Semi-Supervised Method for Countering Batch Effect

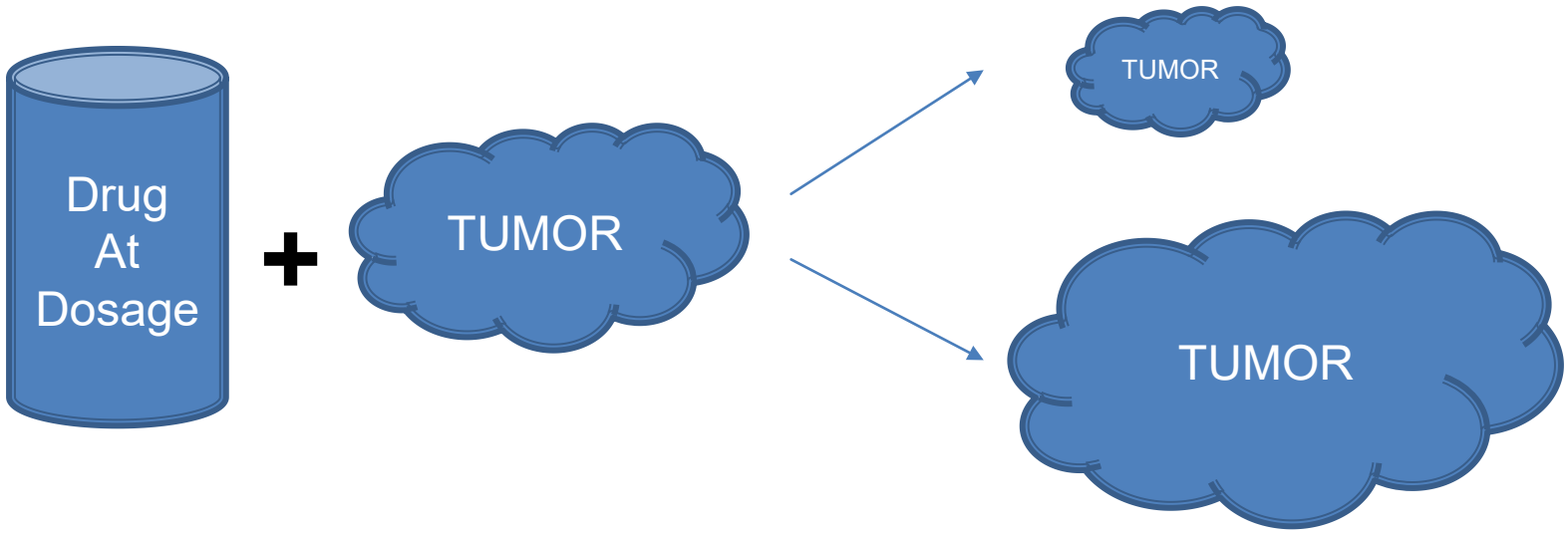
Stewart He, Jonathan Allen, Ya Ju Fan (LLNL)
Judith D Cohn (LANL)
Alex Partin, Fangfang Xia (ANL)

LLNL-PRES-796911
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Prediction Problem

- Given drug features, RNASeq features, and dosage predict tumor reaction.



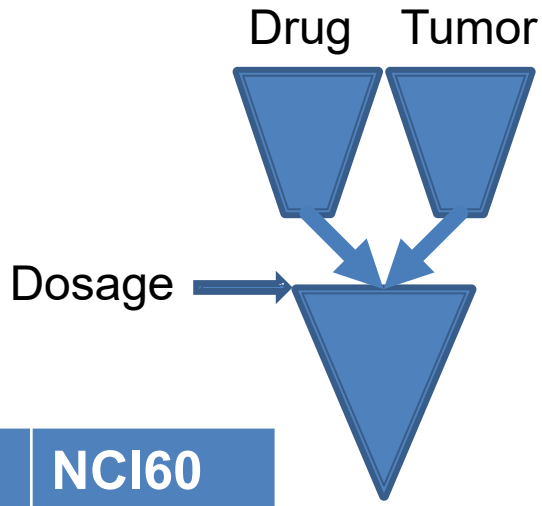
Data

- Over counts since CCLE and GDSC contain many of the same tumors under different IDs

Dataset	# drugs 3820 dims	# tumors 943 dims	Total datapoints
CCLE	24	474	84,444
CTRP	495	812	4,878,603
gCSI	16	357	49,003
GDSC	239	672	1,100,728
NCI60	1006	59	1,097,284

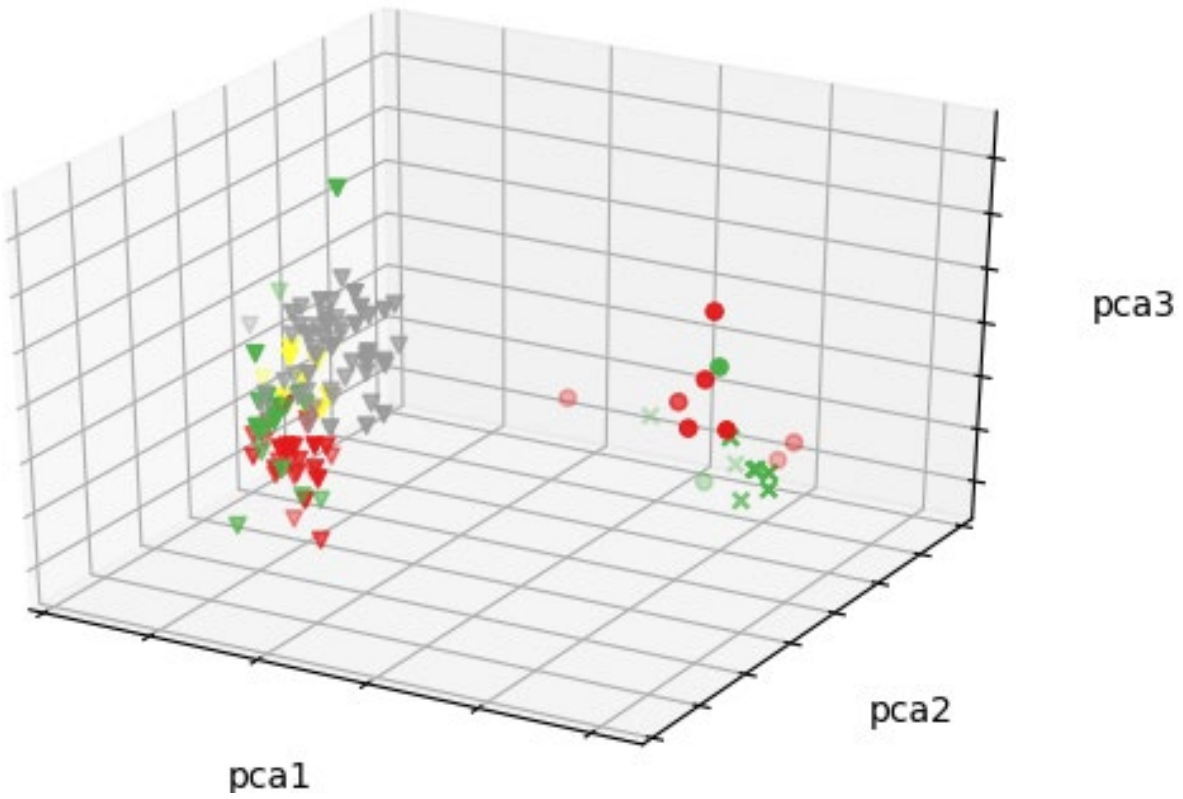
Regression results

- Used forked neural network to do regression
- Perform inter-dataset test.
  - Example: Train/validate on CCLE test on CTRP
- Poor results using R^2



Training \ Testing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	0.912	-3.92	-1.28	-3.91	-3.65
CTRP	0.554	0.896	0.454	-.0462	0.06
gCSI	-1.72	-3.40	0.964	-3.54	-2.97
GDSC	-.576	-1.72	0.206	0.879	-1.64
NCI60	0.443	0.226	0.376	0.06	0.894

PCA across studies

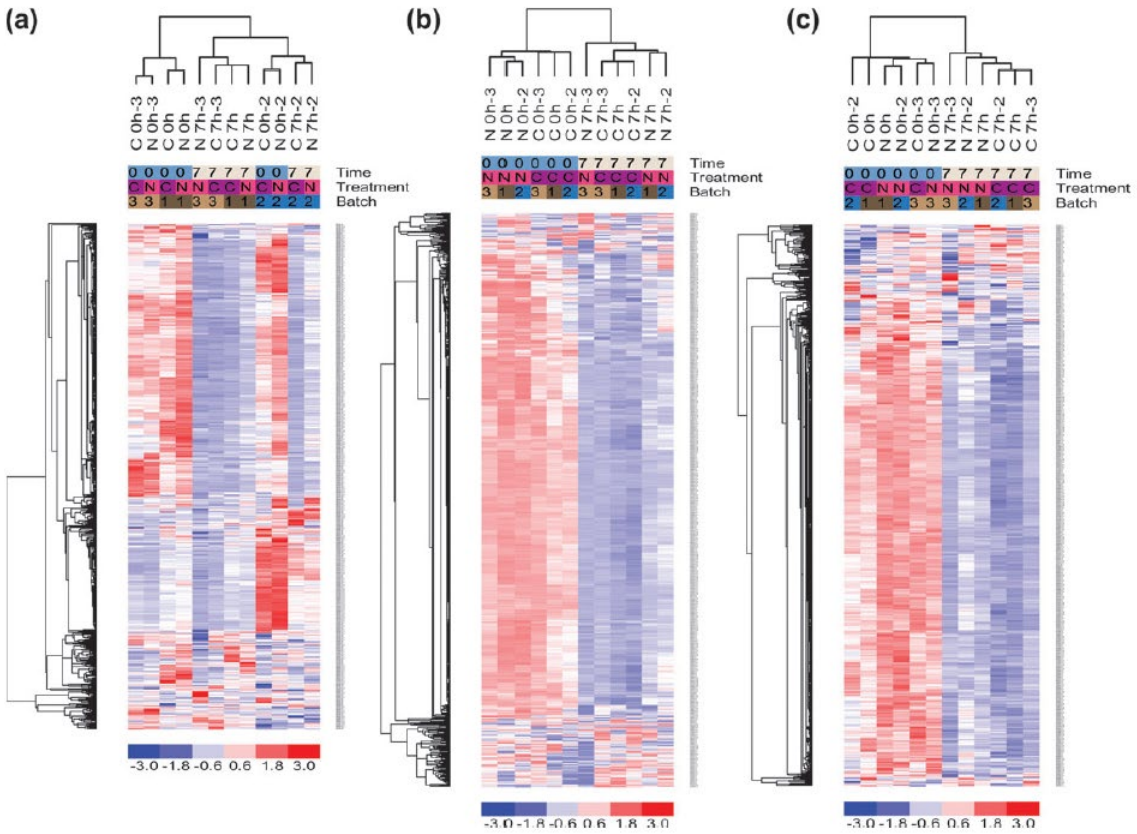


Other study
NCIPDM
GDC

Skin tumor
Pleural tumor
Ovarian tumor
Pancreatic tumor

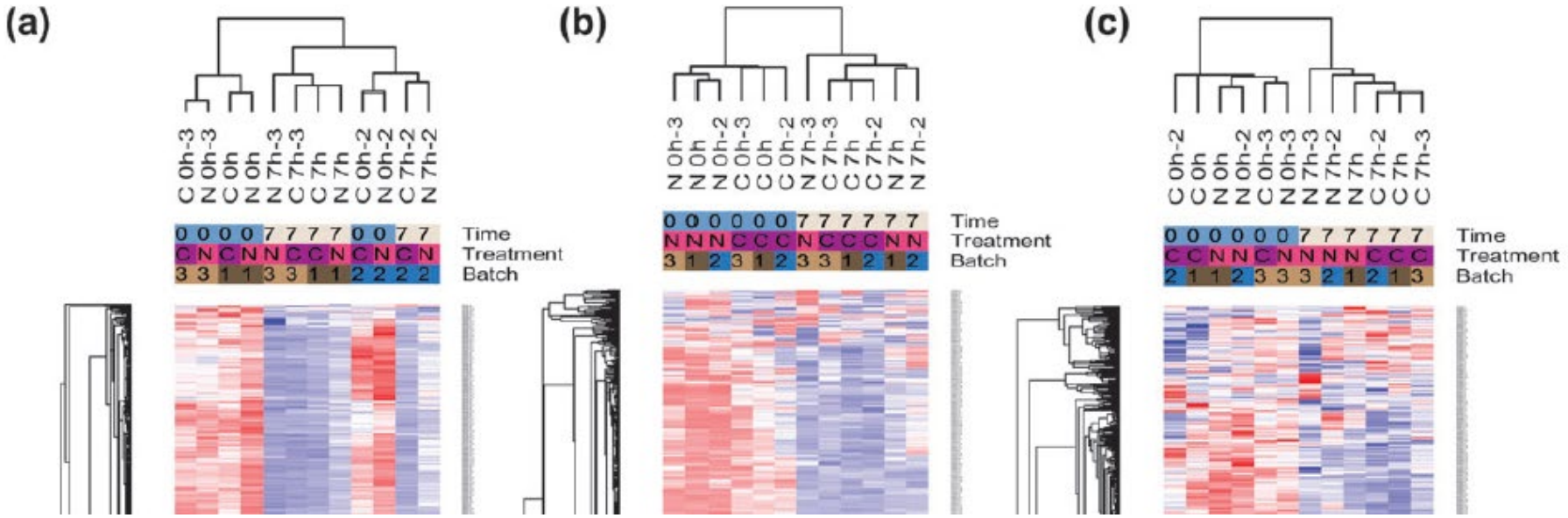
	pca1	pca2	pca3
Explained Variance	0.33	0.12	0.05

ComBat



Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics* **8**(1):118-127.

ComBat



Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics* **8**(1):118-127.

Results: ComBat

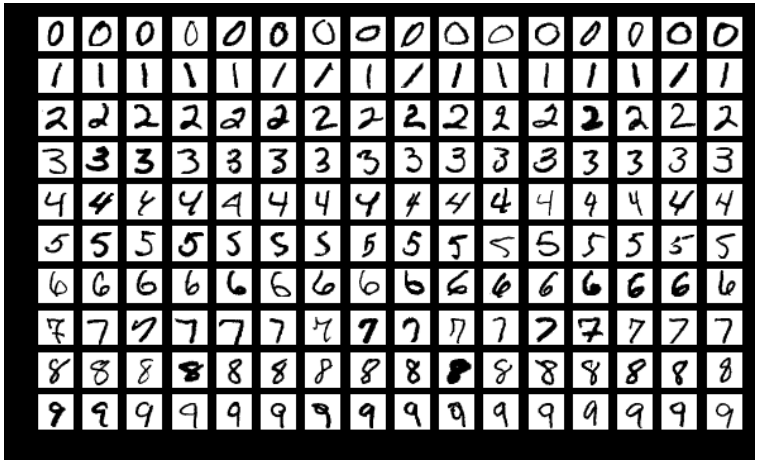
Training \ Testing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	0.912	-3.92	-1.28	-3.91	-3.65
CTRP	0.554	0.896	0.454	-0.046	0.06
gCSI	-1.72	-3.40	0.964	-3.54	-2.97
GDSC	-0.576	-1.72	0.206	0.879	-1.64
NCI60	0.443	0.226	0.376	0.06	0.894

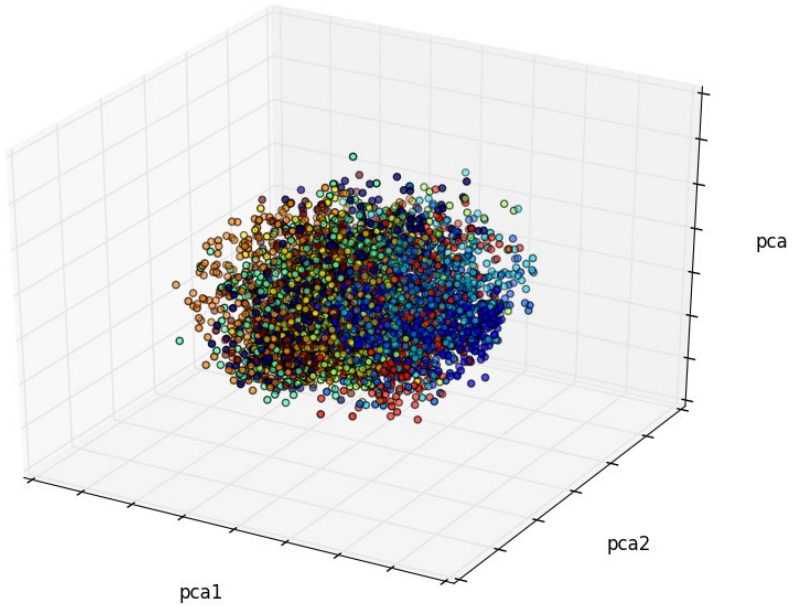
Learn better features

- Very few examples of tumors
- RNASeq originally has 17k features
  - 943 landmark RNASeq genes are hand engineered
- We want better features:
  - More generalized across different types of tumors
  - Learned from unlabeled data
  - Good for regression
  - Use Autoencoders?

The problem with Autoencoders

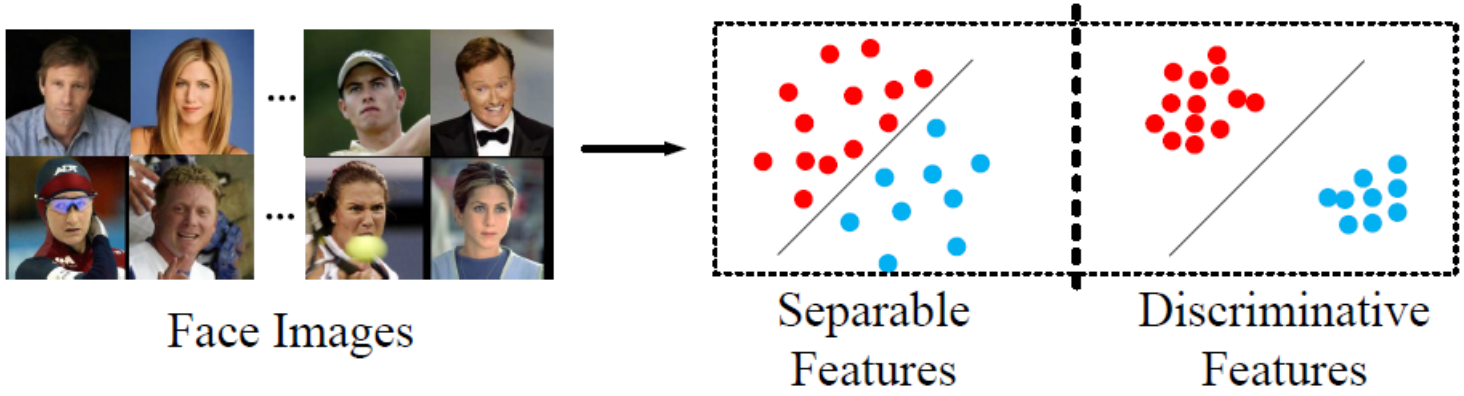
- The latent space is good for reconstruction.
  - That's all the cost function cares about
  - If you're lucky they might be good for other things
- MNIST trained auto encoder
  - First 3 principal components
  - Colored by class





Modify with Center Loss

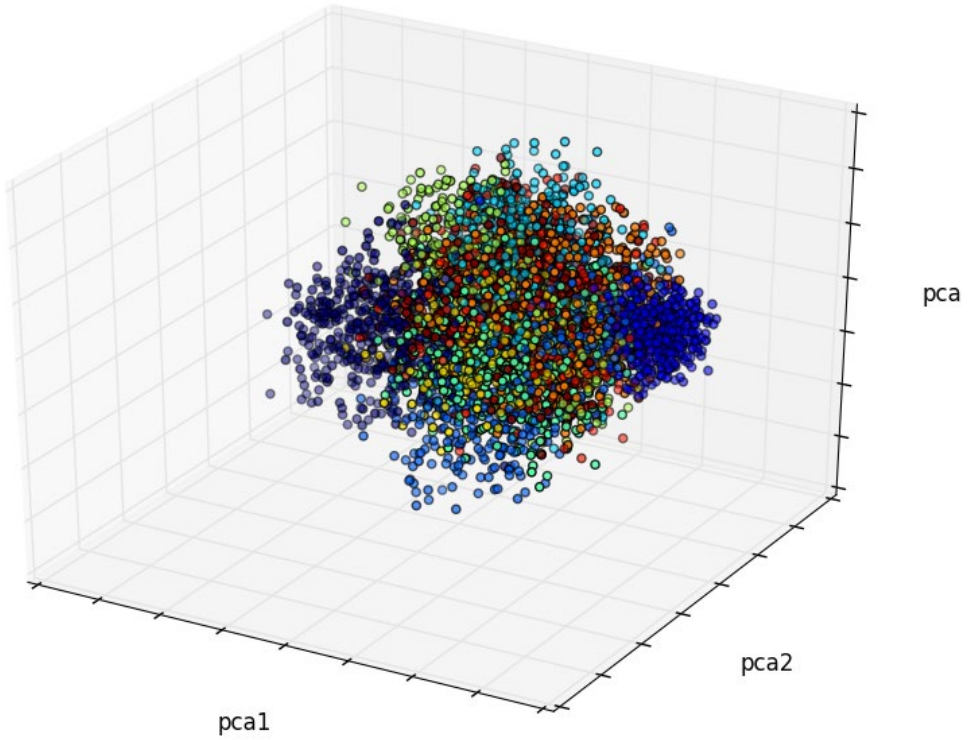
- Center Loss designed to be used for classification
  - *A Discriminative Feature Learning Approach for Deep Face Recognition* Wen et al., 2016



$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

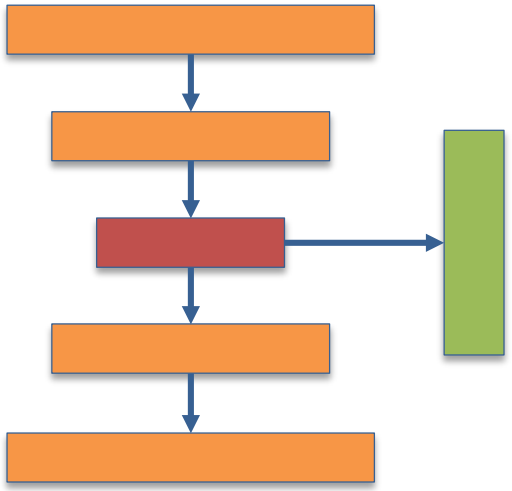
Center Loss + Autoencoders

- Does not play well. MNIST example:
  - Easily falls for trivial solution



Classification + Center Loss + Autoencoder

- Autoencoder Bottleneck Classifier
- Must balance 3 terms in loss function
  1. Reconstruction
  2. Classification
  3. Center loss
- Varied the bottleneck between 943 and 20



Loss comparison between 943 and 20

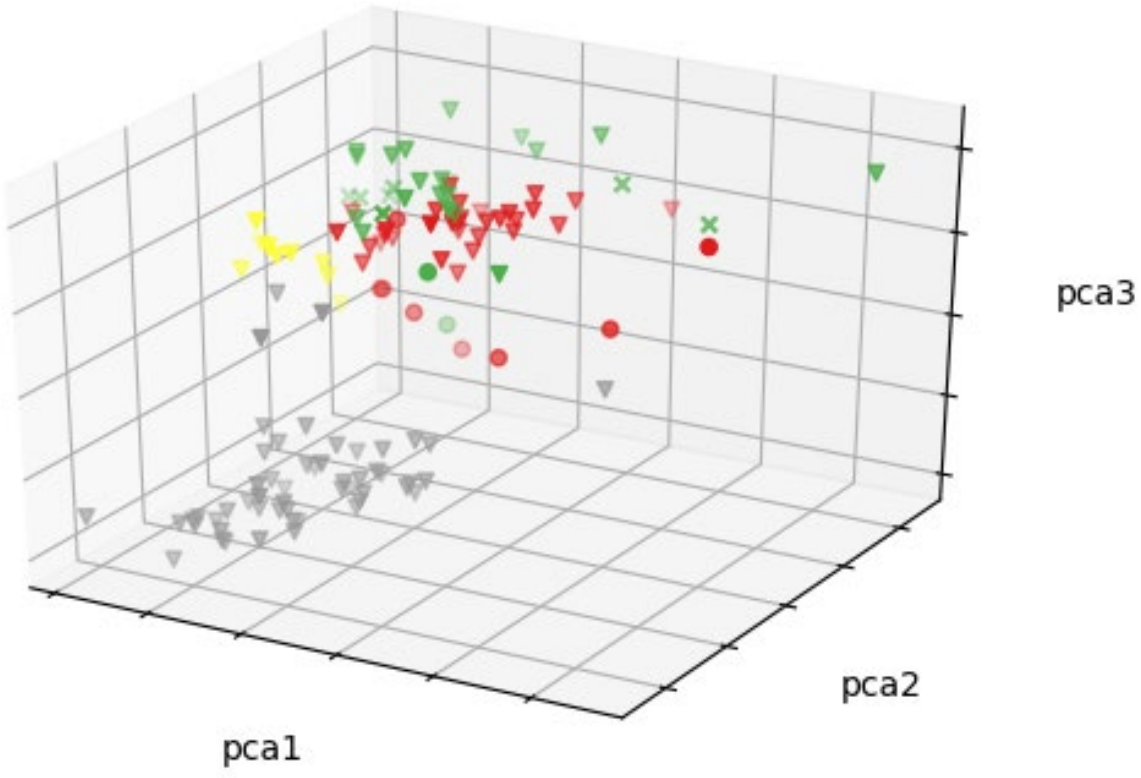
Feature	Reconstruction Error (RMSE)	F1 score
RNASeq encoded 943	0.716	0.803
RNASeq encoded 20	0.761	0.788
ComBat RNASeq encoded 943	0.714	0.796
ComBat RNASeq encoded 20	0.762	0.788

Cluster metrics

Silhouette: between [-1, 1] where 1 is the best score
Calinski-Harabasz: the higher the better
Davis-Bouldin: 0 is the best possible score

Dataset	Silhouette	Calinksi-Harabasz	Davis-Bouldin
RNASeq	-0.035	16.751	2.558
lincs1000	-0.09	15.012	2.772
ComBat	0.03	20.965	2.51
Encoded RNASeq 943	0.115	53.399	1.881
Encoded RNASeq 20	0.147	110.131	1.689
Encoded ComBat 943	0.105	60.84	1.806
Encoded ComBat 20	**0.159**	**113.451**	**1.662**

PCA across studies : AFTER

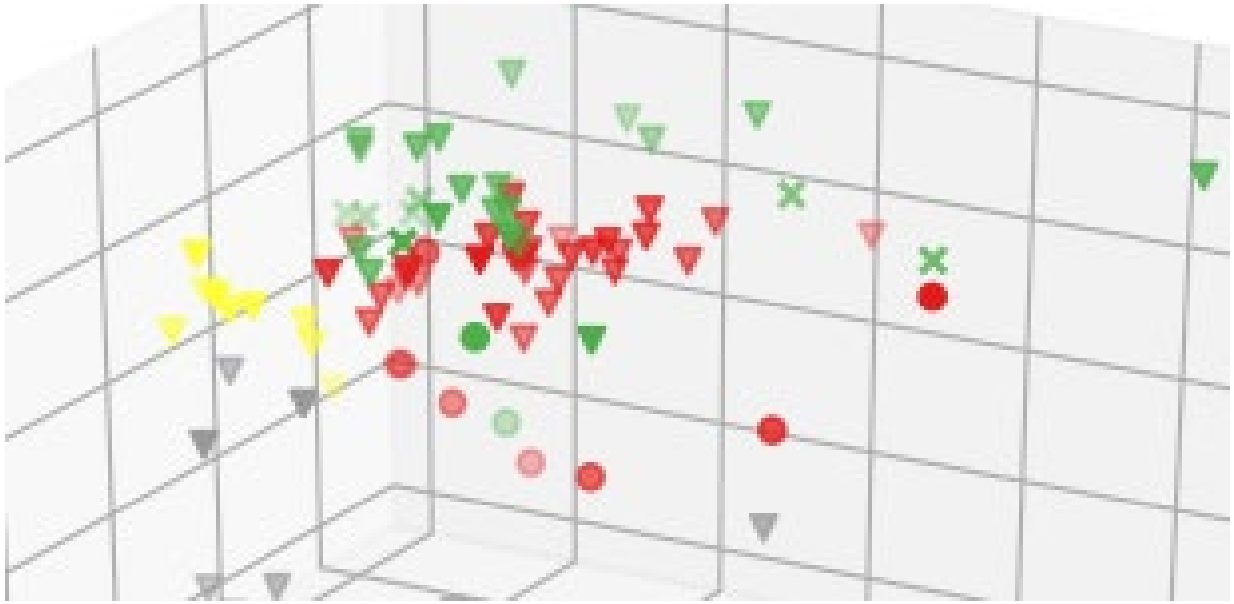


Other study
NCIPDM
GDC

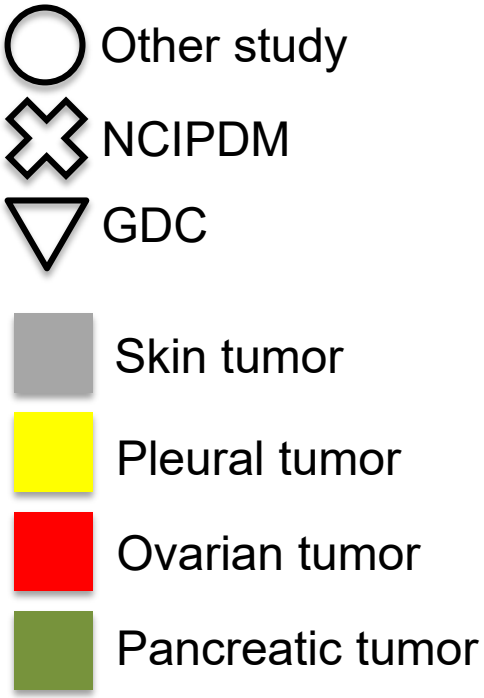
Skin tumor
Pleural tumor
Ovarian tumor
Pancreatic tumor

	pca1	pca2	pca3
Explained Variance	0.15	0.11	0.09

PCA across studies : AFTER



All 3 shapes are now closer together!



	pca1	pca2	pca3
Explained Variance	**0.15**	**0.11**	**0.09**

Results: Encoded RNASeq 943

Training \ Testing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	0.358	0.053	0.304	0.049	0.06
CTRP	0.575	0.681	0.561	0.179	0.366
gCSI	0.048	-0.017	0.206	-0.101	-0.323
GDSC	0.447	-0.049	0.513	0.563	0.13
NCI60	0.435	0.37	0.369	0.23	0.832

Results: Encoded ComBat 943

Training \ Testing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	0.391	-0.025	0.359	0.03	-0.194
CTRP	0.586	0.678	0.566	0.175	0.298
gCSI	0.112	-0.016	0.331	-0.088	-0.489
GDSC	0.419	0.011	0.475	0.549	0.13
NCI60	0.418	0.346	0.388	0.234	0.838

Results: Encoded RNASeq 20

Training \ Testing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	0.727	-0.123	0.359	-0.037	0.044
CTRP	0.599	0.828	0.558	0.132	0.308
gCSI	0.152	0.069	0.37	-0.007	-0.309
GDSC	0.42	-0.104	0.378	0.787	-0.219
NCI60	0.387	0.312	0.405	0.171	0.893

Results: Encoded ComBat 20

Training \ Testing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	0.719	-0.128	0.404	-0.042	-0.051
CTRP	0.574	0.841	0.545	0.082	0.216
gCSI	0.17	0.049	0.376	-0.034	-0.451
GDSC	0.4	-0.072	0.383	0.793	-0.208
NCI60	0.403	0.333	0.377	0.19	0.891

Extension into PDX data

- Possible use case to train on cell line data and apply to PDX data.
  - Spearman rank correlation to relate PDX doubling times and AUC dosage/response rate curve

	ComBat	Encoded RNASeq 943
CCLE	-0.07	-0.07
CTRP	-0.24	-0.304
gCSI	-0.23	-0.261
GDSC	0	-0.033
NCI60	-0.075	-0.077

Acknowledgements

- Maulik Shukla - data organization/access
- Ben Mcmahon - helped Judith with clustering
- Rick Stevens - project PI
- Austin Clyde – performing regression tests
- Jason David Gans – doubling time predictions on PDX dataset


Lawrence Livermore
National Laboratory

Learned features applied to regression

- Modest improvement to regression results

Testing-> Trianing	CCLE	CTRP	gCSI	GDSC	NCI60
CCLE	.913	-1.394	-.061	-1.477	-.640
CTRP	.611	.817	.501	.136	.186
gCSI	.021	-.357	.962	-.201	-.906
GDSC	.380	-.445	.364	.861	-.238
NCI60	.450	.270	.351	.160	.894

Supplemental full image