Paper overview:

Unsupervised feature construction and knowledge extraction from genome-wide assays of Breast cancer with Denoising Autoencoders

JIE TAN, MATTHEW UNG, CHAO CHENG, CASEY S GREENE

Michal Rozenwald

Higher School of Economics
Faculty of Computer Science

November 16, 2017

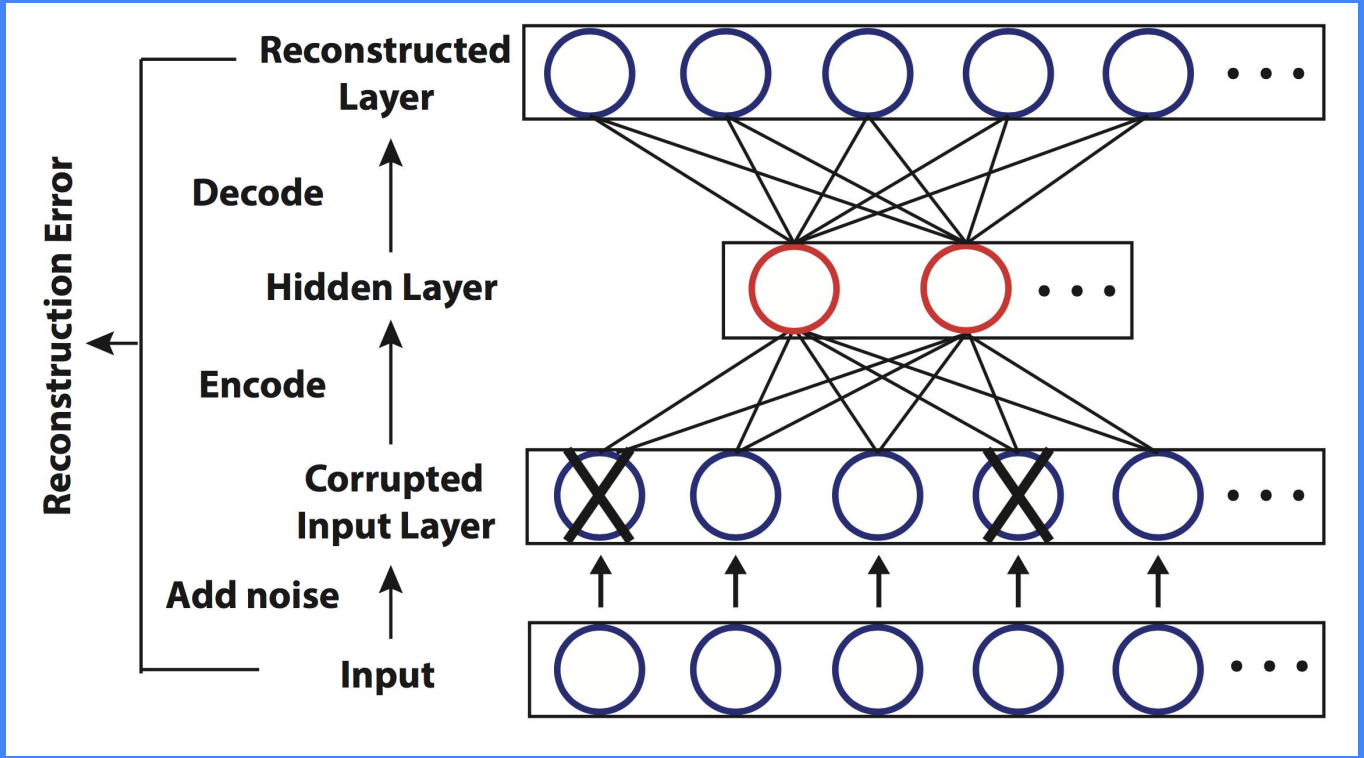
Plan

1. Problem statement
2. Denoising Autoencoders
3. Data
4. Feature interpretation
5. Results
6. Future work
7. Overall review

Bioinformatic Problems

- BIG data
- Many features
- Not as many samples
- Complex biology systems
- Hard to interpret features

Denoising Autoencoders (DAs)



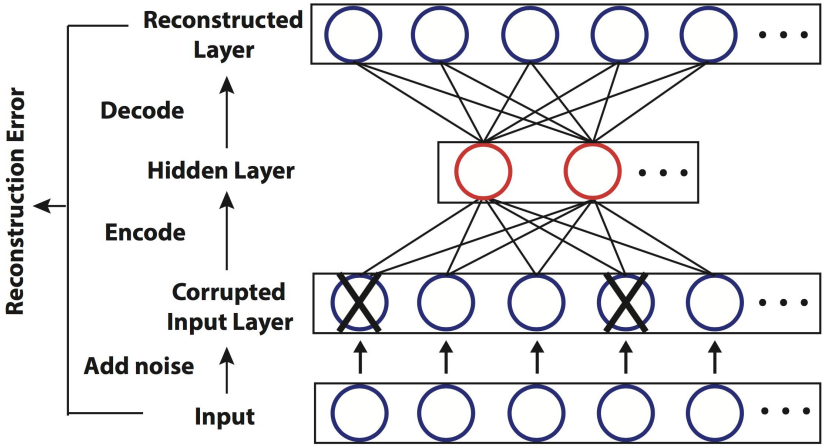
The network structure of Denoising Autoencoder

Denoising Autoencoders (DAs)

Minimize cross-entropy L(x, z)

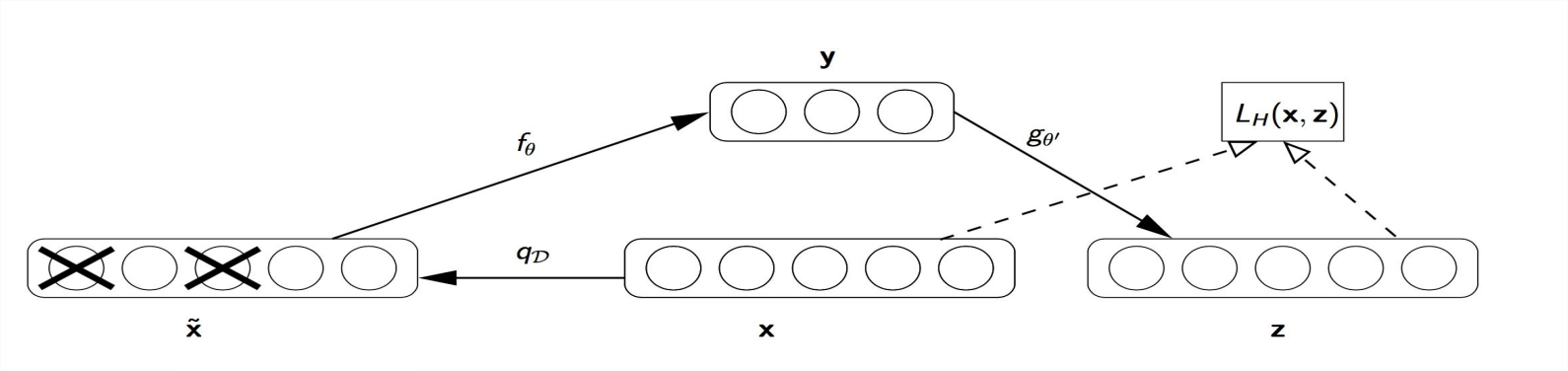
$$y = sigmoid\left(Wx + b\right)$$
$$z = sigmoid\left(W'y + b'\right)$$
$$L_H\left(x, z\right) = -\sum_{k=1}^{d}[x_k \log z_k + (1 - x_k) \log\left(1 - z_k\right)]$$

tied weights: W' = W^T



The network structure of Denoising Autoencoder

Denoising Autoencoders (DAs)



$\mathbf{x} \in [0, 1]^d$ - clean input

$\tilde{\mathbf{x}} \sim q_\mathcal{D}(\tilde{\mathbf{x}}|\mathbf{x})$ - corrupted input partially destroyed

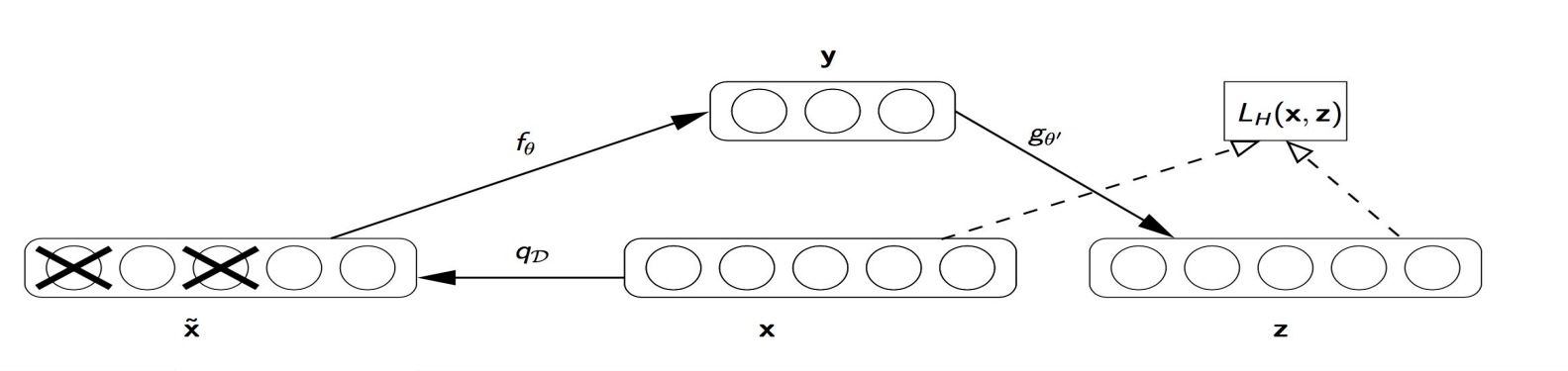
$\mathbf{y} = f_\theta(\tilde{\mathbf{x}})$ - hidden representation

$\mathbf{z} = g_{\theta'}(\mathbf{y})$ - reconstruction

$L_\mathbf{H}(\mathbf{x}, \mathbf{z})$ - cross-entropy “reconstruction error”

$\mathbf{x}$

Denoising Autoencoders (DAs)



$\mathbf{x} \in [0, 1]^d$ - clean input

$\tilde{\mathbf{x}} \sim q_\mathcal{D}(\tilde{\mathbf{x}}|\mathbf{x})$ - corrupted input partially destroyed

$\mathbf{y} = f_\theta(\tilde{\mathbf{x}})$ - hidden representation

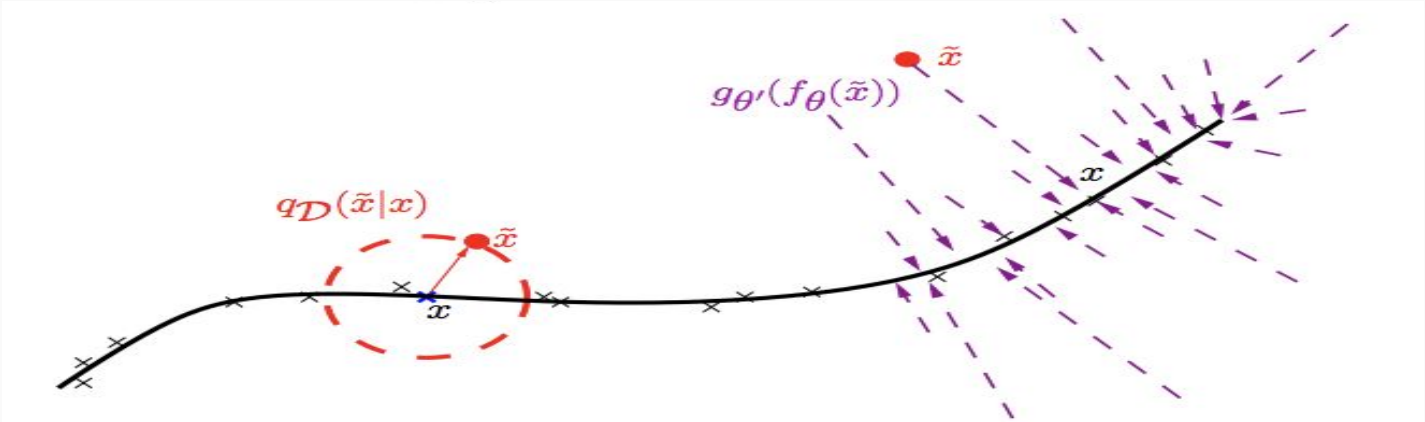
$\mathbf{z} = g_{\theta'}(\mathbf{y})$ - reconstruction

$L_\mathbf{H}(\mathbf{x}, \mathbf{z})$ - cross-entropy “reconstruction error”

x

$$\mathbf{y} = f_\theta(\tilde{\mathbf{x}}) = \text{sigmoid}(\underbrace{\mathbf{W}}_{d' \times d}\tilde{\mathbf{x}} + \underbrace{\mathbf{b}}_{d' \times 1}) \qquad g_{\theta'}(\mathbf{y}) = \text{sigmoid}(\underbrace{\mathbf{W}'}_{d \times d'}\mathbf{y} + \underbrace{\mathbf{b}'}_{d \times 1}).$$

Denoising Autoencoders (DAs)



$\mathbf{x} \in [0,1]^d$ - clean input

$\tilde{\mathbf{x}} \sim q_{\mathcal{D}}(\tilde{\mathbf{x}}|\mathbf{x})$ - corrupted input partially destroyed

$\mathbf{y} = f_\theta(\tilde{\mathbf{x}})$ - hidden representation

$\mathbf{z} = g_{\theta'}(\mathbf{y})$ - reconstruction

$L_{\mathbf{H}}(\mathbf{x}, \mathbf{z})$ - cross-entropy “reconstruction error”

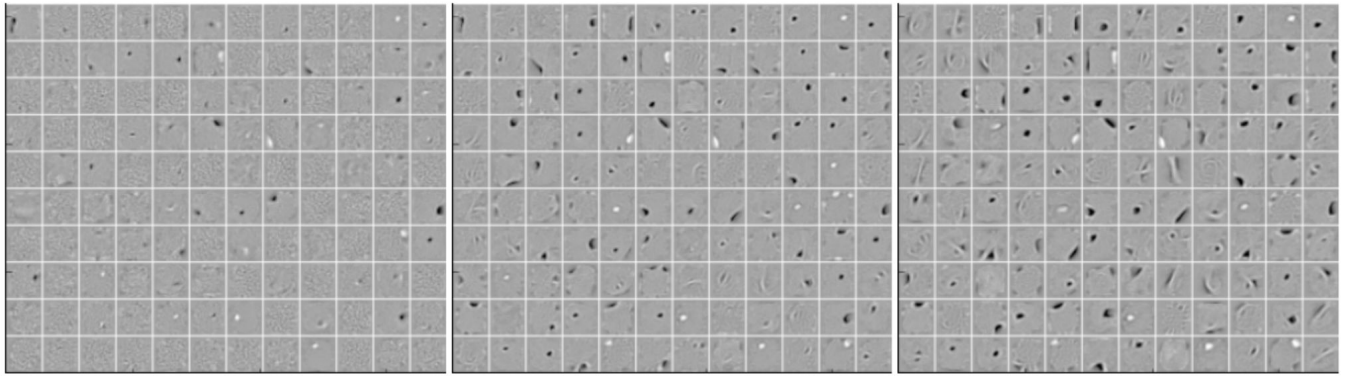
$\mathbf{x}$

$$\mathbf{y} = f_\theta(\tilde{\mathbf{x}}) = \text{sigmoid}(\underbrace{\mathbf{W}}_{d' \times d}\tilde{\mathbf{x}} + \underbrace{\mathbf{b}}_{d' \times 1})$$

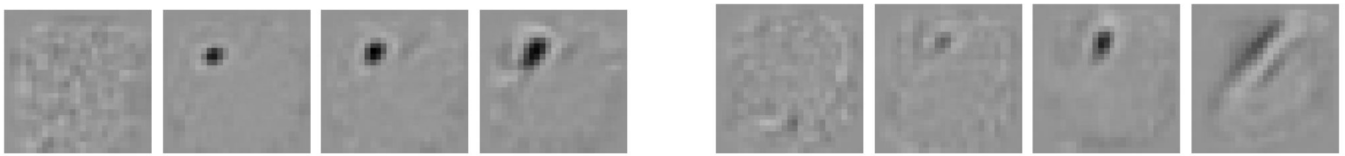
$$g_{\theta'}(\mathbf{y}) = \text{sigmoid}(\underbrace{\mathbf{W}'}_{d \times d'}\mathbf{y} + \underbrace{\mathbf{b}'}_{d \times 1}).$$

DAs for images

Distinctive features from different noise level



(a) No destroyed inputs (b) 25% destruction (c) 50% destruction



(d) Neuron A (0%, 10%, 20%, 50% destruction) (e) Neuron B (0%, 10%, 20%, 50% destruction)

Figure 3. **Filters obtained after training the first denoising autoencoder.**
(a-c) show some of the filters obtained after training a denoising autoencoder on MNIST samples, with increasing destruction levels ν. The filters at the same position in the three images are related only by the fact that the autoencoders were started from the same random initialization point.
(d) and **(e)** zoom in on the filters obtained for two of the neurons, again for increasing destruction levels.
As can be seen, with no noise, many filters remain similarly uninteresting (undistinctive almost uniform grey patches). As we increase the noise level, denoising training forces the filters to differentiate more, and capture more distinctive features. Higher noise levels tend to induce less local filters, as expected. One can distinguish different kinds of filters,

Constructed DA

Parameters: 10-fold cross validation

batch size: **10**	out of 1, 10, 20, 50
epoch size: **500**	out of 100, 200, 500
corruption level: **1%** genes in batch set to 0	out of 0, 0.1, 0.2
learning rate: **0.01**	out of 0.005, 0.01, 0.05
nodes in the hidden layer: **100**	

Loss:

cross-entropy

Data: METABRIC

Training set:
1424 samples

Test set:
712 samples

2520 genes (input size)

Evaluate on independent **TCGA dataset** (no retraining)**: 2520 genes measured for 547 samples**

Genomic data

Genes are:

- NOT linked to their neighbors (images)
- NOT linked temporally (audio)

linked by:

- transcription factors
- pathway membership
- many biological properties

We need:

- features to be linked to **clinical and molecular features** of samples

Feature interpretation

Linking Constructed Features to

- Sample Characteristics
- Patient Survival
- Transcription Factors
- Biological Pathways

Feature interpretation:

Sample Characteristics

Feature interpretation: Sample Characteristics

“Independent evaluation” performance of each node of the hidden layer

Binarized each node activity:

1. Identify each node’s highest and lowest activity values among samples in train set
2. Defined 10 equally spaced activation thresholds between these values
3. Evaluated the balanced accuracy for each node at each threshold to predict the desired sample characteristic
4. Chose nodes with highest balanced classification accuracies (fix thresholds)
5. Repeated 10 times with random partitioning of train/test sets (avoid sampling bias) on METABRIC
6. Average thresholds
7. Evaluate on TCGA

Feature interpretation: Sample Characteristics

Table 1. Performance of hidden nodes in classifying tumor from normal samples.

	METABRIC		TCGA
Node	Discovery	Test	Evaluation
64	0.970	0.968	0.996
99	0.957	0.959	0.998
38	0.879	0.887	0.911
43	0.873	0.873	0.750
69	0.871	0.872	0.906

Table 2. Performance of hidden nodes in classifying ER + from ER - samples.

	METABRIC		TCGA
Node	Discovery	Test	Evaluation
89	0.848	0.833	0.749
30	0.824	0.822	0.856
58	0.808	0.801	0.828
6	0.798	0.799	0.771
69	0.784	0.779	0.820

Classification based on one selected hidden node of DAs compared to sample type and Estrogen Receptor (ER) status.

Measure: Balanced accuracy

Feature interpretation: Sample Characteristics - Molecular Subtype

Table 3. Performance of hidden nodes in classifying each intrinsic subtype.

Subtype	Basal	Her2	LumA	LumB	Normal	LumA/B
Node	30	29	5	66	42	6
METABRIC Discovery	0.929	0.761	0.780	0.755	0.750	0.849
METABRIC Test	0.918	0.741	0.777	0.750	0.748	0.849
TCGA Evaluation	0.992	0.712	0.800	0.717	0.733	0.825

Prediction of subtypes that were developed by Parker et al. as “50-gene subtype predictor (PAM50)” based on gene expression data

Use binary classification and choose 1 best node for each subtype.

Measure: accuracy

Feature interpretation: Sample Characteristics - Molecular Subtype

Table 3. Performance of hidden nodes in classifying each intrinsic subtype.

Subtype	Basal	Her2	LumA	LumB	Normal	LumA/B
Node	30	29	5	66	42	6
METABRIC Discovery	0.929	0.761	0.780	0.755	0.750	0.849
METABRIC Test	0.918	0.741	0.777	0.750	0.748	0.849
TCGA Evaluation	0.992	0.712	0.800	0.717	0.733	0.825

Prediction of subtypes that were developed by Parker et al. as “50-gene subtype predictor (PAM50)” based on gene expression data

Use binary classification and choose 1 best node for each subtype.

Measure: accuracy

Feature interpretation: Sample Characteristics - Molecular Subtype

Table 3. Performance of hidden nodes in classifying each intrinsic subtype.

Subtype	Basal	Her2	LumA	LumB	Normal	LumA/B
Node	30	29	5	66	42	6
METABRIC Discovery	0.929	0.761	0.780	0.755	0.750	0.849
METABRIC Test	0.918	0.741	0.777	0.750	0.748	0.849
TCGA Evaluation	0.992	0.712	0.800	0.717	0.733	0.825

Association with ER status is meaningful:

In METABRIC:
77.6% of ER positive samples are LumA or LumB
3.9% of ER positive samples are Basal subtype
ER negative samples usually Basal subtype

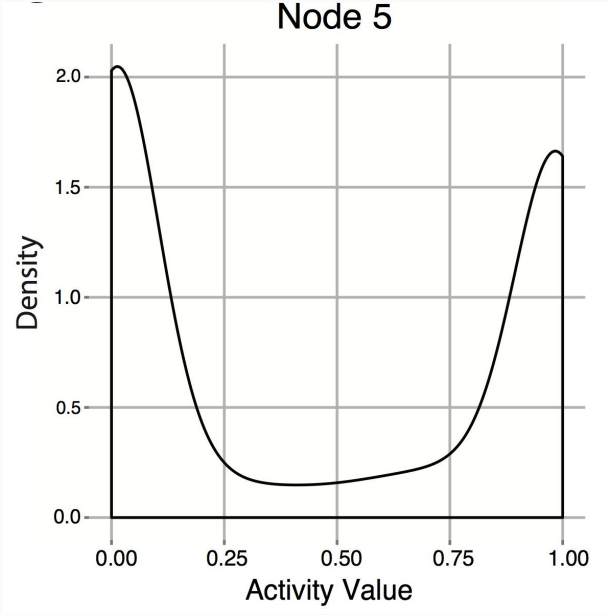
Table 2. Performance of hidden nodes in classifying ER + from ER - samples.

	METABRIC		TCGA
Node	Discovery	Test	Evaluation
89	0.848	0.833	0.749
30	0.824	0.822	0.856
58	0.808	0.801	0.828
6	0.798	0.799	0.771
69	0.784	0.779	0.820

Feature interpretation:

Patient Survival

Feature interpretation: Patient Survival



Active value distribution

METABRIC has <15 years follow-up on patients

For each node:

Bimodal distribution of activity values
one peak close to **0**
another peak close to **1**

Defined **0.5** as the activation **cut-off** to divide samples into **two groups**

Correlate activity values of each node to **patient survival time**

Feature interpretation: Patient Survival

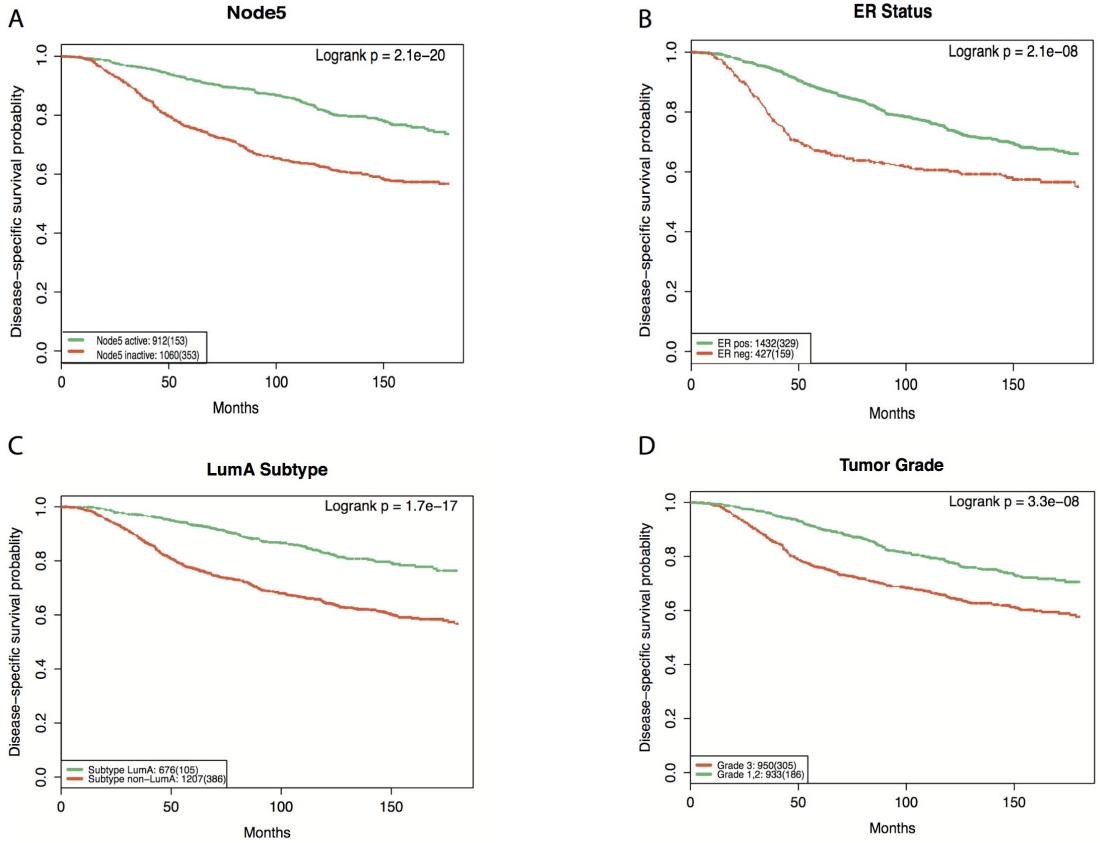


Fig. 2. Kaplan-Meier plots of disease-specific survival for Node5 (A), ER status (B), Luminal A subtype (C),

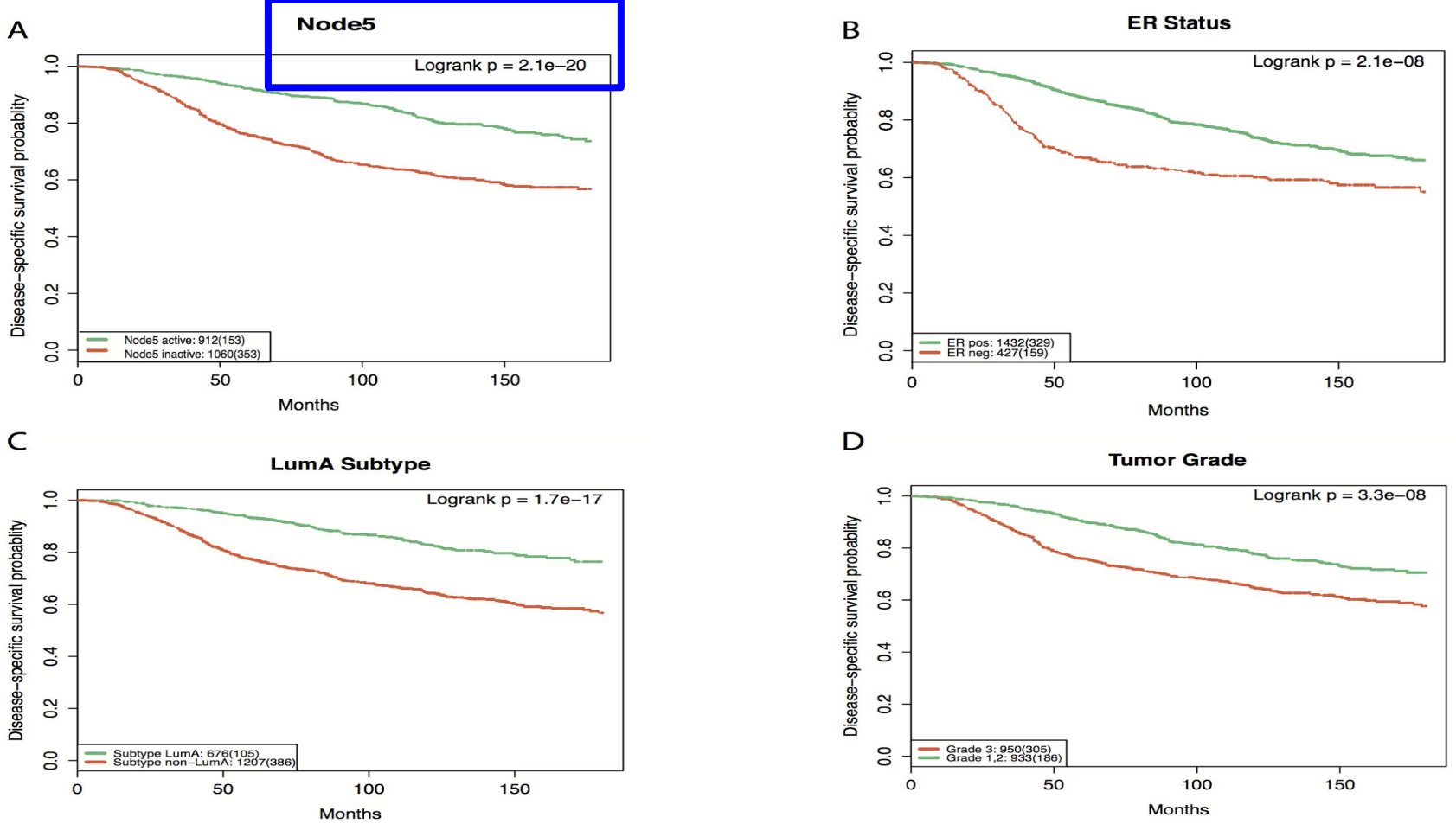


Fig. 2. Kaplan-Meier plots of disease-specific survival for Node5 (A), ER status (B), Luminal A subtype (C), and Tumor Grade (D) demonstrate that the constructed feature outperforms the other predictors.

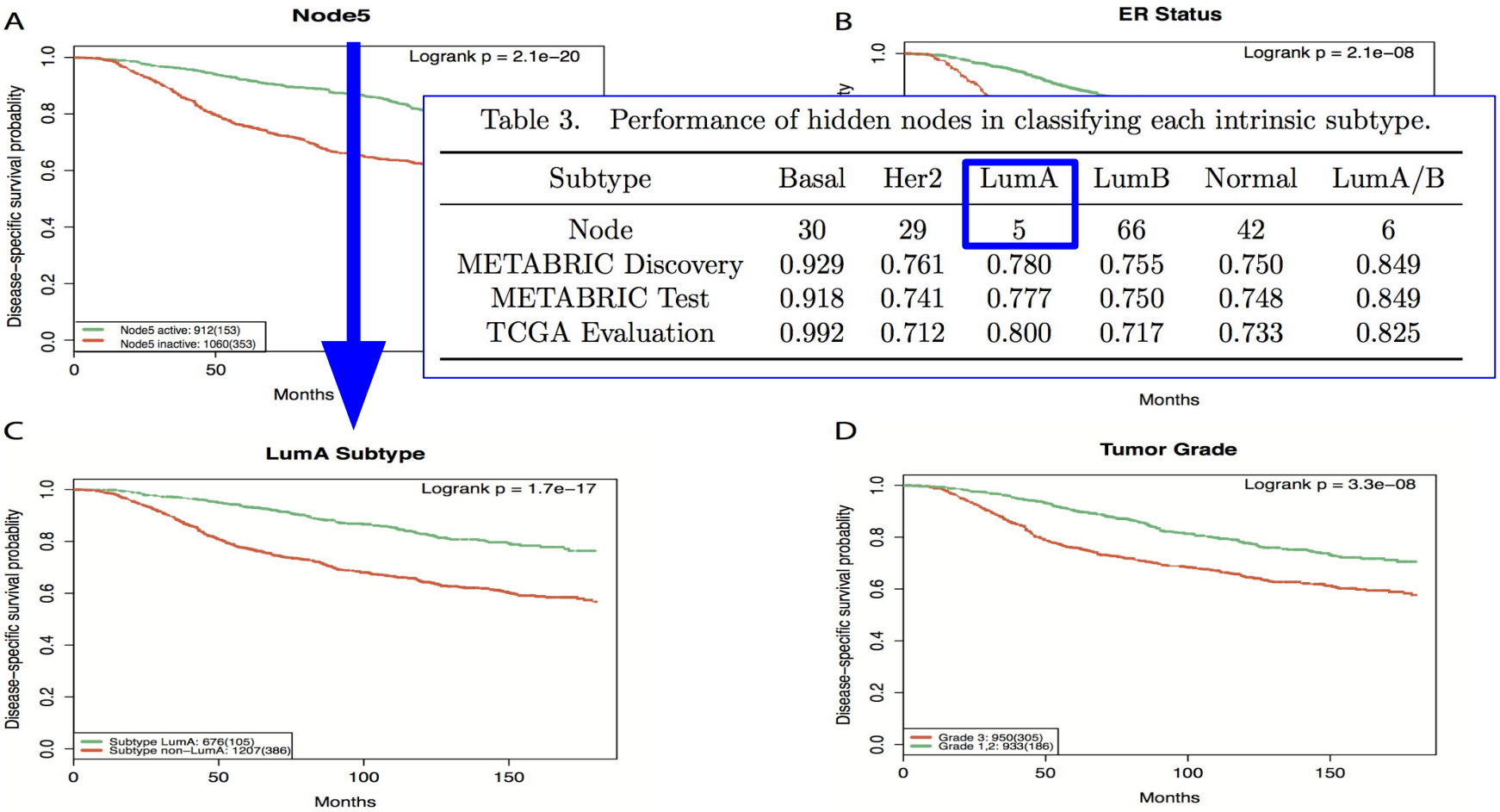


Table 3. Performance of hidden nodes in classifying each intrinsic subtype.

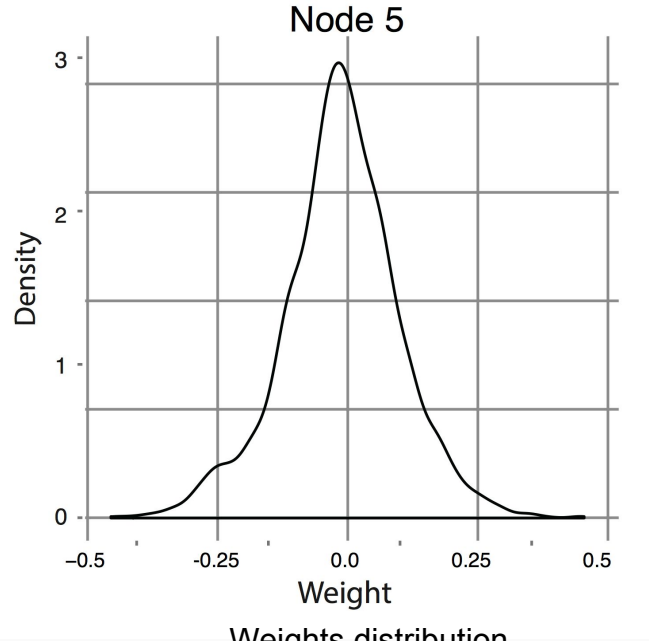
Subtype	Basal	Her2	LumA	LumB	Normal	LumA/B
Node	30	29	5	66	42	6
METABRIC Discovery	0.929	0.761	0.780	0.755	0.750	0.849
METABRIC Test	0.918	0.741	0.777	0.750	0.748	0.849
TCGA Evaluation	0.992	0.712	0.800	0.717	0.733	0.825

Fig. 2. Kaplan-Meier plots of disease-specific survival for Node5 (A), ER status (B), Luminal A subtype (C), and Tumor Grade (D) demonstrate that the constructed feature outperforms the other predictors.

Feature interpretation:

Transcription Factors

Feature interpretation: Transcription Factors



Weights distribution

Transcription factor ChIP-seq data from ENCODE via the UCSC genome browser

For a single node:

~Normal distribution of **weights** centered at **0**

Most genes gave zero or low weight
Small number gave high positive or negative weights

“high-weight genes” - give weight values that lie outside 2 standard deviations from mean of weight distribution

Identified nodes whose high-weight genes were overrepresented by genes bound by one **transcription factor** and calculated the **odds ratio**

Feature interpretation: Transcription Factors

Node 58:

Most active genes ratio for:
FOXA1 and GATA3 Transcription factors

Meaningful for ER status

Biologically meaningful results

Table 2. Performance of hidden nodes in classifying ER + from ER - samples.

	METABRIC		TCGA
Node	Discovery	Test	Evaluation
89	0.848	0.833	0.749
30	0.824	0.822	0.856
58	0.808	0.801	0.828
6	0.798	0.799	0.771
69	0.784	0.779	0.820

Results

Results

Denoising Autoencoders effectively summarize key features in breast cancer data.
Created robust features across datasets.

We looked at interesting approaches for feature analysis

Identified features that stratify:

- tumor/normal samples
- ER+/- samples
- molecular subtypes
- patient survival

Ideas for future work:

- ➔ Copmere the results with existed methods of feature extraction
- ➔ New approaches for interpretation
- ➔ Create new unsupervised features
- ➔ Comparison with existing bio markets
- ➔ Use deep NN and deep Denoising Autoencoders

Questions?

Thank you

Reference

“UNSUPERVISED FEATURE CONSTRUCTION AND KNOWLEDGE EXTRACTION FROM GENOME-WIDE ASSAYS OF BREAST CANCER WITH DENOISING AUTOENCODERS”

JIE TAN, MATTHEW UNG, CHAO CHENG, CASEY S GREENE

Pac Symp Biocomput. 2015; 20: 132–143.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299935/

“Extracting and Composing Robust Features with Denoising Autoencoders”

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.:
Neural Information Processing Systems, NIPS (2008)

http://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf

http://videolectures.net/icml08_vincent_ecrf/

“Kaplan-Meier Survival Curves and the LogRank Test”

Staub L., Alexandros G.

https://stat.ethz.ch/education/semesters/ss2011/seminar/contents/presentation_2.pdf