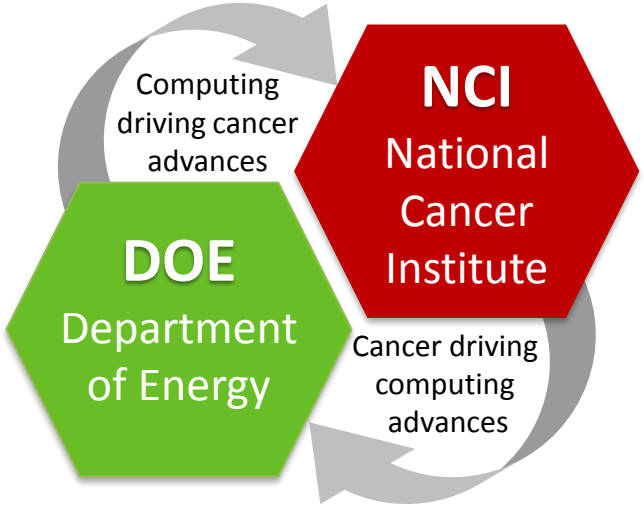
# Pilot 1: Predictive Modeling for Pre-Clinical Screening

*Yvonne Evrard*
*NCI Patient-Derived Models Repository*
*Frederick National Laboratory for Cancer Research*

*Rick Stevens*
*Argonne National Laboratory*
*University of Chicago*




October 18 2017
**Presented to: FPOC**

U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

# Cross Laboratory Team: ANL, LANL, LLNL, ORNL, NCI

**ANL**: **Rick Stevens**, Frank Alexander, Cori Anderson, Jillian Aurisano, Prasanna Balaprakash, Tom Brettin, Jim Davis, Emily Dietrich, Nicoli Dryden, Veronika Dubinkina, Hal Finkel, Ian Foster, Monisha Ghosh, Ushma Kriplani, Ravi Madduri, Sergei Maslov, Bob Olson, Dan Olson, Mike Papka, Lorenzo Pesce, Alex Partin, John Santerre, Maulik Shukla, Venkat Vishwanath, Fangfang Xia, Harry Yoo

**LANL**: **Marian Anghel**, Tanmoy Bhattacharya, Judith Cohn, Paul Dotson, Will Fischer, Kumkum Ganguly, Jason Gans, Cristina Garcia-Cardona, Nick Hengartner, William Hlavacek, John Hogden, Patrick Kelly, Miranda Lynch, Ben Mcmahon

**LLNL: Jonathan Allen**, Ya Ju Fan, Stewart He

**ORNL**: **Fernanda Foertter**, Arvind Ramanathan

**NCI: James Doroshow, Yvonne Evrard, Susan Holbeck**, **Eric Stalhberg**, George Zaki

P1 Hackathon 2

CANDLE Hackathon 1

P1 Hackathon 3

# Pilot 1: Predictive Models for Preclinical Screening

*Terabytes* **Cell Line Compound Screens**

*Terabytes* **Compound Databases**

*Terabytes* **Cell Line Molecular Assays**

*Petabytes* **PDX Features (DNA, RNA, Images)**

*Terabytes* **PDX Compound Screens**

*Petabytes* **Protein-Protein Interactions**

*Terabytes* **Cancer Pathway Databases**

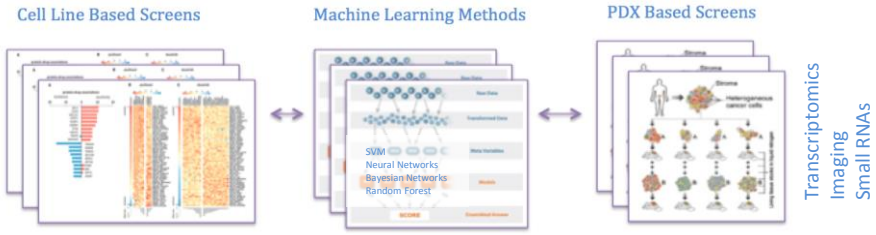
*Petabytes* **Cancer Genome Atlas**

*Petabytes* **Gene Expression Omnibus**

*Petabytes* **Cancer Data Commons**

## Machine Learning Based Predictive Models

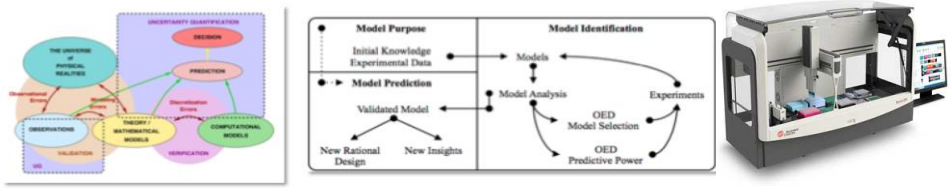
Transcriptomics
Proteomics
Variation




Transcriptomics
Imaging
Small RNAs

Feature Engineering, Cross Validation, Scalable Compute on CORAL

## Uncertainty and Optimal Experiment Design

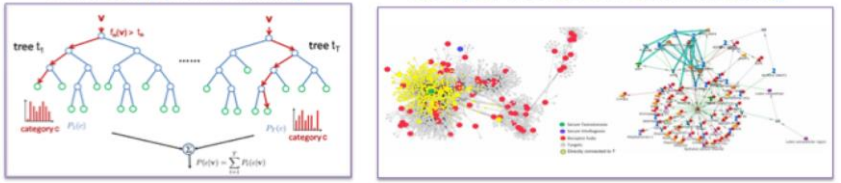



UQ Analysis, Model Selection, Model Improvement, Proposed Experiments

## Hypotheses Formation and Mixed Modeling

Feature Importance Mining

Biological Interaction Network Modeling



Integration of Mechanistic, Statistical and Inferential Modeling

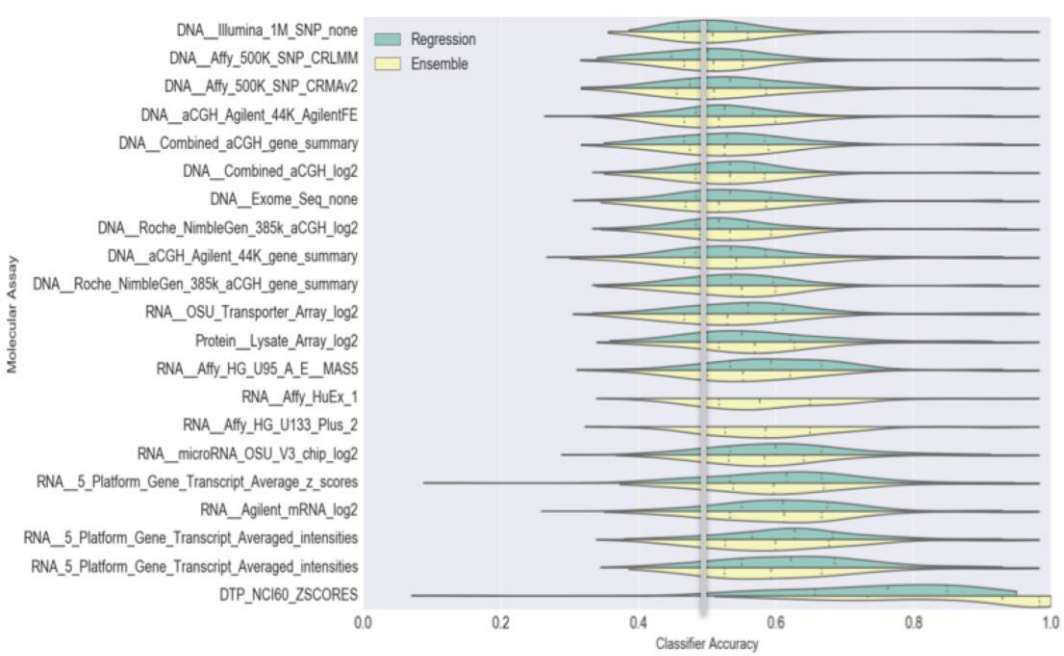
# Aims for Preclinical Screening Pilot

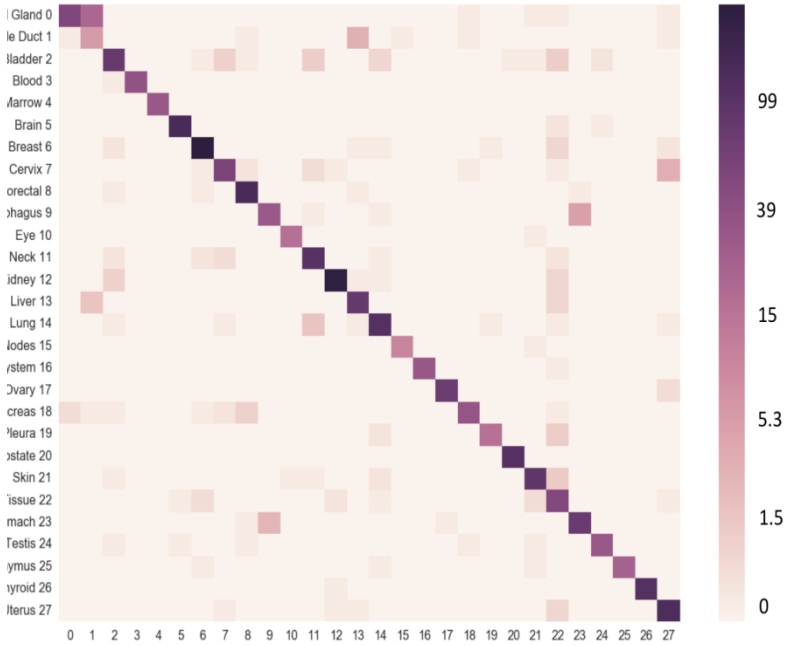
- **Reliable machine learning based predictive models of drug response that enable the projection of screening results from and between cell-lines and PDX models**
- **Uncertainty quantification and optimal experimental design to assert quantitative limits on predictions and to recommend experiments that will improve predictions**
- **Improved modeling paradigms that support the graded introduction of mechanistic models into the machine learning framework and to rigorously assess the potential modeling improvements obtained thereof**

# Pilot 1: Relevant Datasets

| Datasets | NCI-60 | CLDX | PDX | Sarcoma | SCLC | CCLE | GDSC | GDC |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| Cell lines / Samples | 61 | 49 x 4 | 1000 | 74 | 76 | 504 | 1074 | 14,531 |
| Cancer types | | | | | | | | |
| Compunds Tested | 92,691 | 1000 | | 445 | 525 | 24 | 265 | NA |
| | | | | | | | | |
| Dose Response Data | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | |
| DNA Copy Number Variation - Agilent 44K aCGH | ✓ | | | | | | | |
| DNA Copy Number Variation - NimbleGen 385k aCGH | ✓ | | | | | | | |
| DNA Copy Number Variation - Affy HU SNP Array 6.0 | | | | | | ✓ | ✓ | ✓ |
| DNA Methylation - Illumina Methylation 450k | ✓ | | | | | | ✓ | ✓ |
| DNA Methylation - Illumina Methylation 27 | | | | | | | | ✓ |
| SNP - Affy 500k | ✓ | | | | | | | |
| SNP - Affy HU SNP Array 6.0 | | | | | | ✓ | | ✓ |
| SNP - Illumina 1M | ✓ | | | | | | | |
| SNP - Exome Seq | ✓ | | ✓ | | | | ✓ | ✓ |
| SNP - OncoMap 3.0 | | | | | | ✓ | | |
| SNP - Hybrid capture sequencing | | | | | | ✓ | | |
| Gene Expression - Affy HG U133 plus 2 | ✓ | ✓ | | | | ✓ | | |
| Gene Expression - Affy HG U133 A-B | ✓ | | | | | | | |
| Gene Expression - Affy HG U95 A-E | ✓ | | | | | | | |
| Gene Expression - Affy HG U219 | | | | | | | ✓ | |
| Gene Expression - Affy Human Exome Array 1.0 | ✓ | | | ✓ | ✓ | | | |
| Gene Expression - Agilent mRNA | ✓ | | | | | | | |
| RNA-seq - gene expression | ✓* | | ✓ | | | ✓ | ✓ | ✓ |
| RNA-seq – matched normal vs tumor samples | | | | | | | | ✓ |
| miRNA Expression - Agilent miRNA | ✓ | | | | | | | |
| miRNA Expression - NanoString | | | | ✓ | ✓ | | | |
| miRNA-seq – Illumina | | | | | | | | ✓ |
| Proteome and kinome | ✓ | | | | | | | |
| Histology images | | | ✓ | | | | | |

# NCI-60 Drug Response Prediction Sweeps on Assays and Machine Learning Algorithms




# Cancer Site/Type Classification Prediction from Expression Profiles: 98% Accuracy

# Pilot 1 Modeling Efforts are Constrained by Small Sample Sizes

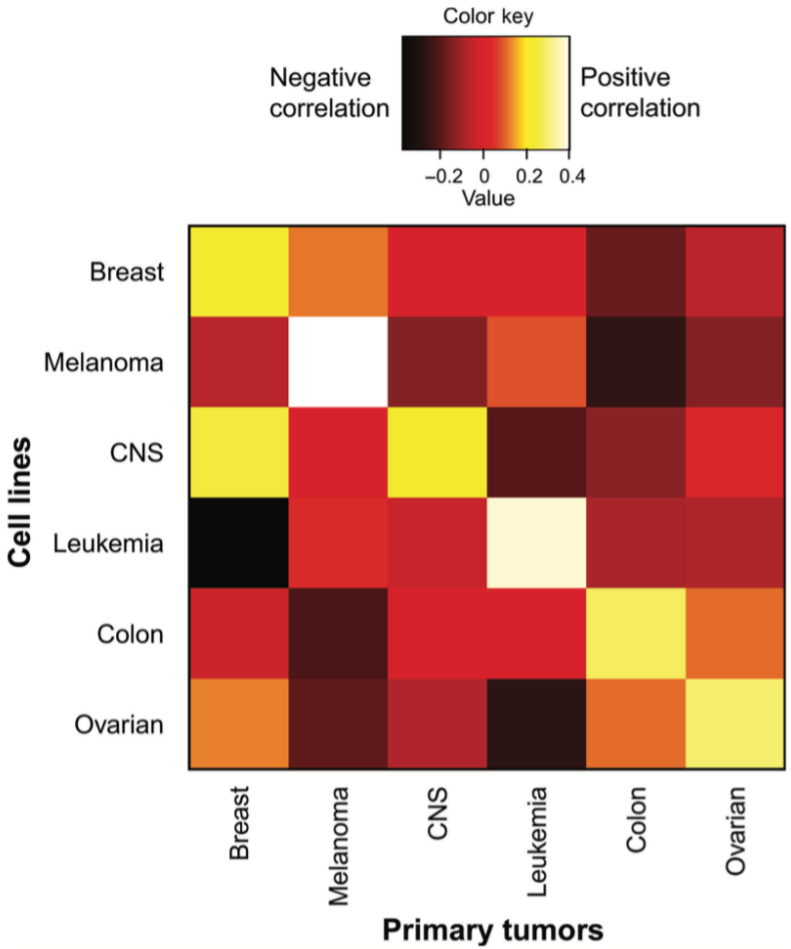
Effort is aimed at understanding how to address this problem

1) By aggregation of data from multiple sources (NCI-60, CCLE, GDSC, PDX , **GDC**, etc.)
2) By data augmentation, oversampling and synthetic data generation (GANs, VAEs, etc.)
3) By using semi-supervised learning (augmenting labeled data with unlabeled data)
4) By Transfer learning (training predictors on related data to improve performance on given data)
5) By Multi-task learning (where several objectives are learned at once leveraging each other)

| Sample size | 30 | 100 | 300 | 1000 |
|---|---|---|---|---|
| Confidence bounds | ±15% | ±10% | ±6% | ±3% |

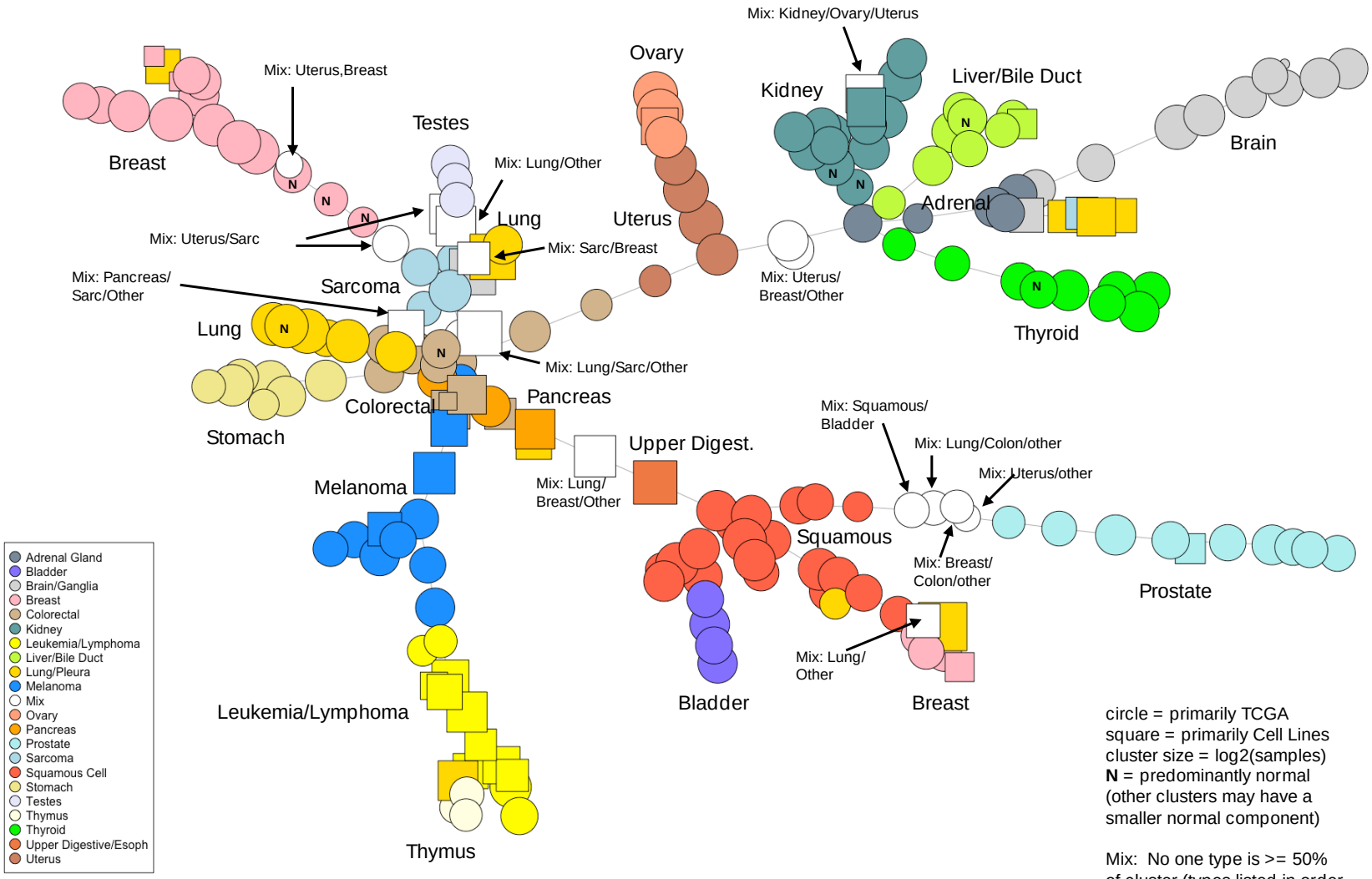
Table 1: **Confidence bounds to be expected for a binary classification**, summarizing experiments and simulations in Figure 1. Actual confidence bounds may be significantly larger in adverse situations such as with correlated observations or very unstable classifiers.

# Cell Line Expression vs. Primary Tumor




JNCI_J_Natl_Cancer_Inst_2013_Gillet.pdf

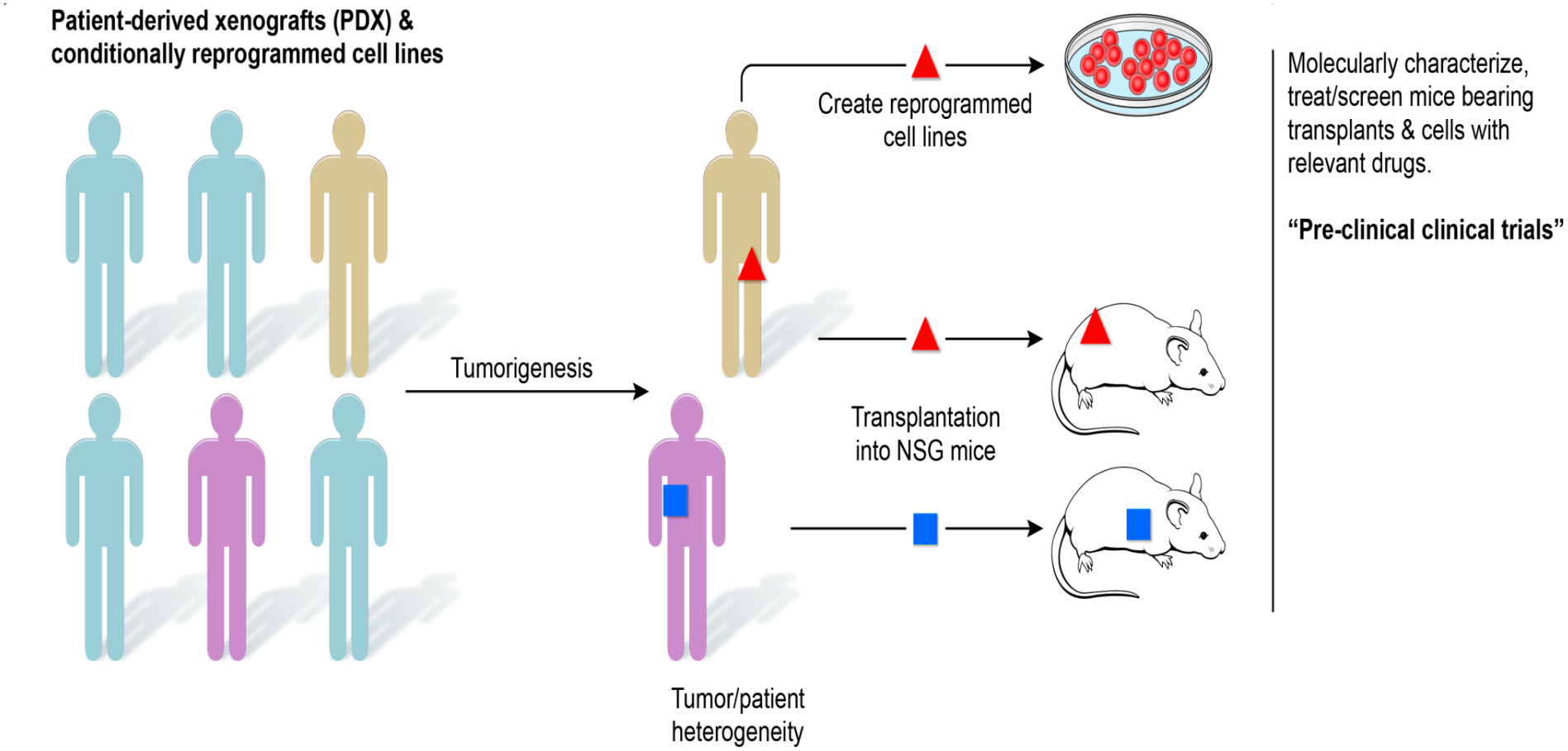
# 11,081 Tumors, 2401 Cell Lines, k=200




circle = primarily TCGA
square = primarily Cell Lines
cluster size = log2(samples)
**N** = predominantly normal (other clusters may have a smaller normal component)

Mix: No one type is >= 50% of cluster (types listed in order of prevalence

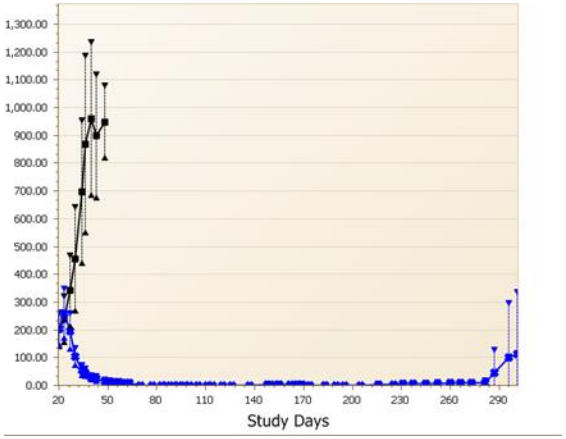
# Patient Derived Xenograft Models

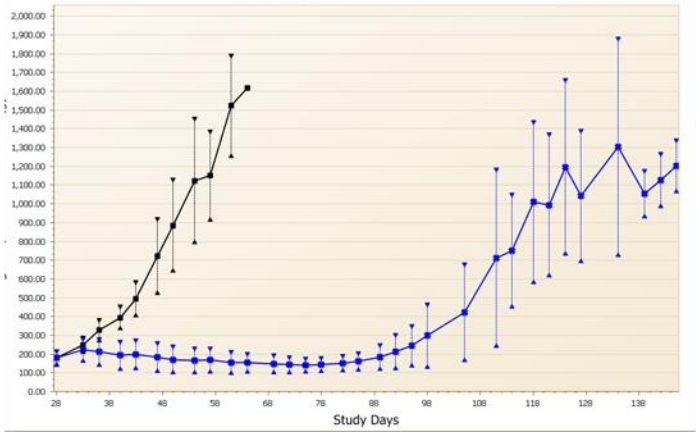


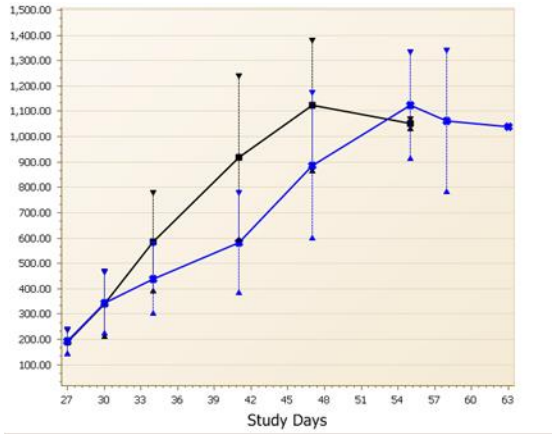

Nature Rev. Clin. Oncol. 11: 649-662, 2014.

# RNASeq Comparisons of models with Complete Response and No Response to ABT-888 + Temozolomide

Preclinical Trial with ABT-888 (PARP inhibitor) + Temozolomide across multiple histologies.
Several models had complete responses to the combination with no measurable tumor out to 300-d post-treatment
Collected samples at multiple time points
Performing RNASeq on 6 models as a pilot study using pre-treatment and post-dose day 5.
3 Bladder Ca models and 3 Colon Adenocarcinoma models. In both cases had a Complete Response model, No Response model, and a model with an intermediary response.
**Can differences in expression that contribute to response, or lack of response, be identified?**



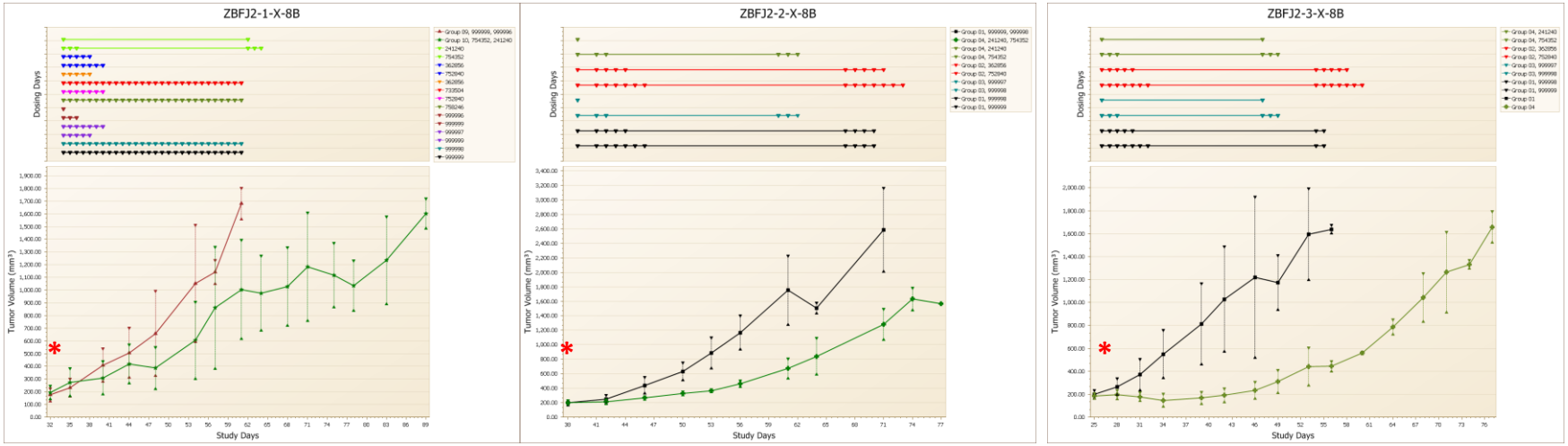



# Loss of Carboplatin Resistance in PDX

Patient:  Metastatic Lung Adenocarcinoma.
April 2014: Carboplatin, Paclitaxel: Stable Disease
August 2014: Carboplatin, Paclitaxel: Disease Progression (scalp met resistant)
March 2015: Sample for PDMR (scalp met)

- **Vehicle Control**
- **Carboplatin + AZ1775 (Wee1 inhibitor)**




Passage 3 PDX  Passage 6 PDX  Passage 9 PDX

***PDMR preparing samples for RNASeq of baseline PDX tumor material at each passage.  Are there differences in expression that contribute to loss of sensitivity to carboplatin treatment?**

# Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution

**Massive Heterogeneity over 100M Mutations in one Tumor**

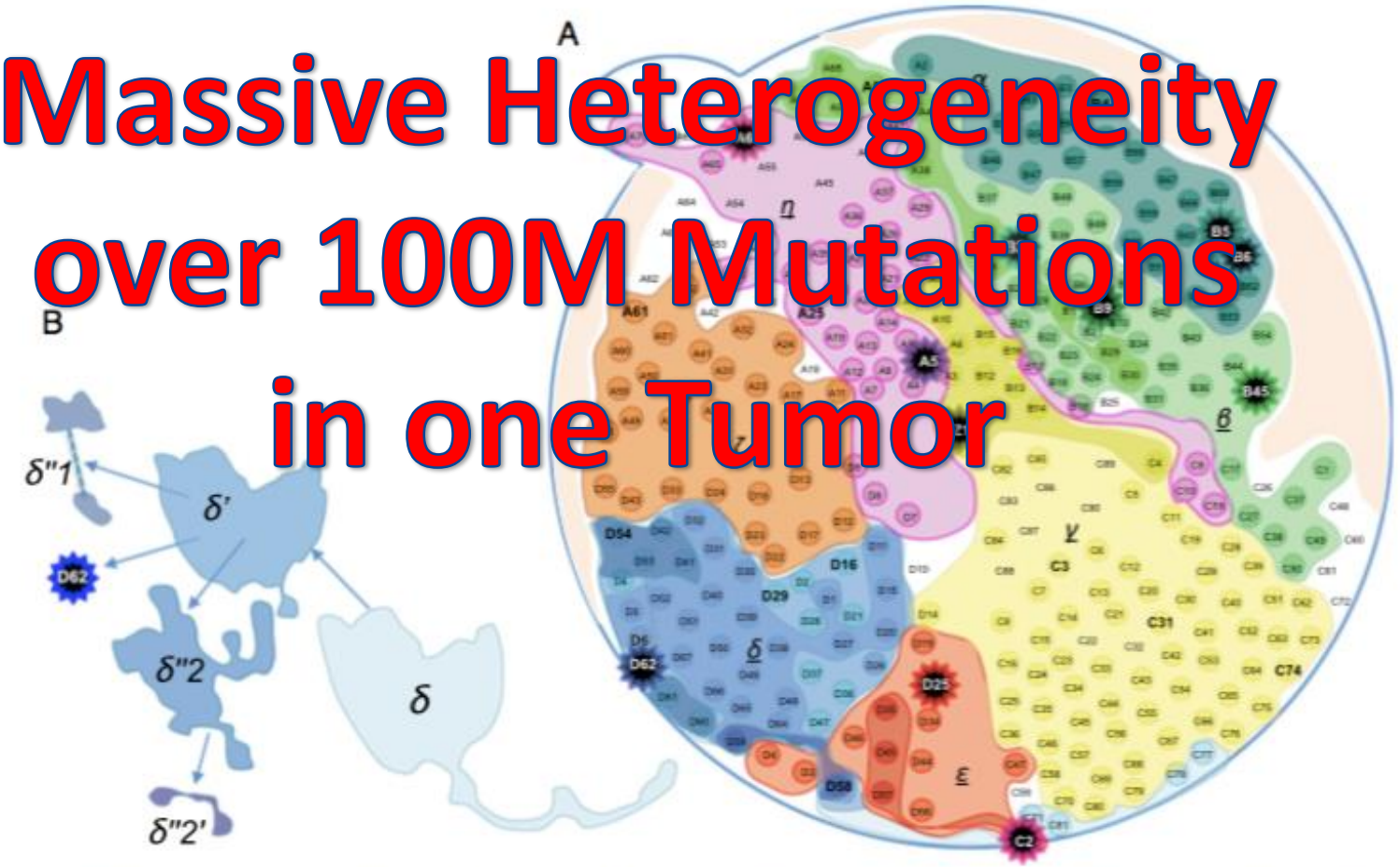
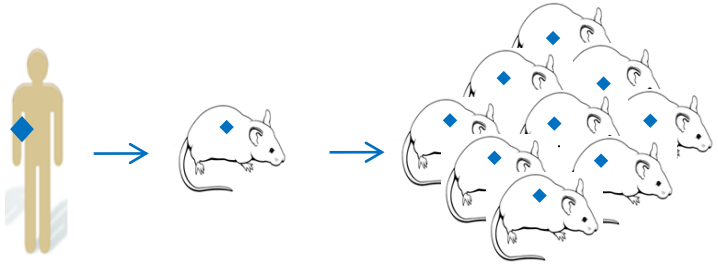


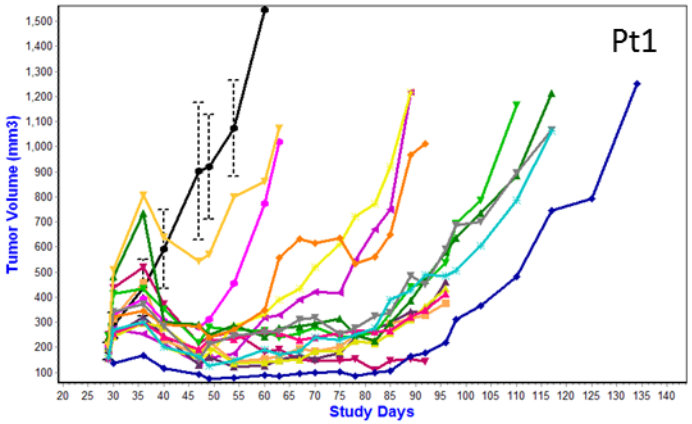

Fig. 2. Map of the mutation clones of HCC-15. A mutation clone is the aggregate of all samples carrying that mutation (main text). Hence, subclones (with increasingly darker hues) are nested within their parent clones. (*A*) Each star symbol indicates a singleton clone, represented by one sample. The clonal boundaries are delineated by the genotypes of all 286 samples. Many samples straddle two clones (including A3, B17, B19, B20, C78, D6, D9, and Z1). In this "sectoring" pattern of growth, δ′ grew outward from δ and, subsequently, δ″s (−1, −2) grew outward from δ′. Note that tumors grew in three-dimensional (3D) space but the observations made were on a two-dimensional (2D) plane. This was apparent in the "northeast" direction, along which both the α and β clones were extending from the interior toward the periphery. It appears that α grew above or below β in their expansion toward the periphery. (*B*) The δ lineage clones are pulled out to display the overlaying pattern of mutation clones. The clonal map was also used to compute the mutation frequency spectrum, $\xi_i$, which is the number of sites where the frequency of the mutation was between $(i-1)/23$ and $i/23$ from the 286 samples. We kept the number of frequency bins at 23 because the mutations discovered remained based on the initial 23 samples. The spectrum, as given in the text, is $[\xi_i = 26, 7, 1, 1, 0, 0, \ldots]$ for $i = 1$–22 (*Materials and Methods*, section 9 and Dataset S8).

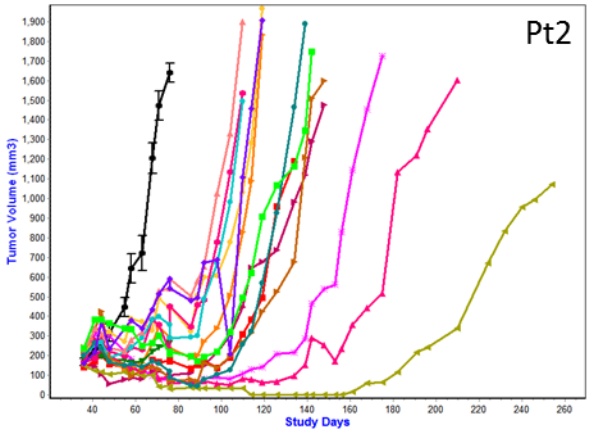
Ling et al. PNAS 11 Nov 2015

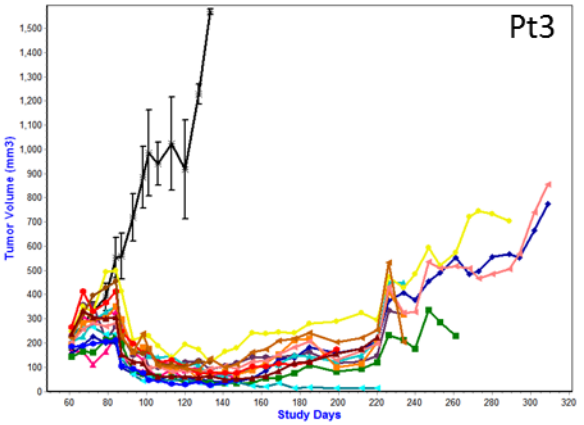
# Individual Patient Response Heterogeneity in Patient-Derived Xenografts

- Vehicle Control (Black line, median)
- 15 individual mice implanted with the same patient's tumor material treated with Combination Therapy



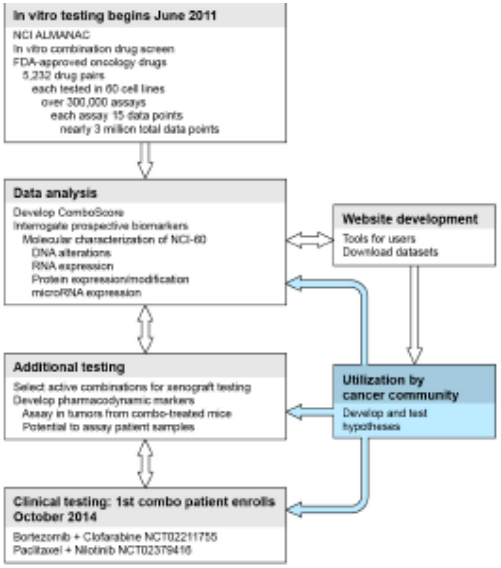



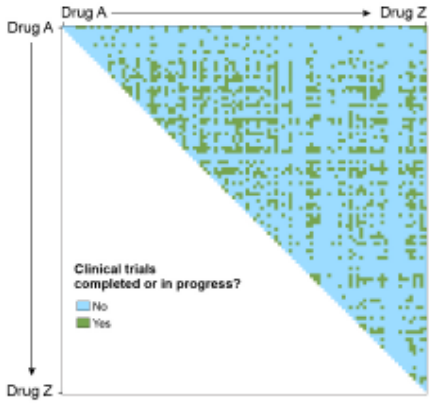
# The NCI ALMANAC: Testing All Pairwise Combinations of Approved Cancer Drugs

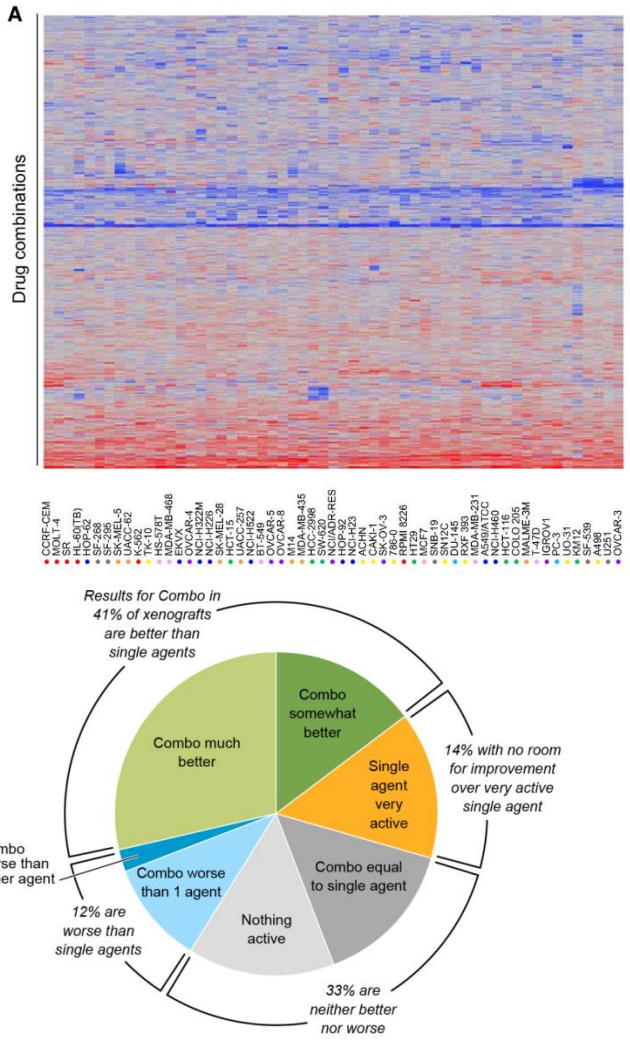
- The NCI ALMANAC (**A** **L**arge **M**atrix of **A**nti**N**eoplastic **A**gent **C**ombinations)
- Currently just over 100 small molecule oncology drugs are FDA-approved.
- Test all possible pairwise combinations: ~5000 drug pairs
- Test each drug pair in each of the cell lines in the NCI-60 panel:
  - ~300,000 experiments
  - ~4.3 million wells
- Screen run at Frederick National Labs & 2 contract locations

Holbeck, et al., 2017
Cancer Research 77

# NCI-60 Combination Data



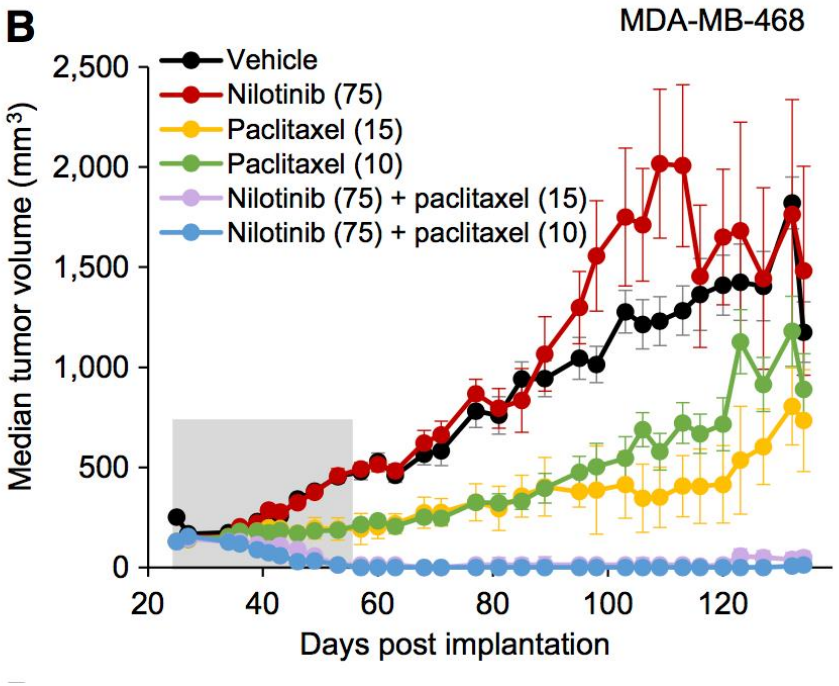



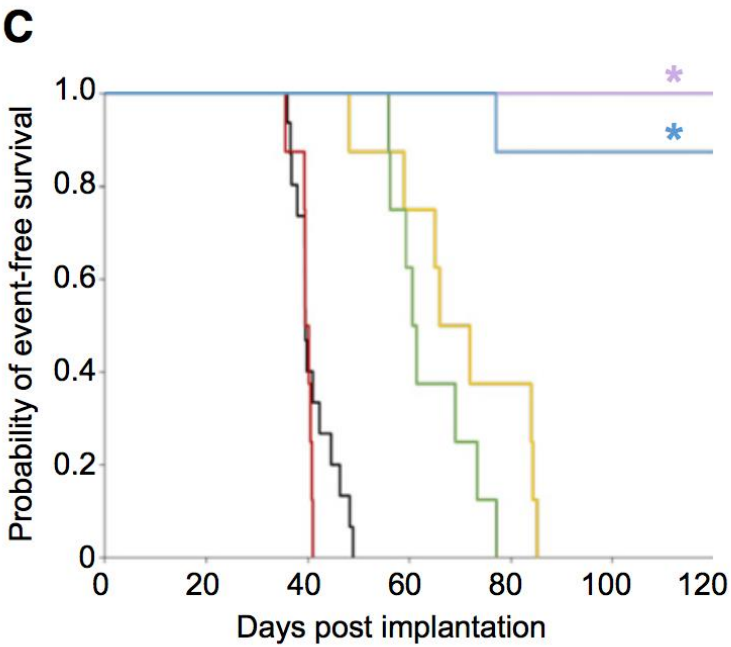




In addition, 92 xenograft experiments have been completed with at least 80% of control mice reaching 1 doubling “event” for 41 drug pairs. These drug pairs all had a good ComboScore in the corresponding cell line.

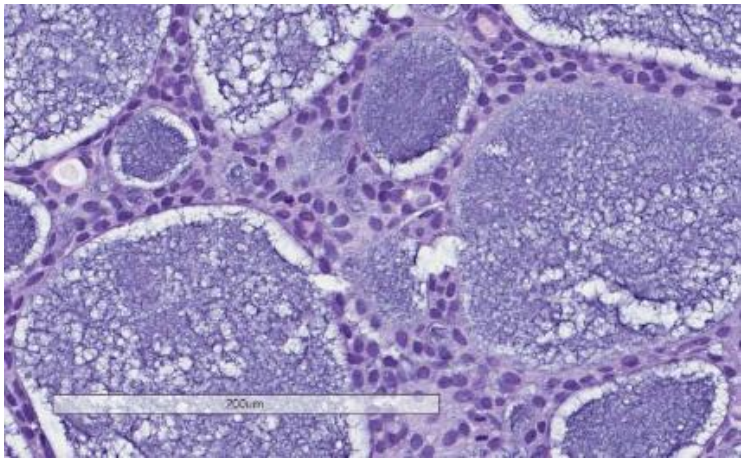
**Holbeck, et al., 2017**
**Cancer Research 77**

# Drug Pair Synergy



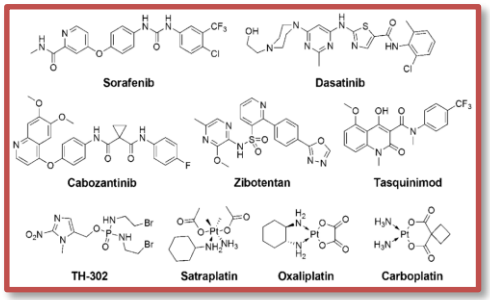



# Problem: Modeling Drug Response

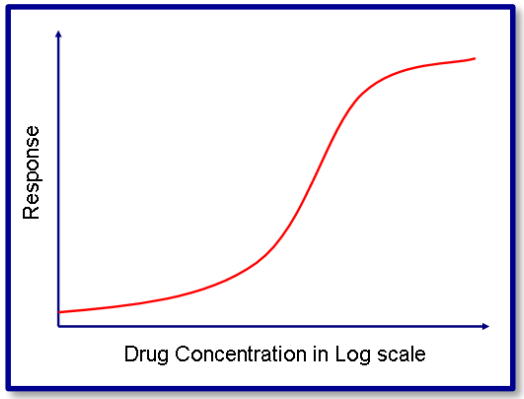


**Drug (s)**

- descriptors
- fingerprints
- structures
- SMILES
- dose




$$\mathcal{R} = f(\mathcal{T}, \mathcal{D})$$

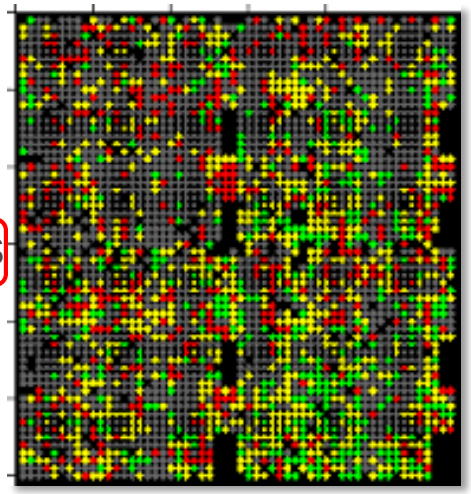



- IC50
- GI50
- % growth
- Z-score

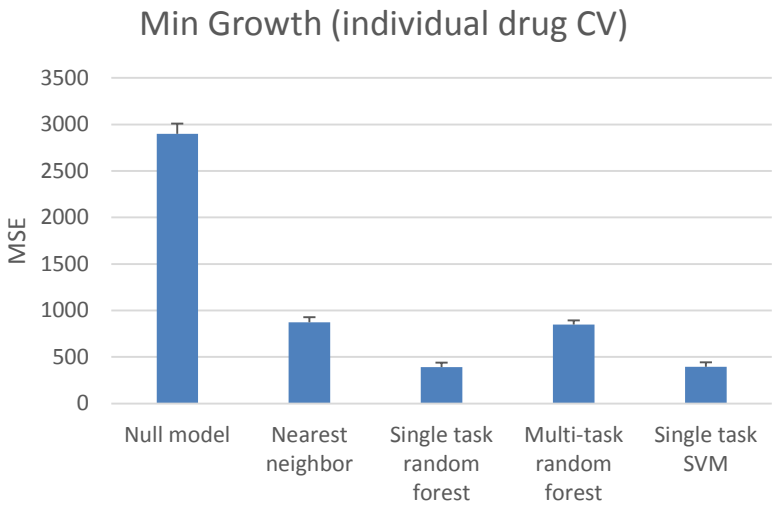
**Response**

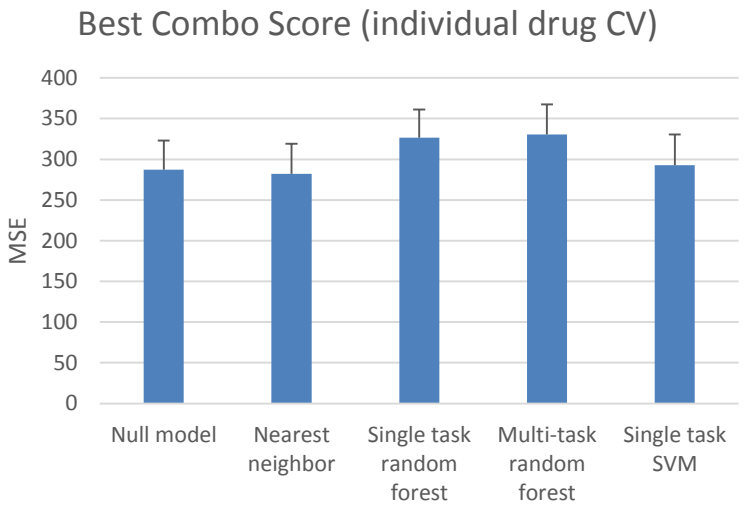
- gene expression levels
- SNPs
- protein abundance
- microRNA
- methylation

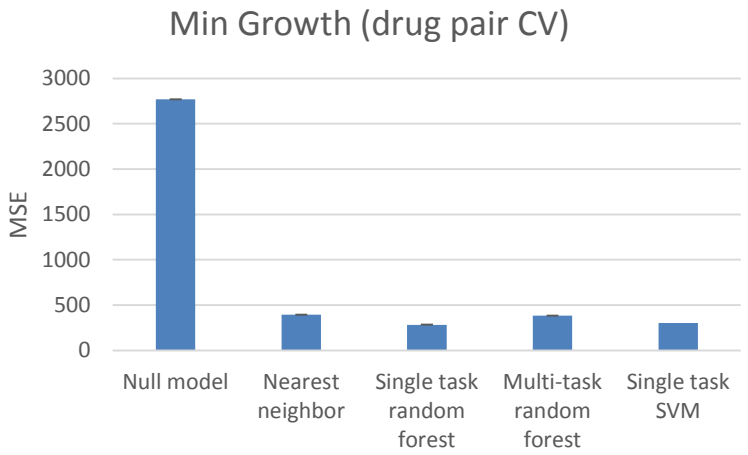
**Tumor**

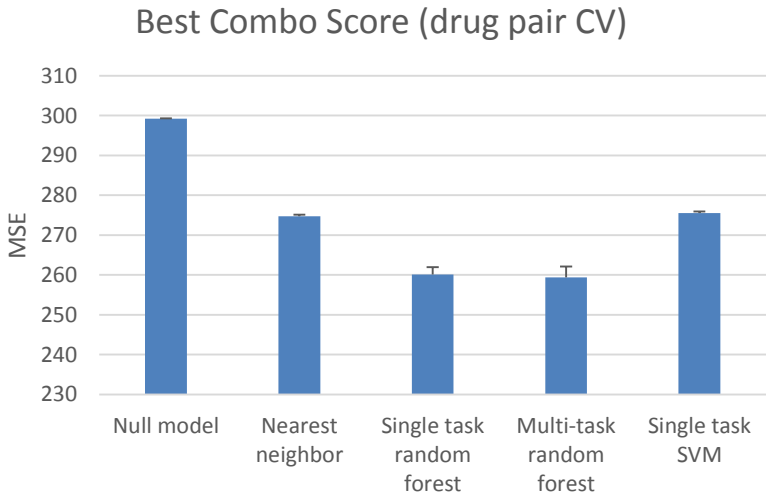


# Drug Pair Response Prediction



# Deep Learning Model for Drug Pair Response

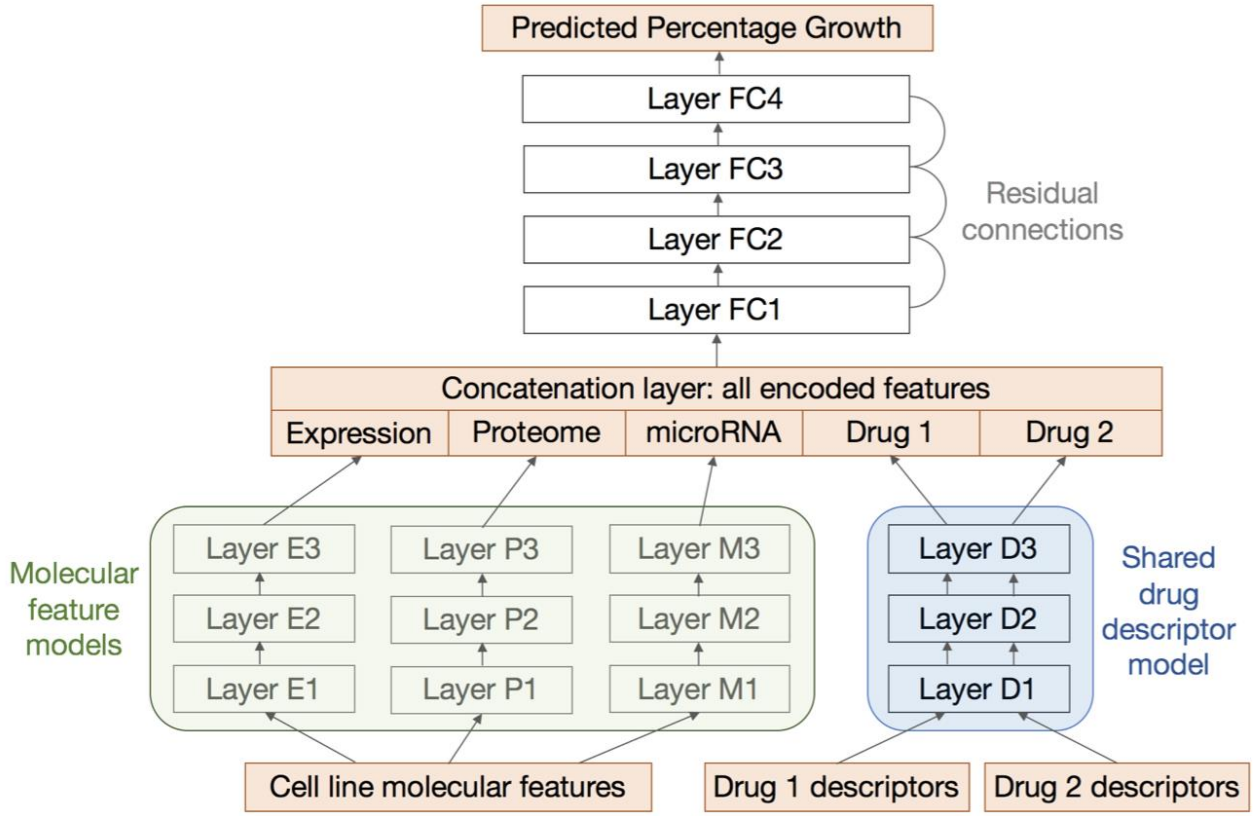



Fig. 2. **Neural network architecture.** The orange square boxes, from bottom to top, represent input features, encoded features, and output growth values. Feature models are denoted by round shaded boxes: green for molecular features and blue for drug features. There are multiple types of molecular features that are fed into submodels for gene expression, proteome, and microRNA. The descriptors for the two drugs share the same descriptor model. All encoded features are then concatenated to form input for the top fully connected layers. Most connecting layers are linked by optional residual skip connections if their dimensions match.

# Deep Learning Model for Drug Pair Response

TABLE I
CROSS VALIDATION RESULTS FROM FEATURE COMBINATION EXPERIMENTS

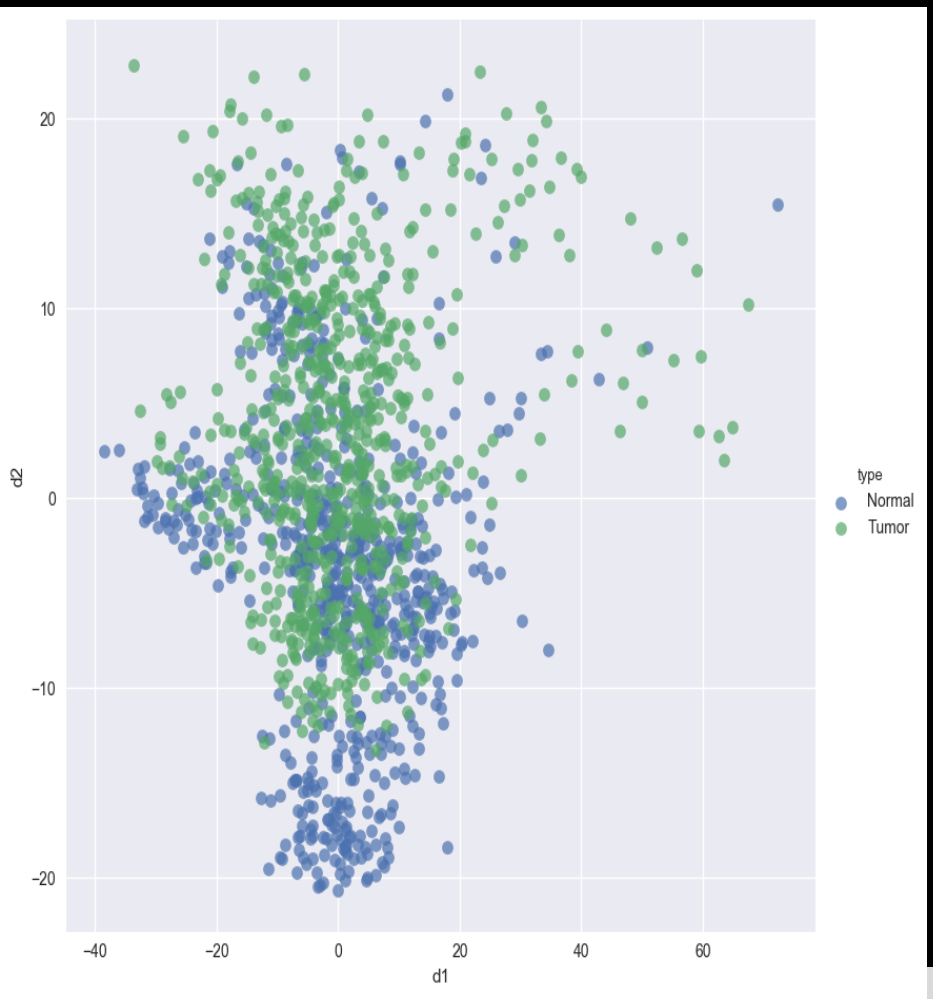
| Molecular Features | Drug Features | MSE | MAE | $R^2$ |
|---|---|---|---|---|
| baseline | baseline | 0.5253 | 0.5709 | -1.001 |
| one-hot encoding | one-hot encoding | 0.2448 | 0.3997 | 0.1269 |
| gene expression | one-hot encoding | 0.2447 | 0.3999 | 0.1272 |
| gene expression | 500-dimensional noise | 0.2450 | 0.4008 | 0.1271 |
| one-hot encoding | Dragon7 descriptors | 0.0292 | 0.1086 | 0.8892 |
| proteome | Dragon7 descriptors | 0.0303 | 0.1117 | 0.8844 |
| microRNA | Dragon7 descriptors | 0.0275 | 0.1050 | 0.8952 |
| gene expression | Dragon7 descriptors | 0.0261 | 0.1014 | 0.9005 |
| gene expression, microRNA, proteome | Dragon7 descriptors | **0.0209** | **0.0956** | **0.9208** |

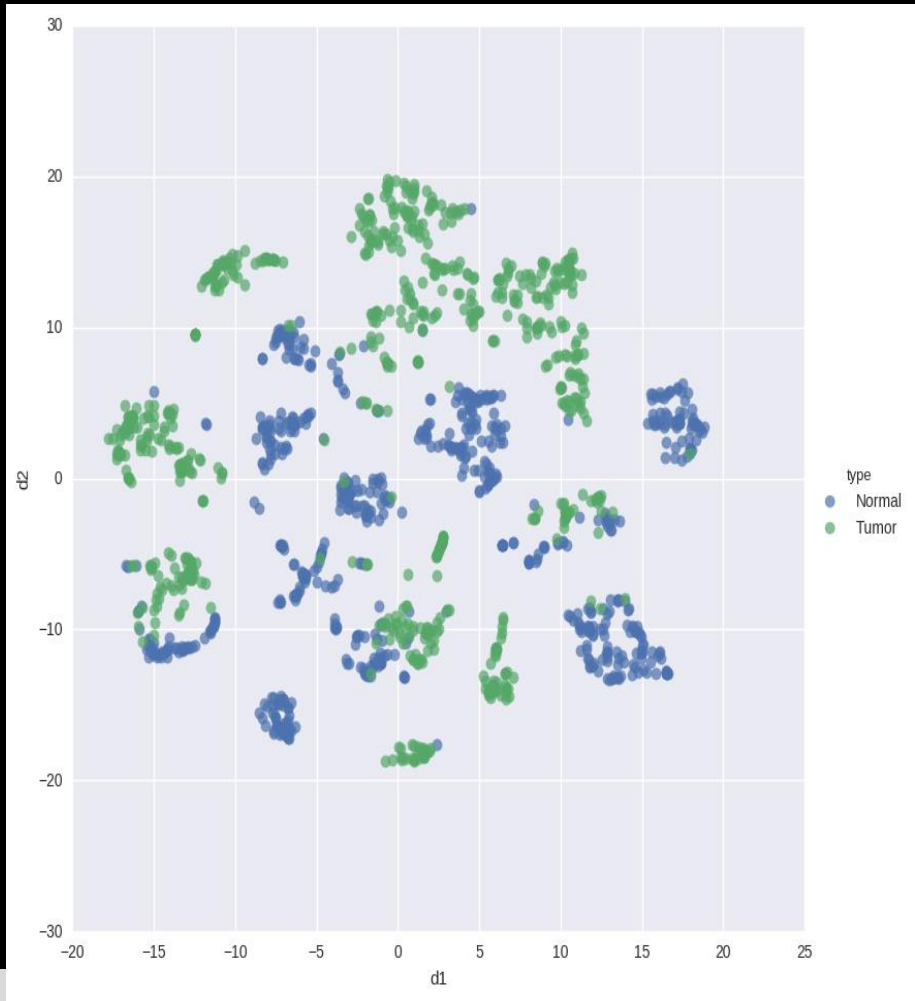
DNN Model explains 92% of the variance

# Progress in Deep Learning for Cancer

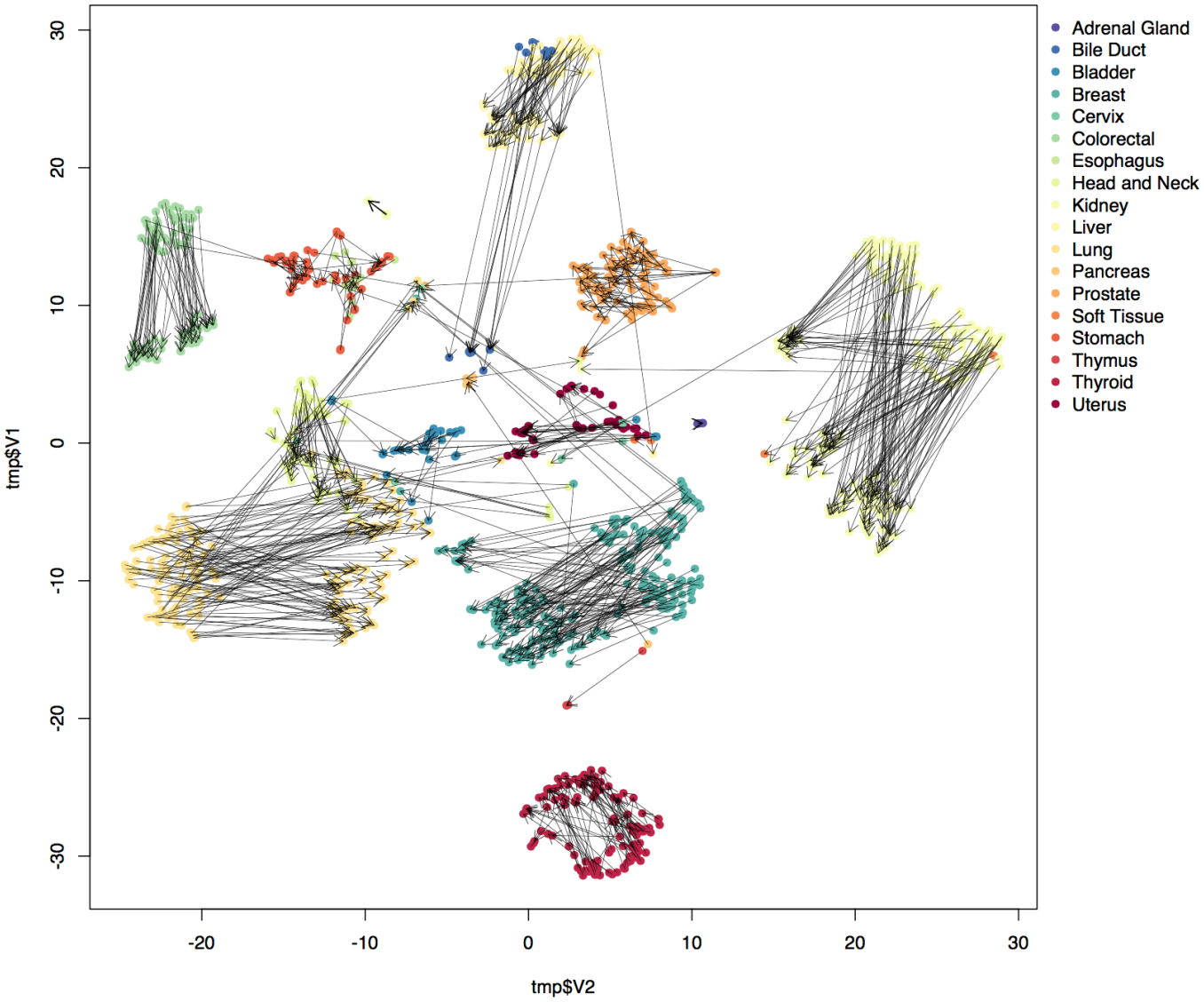
- **AutoEncoders** – learning data representations for classificaiton and prediction of drug response, molecular trajectories
- **VAEs and GANs** – generating data to support methods development, data augmentation and feature space algebra, drug candidate generation
- **DNN/CNNs** – type classification, drug response, outcomes prediction, drug resistance
- **RNNs** – sequence, text and molecular trajectories analysis
- **Multi-Task Learning** – terms (from text) and feature extraction (data), data translation (RNAseq <-> uArray)

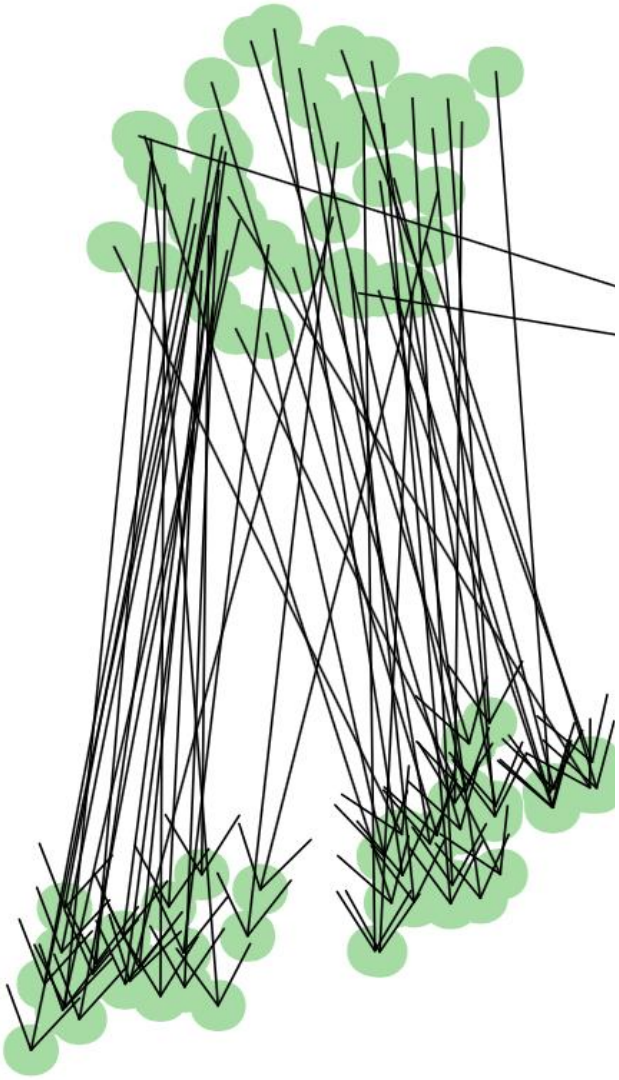
# PCA vs Deep Learning Autoencoder Clustering Normal vs Tumor Samples



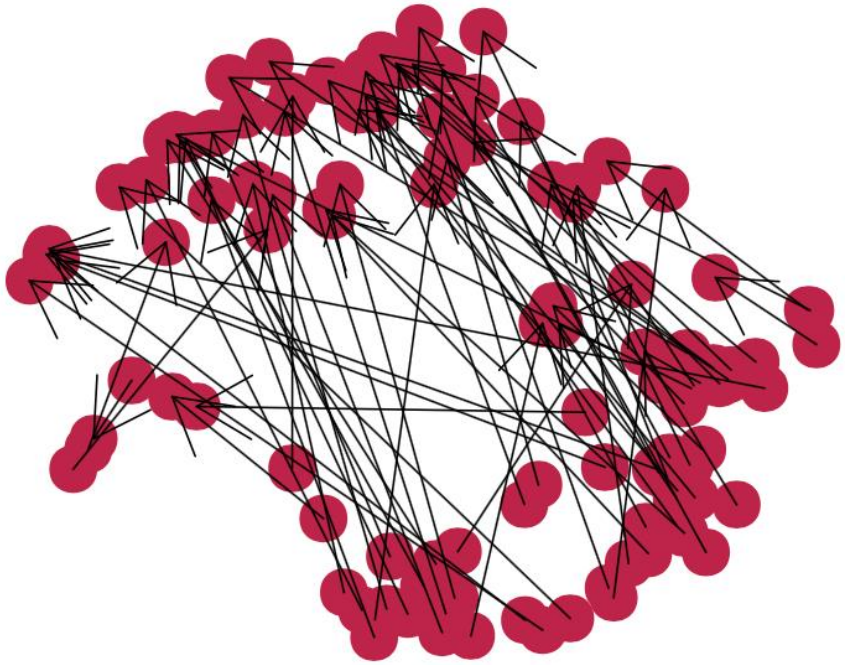
# t-sne Plot of Matched Normal Pairs Showing Translation in Features Space

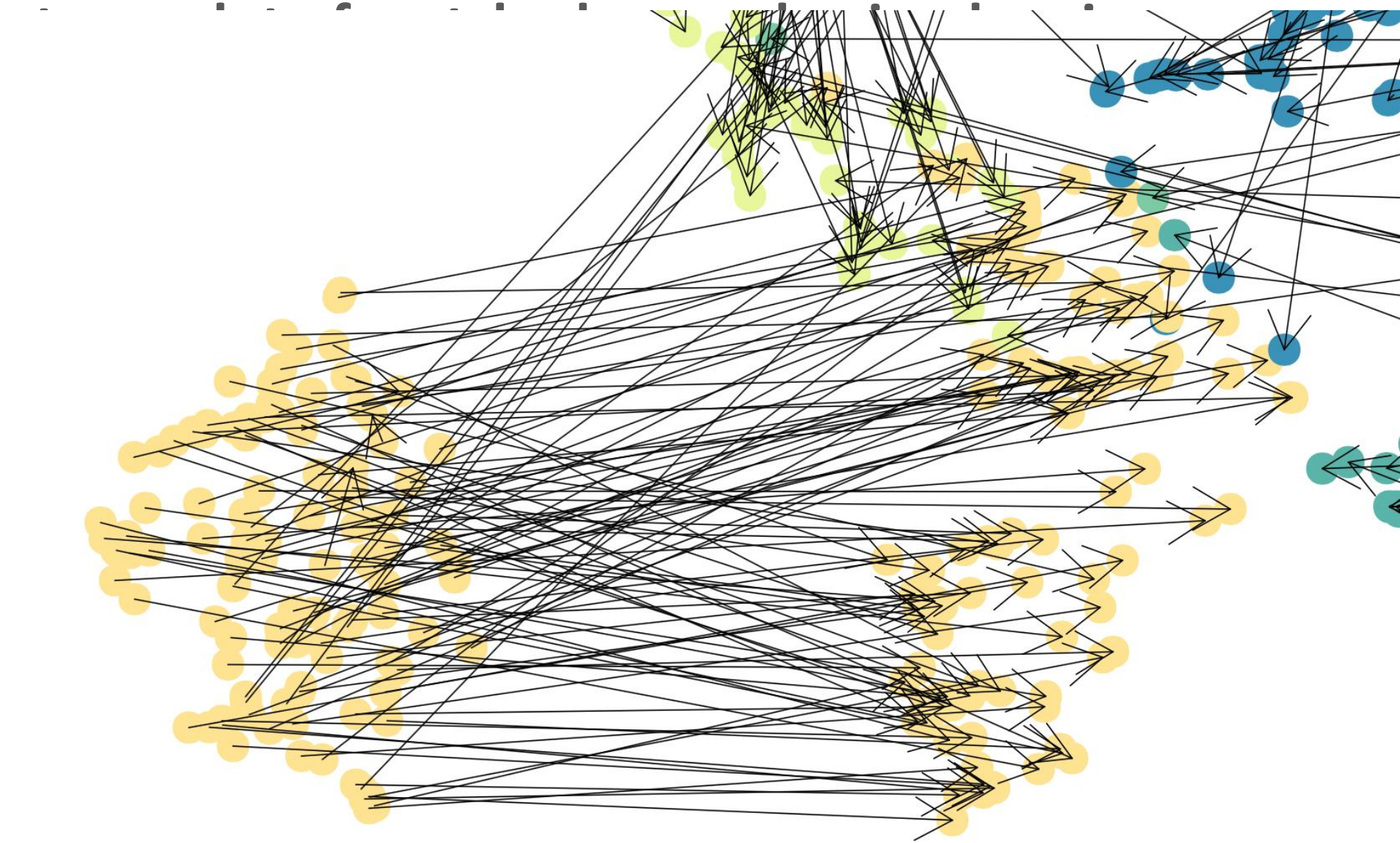




Colon-Rectal

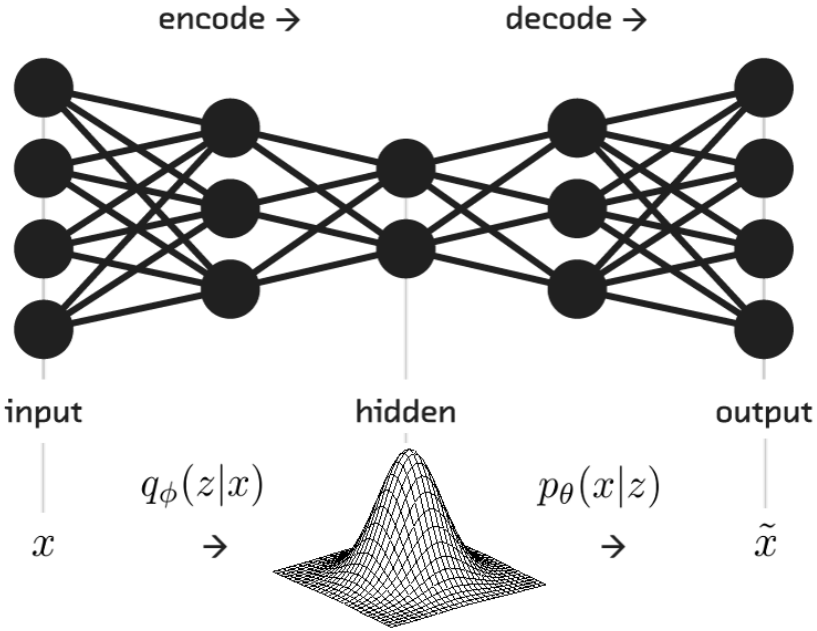
Uterus

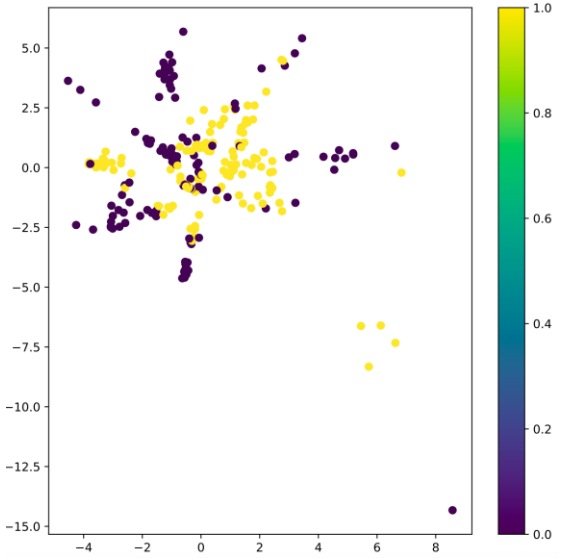


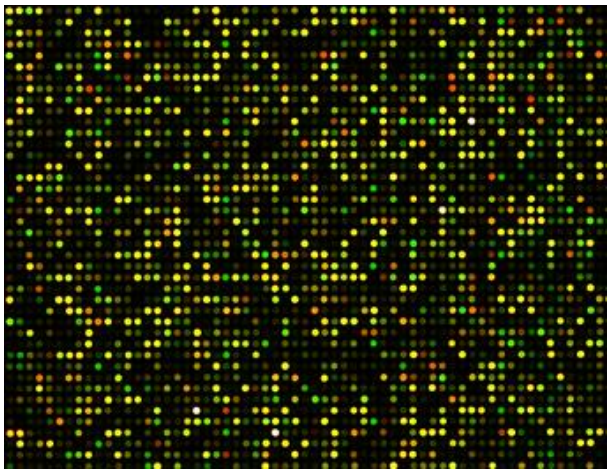


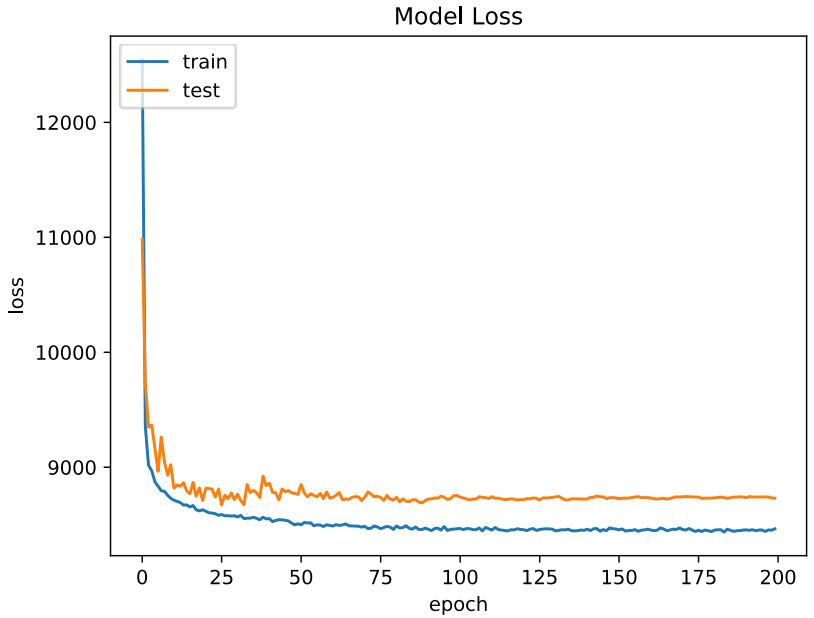
# Lung Cancer

# Variational Auto Encoder





# Hybrid Models in Cancer

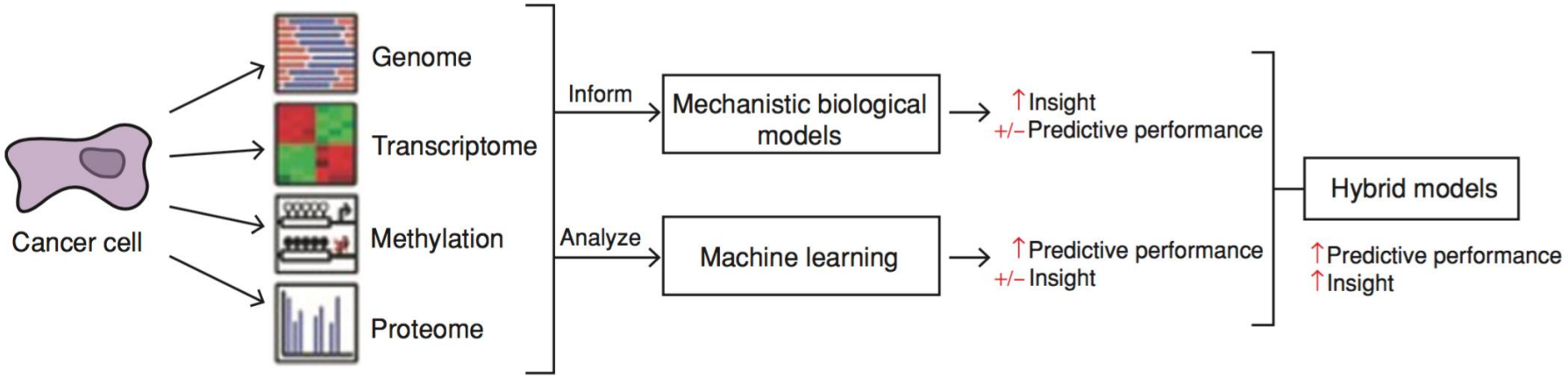



**Figure 1.** In two DREAM challenges, high throughput data characterizing cancer cells are used to build predictive models. Mechanistic models provide insight into the underlying biology, but do not take full advantage of the information within the data to achieve high performance. Machine learning methods are associative and extract maximum predictive value from the data, but do not always provide insight about mechanism. The future may bring hybrid models that combine the best of both approaches.

## Predicting Cancer Drug Response: Advancing the DREAM

Russ B. Altman

**Summary:** The DREAM challenge is a community effort to assess current capabilities in systems biology. Two recent challenges focus on cancer cell drug sensitivity and drug synergism, and highlight strengths and weaknesses of current approaches. *Cancer Discov; 5(3); 237–8. ©2015 AACR.*

# DOE and NIH Partnerships In Predictive Oncology

NIH NATIONAL CANCER INSTITUTE

U.S. DEPARTMENT OF ENERGY