Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Improving cell-specific drug connectivity mapping with collaborative filtering

Rebecca Newman *,[1], **Christopher Michael Pietras** *,[1], **Fangfang Qu**[1], **Diana Sapashnik**[1], **Lior Kofman**[1], **Sean Butze**[1], **Faith Ocitti**[1], **Inbar Fried**[2], **Donna K. Slonim**[1,3]

[1]Department of Computer Science, Tufts University, Medford, MA, 02155
[2]Department of Medicine, University of North Carolina, Chapel Hill, NC, 27516
[3]Department of Immunology, Tufts University School of Medicine, Boston, MA, 02111

*These authors contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Drug re-positioning allows expedited discovery of new applications for existing compounds, but re-screening vast compound libraries is often prohibitively expensive. "Connectivity mapping" is a process that links drugs to diseases by identifying drugs whose impact on expression in a collection of cells most closely reverses the disease's impact on expression in disease-relevant tissues. The high throughput LINCS project has expanded the universe of compounds, cellular perturbations, and cell types for which data are available, but even with this effort, many potentially clinically useful combinations are missing. To evaluate the possibility of finding disease-relevant drug connectivity despite missing data, we compared methods using cross-validation on a complete subset of the LINCS data.
Results: Modified recommender systems with either neighborhood-based or SVD imputation methods were compared to autoencoders and two naive methods. All were evaluated for accuracy in prediction of both expression signatures and connectivity query responses. We demonstrate that cellular context is important, and that it is possible to predict cell-specific drug responses with improved accuracy over naive approaches. Neighborhood-based collaborative filtering was the most successful, improving prediction accuracy in all tested cells. We conclude that even for cells in which drug responses have not been fully characterized, it is possible to identify drugs that reverse the expression signatures observed in disease.
Contact: donna.slonim@tufts.edu
Supplementary information: bcb.cs.tufts.edu/cmap

1 Introduction

Connectivity mapping (Lamb *et al.*, 2006) refers to the process of drug repositioning by finding candidate drugs that best reverse the expression changes caused by a given disease or condition. The original Connectivity Map database used microarrays to profile gene expression changes in up to four cancer cell lines treated with 164 perturbagens, many of them FDA-approved drugs. An expression "signature" from a given disease state, essentially sets of up and downregulated genes in relevant tissues from patients with the disease compared to normal controls, could then be used to find database compounds whose effect on gene expression was negatively correlated with the expression changes caused by the disease. Some compounds identified in this way were shown in further studies to have high potential for therapeutic efficacy in disease (Lamb *et al.*, 2006; Wei *et al.*, 2006; Zhang *et al.*, 2012).

Recently, the LINCS Consortium has dramatically scaled up the connectivity map database using the L1000 assay, which measures expression of 978 genes at a much lower cost. The LINCS connectivity map includes a much larger set of compounds, small molecules, and cellular perturbations across a wider range of cell types (Subramanian *et al.*, 2017). However, while there are now many more cells profiled in the connectivity database, the data matrix is still very sparse, with most drug profiles in a small set of cancer cell lines. Yet recent work has shown that even different breast cancer cell lines can have different, context-specific responses to perturbation (Niepel *et al.*, 2017). We observe that variation in primary cells' responses is even greater.

Further, gene expression profiles in primary cells are a relatively small part of the LINCS data set, and profiling all possible cell types and states is impractical even with more efficient assays. Thus, candidate drugs identified through connectivity mapping may have very different effects *in vivo*. For a precision medicine approach to connectivity, particularly outside the realm of oncology, the ability to impute missing connectivity data will improve scalability and relevance. We therefore aim to determine whether accurate imputation of context-specific connectivity query results is possible despite missing data.

Our first step is imputing expression values for drug/cell combinations that lack experimental data. Since the early days of microarrays, there have been efforts to impute missing microarray expression values caused by array defects or hybridization issues (e.g., array scratches, localized manufacturing defects, reagent spatters). The naïve approach, averaging over expression values for a given gene in other arrays in the data set, was quickly improved upon by more principled methods, including k-nearest neighbors and SVD (Troyanskaya *et al.*, 2001), local least squares optimization (Kim *et al.*, 2005), and Bayesian prediction (Oba *et al.*, 2003). Several methods made use of time series information when available (Troyanskaya *et al.*, 2001; Saha *et al.*, 2013; Bar-Joseph *et al.*, 2003). Collaborative filtering methods have even been applied to this problem (Saha *et al.*, 2016; Wang and Tseng, 2012).

However, nearly all of these approaches address a different problem from that we handle here – one in which there is a limited fraction (e.g. under 20%) of missing genes for a given sample, and in which the missing data points are not correlated across samples. A few approaches try to fit some characteristics of the random missing data, such as the observed histogram of the fraction of missing genes per sample (Oba *et al.*, 2003). Only some of the time series methods (e,g, (Bar-Joseph *et al.*, 2003)) deal with imputing whole expression profiles.

Furthermore, in order to impute entirely missing expression profiles, one *must* incorporate additional domain information, such as that from nearby time points, or functional relationships between genes' expression patterns. In another example, transfer learning has been used to impute entire bulk RNA-sequencing profiles when methylation profiles for the same samples are available (Zhou *et al.*, 2020). In our case, we use expression profiles of related drugs and cells.

Explicitly imputing connectivity data for unassayed drugs and cells is thus a novel problem. Further, connectivity expression matrices can be very sparse, with well over half of the potential cell by drug matrix consisting of missing values. The novelty of our approach lies in using what data we do have about how specific drugs perform in other cells, and how various cells respond differently to other drugs, to infer missing data and then to assess efficacy specifically with respect to drug connectivity inference. The closest prior work we have seen is that of (Gottlieb *et al.*, 2017), which uses (a limited number of) expression values inferred from eQTL and patient cohort data to predict appropriate doses for warfarin. Such methods have great potential, but cannot be used without data on sequence variation.

2 Methods

2.1 Overview of approach

To assess our ability to determine connectivity with missing data, we need a data set where we know the right answers. We create such a data set by taking a complete subset of the LINCS data and performing five-fold cross-validation. Each drug/cell combination is removed from the data set in one fold and predicted from the remaining data. We then assess prediction accuracy of both the expression vectors and of connectivity queries performed with genes characterizing each drug's impact on cells. Specifically, we start with a drug-cell combination of interest, which we denote as drug d_i and cell c_j, and for which the expression profile is unavailable. Our key question is how well we can impute expression and thus connectivity for d_i and c_j, and whether we can do so more accurately by taking cell type into account.

2.2 Connectivity mapping data and queries

Let C be a set of c cell types, and let D be a collection of d drugs, compounds, or cellular perturbations. Perturbations in LINCS include gene overexpression or knockdowns, but for the purposes of this study we consider primarily treatment of cells with named compounds or chemicals.

We start with a matrix M of gene expression profiles for a common set of n genes in some subset of $D \times C$. The matrix of cells and drugs is sparse, but the expression profiles are all-or-nothing; if we have expression data for some cell/drug combination, we have expression values for all n genes in the treated cell compared to the untreated cell.

As in the paper describing the L1000 connectivity data set (Subramanian *et al.*, 2017), these expression changes are represented by z-scores. Specifically, we use the published "Level 5" z-score data, which include z-scores of expression changes in drug-treated cells relative to controls, averaged over at least three replicates. When a cell / drug combination appears multiple times in the LINCs database, usually because that combination has been tested at different dosage levels or had expression profiles taken at different times after application, the expression profiles were combined into a single consensus profile, using Stouffer's method for combining Z-scores (Stouffer, 1949).

A *query signature* of a particular biological or disease state is defined to consist of two sets, one containing the k most up-regulated, and the other the k most down-regulated, genes in that state compared to a suitable control. So for example, a query signature for prostate cancer might consist of the 50 most up- and down-regulated genes in tumors from patients compared to normal prostate tissue.

A connectivity map query is performed using the query signature in the following way, as described in more detail in (Subramanian *et al.*, 2017). Given the query signature q_u containing the k most upregulated and q_d containing the k most downregulated genes for a given cell type and drug pair c, d and a reference profile r, we compute the weighted connectivity score (WTCS) as:

$$\text{WTCS} = \begin{cases} \frac{\text{ES}(q_u,r) - \text{ES}(q_d,r)}{2} & \text{if } \text{sgn}(\text{ES}(q_u,r)) \neq \text{sgn}(\text{ES}(q_d,r)) \\ 0 & \text{otherwise} \end{cases}$$

where $\text{ES}(q,r)$ is the weighted Kolmogorov-Smirnov enrichment statistic (ES) described in (Subramanian *et al.*, 2005),

and captures the enrichment of the set of genes q in profile r. WTCS ranges from -1 to 1. A score of 1 represents high positive connectivity, meaning that the drug's effect on the given cell appears to be similar to that in the query signature, while a score of -1 represents high negative connectivity, or a drug/cell combination that up-regulates the down-regulated genes from the query signature and down-regulates its up-regulated genes.

Our data are derived from the LINCS data published in the Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002) as GSE70138 and GSE92742. Together these contain 591,699 expression profiles for 98 cell types and 29,668 perturbagens spread over 189,173 unique cell/perturbagen combinations. The data were downloaded on Feb. 28, 2018.

For assessment purposes, we identified a complete data sub-matrix containing 12 cell types: 8 cancer cell lines, A375, A549, HCC515, HEPG2, HT29, MCF7, PC3, & VCAP; HA1E, an immortalized normal kidney cell line; and three primary cell types: ASC (adipose cells), NPC (neural progenitor cells partially differentiated from iPSCs) and NEU (fully differentiated neurons).

To create a complete and interpretable data matrix, We also selected a subset of 450 drugs that have "real" names (i.e., they are not just numbered compounds in development) and for which there is expression data for all 12 of the cell types above. We then created two versions of that 12x450 data set: one using the 978 "landmark" genes from the L1000 assay, and a second with all 12,328 genes in the full LINCS data set, most of which are inferred from the expression levels of the 978 landmark genes. All experiments described in this manuscript use just the landmark gene set, except when we specifically discuss assessing performance on the larger gene set.

2.3 Data imputation methods

2.3.1 Baseline methods

In the original connectivity map paper (Lamb *et al.*, 2006), connectivity scores were computed without consideration of cell type, essentially averaging across all cells. Given that the vast majority of these profiles were in a single cell line (MCF7), ignoring cellular context made sense. Even the current connectivity tool, using the much larger and more varied LINCS data set, reports averaged "summary" profiles (Subramanian *et al.*, 2017). This informs the idea behind our baseline imputation methods.

Tissue-agnostic: A good baseline prediction of drug d_i's performance in cell c_j might be simply to look at what drug d_i does to a cell, regardless of what type of cell it is. Assume that we have expression profiles for drug d_i on other cell types. By taking the median gene expression profile over of that drug all the other cells for which we do have data (the highlighted row in Figure 1a), we arrive at a prediction of what drug d_i "usually does" to a cell. We call this the *tissue agnostic* imputation method, and compare other results to this.

Two-way average: We might additionally want to include tissue-specific information in a very straightforward way. We can accomplish this by averaging the tissue-agnostic prediction for drug d_i across all other cells with the analogous "drug-agnostic" average for c_j (the highlighted column in Figure 1a) that tells us how expression of c_j after perturbation typically differs from expression in other cells. We refer to this as *two-way average* prediction.

2.3.2 Collaborative filtering

Collaborative filtering is an approach used in recommender systems to impute missing rating values and thereby recommend new products to users based both on information from similar users and other items that user has rated. Calculations are typically based on sparse databases with m users and n items containing those users' ratings for $\leq n$ of those items (Su and Khoshgoftaar, 2009). These ratings can represent any kind of relationship between users and items. In applications such as movie or purchase recommendations, the ratings might be represented by integers in the range [1,5]. But in other applications, ratings might be real numbers or categorical variables. This approach is of particular interest for imputation of connectivity data because of the sparsity of the database.

Neighborhood approach: One approach to collaborative filtering is to rely on the closest neighbors of a particular sample as a model. An average, weighted by similarity of the neighbors' ratings, approximates the rating for the sample of interest. A critical change was necessary to use this approach with our data, because each "rating" in the connectivity matrix is actually a vector of gene expression values. To avoid loss of information, we view the gene expression values as a multi-part rating of the same item. In the user/movie-rating metaphor, these values would represent a rating that perhaps specified an overall rating, as well as ratings of the movie's acting, cinematography, and soundtrack. If two users rate a movie similarly but one rates the cinematography poorly and other rates it highly, these users might actually be more different than they first appear.

From the data matrix M defined in section 2.2, we then compute a "ratings" matrix R by mean centering the row of a drug's "rating vectors" for all cells. Specifically, let μ be the mean of all non-missing values in the row, and then replace each existing value v by $(v - \mu)$ and each missing value by $(0 - \mu)$.

We then calculate similarities by taking the cosine similarity between all pairs of rows in R, yielding a symmetric matrix.

For each missing expression profile (vector), we then predict that profile by finding the top x "neighbors" of the row containing the missing value. We compute the predicted profile P_{ij} for drug c_i and cell c_j, as

$$P_{ij} = r_{ij} + \frac{\sum_{l=1}^{x} \cos(r_i, r_l) * r_{lj}}{\sum_{l=1}^{x} \cos(r_i, r_l)}$$

where cos refers to the cosine distance between two ratings vectors, r_{ij} refers to the ratings vector in row i and column j, r_i is the average value of all values present in row i, and r_{lj} is the value corresponding to the current drug-cell-gene combination in the neighbor row l.

Matrix decomposition: The goal of this approach is to account for latent subclasses within the drugs or cells. We used the FunkSVD package, which applies stochastic gradient descent SVD optimization to build an approximation of an original input matrix M. This specific methodology was shown to be particularly effective in predicting Netflix movie ratings (Bennett and Lanning, 2007). The FunkSVD method does not require a complete matrix to run, and effectively "overlooks" missing or unknown ratings (Funk, 2006). Note that this is not the case for a simple SVD decomposition, which would require some initial "guess" for the missing entries (Sarwar *et al.*, 2000). Therefore, we worked directly with the raw values in M. We still consider M to be a multi-part ratings matrix, although it is not mean-centered like R.

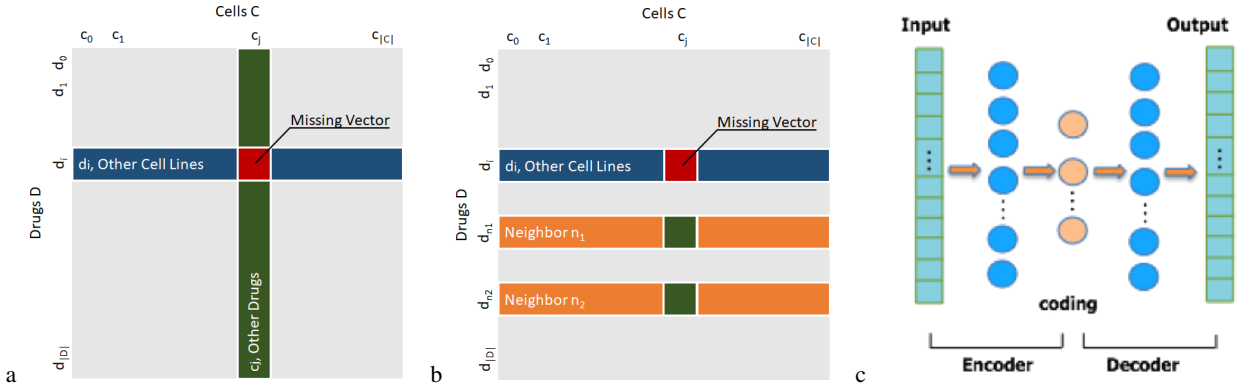

Fig. 1. Fig. 1. Methods overview. a) In tissue-agnostic imputation, the median of the expression profiles (vectors) for drug d_i on cells other than c_j (in blue) is used to predict the unknown expression profile for drug d_i on cell c_j. The two-way average is the average of two sets of expression profiles: the median of the expression profiles for drug d_i on cells other than c_j (in blue) and the median of the expression profiles for drugs other than d_i on cell c_j (in green). b) In the neighborhood approach to collaborative filtering imputation, neighbors are drugs whose expression profiles on cells other than c_j (in orange) are most similar to d_i's profile on cells other than c_j (in blue). In this example with $x = 2$ neighbors, the expression profile for cell c_j on neighbor drugs d_{n1} and d_{n2}, shown in green, are used to impute the profile for drug d_i on cell c_j. c) In the autoencoder approach, a lower-dimensional coding representation of cell and drug specific expression profiles are learned through an encoder and then decoded back to reconstruct the original data or to predict the expression profiles of missing cell and drug combinations.

FunkSVD decomposes the matrix into component matrices U and V with singular values folded into those matrices. The parameters of this function determine the output rank of the approximation. (We tuned the rank parameter k on a smaller and older data set consisting of 200 drugs and 6 cells; we did not find the results to be highly sensitive to changes in k.) A lower-rank approximation can then be obtained by reconstructing the matrix with $M = UV'$. We then predicted the values of missing data per row with the approximation $r_{\text{new}} = uV'$ where u is a row in M containing an unknown multi-part rating.

2.3.3 Autoencoder

An autoencoder uses neural networks to learn a lower-dimensional representation of the data and handles sparse or missing data well. It has been successfully applied in image processing and speech tasks as a latent factor model. The method maps (encodes) an input to a hidden representation (code) via an encoder (Baldi, 2012; Bourlard and Kamp, 1988). The coded representation is then decoded to the target output with the same dimension as the input via a decoder. Both the encoder and decoder are feed-forward artificial neural networks. Non-linear hidden units and additional hidden layers enable autoencoders to learn more complex hidden structures in data (Chicco *et al.*, 2014).

Here, we first train an autoencoder to learn the hidden structure of expression profiles of drug-cell combinations, and then we impute the missing profile given the cells' and drugs' two-way average profile.

During the training phase, the parameters of the encoder and decoder are learned by minimizing the reconstruction error, specifically, the Mean Square Error (MSE) with an L2 regularization term to avoid overfitting. During the imputation stage, the expression profiles of specific missing cell and drug combinations are first computed using the two-way average described in section 2.3.1, then fed into the trained autoencoder to reconstruct the imputed profiles.

In this work, we did not experiment with changing the original network architecture, instead choosing a default architecture having 3 hidden layers with 100 units coding representation, corresponding to a reducing factor of 9.78, as shown in Figure 1c. We use the Rectified Linear Unit (reLu) as a non-linear activation function (Nair and E. Hinton, 2010). Adadelta (Zeiler, 2012) was used as our optimization method, and the running average parameter ρ was set to 0.95 as suggested in (Zeiler, 2012). The regularization parameter λ was set to 0.01. Using a different data set, we also examined whether dropout would make any improvements and found that applying dropout on hidden layers did not improve the performance.

2.4 Evaluation

2.4.1 Cross-validation

To assess our performance, we need data on which we know the right answers. We therefore started with the complete data matrix described in section 2.2. From this, we created a cross-validation data set in the following way.

Each cell / drug combination is randomly and independently assigned to one of five folds. We then verify that the candidate set of fold assignments has no fold where more than 75 percent of the cells for a given drug, or 75 percent of the drugs for a given cell, are assigned to that fold, ensuring that any method would be able to produce an imputed expression profile for any missing cell/drug combination. If this requirement was violated, fold assignments were completely regenerated until the requirement was met.

For each fold, a given method is provided *only* the z-score normalized gene expression profiles for cell / drug combinations not in the fold, and must impute the expression profiles for cell / drug combinations that are in the fold. Over all five folds, a given method will produce a single imputed profile for each cell / drug combination. We then compared the imputed profile to the true profile for that cell / drug combination, using the various scoring metrics described below.

To address any variance due to the randomness in this cross-validation procedure, we created five independent instances ("runs") of cross validation data sets. We summarize each of the scoring metrics across those five runs.

2.4.2 Expression prediction

Once we have a predicted expression profile for a drug-cell combination, we must assess how accurate our prediction is

through some comparison to the known true expression profile. The simplest way to do this is through direct comparison of the predicted and true z-score normalized expression profiles, which we do using Spearman rank correlation. Note that such correlations are expected to be relatively low, because they consider expression changes in *all* the genes in the data set. If many genes are basically not changing with drug treatment, these may appear in essentially random order in the middle of the predicted list, but their correlation with the true near-zero expression changes influences the correlation score.

2.4.3 Connectivity prediction

Since our aim is to infer connectivity for cell-drug pairs for which we lack data, a better evaluation method would compare the connectivity results from the withheld true data with those obtained using the imputed data.

There are two parts of a list of drugs returned by a connectivity query that are of interest. Drugs with the most positive connectivity scores are drugs that replicate the query signature; these are sometimes used to identify similar drugs, or compounds that might mimic the query expression profile change as an adverse event. Drugs with the most negative connectivity scores are those that reverse the observed query signature; these are candidate therapeutics for an observed disease signature. Accordingly, we want to assess primarily whether the most-positive, or most-negative, connectivity results from the imputed data replicate those from the withheld true data.

To do this we use the Weighted Spearman Rank Correlation measure, defined by (Shieh *et al.*, 2000). We use a weight function defined as

$$w(r) = 2\phi(r|\mu = 0, \sigma = N\epsilon),$$

where r is the result rank, ϕ is the normal distribution of the form $\phi(x|\mu, \sigma)$, and ϵ is a parameter with value between 0 and 1 controlling the "aggressiveness" of the curve.

In this work, we chose $\epsilon = 0.01$, which applies weights of significant magnitude to approximately the top 20 results, after which weights rapidly approach zero. We tested the effect of doubling or halving ϵ, but saw minimal changes in the results.

2.5 ATC Code Matching

The Anatomical Therapeutic Chemical (ATC) drug classification system, developed by the World Health Organization, hierarchically classifies drugs based on therapeutic or pharmacological properties, or by the organ system in which the drug acts. The classifications are represented by a code, such as C03CA01. Each letter or two-digit number in an ATC code indicates a classification at one of five levels, with the first level being the most general and the fifth level uniquely identifying a drug.

We therefore might expect that a connectivity query made with a query signature derived from treatment with drug d_i would show high positive connectivity scores for drugs with ATC codes that match the ATC code of d_i at some level. Here, an ATC code match at level X means that the first X letters or two-digit numbers of the ATC codes are identical.

In our data set, 201 of the 450 drugs have ATC codes. We can use the imputed expression profile for one of these drugs, d, to create a query signature with the top and bottom 50 genes (the number recommended by (Subramanian *et al.*, 2017)). We can then use that query signature to query the connectivity database and identify the ranks of any other drugs that have a level X match with the ATC code of the query drug d. Lower ranks indicate that these similar drugs appear closer to the top of the positive query results. For this profile, we compute the average rank at which these matching drugs appear. We compare the distributions of these average ranks over all imputed profiles to the distributions of the ranks obtained if the ranking of drug results from a connectivity query were randomly assigned instead of determined by the WTCS values.

3 Results

3.1 Using cell-specific data improves imputation

Figure 2a shows the Spearman correlations between the predicted and true expression profiles for each of the 12 cell lines and each of the five imputation methods. Figure 2b shows the percent change for each of the methods compared to the tissue-agnostic baseline. In all cells, we found that it was possible to improve imputation over the tissue-agnostic method by at least 20%.

The neighborhood collaborative filtering approach improved performance the most of all the methods we tried, in all cells. Further, in all cells both SVD and the two-way average also outperform the tissue-agnostic approach. Our autoencoder method also improves upon the performance of the tissue-agnostic method in all but the HEPG2 cell line.

The images in Figure 2 are arranged to show the cancer cell lines and the immortalized normal kidney line HA1E in the first two rows, while the primary cell types NEU, NPC, and ASC are all on the bottom, along with VCAP, a somewhat atypical-looking cancer cell line derived from metatstatic prostate cancer.

Performance in the immortalized lines tends to differ from that in primary cells. The primary cells have lower correlations overall, but a greater improvement using Neighborhood imputation. The greatest improvement is in NEU cells (neurons differentiated from iPSCs), where the correlation of the Neighborhood-imputed predictions to the true withheld results was 127% higher than that of the tissue-agnostic baseline method. This is, admittedly, a change from a correlation of only 0.09 to 0.22, but incorporating tissue-specific information via collaborative filtering can bring the observed correlation for this unique cell type into the range typical of the tissue-agnostic method on most cancer cell lines. These results clearly demonstrate that cell type is an important aspect of connectivity mapping, and that using cell-specific information improves outcomes most dramatically in cells that are neither malignant nor immortalized.

Figures 3 and 4 show the average weighted Spearman correlations (over all drugs) between true and predicted connectivity results (a), and the percent changes for each over the tissue agnostic method (b). These are based on connectivity queries with query signatures containing the most dysregulated genes for each drug in the treated cell line compared to control. Specifically, the query signatures consist of the 50 most up- and down-regulated genes for a given drug. Positive connectivity scores identify the drugs most similar to the query drug. Negative connectivity scores are intended to identify drugs whose effects on a cell are roughly the opposite of the query drugs.

The positive imputation results show little improvement over baseline for all but the primary cell types. Indeed, the autoencoder and SVD methods are actually worse than tissue-agnostic imputation here. But the neighborhood collaborative filtering and two-way average approaches, both of which use cell specific information, are at least roughly as good as the

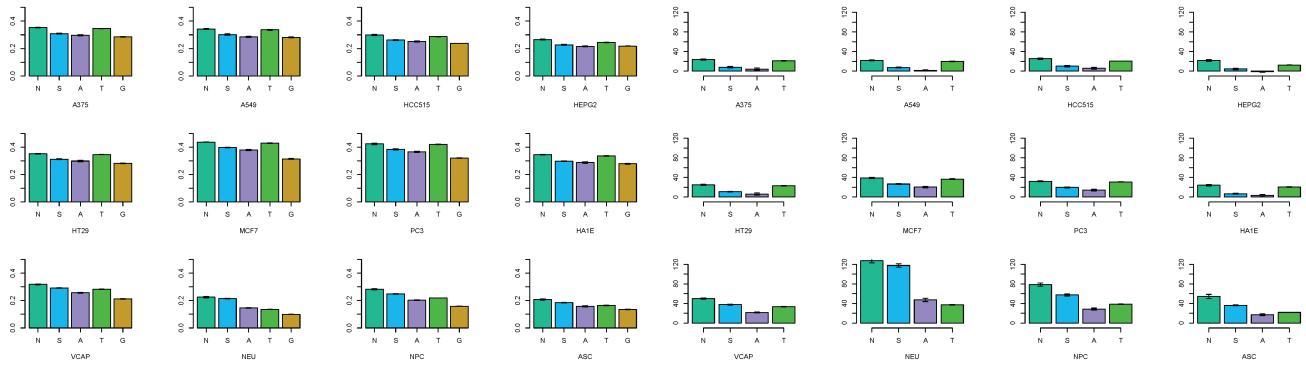

Fig. 2. a) Spearman correlation across all genes between the true and imputed expression z-scores, for each of the 12 cell lines in the data set. Methods are denoted by single-letter labels: N: neighborhood collaborative filtering; S: collaborative filtering using SVD; A: autoencoder; T: two-way average; G: tissue-aGnostic (baseline method). b) Percent change in Spearman correlation compared to that of the tissue-agnostic method; labels are the same as in part a. Method colors match those in Figure 5.

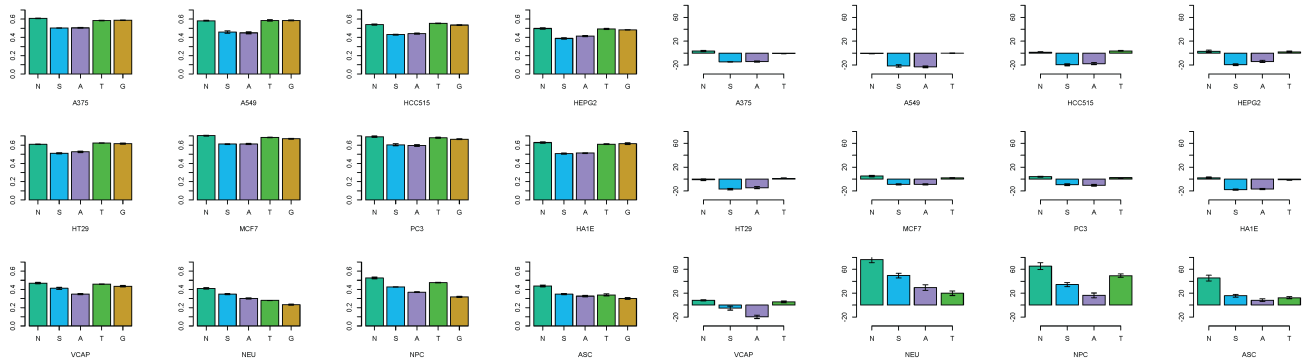

Fig. 3. a) Positive weighted connectivity correlation across all genes, for each of the 12 cell lines. Error bars show variation across cross validation runs. Labels are the same as in Figure 2. b) Percent change in positive weighted connectivity correlation compared to the tissue-agnostic method.

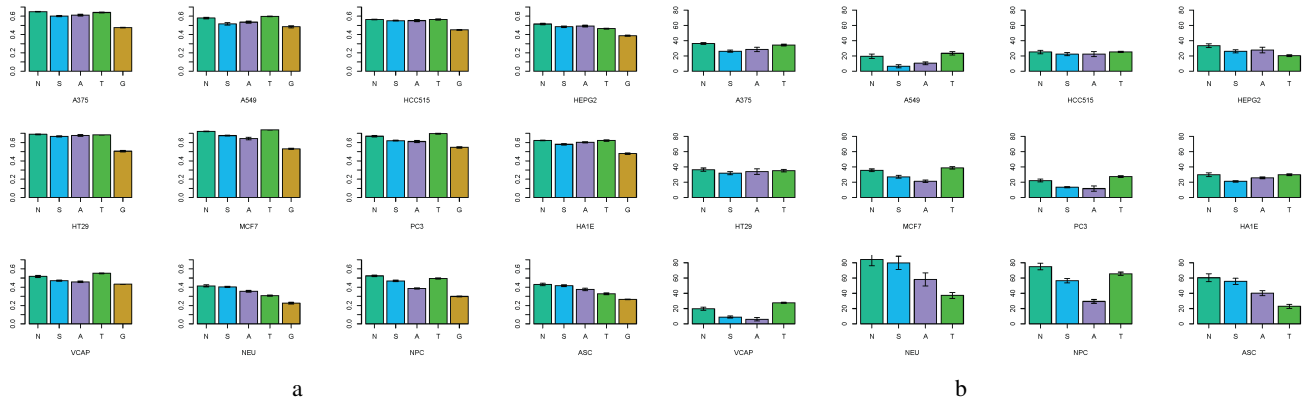

Fig. 4. a) Negative weighted connectivity correlation across all genes, for each of the 12 cell lines. Error bars show variation across cross validation runs. Labels are the same as in Figure 2. b) Percent change in negative weighted connectivity correlation compared to the tissue-agnostic method.

tissue-agnostic approach for almost all cancer cells. For the primary cells, the neighborhood approach still shows a substantial improvement, with two-way average in second place. The tissue-agnostic weighted correlation for NEU cells, the lowest, is 0.23, with the other primary cells in the low 0.30s, and most cancer lines showing a weighted correlation between 0.5 and 0.63. Thus, improving on these much higher scores may be harder.

For negative connectivity, which is the most commonly envisioned use case of the connectivity map, improvements are more robust under almost all methods, with neighborhood collaborative filtering again leading the pack. Improvements in primary cell types are again especially large, but even in typical cancer lines, an improvement of 20-35% over baseline is possible by using cell-specific data.

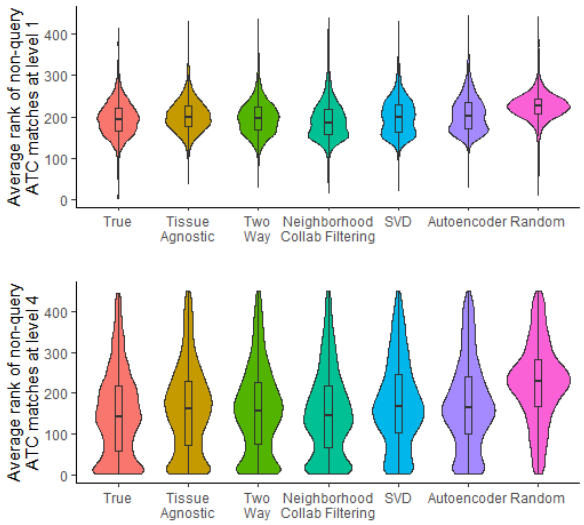

Fig. 5. Violin plot showing distribution of average ranks at which a query signature derived from the imputed expression profile (or the true expression profile) finds drugs different from the query drug but that match the ATC code of the query drug at levels 1 and 4. Presented for comparison are the average ranks if the query results were random. Plots for all ATC levels are available as supplemental figures.

3.2 Finding drugs with similar properties

Figure 5 shows the distribution of average ranks at which a query for the profile imputed by each method has matches at ATC code levels 1 and 4 (the plots for all levels appear as supplemental data). A rank of 1 means the drug is the most positively connected one to the query, suggesting that its expression patterns are highly reflective of the expression profile of the query drug.

For example, suppose the query signature is derived from the expression changes observed on treating cells with cerivastatin, which has ATC code C10AA06. The expectation is that other drugs that are level 4 ATC code matches with cerivastatin, such as C10AA04 (fuvastatin), would have disproportionately low-numbered ranks, putting them near the top of the list. However, level 3 matches, such as C10AD01 (niceritrol, a lipid lowering agent, but not a statin) would on average have slightly higher numbered ranks. Presented for comparison are the rank distributions for a query of the true expression profile - which reflects the best-case distribution - and a query where the results of the query are randomized - which reflects the worst-case distribution.

As we would expect, the average rank of ATC code matches for the true profile decreases as ATC level - and, therefore, similarity of matching drugs - increases. All imputation methods follow this pattern. We can also use these plots as another way of comparing the various imputation methods. Neighborhood collaborative filtering performs the best, with the lowest median and distribution reflecting a larger number of low-ranked (more similar) drugs from the same ATC code.

3.3 Drugs for which imputation is most helpful

One important question is whether there are specific types or properties of drugs for which imputation of missing data is most or least effective.

Table 1 shows the ten drugs whose improvement in overall expression correlation is largest, and smallest, for the neighborhood collaborative filtering method. (Other methods show somewhat overlapping lists.) A brief description of each drug's overall mechanism of action or indication is given, along with ATC codes when available.

What stands out is that the compounds that seem most improved by imputation have specific tissues in which they are likely to be active: they act on the central nervous system (CNS), respiratory system, muscles, or metabolism. Those that are least improved are those that are general cellular disruptors. They inhibit protein synthesis, essential signaling or growth mechanims, or are toxins of unknown mechanism. While few of these particular compounds have ATC codes, many of them have anti-cancer properties, so they are likely to function by killing cells, regardless of cell type. Thus, imputation using the tissue-agnostic approach is likely to be sufficient for these compounds, whereas for highly tissue-specific drugs, using cellular context to improve our prediction of connectivity is more important.

The table's one apparent exception to this rule is ergocornine, which seems to act as a dopamine receptor agonist yet is one of the compounds that least benefits from cell-specific information. However, closer inspection reveals that ergocornine has many properties of unknown mechanism, and that it acts in other tissues beyond the brain. In fact, this compound *does* show improvement with cell-specific imputation in neurons and neural progenitor cells. However, the tissue agnostic method does so much better at predicting this compound's expression in most of the cancer cells that *on average*, it is one of the compounds showing the least improvement.

3.4 Comparing landmark and full gene sets

We repeated our main experiment using not just the 978 landmark genes, but the full imputed data set with 12,328 genes. Note that the remainder of these genes are estimated from the expression levels of the 978 landmark genes that are directly measured. There is evidence not only that imputation of gene expression levels is successful for these genes, but that connectivity mapping has greater power using these than not (Subramanian *et al.*, 2017).

At first glance, our results seem to contradict this claim. That is, our Spearman correlation coefficients across the entire gene set are universally lower for the larger gene set - something that is not too surprising given that the correlation is being measured across a much larger number of genes. Not only are the correlations lower, but the percent improvement in the correlation metric using neighborhood collaborative filtering or any of the other methods appears lower with the larger number of genes (see supplemental figures). That is, which method comes out ahead changes with the number of genes in some cases.

However, when we looked at connectivity performance, the 12,328-gene results were fairly similar to those with only the landmark genes. Again, neighborhood collaborative filtering and two-way average vie for supremacy, with collaborative filtering winning by large margins on primary cell types (and improving over baseline methods by larger margins than for the landmark gene data for most primary cell / method combinations).

More exploration of results from the 12,328 gene set is in order, but our initial assessment suggests that using it for connectivity mapping will not be harmful and may be beneficial. Further, it seems that connectivity is established mostly by the most extreme

Most Improved	Description	ATC code
hydrastinine	alkaloid, haemostatic (CNS)	
beclomethasone-dipropionate	steroid	
flumetasone	steroid	D07AB03
budesonide	steroid	D07AC09
testosterone	steroid	G03BA03
isocarboxazid	monoamine oxidase inhibitor (CNS)	N06AF01
denbufylline	phosphodiesterase inhibitor (respiratory)	R03DA10
tubocurarine	alkaloid, anesthetic (muscle)	M03AA02
buspirone	anxiolytic (CNS)	N05BE01
tolazamide	sulfonylurea (metabolic)	A10BB05
Least Improved	**Description**	**ATC code**
brefeldin-a	protein transport inhibitor	
cycloheximide	protein synthesis inhibitor	
farnesylthioacetic-acid	calcium influx inhibitor	
palbociclib	CDK inhibitor	L01XE33
linifanib	RTK, VEGF, PDGF inhibitor	
ergocornine	alkaloid, dopamine receptor agonist (CNS, muscle)	
ochratoxin-a	toxin	
ebelactone-b	urinary kinase inhibitor	
kenpaullone	CDK inhibitor	
verrucarin-a	protein synthesis inhibitor	

Table 1. Drugs whose expression correlation was most or least improved by the neighborhood imputation method.

expression changes, so overall expression correlation through the entire gene list may not be as predictive of connectivity performance as seems likely at first.

4 Discussion

We have demonstrated that context-specific connectivity data can be used to infer missing data in a connectivity data matrix. This has applications for the full LINCS data set, where more than 80 cell types in the two GEO data sets we used have experimental data characterizing at least 10 drugs), and even beyond. An important question for future work is how well imputation works as the training data contains smaller numbers of drugs and/or cells. In other words, what sort of data do we need to generate in a new cellular context in order to predict connectivity effectively? How widely do we need to profile a new drug?

One immediately apparent observation about our results is the distinction between the cancer cell lines that make up the majority of the connectivity data, and the primary cells. The cancer lines have tissue-agnostic expression correlations from .21 to .32, and improve by 21 to 50 percent (by this metric) with the neighborhood collaborative filtering method. However, the primary cell types ASC, NEU, and NPC look fundamentally different. Tissue-agnostic prediction is much worse for these primary cells, ranging from .09 to .16, but the percent improvement is correspondingly greater, ranging from 54 to 127 percent. These improvements bring the overall expression correlation into the range typically seen using tissue-agnostic methods for the cancer cell lines.

Interestingly, the immortalized normal kidney cell line HA1E looks more like the cancer cell lines, suggesting that transformed normal cells look more like each other than like primary cells. Also unusual is that the prostate cancer cell line VCAP looks more like a cross between a cancer cell line and a primary cell, suggesting that it is less like the other cancer lines. It is known to come from a prostate tumor that metastasized to bone, so perhaps it is atypical both because of its hormonal signature and metastatic potential.

Regarding method comparisons for imputation, the neighborhood collaborative filtering approach has been most successful in this study. That said, there is the possibility of improving SVD by using higher-order methods that better capture the three-dimensional structure of the data. Such approaches require considerable parameter tuning on a distinct data set. Similarly, we have not yet worked on tuning or trying other architectures for autoencoders, so there is more that can be done here as well.

The question about whether or not to use the full set of genes when only 978 genes are measured directly in the L1000 assay is an important question. Prior work suggests that connectivity performance improves by using these genes (Subramanian *et al.*, 2017). However, what we found is that connectivity performance remains more or less unchanged, while prediction of actual expression z-scores is considerably worse. Future work should address how best to reliably use the additional gene information.

Finally, we have demonstrated that cellular context is critical for accurate connectivity mapping. This was apparent in prior work using multiple different breast cancer cell lines (Niepel *et al.*, 2017), but becomes even more important across more varied contexts. Both for precision cancer medicine purposes and for use in applications beyond oncology, taking context into account is therefore essential.

Acknowledgements

We thank Ted Natoli and Aravind Subramanian for their help accessing and interpreting an earlier version of the data, and for feedback on earlier versions of this work. We also thank members of the Tufts BCB research group, especially Dan Meyer, and Liping Liu and Mike Hughes, for helpful suggestions.

Funding

This work was supported by NIH R01 HD076140 to Dr. Slonim and a pilot grant from NCATS award UL1TR002544. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, pages 37–50.

Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S., and Simon, I. (2003). Continuous representations of time-series gene expression data. *J. Comput. Biol.*, **10**(3), 341 – 356.

Bennett, J. and Lanning, S. (2007). The netflix prize.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, **59**, 291–4.

Chicco, D., Sadowski, P., and Baldi, P. (2014). Deep autoencoder neural networks for gene ontology annotation predictions. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 533–540, New York, NY, USA. ACM.

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

Funk, B. W. S. (2006). Netflix update: Try this at home. https://sifter.org/simon/journal/20061211.html.

Gottlieb, A., Daneshjou, R., DeGorter, M., Bourgeois, S., Svensson, P. J., Wadelius, M., Deloukas, P., Montgomery, S. B., and Altman, R. B. (2017). Cohort-specific imputation of gene expression improves prediction of warfarin dose for African Americans. *Genome Med*, **9**, 98.

Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: Local least squares imputation. *Bioinformatics*, **21**, 187–198.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J. P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**(5795), 1929–1935.

Nair, V. and E. Hinton, G. (2010). Rectified linear units improve restricted boltzmann machines vinod nair. volume 27, pages 807–814.

Niepel, M., Hafner, M., Duan, Q., Wang, Z., Paull, E. O., Chung, M., Lu, X., Stuart, J. M., Golub, T. R., Subramanian, A., Ma'ayan, A., and Sorger, P. K. (2017). Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nat Commun*, **8**(1), 1186.

Oba, S., Sato, M., Takemasa, I., Monden, M., K., M., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**(16), 2088–2096.

Saha, S., Dey, K., Dasgupta, R., Ghose, A., and Mullick, K. (2013). Missing value estimation in DNA microarrays using B-splines. *Journal of Medical and Bioengineering*, **2**(2), 88=92.

Saha, S., Ghosh, A., and Nath Dey, K. (2016). An improved fuzzy based approach to impute missing values in DNA microarray gene expression data with collaborative filtering. In *International Conference on Advances in Computing, Communications and Informatics*, ICACCI '16, pages 911–916. IEEE.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Application of dimensionality reduction in recommender system - a case study. Technical report, University of Minnesota.

Shieh, G. S., Bai, Z., and Tsai, W.-Y. (2000). Rank tests for independence – with a weighted contamination alternative. *Statistica Sinica*, **10**(2), 577–593.

Stouffer, S. (1949). A study of attitudes. *Sci. Am.*, **14**(5), 11–15.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, pages 4:2–4:2.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O. M., Piccioni, F., Johnson, S. A., Lyons, N. J., Berger, A. H., Shamji, A. F., Brooks, A. N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D. Y., Hu, R., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Greenside, P., Gray, N. S., Clemons, P. A., Silver, S., Wu, X., Zhao, W. N., Read-Button, W., Wu, X., Haggarty, S. J., Ronco, L. V., Boehm, J. S., Schreiber, S. L., Doench, J. G., Bittker, J. A., Root, D. E., Wong, B., and Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, **171**(6), 1437–1452.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**(6), 520–5.

Wang, B.-W. and Tseng, V. S. (2012). Improving missing-value estimation in microarray data with collaborative filtering based on rough-set theory. *International Journal of Innovative Computing, Information and Control*, **8.3**(B), 2157–2172.

Wei, G., Twomey, D., Lamb, J., Schlis, K., Agarwal, J., Stam, R., Opferman, J., Sallan, S., den Boer, M., Pieters, R., Golub, T., and Armstrong, S. (2006). Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell*, **10**(4), 331–42.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR*, (abs/1212.5701).

Zhang, C., Ryu, Y., Chen, T., Hall, C., Webster, D., and Kang, M. (2012). Synergistic activity of rapamycin and dexamethasone in vitro and in vivo in acute lymphoblastic leukemia via cell-cycle arrest and apoptosis. *Leuk Res*, **36**(3), 342–9.

Zhou, X., Chai, H., Zhao, H., Luo, C., and Yang, Y. (2020). Imputing missing rna-sequencing data from dna methylation by using a transfer learning-based neural network. *Gigascience*, **9**(7), giaa076.