Similarity bias from averaging signatures from the Connectivity Map

Ian Smith[1,2] and Benjamin Haibe-Kains[1,2,3,*]

[1]Department of Medical Biophysics, University of Toronto, Toronto, Canada
[2]Princess Margaret Cancer Center, University Health Network, Toronto, Canada
[3]Vector Institute, Toronto, Canada
[*]Corresponding author: bhaibeka@gmail.com

January 2022

Arising from: J. Zhu, *Nature Biotechnology* https://doi.org/10.1038/s41587-021-00946-z (2021).

Abstract

The Next Generation L1000 Connectivity Map (L1000) is a massive, high-throughput dataset measuring transcriptional changes - or signatures - in cancer cell lines due to chemical and genetic perturbation[1]. Zhu et al [2] recently presented Deep Learning-Based Efficacy Prediction System (DLEPS), a method that models averages of many signatures (CTPs or meta-signatures) of the same compound across conditions. However, our analysis shows that averaging increasing numbers of signatures from L1000 results in a substantial positive increase in signature similarities even among entirely unrelated perturbations, drastically shifting the baseline threshold for similarity. Consequently, DLEPS overstates how informative the model is about each compound, highlighting the need to account for this similarity bias in any L1000 Connectivity Map analysis.

A perturbation is a combination of a compound or genetic reagent, a dose, and a time point, while a signature measures the effect of that perturbation in a particular cell line. Each signature in the L1000 dataset is a vector of z-scores representing changes in gene expression due to perturbation in measured and inferred genes[1]. We define a meta-signature as the average of a set of signatures, usually with the same perturbation. The typical connectivity map paradigm employed by several studies [1, 3–5] is to apply a similarity metric to pairs of signatures from the dataset to identify perturbations that produce similar biological effects. Alternatively, the signatures can be modeled, as in Zhu et al and others [6, 7], to predict the perturbational signatures using additional data. In particular, DLEPS uses a deep learning model to predict meta-signatures from compound molecular structures and evaluates predictions by Pearson's correlation of the prediction with the measured meta-signature.

To illustrate the potential pitfalls of meta-signatures or averaging multiple signatures, we explore the correlation of averages of signatures in L1000. Let a k-signature be the average of k signatures. We consider the correlation of pairs of disjoint k-signatures from three classes: (1) signatures of the same compound treatment, (2) all compound signatures, and (3) control signatures of DMSO treatment (Figure 1a). We then compute Pearson's correlation (PCC) between pairs of k-signatures for the three classes, as this metric was used in Zhu et al, and sample disjoint k-signatures 100 times for each value of k to produce a statistical sample. The null correlation introduced by averaging is shown by sampling from all compounds and DMSOs - unrelated perturbations and negative controls, respectively - which have no meaningful biological association. We find that while k-signatures from the same compound tend to correlate better as more signatures are averaged (increasing k), the null cases of unrelated signatures and DMSOs also show increases in positive correlation with k ('autocorrelation', Figure 1b). While some compounds have k-signatures with excellent

agreement for even small k, indicating strong signal, other compounds correlate only marginally better than the unrelated perturbations or negative controls (Figure 1c). Importantly, the increase in null autocorrelation does not plateau, and it is substantial - with unrelated k-signatures exceeding a Pearson's correlation of 0.5 with k as low as 10. The conclusion is that meta-signatures from the L1000 data tend to look more similar to each other as k increases irrespective of phenotypic similarity of the treatments.

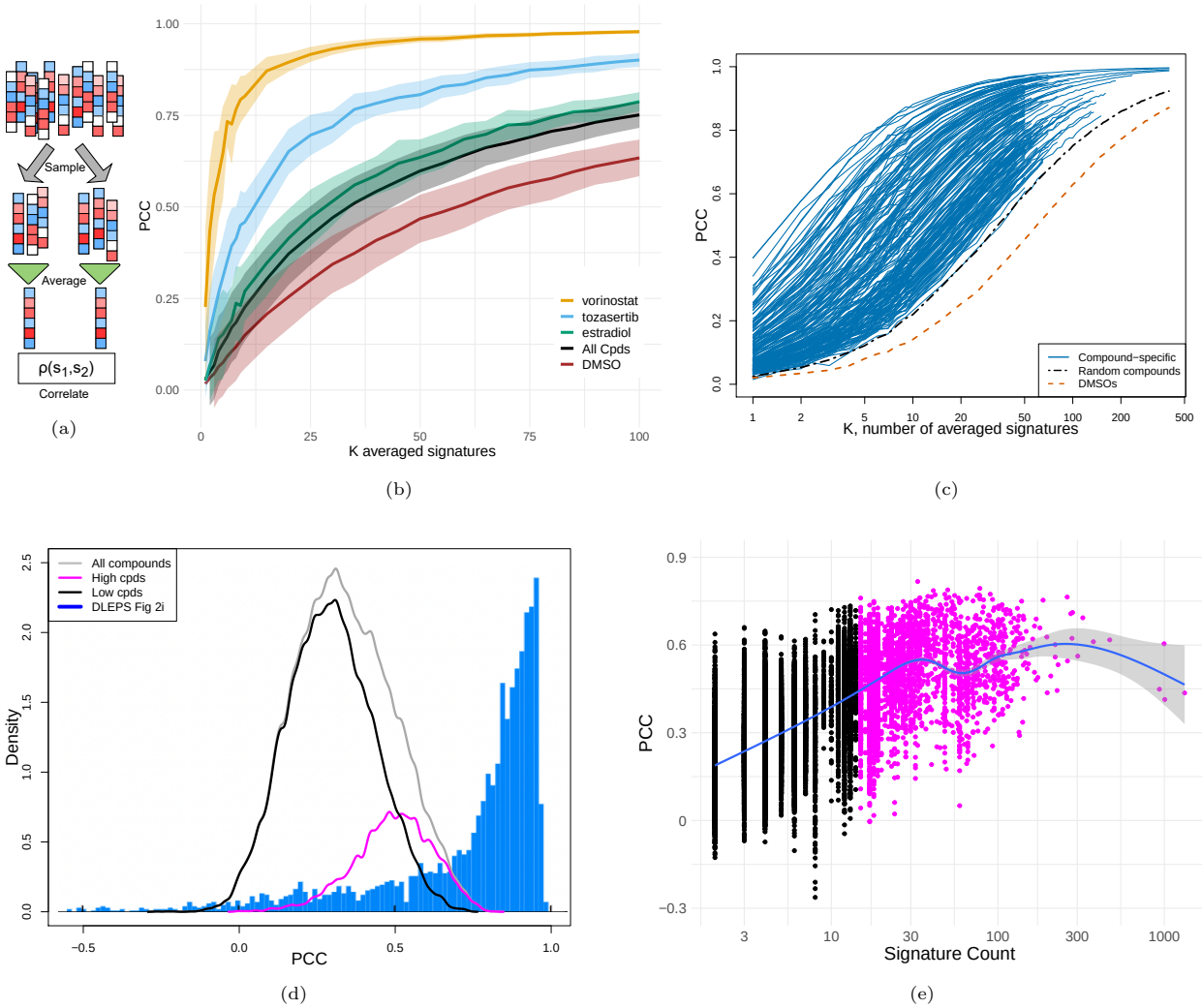

Figure 1: Correlations between meta-signatures increase as the number of components grows. (a) From a pool of signatures, two disjoint sets of k signatures are drawn, averaged gene-wise to produce two k-signatures, and correlated. (b) The correlation (y-axis) between k-signatures increases with k (x-axis) regardless of whether the signatures were of same compound (gold, blue, green), of unrelated compound signatures (black), or of negative control DMSOs (red). The line shows the mean correlation, and the ribbon shows +/- 1 standard deviation for 100 samples. The three compounds here - vorinostat, tozasertib, and estradiol - are examples of highly, modestly, and weakly correlated compounds. (c) The increase in autocorrelation for the 200 most-assayed compounds in the L1000 dataset. Each blue line is the mean correlation for that compound. The red and black dotted lines show random compounds and DMSOs. (d) The PCCs (x-axis) between compound metasignatures and that of the entire dataset (grey) approach that of DLEPS's test set (blue). Those compounds with 15+ signatures (magenta) show particularly high correlation compared to those with fewer than 15 (black). (e) A scatter plot of the data in (d), where each point is a compound, x-axis is the log10 of the number of component signatures, y-axis is PCC, and the blue line is the smoothed trend line.

This result is an important confounder in the Zhu et al. analysis. To illustrate the impact of spurious autocorrelation of meta-signatures on the DLEPS analysis, we generated a meta-signature for each compound by averaging all of that compound's signatures. We then correlated the compound's meta-signature with the average of all other compound signatures in the dataset to create a null correlation. The whole dataset meta-

signature is uninformative because it uses no information about the compound of interest, and any non-zero correlation is spurious and reflective of the autocorrelation phenomenon. We evaluated null correlations for those 17250 compounds with 5+ replicates as with DLEPS. The distribution of PCCs with this null model are much greater than 0, have an average PCC of 0.3355, achieve PCCs in excess of 0.75, and overlap with the DLEPS test predictions (Figure 1d, Zhu Figure 2i). As before, those meta-signatures with the most component signatures show the greatest increase in null correlation (Figure 1e). For example, those compounds with 15 or more signatures ("High", magenta) have an average null PCC of 0.4925, while those with fewer than 15 signatures ("low", black) have an average null PCC of 0.2919. This analysis demonstrates that due to the autocorrelation phenomenon, a model does not require any information about the compound of interest to achieve large Pearson correlations and suggests that the DLEPS approach is overstating its predictive power.

Furthermore, the impact of averaging multiple signatures together poses a challenge for conventional Connectivity Map querying as well as modeling. To assess this, we computed the similarity between a set of 500 randomly chosen L1000 compound signatures with the set of all L1000 compound signatures, producing a matrix of dimension 205034 x 500. We used each of four similarity metrics: Pearson's correlation, Spearman's correlation, cosine similarity, and weighted connectivity score (wtcs), the standard L1000 metric. Each of the 205034 level 5 compound signatures is the average of a number of level 4 "instances", and this quantity is recorded in the metadata as "nsample"[1]. Most signatures are composed of 1-6 instances; to simplify the analysis, we binned nsample into 1-2 (27.3% of the data), 3 (45.2%), and 4 or more (27.5%). Those signatures with more component instances are systematically more similar to query signatures than those with fewer, and this is most apparent in the top query hits. For each metric, we considered what proportion of signatures from each bin occurred in the top 1% of each of the 500 queries, corresponding to the top 2050 most similar signatures, and found that signatures in the top bin were between 1.19 and 2.2 times more likely to be in the top 1% (Figure 2). This bias in similarity will result in signatures with more component instances preferentially occurring in the tails of the query regardless of actual biological association in all L1000 queries.

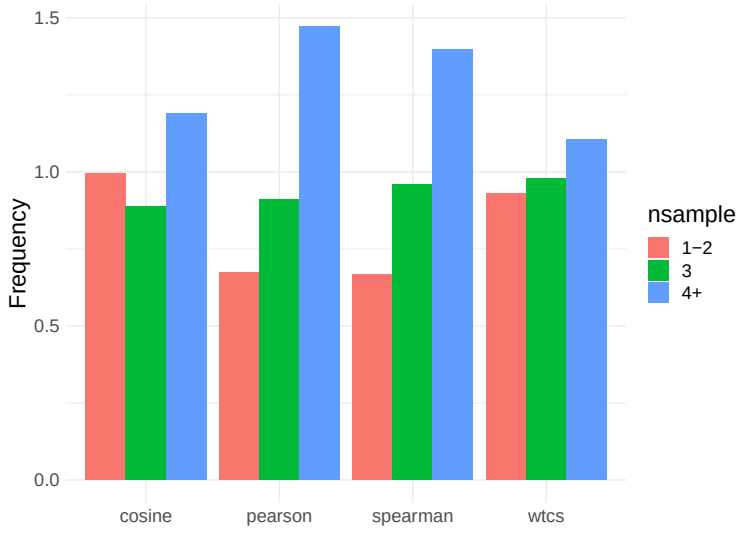

Figure 2: Querying is impacted by spurious autocorrelation. For each of four metrics and for each bin of component instances, this bar plot shows the probability of being in the top 1% of query results. Those signatures with more component instances are more likely to appear at the top of the query.

The spurious increase in autocorrelation with averaging signatures presents a challenge for any analysis using the L1000 Connectivity Map data. We do not claim that increasing autocorrelation invalidates Zhu et al's predictive model, only that it substantially inflates the apparent performance of the DLEPS model architecture. Using correlation with a meta-signature as a loss overstates the information gain from compound molecular structure. An uninformative model can exploit this similarity bias to appear to fit the data well,

resulting in an excess of false positives when evaluating model performance. The true model impact can be controlled by assessing the gain in predictive accuracy from a model over an uninformative, background model. This phenomenon must be recognized and accounted for whenever high-dimensional biological data is analyzed and particularly when averaging elements from a high dimensional space, like gene expression data.

Acknowledgments

We thank the Broad Institute and the LINCS Consortium for sharing the L1000 data with the scientific community.

References

1. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171,** 1437–1452. ISSN: 00928674. http://www.ncbi.nlm.nih.gov/pubmed/29195078%20http://linkinghub.elsevier.com/retrieve/pii/S0092867417313090 (Nov. 2017).
2. Zhu, J. *et al.* Prediction of drug efficacy from transcriptional profiles with deep learning. *Nature Biotechnology,* 1–9. ISSN: 15461696. http://www.nature.com/articles/s41587-021-00946-z (June 2021).
3. Lamb, J. *et al.* The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935. ISSN: 00368075 (Sept. 2006).
4. Abelin, J. G. *et al.* Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. *Molecular and Cellular Proteomics* **15,** 1622–1641. ISSN: 15359484. /pmc/articles/PMC4858944/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858944/ (May 2016).
5. Way, G. P. *et al.* Predicting cell health phenotypes using image-based morphology profiling. *Molecular Biology of the Cell* **32,** 995–1005. ISSN: 19394586. https://www.molbiolcell.org/doi/abs/10.1091/mbc.E20-12-0784 (Apr. 2021).
6. Smith, I. *et al.* Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biology* **15** (ed Freeman, T.) e2003213. ISSN: 15457885. http://dx.plos.org/10.1371/journal.pbio.2003213 (Nov. 2017).
7. Szalai, B. *et al.* Signatures of cell death and proliferation in perturbation transcriptomics data - from confounding factor to effective prediction. *Nucleic Acids Research* **47,** 10010–10026. ISSN: 13624962. https://portals.broadinstitute. (2019).