Nate Evans

CS652: Deep Learning

Meysam Asgari

10/4/2019

# Functional Drug Response Deep Neural Network Model using Gene Dependency Screens

"For decades now, we have been able to predict with precision the behavior of an electronic integrated circuit in terms of its constituent parts—its interconnecting components, each responsible for acquiring, processing, and emitting signals according to a precisely defined set of rules. Two decades from now, having fully charted the wiring diagrams of every cellular signaling pathway, it will be possible to lay out the complete “integrated circuit of the cell” upon its current outline. We will then be able to apply the tools of mathematical modeling to explain how specific genetic lesions serve to reprogram this integrated circuit in each of the constituent cell types so as to manifest cancer." [9]

The heterogeneity in tumor drug response makes it difficult for oncologists to develop robust umbrella treatment strategies. *Precision oncology* aims to provide treatments optimized for individual patients. A core tenet of this philosophy is that patient specific features such as tumor mutations, expression patterns, and immune response contribute to the variation in drug response. Therefore, by measuring these features and linking them to drug response, clinicians will be able to provide targeted treatment strategies. Dose-response assays are a high-throughput functional data screening, which use colorimetric assays to quantify cell-viability over a range of drug concentrations. From this data, summary metrics such as *area under the curve (AUC)* or *IC50* act as a surrogate for patient drug sensitivity. Another functional drug response data type is *gene knockdown* by *shRNA* and *gene knockout (KO)* by CRISPR data in pooled high throughput experiments. Using these methods, biologists can test a hundreds of cell line’s gene dependencies for tens of thousands of genes. The Broad institute maintains the *Cancer Dependency Map* (DepMap), which aggregates a wide number of functional drug data, performed on patient derived cancer cell lines. This resource provides cell line characteristics (RNAseq, DNAseq, CNV), cell line dependencies (siRNA, CRISPR) and cell line drug response assays (colorimetric dose-response, pooled barcoded cell-line). An important feature of drug response predictions rely on the known drug targets and a recent paper, proposing the *Cancer Targetome* [11], outlines the resources and methods for drug target information as well as aggregates a subset of drug targets in a highly useable format. BeatAML project has publicly recently released a significant dataset (>600 patients, >100 drugs) of patient derived drug screens, and represents a valuable source of cancer drug response information [5]. There have been several deep learning models developed previously that use various 'omics to predict drug response [1,7,8]. One competitive model, coined *DeepDSC* [1] is trained from Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity (GDSC) data and uses genomic features to predict IC50.

A major limitation of training DNN's are the availability and quality of data that generalizes well to target populations. For drug response models, functional drug screens are a gold standard for this training data and publicly available dataset comprise ~330,000 observations (DepMap: GDSC, CCLE). A datatype that, to our knowledge, has not been used to train these models is gene dependency information (siRNA & CRISPR pooled screens). There are several DepMap resources (RNAi, Achilles) that have performed high throughput pooled gene dependency screens on several hundred cell lines. These screens use either shRNA or CRISPR, delivered in a lentivirus to infect a cell line, to knockout specific genes. By measuring the number of shRNA integrated DNA segments (by probe array or NGS DNAseq) after many doublings (~16) this assay can detect the fold change of a given shRNA; These dependency scores are log2 transformed meaning negative values correlate to essentiality of the given gene, positive values correlate with suppressive gene behavior. These assays typically measure around 18 thousand genes in over 12 hundred cell lines. This data is highly analogous to drug data in the target-response format. It is not well understood if this data could be used to train a DNN model and inform target sensitivity linked to genomic features. Together Achilles and a combined RNAi dataset have around 23 million observations. DepMap has also made the *PRISM* dataset available [10] which is a recently developed high throughput pooled drug screen. The incorporation of this dataset into a functional drug response DNN, to our knowledge, has not been attempted. Additionally, the response metric IC50 has been found to be less robust than AUC [6] and yet previous models routinely predict IC50. Last, recently the recently published beatAML dataset and OHSU internal HNSCC project offers drug-response data on tumor patient samples, and represents a gold-standard test of the generalizability and usefulness of a given DNN. There is immediate need for improved models of genome specific functional drug response. With the rise of precision oncology, projects like OHSU's SMMART (Mills, Gray, Raymond) and HNSCC (McWeeney, Kullez-Martin) are rate limited by current in-vitro drug-response screens. Robust drug-response models can provide in-silico prioritization of in-vitro screens, clinical treatments, or inform drug research.

The long-term goal of this project is to provide a method and pipeline for in-silico prioritization of drugs based on efficacy, therapeutic window and synergies. For our immediate objective, we propose to train a DNN using drug response data (CCLE, GDSC, PRISM, beatAML) and gene dependency data (RNAi combined: Broad, Novartis, Marcotte, CRISPR: AVANA) on roughly one thousand cell lines and six hundred patient tumor samples. We propose to use exogenous variables of drug name, drug family, drug target, expression data (RNAseq), haplotype, mutation data and copy number variation (CNV) data (DNAseq) to predict in-vitro drug response (AUC). Additionally, we propose to use resources outlined-in or aggregations-from the *Cancer Targetome* [11] to curate drug target data. By using previously unavailable datasets and siRNA/CRISPR screen data we can significantly improve predictive value of the model. Additionally, AUC has been shown to be a more robust response metric and will improve dataset intersection [6]. We hypothesize that the relations between cancer genomic features and gene dependency will improve the predictive power of a complex deep learning model. An improved model of cancer genomic feature specific drug response can provide vast capacity of in-silico drug screens and ultimately inform clinical treatment of cancer, prioritize in-vitro drug screens, or guide research.

**Aim 1**: Train a predictive model using Deep Neural Network architecture to use cancer 'omic and drug features to predict cancer cell line response. We will approach this by using functional drug data from CCLE and GDSC and associated 'omic features. Use drug features provided by resources outlined in the cancer targetome [11].

**Aim 2:** Convert gene dependency data into functional drug response form. We will approach this by using the shRNA and CRISPR screens, we will map the 16-generation mutant-gene presence fold change to an AUC analog, and use the ~11/12 million screen, ~600/500 cell-line dataset to do further training. Use test set from AIM 1 to validate model performance.

**AIM 3:** Apply our trained model on the HNSCC and beatAML datasets. We will approach this by quantifying model performance on the beatAML and HNSCC dataset to explore the generalizability of our model. Using HNSCC data, test all known (training set) drug factors and rank by AUC. This will function as model validation / generalizability and use case to prioritize untested but potentially effective inhibitors for future drug panel screens.

While a variety of DNN models have been applied to functional drug response prediction, few of them predict the AUC summary metric, which has been shown to have better correlation between datasets and thus should add robustness to the model. Additionally, to our knowledge, no one has used siRNA/CRISPR gene screen data to inform the model of cancer dependencies. This source of information may represent a significant improve in predictive power. We expect to release a functional DNN model, trained on greater data volume then has been previously used for these purposes and validated on real-world data to estimate model accuracy and use. An accurate and generalizable model that effectively uses genomic features to predict drug-response will empower widespread in-silico drug testing, and improve patient outcome by prioritizes treatments, and guiding research.

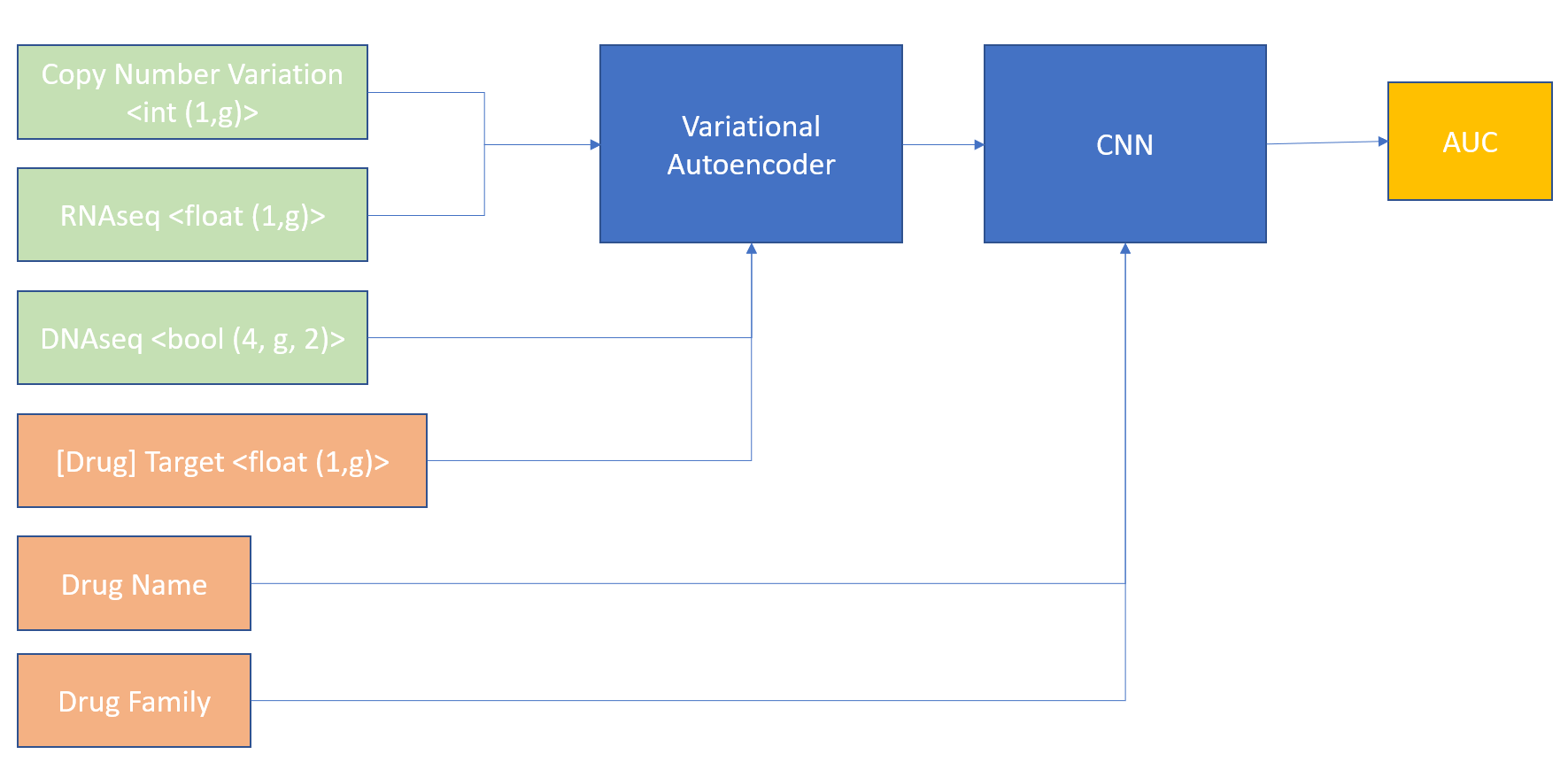
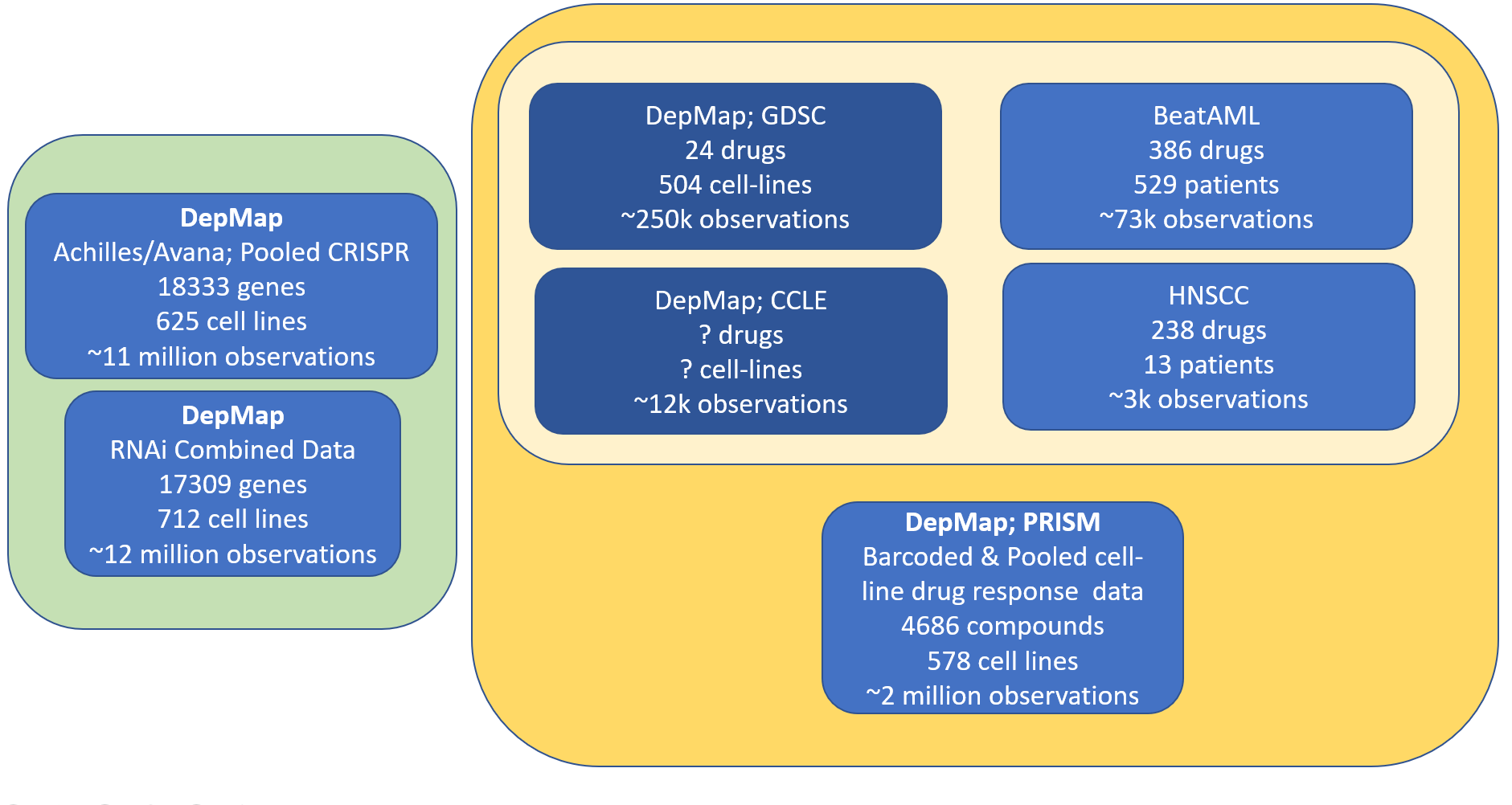
 

Figure 2: Overview of the data sets available for training our model. To our knowledge, previous DNN models have only utilized GDSC, CCLE for training data (dark blue). Green indicates gene dependency data. Light gold indicates traditional colorimetric dose-response data. Dark Gold indicates drug-response data. Only beatAML and HNSCC datasets use fresh patient derived tumor samples for screens.

Figure 1: The architecture overview and input outputs. The variational autoencoder will be trained using additional data sources such as TCGA.

# References

[1] Li, M. *et al.* DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Trans Comput Biol Bioinform* (2019). doi:[10.1109/TCBB.2019.2919581](https://doi.org/10.1109/TCBB.2019.2919581)

[2] Defining a Cancer Dependency Map: Cell. Available at: <https://www.cell.com/cell/fulltext/S0092-8674(17)30651-7?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867417306517%3Fshowall%3Dtrue>. (Accessed: 3rd October 2019)

[3] Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

[4] El-Brolosy, M. A. *et al.* Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**, 193–197 (2019).

[5] Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences* **105**, 20380–20385 (2008).

[6] Pozdeyev, N. *et al.* Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget* **7**, 51619–51625 (2016).

[7] Chiu, Y.-C. *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics* **12**, 18 (2019).

[8] Xia, F. *et al.* Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* **19**, 486 (2018).

[9] Weinberg, R. A. *The Biology of Cancer, 2nd Edition*. (W. W. Norton & Company, 2013).

[10] Non-oncology drugs are a source of previously unappreciated anti-cancer activity | bioRxiv. Available at: <https://www.biorxiv.org/content/10.1101/730119v1>. (Accessed: 4th October 2019)

[11] Blucher, A. S., Choonoo, G., Kulesz-Martin, M., Wu, G. & McWeeney, S. K. Evidence-Based Precision Oncology with the Cancer Targetome. *Trends Pharmacol. Sci.* **38**, 1085–1099 (2017).