* Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood* 2008;111(12):5654-5662.

Hu, G. and Agarwal, P. Human disease-drug network based on genomic expression profiles. *PloS one* 2009;4(8):e6536.

Iwata, M.*, et al.* Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Scientific Reports* 2017;7:40164.

Kangas, J.D., Naik, A.W. and Murphy, R.F. Efficient discovery of responses of proteins to compounds using active learning. *BMC bioinformatics* 2014;15(1):1.

Khan, S.A.*, et al.* Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics* 2014;30(17):i497-i504.

Kolda, T.G. and Bader, B.W. Tensor decompositions and applications. *SIAM review* 2009;51(3):455-500.

Lagunin, A.*, et al.* DIGEP-Pred: web service for in silico prediction of drug-induced gene expression profiles based on structural formula. *Bioinformatics* 2013:btt322.

Lamb, J.*, et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;313(5795):1929-1935.

Liu, J.*, et al.* Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2013;35(1):208-220.

Melé, M.*, et al.* The human transcriptome across tissues and individuals. *Science* 2015;348(6235):660-665.

NIH, B.I. NIH LINCS Program. In.; 2014.

Parkkinen, J.A. and Kaski, S. Probabilistic drug connectivity mapping. *BMC bioinformatics* 2014;15(1):113.

Parrales, A., Lopez, E. and Lopez-Colome, A. Thrombin activation of PI3K/PDK1/Akt signaling promotes cyclin D1 upregulation and RPE cell proliferation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 2011;1813(10):1758-1766.

Qu, X.A. and Rajpal, D.K. Applications of Connectivity Map in drug discovery and development. *Drug discovery today* 2012;17(23):1289-1298.

Tomioka, R., Hayashi, K. and Kashima, H. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789* 2010.

Troyanskaya, O.*, et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;17(6):520-525.

Wang, J.*, et al.* Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *The American Journal of Human Genetics* 2016;98(4):697-708.

Wei, G.*, et al.* Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer cell* 2006;10(4):331-342.

Xu, Y.*, et al.* Parallel matrix factorization for low-rank tensor completion. *arXiv preprint arXiv:1312.1254* 2013.