A New Approach to Batch Effect Removal Based on Distribution Matching in Latent Space

Huaqing Li, Haluk Dogan, and Juan Cui*

Systems Biology and Biomedical Informatics Laboratory
https://sbbi.unl.edu/

Department of Computer Science and Engineering
University of Nebraska-Lincoln

November 20, 2019

Introduction

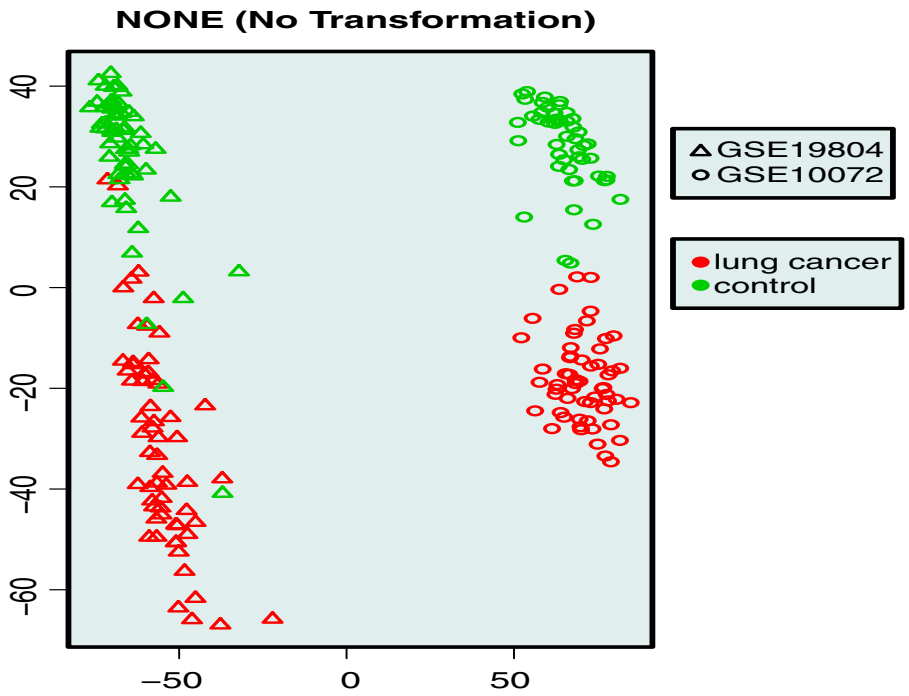
Batch effect (BE) exists when:

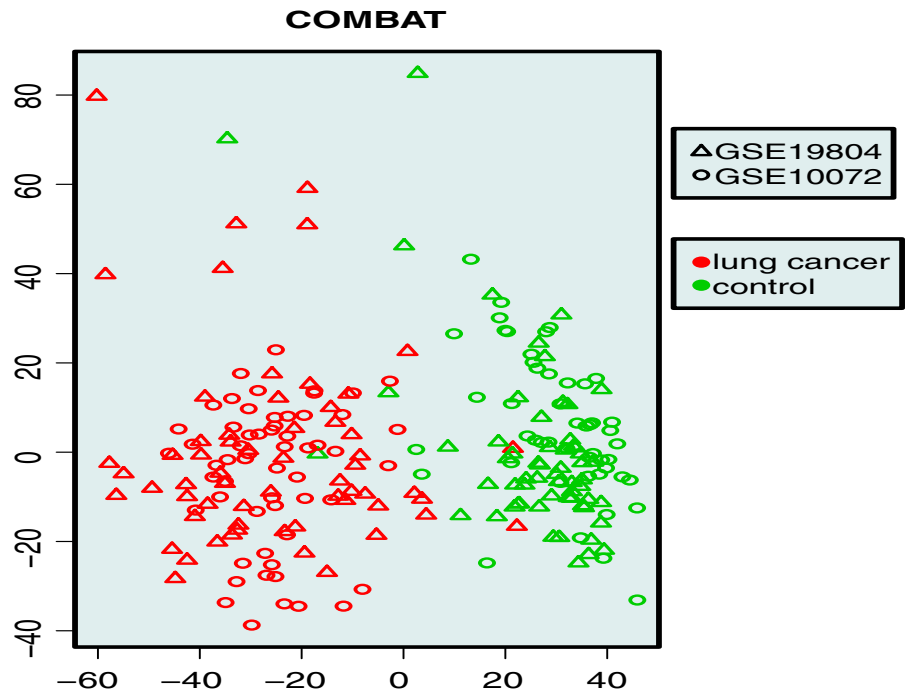
- Measurements from multiple subjects
  - different patients
- Various experimental conditions
  - treatments
- Data augmentation by combining dataset from various sources

BE can be caused by various factors:

- Instrument variation
- Machine calibration
- Human handling

Introduction (cont'd)





Taminau, Meganck, Lazar, Steenhoff, Coletta, Molter, Duque, Schaetzen, Solís, Bersini, and Nowé 2012

- Systematic errors
- Forming distinct groups
- Larger than biological variation

- Remove unwanted between-batch variations
- Preserve in-batch biological variability

Introduction (cont'd)

Existing methods based on statistical modeling:

- BMC: **B**atch **M**ean **C**entered (Sims, Smethurst, Hey, Okoniewski, Pepper, Howell, Miller, and Clarke 2008)
- ComBat: Location/Scale modeling with parametric/non-parametric empirical Bayes (Johnson, C. Li, and Rabinovic 2006)

Pros:

- Simple but easy to interpret

Cons:

- Strong assumptions and constraints
  - Normal distribution
  - Similar priors
- Lose biological signal

Challenges

- Difficulty in identifying real BE from measuring signal
- Complex data
  - Non-linear
  - Non-uniform
- Unreported cause/generating factors
  - Technical error
- Association or correlation in various factors
  - Disease
  - Age

Outline

1 Introduction

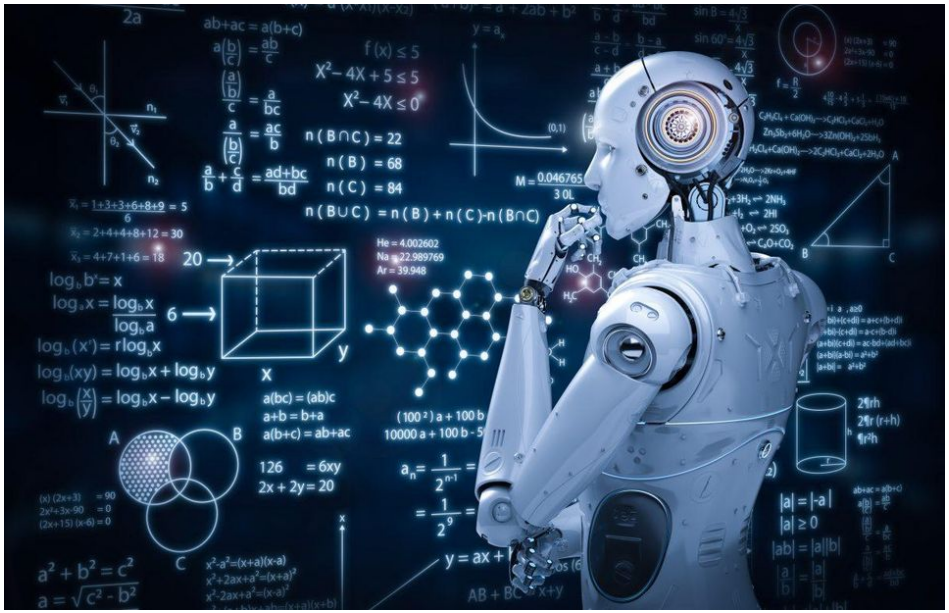
2 Methodology

3 Results

4 Conclusion

Modeling Batch Effect Using Machine Learning

- Data-driven
- Powerful in modeling complex data without making strong assumption
- Able to learn meaningful features and denoising data



Existing Methods Using Machine Learning

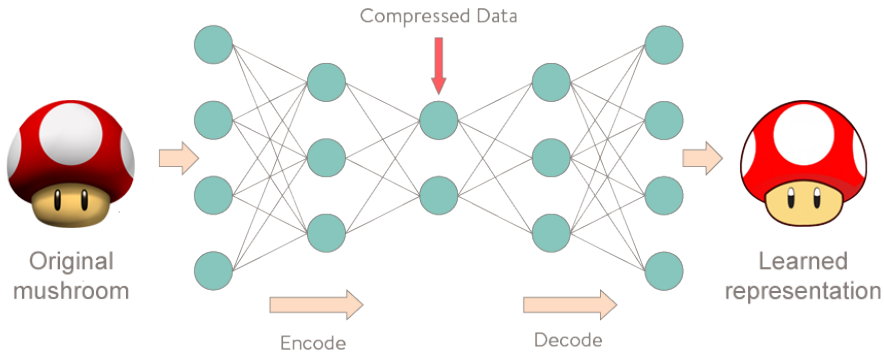
BE removal using autoencoder by Amodio et al. 2018

Advantages:

- Ability to model batch effect in latent space
- Discriminate the factors that are associated with batch effect

Disadvantages:

- Statistical alignment (percentile) of neurons distributions one by one
- Strong assumption: batch effect in each feature is uncorrelated



https://www.curiousily.com/posts/data-imputation-using-autoencoders/

1. Train an autoencoder to extract hidden layer
2. Identify batch effect related neurons
3. Edit the neurons to correct batch effect

Existing Methods Using Machine Learning (cont'd)

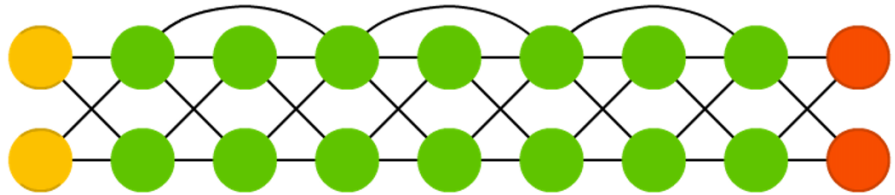
BE removal using residual network by Shaham et al. 2017

Advantages:

- Non-parametric
- More precise alignment

Disadvantages:

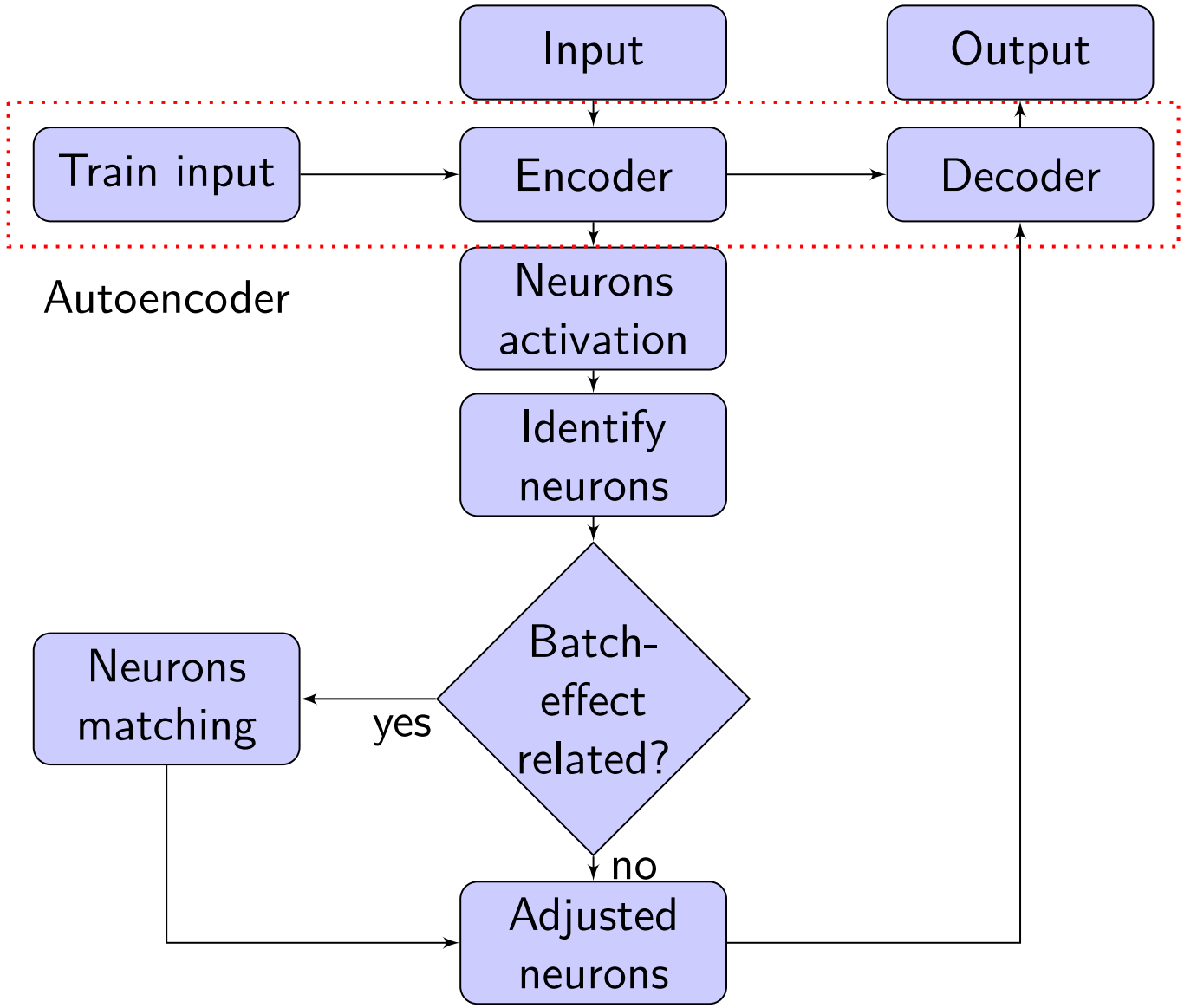
- Potential loss of biological signal by mapping to the entire feature space
- Computationally intensive



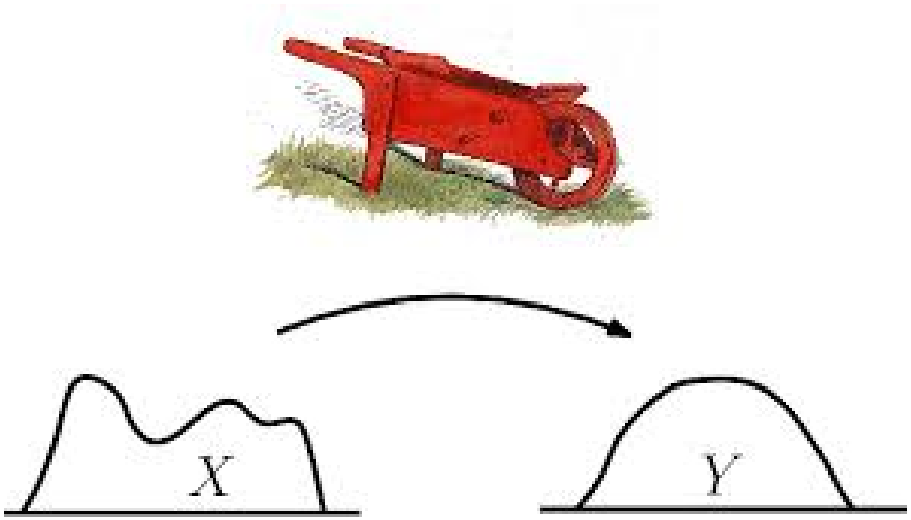
https://www.asimovinstitute.org/neural-network-zoo/

1. Directly train the network to learn a data distribution
2. Samples from different populations have a similar likelihood

Our Proposed Approach

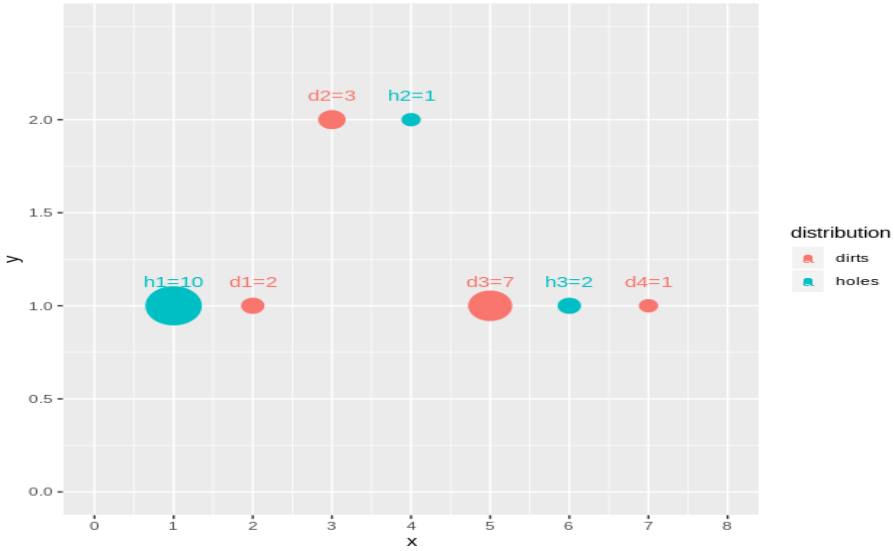


Identify Batch Related Neurons



Earth Mover's Distance (EMD):

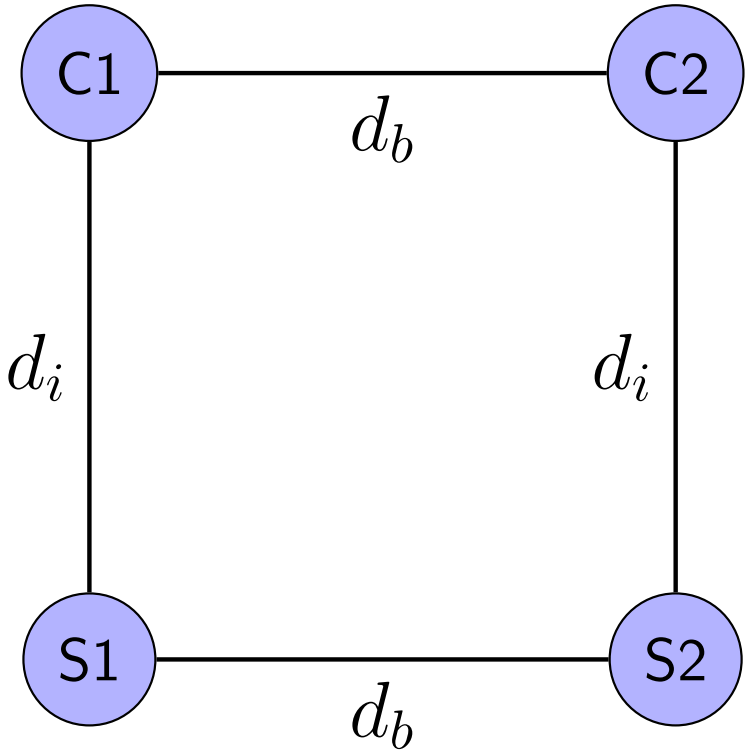
$$\text{EMD}(X,Y) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}$$



from	to	flow	dist	work
d1	h1	2	1.00	2.00
d2	h1	3	2.24	6.71
d3	h1	5	4.00	20.00
d3	h2	1	1.41	1.41
d3	h3	1	1.00	1.00
d4	h3	1	1.00	1.00
		–		–
		13		32.12

EMD = 32.12 / 13 = 2.47

Identify Batch Related Neurons (cont'd)



- d_b: between batch EMD distance
- d_i: in-batch EMD distance

If any $\frac{d_i}{\min(d_b)} < 1$:

BE has significant impact in that neuron

Deep Learning Architectures

Autoencoder:

- Layers: 54675, 512, 256, 128
- Activation: ReLU

Hyperparameter	Value
Batch size	64
Learning rate	0.001
Epoch	20
Regularizer	ℓ_2
Optimizer	Adam
Loss	MSE

ResNet:

- 3 Blocks
- Activation: ReLU

Hyperparameter	Value
Batch size	32
Initial learning rate	0.001
Early stopping epoch	50
Regularizer	ℓ_2
Optimizer	RMSPROP
Loss	MMD

Outline

1 Introduction

2 Methodology

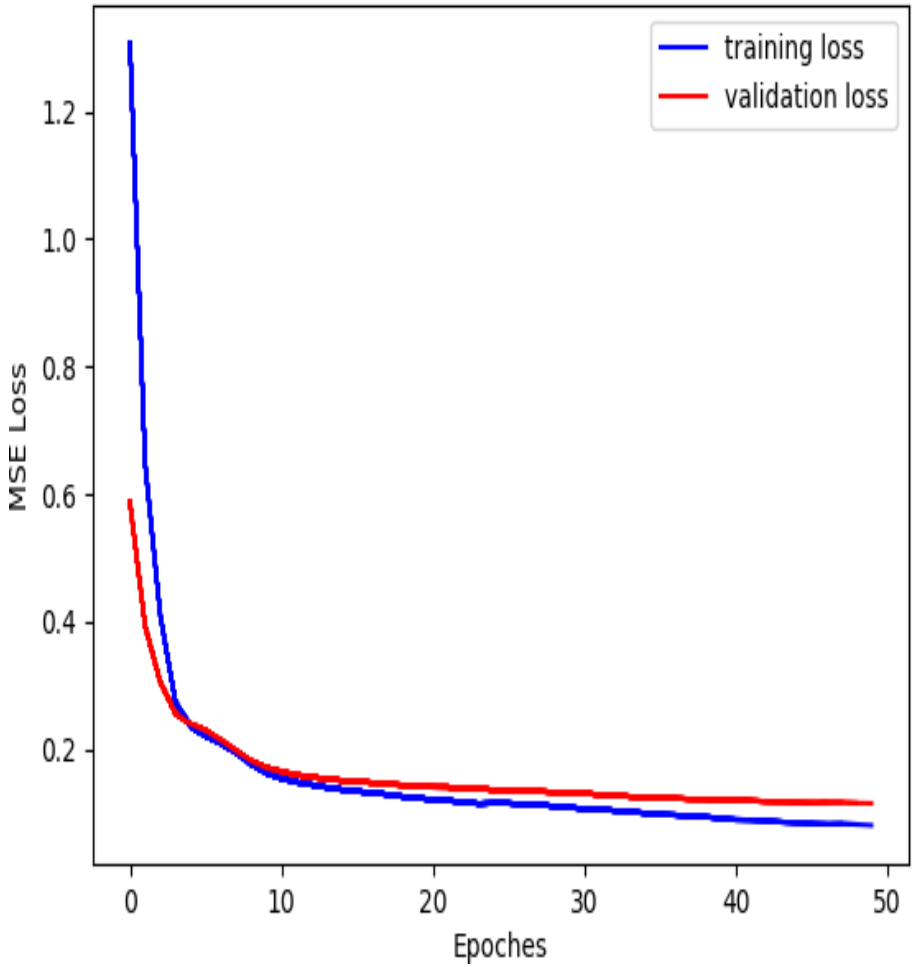
3 Results

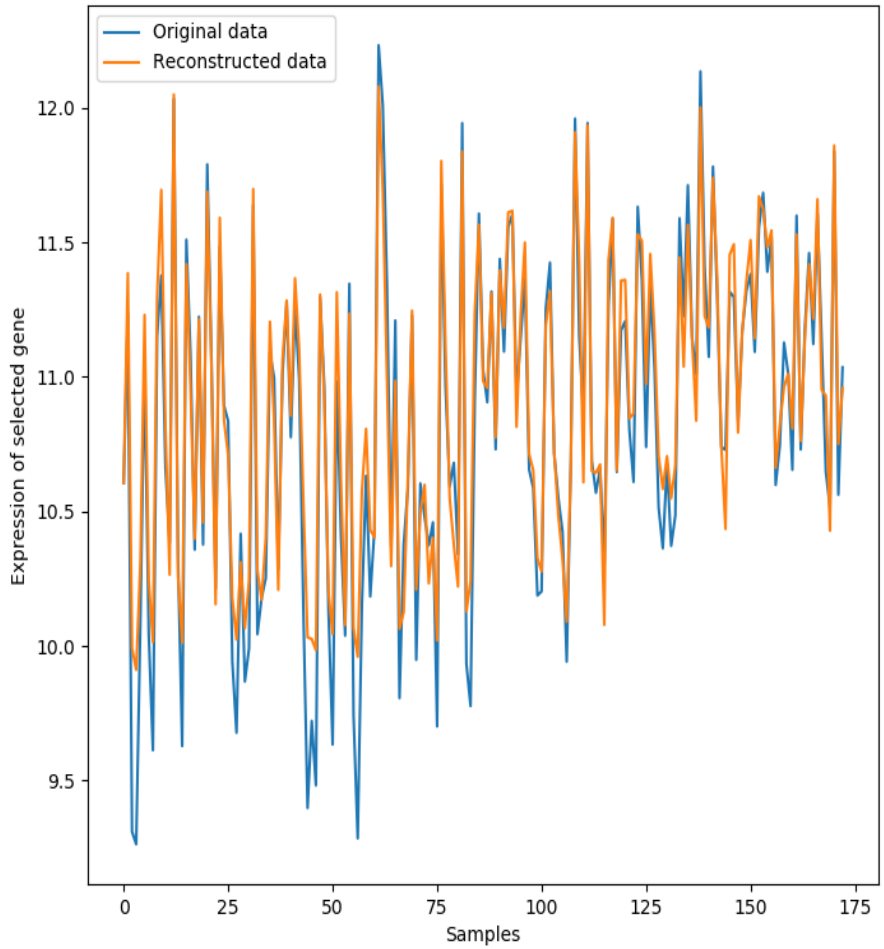
4 Conclusion

Dataset

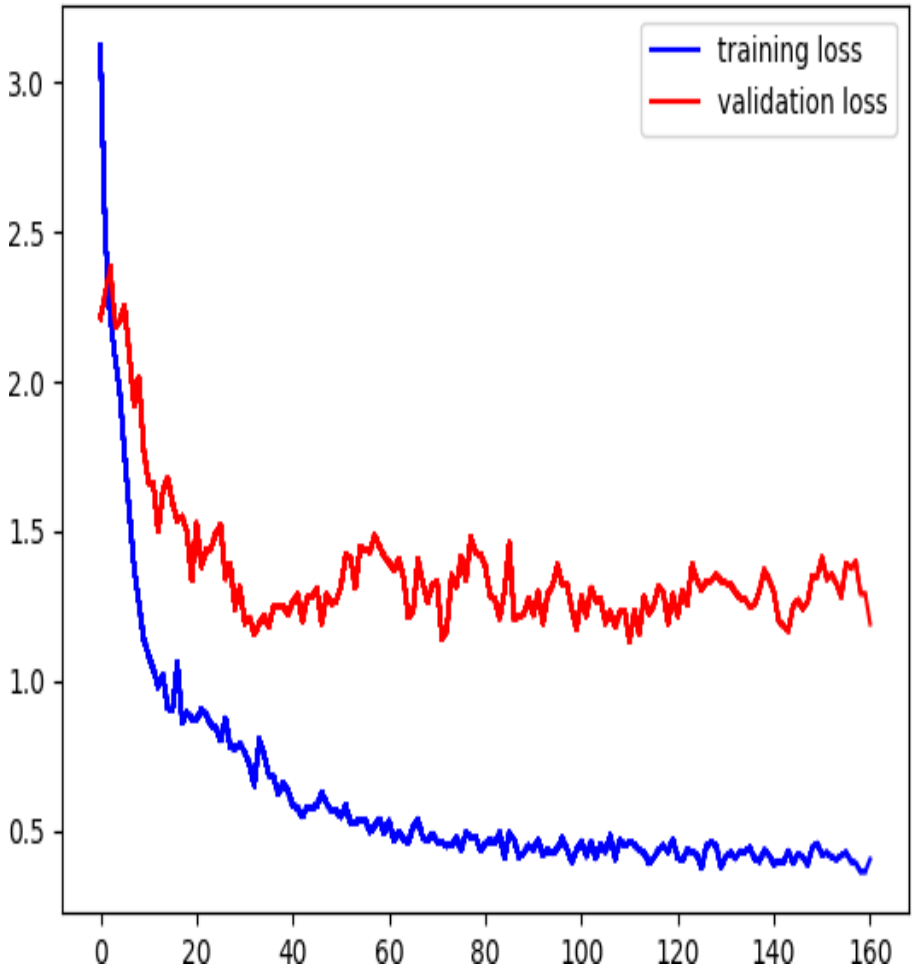
Dataset	Batch Number	Normal	Alzheimer	NO/AD	Total
GSE48350	#1	64	189	0.34	253
GSE5281	#2	74	87	0.85	161
Total	2	148	276	0.54	414

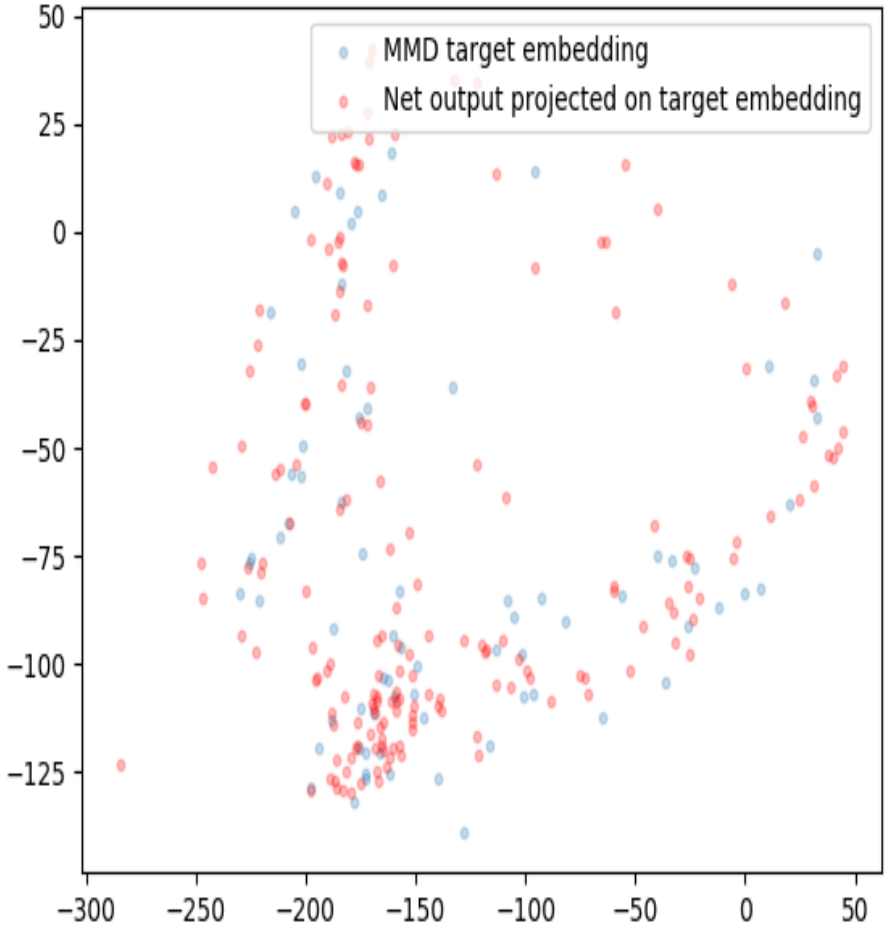
Training Autoencoder



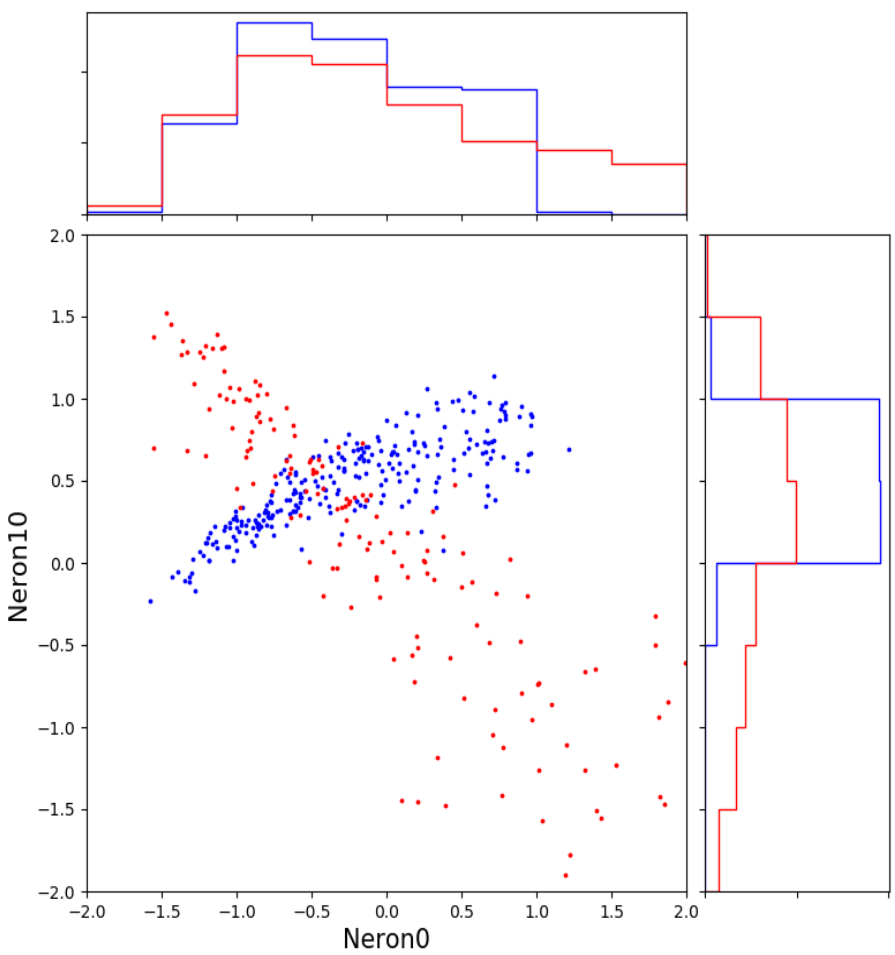


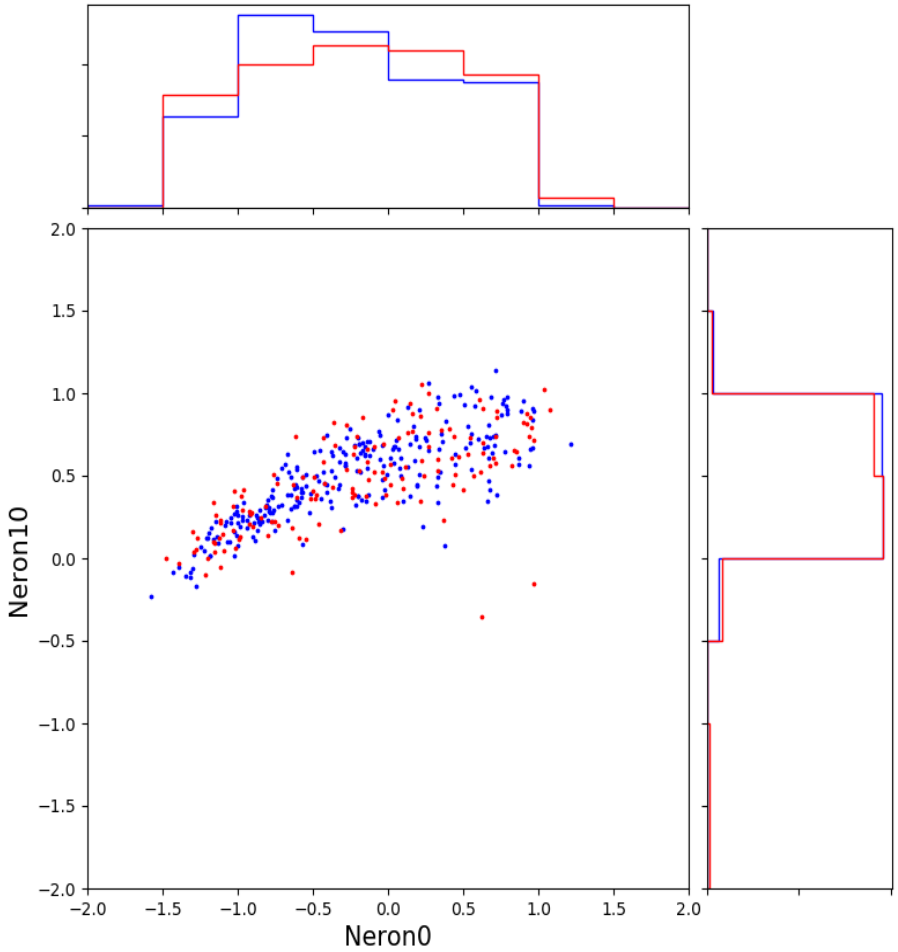
Training ResNet



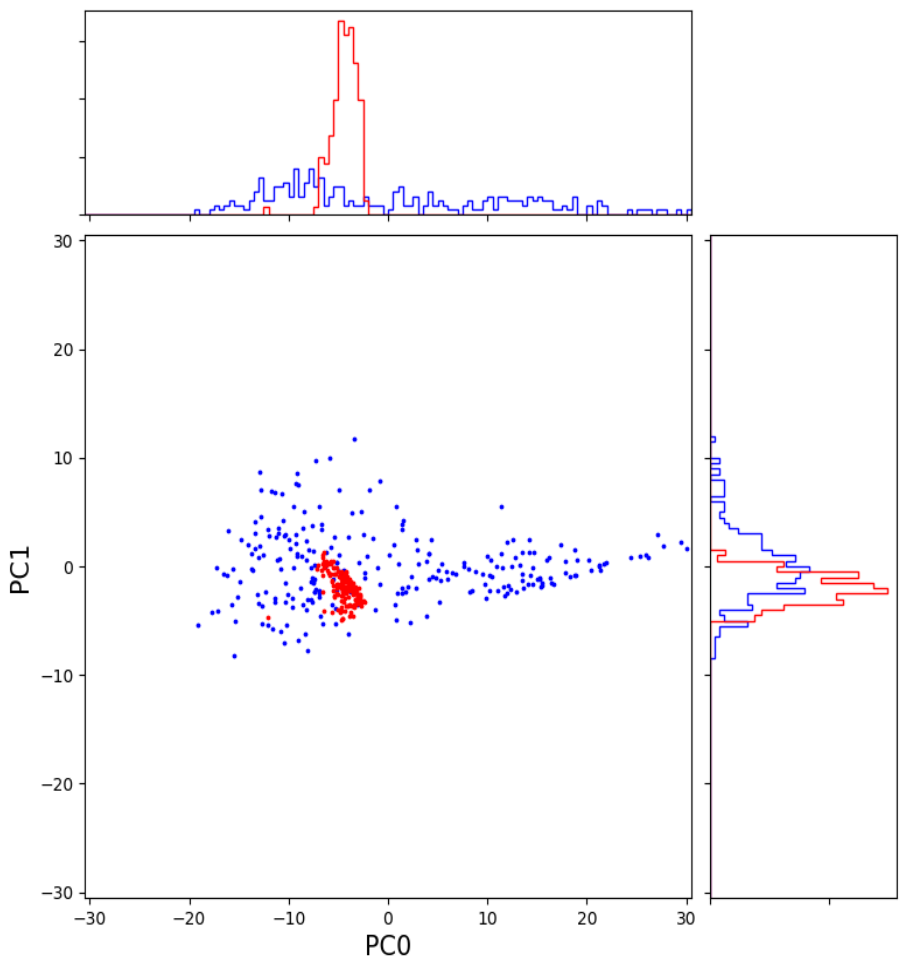


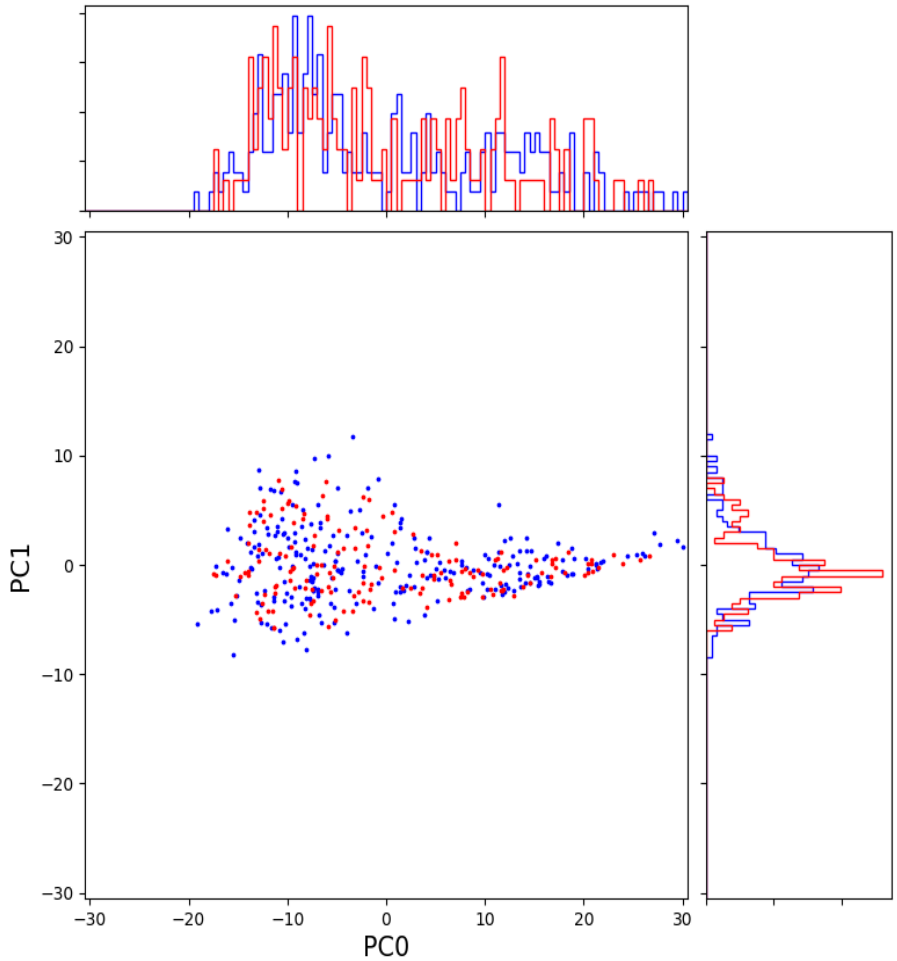
Neurons Adjustment



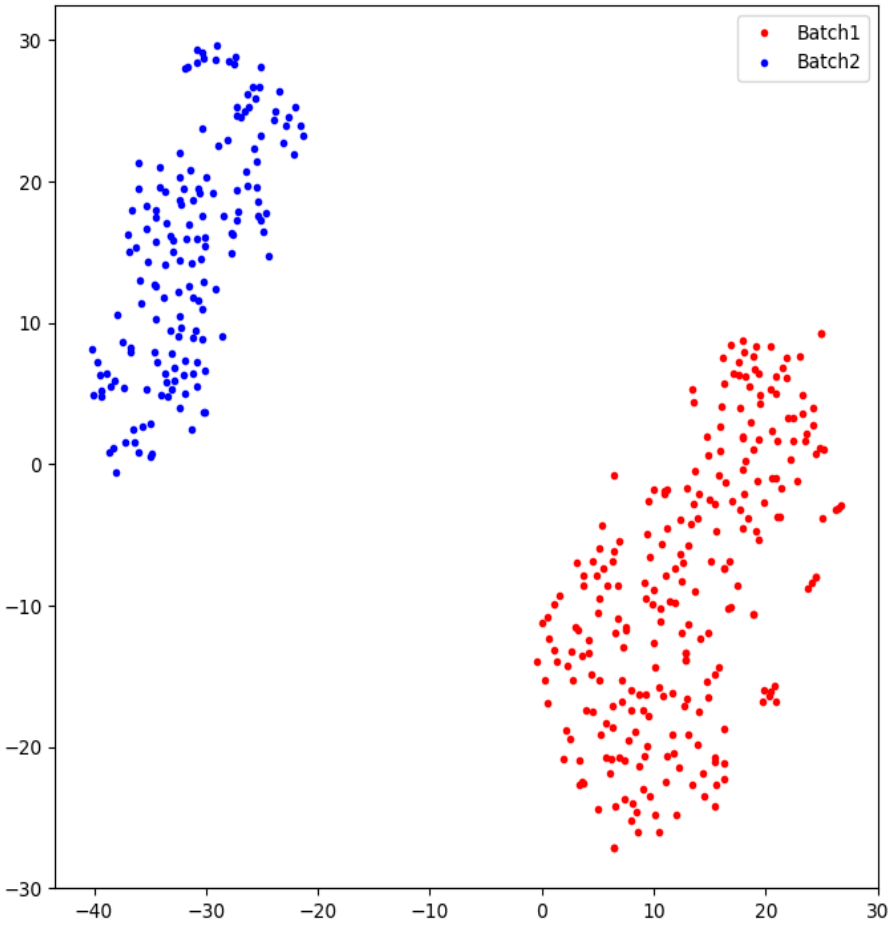


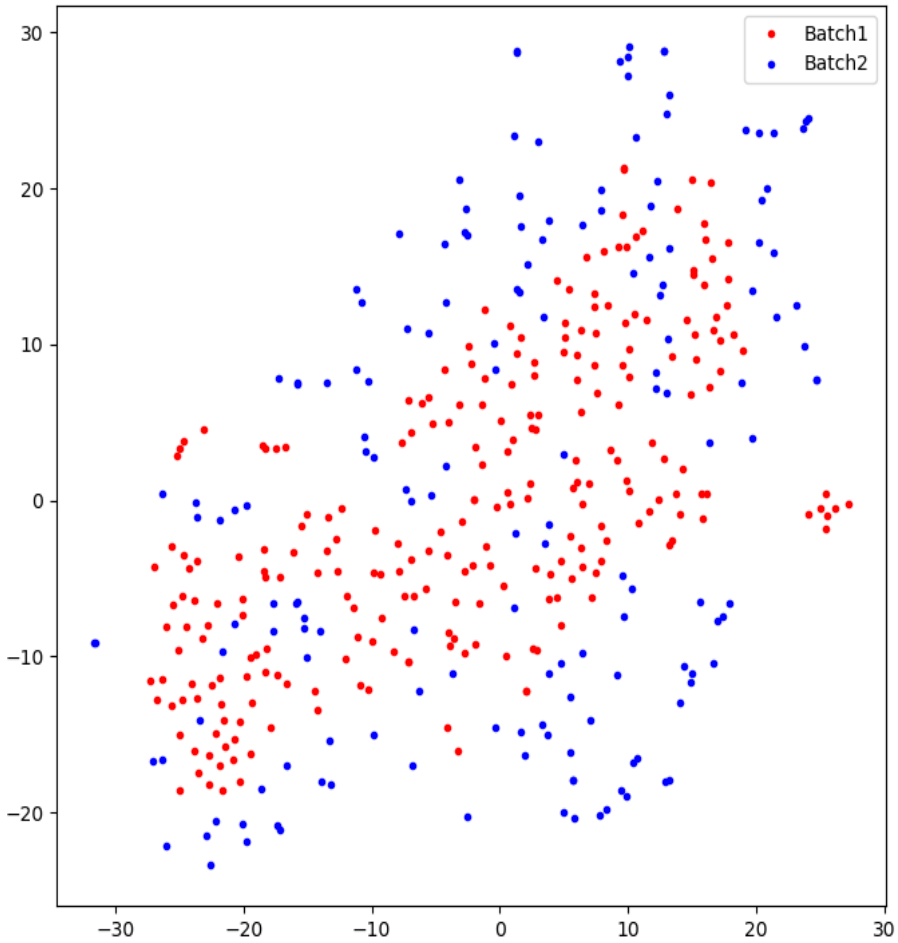
Neurons Adjustment (cont'd)



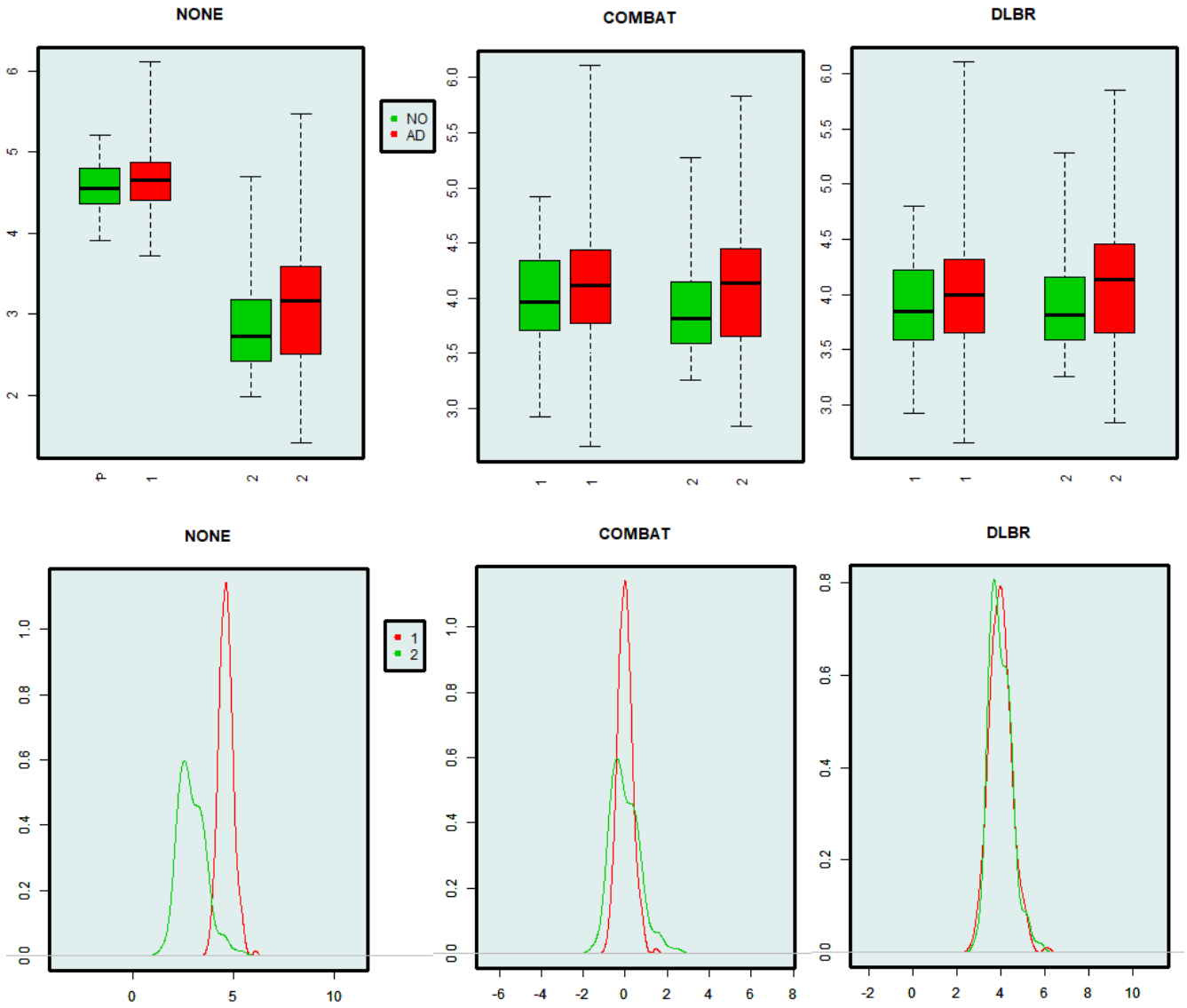


Neurons Adjustment (cont'd)

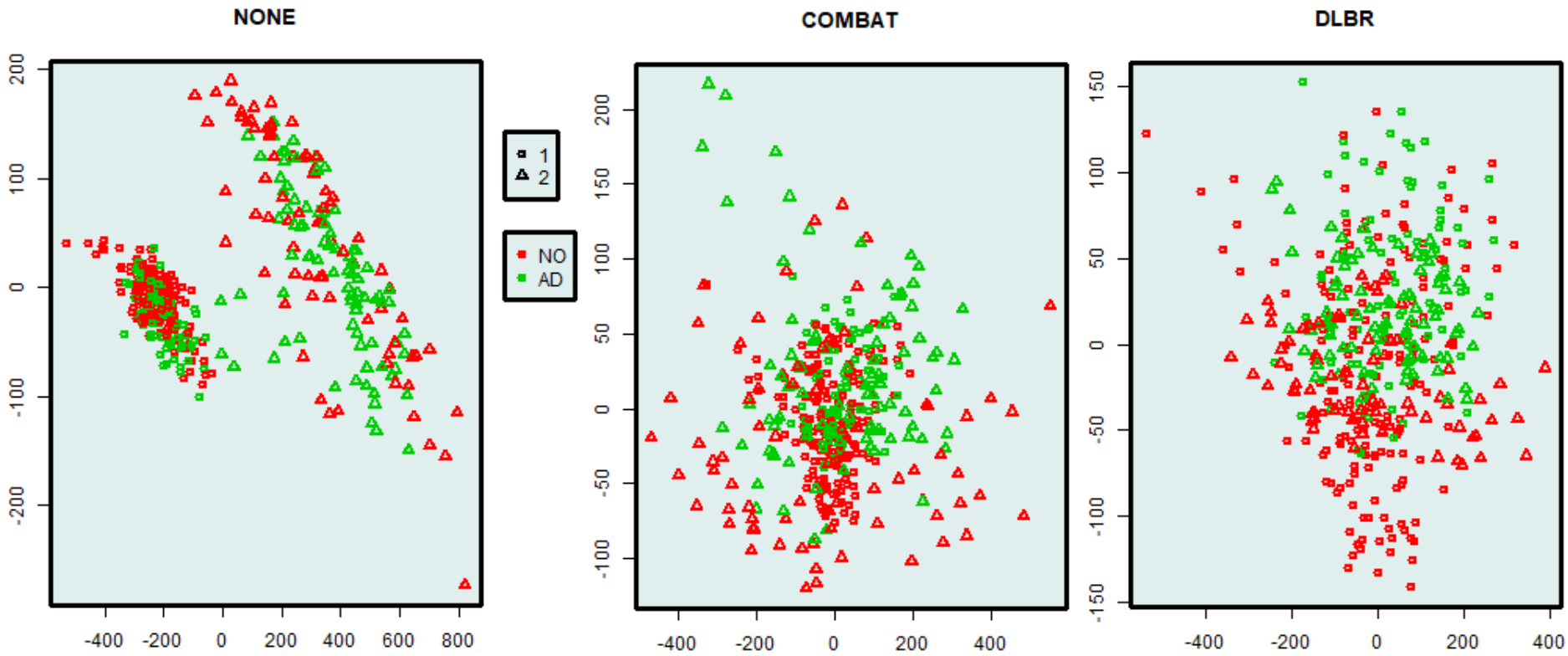




Comparison

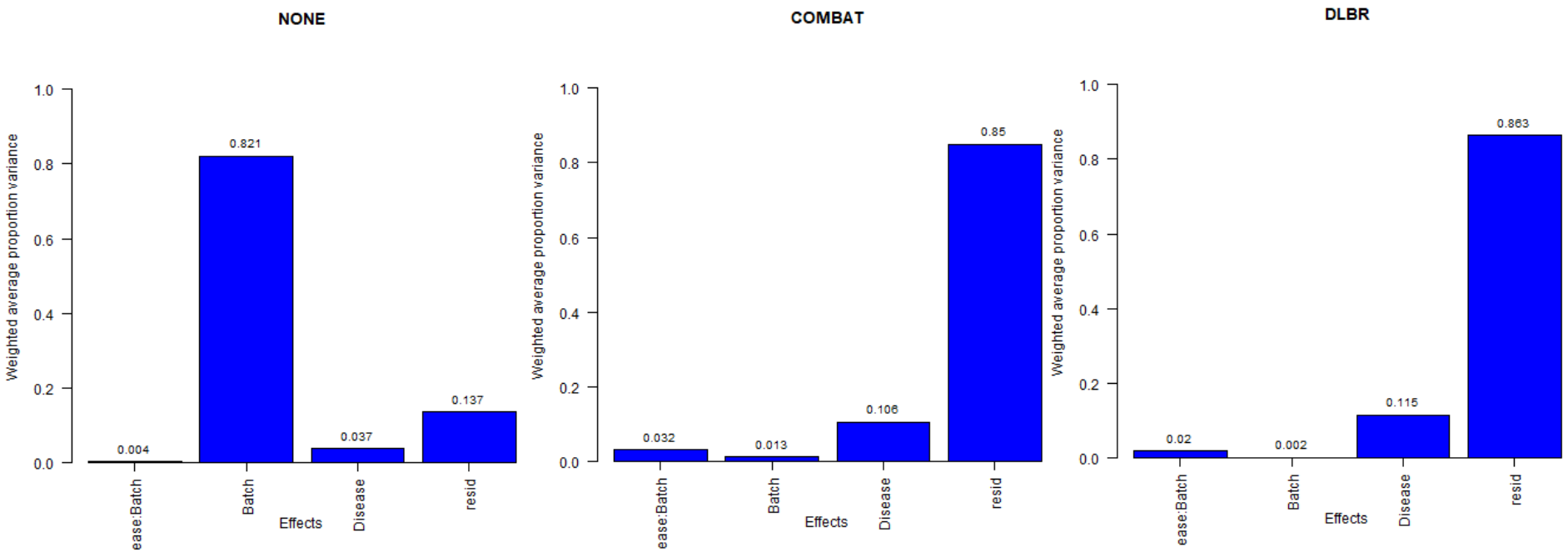


Comparison (cont'd)



Sources of Variability Analysis

PCVA R package by Bushel 2012



Performance Evaluation

- SVM
- 60% Training, 40% Testing

Metrics	None	COMBAT	DLBR
Sample asymmetry[1]	0.079	0.003	0.003
Sample overlap[1]	340.850	248.389	225.785
Samples correlation	1.000	0.918	0.924
Gene overlap	0.837	0.112	0.109
Gene correlation	1.000	0.967	0.982

Metrics	None	COMBAT	DLBR
Sensitivity	0.65	0.70	0.81
Specificity	0.82	0.74	0.66
ACC	0.75	0.72	0.72
AUC	0.73	0.72	0.73
MCC	0.47	0.43	0.46

[1] Lower is the better

Outline

1 Introduction

2 Methodology

3 Results

4 Conclusion

Contributions

- A new machine learning method for batch effect removal
- Combination of autoencoder and ResNet in modeling batch effect
- Dimensionality reduction and adjusting the neurons precisely
- Evaluate the performance on real microarray data

Acknowledgement

NIH-funded COBRE grant (1P20GM104320)
NIH [1R01DK107264]/NIFA [2016-67001-06314]

Questions

Questions?

References I

- Amodio, M., R. Montgomery, J. Pappalardo, D. Hafler, and S. Krishnaswamy (2018). "Neuron interference: Evidence-based batch effect removal". In: *arXiv preprint arXiv:1805.12198* (cit. on p. 8).
- Bushel, P. (2012). "pvca: Principal variance component analysis (PVCA)". In: *vol. R Package Version* 1.0 (cit. on p. 23).
- Johnson, W. E., C. Li, and A. Rabinovic (Apr. 2006). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127. DOI: 10.1093/biostatistics/kxj037 (cit. on p. 4).
- Shaham, U., K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger (Apr. 2017). "Removal of batch effects using distribution-matching residual networks". In: *Bioinformatics* 33.16. Ed. by J. Wren, pp. 2539–2546. DOI: 10.1093/bioinformatics/btx196 (cit. on p. 9).

References II

Sims, A. H., G. J. Smethurst, Y. Hey, M. J. Okoniewski, S. D. Pepper, A. Howell, C. J. Miller, and R. B. Clarke (Sept. 2008). "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis". In: *BMC Medical Genomics* 1.1. DOI: 10.1186/1755-8794-1-42 (cit. on p. 4).

Taminau, J., S. Meganck, C. Lazar, D. Steenhoff, A. Coletta, C. Molter, R. Duque, V. de Schaetzen, D. Y. W. Solís, H. Bersini, and A. Nowé (Dec. 2012). "Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages". In: *BMC Bioinformatics* 13.1. DOI: 10.1186/1471-2105-13-335 (cit. on p. 3).