



# Novel Feature Learning Method for Gene Expression Data Based on An Optimized Denoising Autoencoder

Leo Dong, Middlesex School

Mentor: Gil Alterovitz

8<sup>th</sup> Annual PRIMES CS Conference

October 13, 2018



# I Introduction and Motivation

1. What is Gene Expression Data?
2. Gene expression data in cancer research
3. The “Curse of Dimensionality”
4. Current dimensionality reduction methods
5. Current state-of-the-art method

# 1. What is Gene Expression Data?

# Gene Expression

---

Expression level of a gene = how much does this gene *contribute* to the final *gene product* (either proteins or functional RNAs)?

---

Biomedical sequencing technology allows for simultaneous measurements of genome-wide gene expressions in organism tissues

---

Genome-wide gene expression levels are different for healthy subjects and diseased subjects



## 2. Gene Expression Data for Cancer Research

# Cancer diagnosis & identifying key genes to tumor formation

## Multiclass cancer diagnosis using tumor gene expression signatures

**Sridhar Ramaswamy\*<sup>†</sup>, Pablo Tamayo\*, Ryan Rifkin\*<sup>‡</sup>, Sayan Mukherjee\*<sup>‡</sup>, Chen-Hsiang Yeang\*<sup>§</sup>, Michael Angelo\*, Christine Ladd\*, Michael Reich\*, Eva Latulippe<sup>¶</sup>, Jill P. Mesirov\*, Tomaso Poggio<sup>‡</sup>, William Gerald<sup>¶</sup>, Massimo Loda<sup>¶</sup>, Eric S. Lander\*<sup>•••</sup>, and Todd R. Golub\*<sup>††‡‡</sup>**

\*Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02138; Departments of <sup>†</sup>Adult and <sup>††</sup>Pediatric Oncology, Dana–Farber Cancer Institute/Harvard Medical School, Boston, MA 02115; <sup>‡</sup>Department of Pathology, Brigham and Women’s Hospital, Boston, MA 02115; <sup>¶</sup>Department of Pathology, Memorial Sloan–Kettering Cancer Center, New York, NY 10021; and Departments of <sup>••</sup>Biology, <sup>•</sup>McGovern Institute, Center for Brain and Computational Learning, and <sup>§</sup>Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139

Contributed by Eric S. Lander, October 23, 2001

(Ramaswamy *et al.*, 2001, *PNAS*)



# Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study



Edward C Schwalbe, Janet C Lindsey, Sirintra Nakjang, Stephen Crosier, Amanda J Smith, Debbie Hicks, Gholamreza Rafiee, Rebecca M Hill, Alice Iliasova, Thomas Stone, Barry Pizer, Antony Michalski, Abhijit Joshi, Stephen B Wharton, Thomas S Jacques, Simon Bailey, Daniel Williamson, Steven C Clifford

(Schwalbe *et al.*, 2017, *The Lancet*)

---

# Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes

Mark D M Leiserson<sup>1,2,14</sup>, Fabio Vandin<sup>1,2,13,14</sup>, Hsin-Ta Wu<sup>1,2</sup>, Jason R Dobson<sup>1-3</sup>, Jonathan V Eldridge<sup>1</sup>, Jacob L Thomas<sup>1</sup>, Alexandra Papoutsaki<sup>1</sup>, Younhun Kim<sup>1</sup>, Beifang Niu<sup>4</sup>, Michael McLellan<sup>4</sup>, Michael S Lawrence<sup>5</sup>, Abel Gonzalez-Perez<sup>6</sup>, David Tamborero<sup>6</sup>, Yuwei Cheng<sup>7</sup>, Gregory A Ryslik<sup>8</sup>, Nuria Lopez-Bigas<sup>6,9</sup>, Gad Getz<sup>5,10</sup>, Li Ding<sup>4,11,12</sup> & Benjamin J Raphael<sup>1,2</sup>

(Leiserson *et al.*, 2014, *Nature Genetics*)


OPEN

## Machine Learning-Assisted Network Inference Approach to Identify a New Class of Genes that Coordinate the Functionality of Cancer Networks

Received: 9 January 2017

Accepted: 27 June 2017

Published online: 1 August 2017

Mehrab Ghanat Bari<sup>1</sup>, Choong Yong Ung<sup>1</sup>, Cheng Zhang<sup>1</sup>, Shizhen Zhu<sup>2</sup> & Hu Li <sup>1</sup>

(Bari *et al.*, 2017, *Scientific Reports*)

# 3. The Curse of Dimensionality

# High dimensions of gene expression data is problematic...

- Dimensions of gene expression data is too high!
  - Each sample has thousands of genes describing it (i.e. “thousands-dimensional”)
- When the sample size is significantly smaller than the dimension size, there is no way to gain any useful information from the data

# Dimensionality Reduction Strategies

- Feature (gene) Selection
  - e.g. Differential analysis
  - You still keep genes as features
- Feature Learning/feature extraction
  - e.g. PCA
  - You construct artificial features from combinations of genes



This research follows the feature learning approach to extract both computationally and biologically meaningful features from gene expression data.



## II Methodology

1. Related Works
2. netDAE: An Overview
3. Denoising Autoencoders (DAE)
4. Network Modularity
5. Putting it all together!

# Disclaimer

- The following slides contain unpublished contents of a new algorithm.
- Please refrain from taking pictures or notes, and please do not post anything online about this algorithm before it is formally published.
- Thank you so much for your understanding!

# 1. Related Works (state-of-the-art)

# State-of-the-art: ADAGE

- **A**nalysis using **D**enoising **A**utoencoders of **G**ene **E**xpression
- Developed by Tan *et al.* at Geisel School of Medicine at Dartmouth and Perelman School of Medicine at University of Pennsylvania
- Tan *et al.*, 2015, *Pac. Symp. Biocomput.*: First use of DAE for feature learning of gene expression data; tested on METABRIC and TCGA-BRCA breast cancer datasets
- Tan *et al.*, 2016, *mSystems*: GE feature learning with DAE generalized into framework called ADAGE; tested on *Pseudomonas aeruginosa*
- Tan *et al.*, 2017, *Cell Systems*: Ensemble version of ADAGE, eADAGE, was developed; tested on *Pseudomonas aeruginosa*

Published in final edited form as:

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

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

**JIE TAN, MATTHEW UNG, CHAO CHENG, and CASEY S GREENE\***

Department of Genetics Institute for Quantitative Biomedical Sciences Norris Cotton Cancer  
Center The Geisel School of Medicine at Dartmouth Hanover, NH 03755, USA

(Tan *et al.*, 2015, *Pac. Symp. Biocomput.*)



# ADAGE-Based Integration of Publicly Available *Pseudomonas aeruginosa* Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions

Jie Tan,<sup>a</sup> John H. Hammond,<sup>b</sup> Deborah A. Hogan,<sup>b</sup>  Casey S. Greene<sup>a,c</sup>

Department of Genetics<sup>a</sup> and Department of Microbiology and Immunology,<sup>b</sup> Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA; Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA<sup>c</sup>

(Tan *et al.*, 2016, *mSystems*)

# Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks

Jie Tan,<sup>1,6</sup> Georgia Doing,<sup>2,6</sup> Kimberley A. Lewis,<sup>2</sup> Courtney E. Price,<sup>2</sup> Kathleen M. Chen,<sup>3</sup> Kyle C. Cady,<sup>4,5</sup> Barret Perchuk,<sup>4,5</sup> Michael T. Laub,<sup>4,5</sup> Deborah A. Hogan,<sup>2</sup> and Casey S. Greene<sup>3,7,\*</sup>

<sup>1</sup>Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

<sup>2</sup>Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

<sup>3</sup>Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA

<sup>4</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>5</sup>Howard Hughes Medical Institute, Cambridge, MA, USA

<sup>6</sup>These authors contributed equally

<sup>7</sup>Lead Contact

\*Correspondence: [csgreene@upenn.edu](mailto:csgreene@upenn.edu)

<http://dx.doi.org/10.1016/j.cels.2017.06.003>

(Tan *et al.*, 2017, *Cell Systems*)



# The ADAGE Framework

- ADAGE is essentially a framework of using denoising autoencoders (DAE) for feature learning of gene expression data
  - Provided very innovative ways to interpret the learned features
  - Able to retrieve a *gene subset* even with a *feature learning* method
  - Learned features are biologically meaningful
- My method is inspired by and based on ADAGE
- I only compare with ADAGE, not the ensemble version eADAGE, for the sake of simplicity (my method can also be extended to an ensemble version, so here I only compare the base version)

## 2. netDAE: An Overview

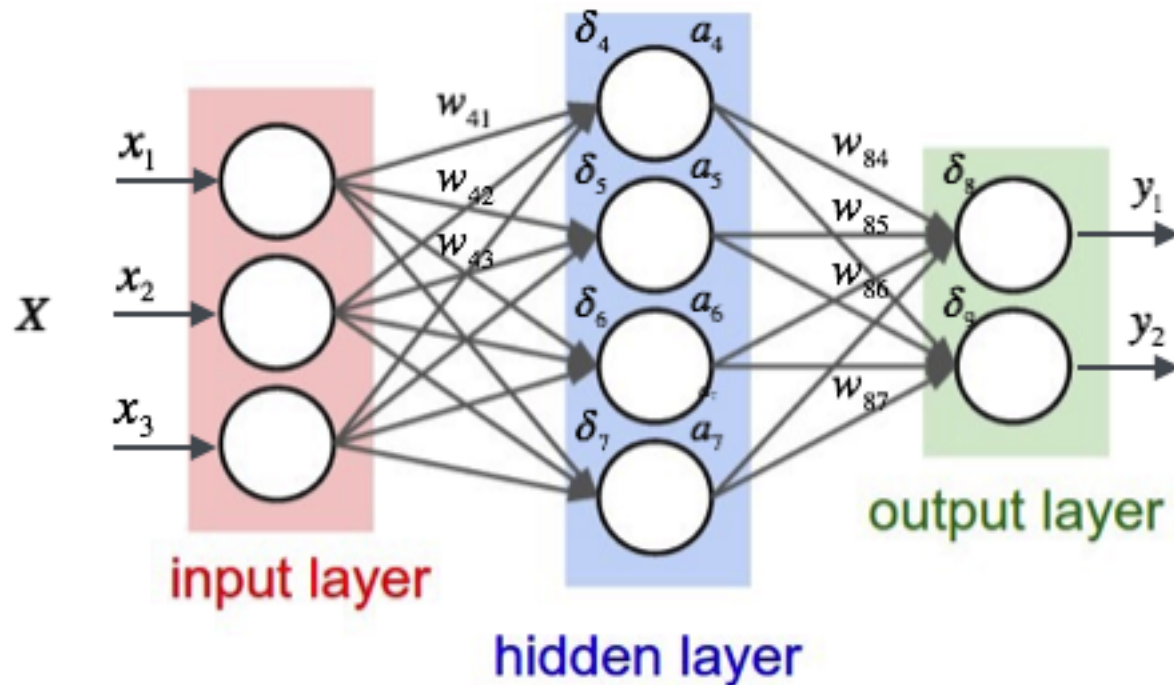
# It is very simple: net + DAE = netDAE!!

- netDAE is essentially a variation of denoising autoencoders (DAE) that, for the first time ever, incorporates complex network measures
- Preserves the advantages of ADAGE: you are still able to retrieve an important *gene subset* from a *feature learning* method
- Task/trait-specific (not really “supervised”)
  - Needs to know the clinical traits of the samples
- Superior accuracies
- I will explain the “DAE” part and the “net” part separately and talk about how they connect to form netDAE

# 3 Denoising Autoencoder (DAE)

# Neural Networks: From input to output

- A feedforward network structure that learns the relationship between the input  $\mathbf{X}$  and the output  $\mathbf{Y}$  through nonlinear combinations of input features (using *weights* and *biases*).

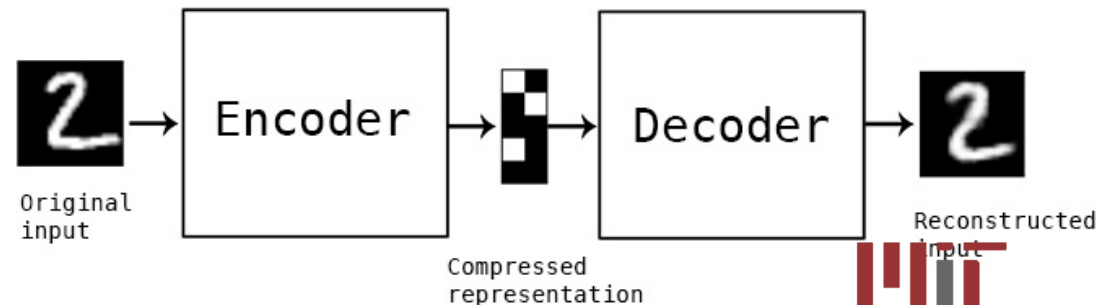
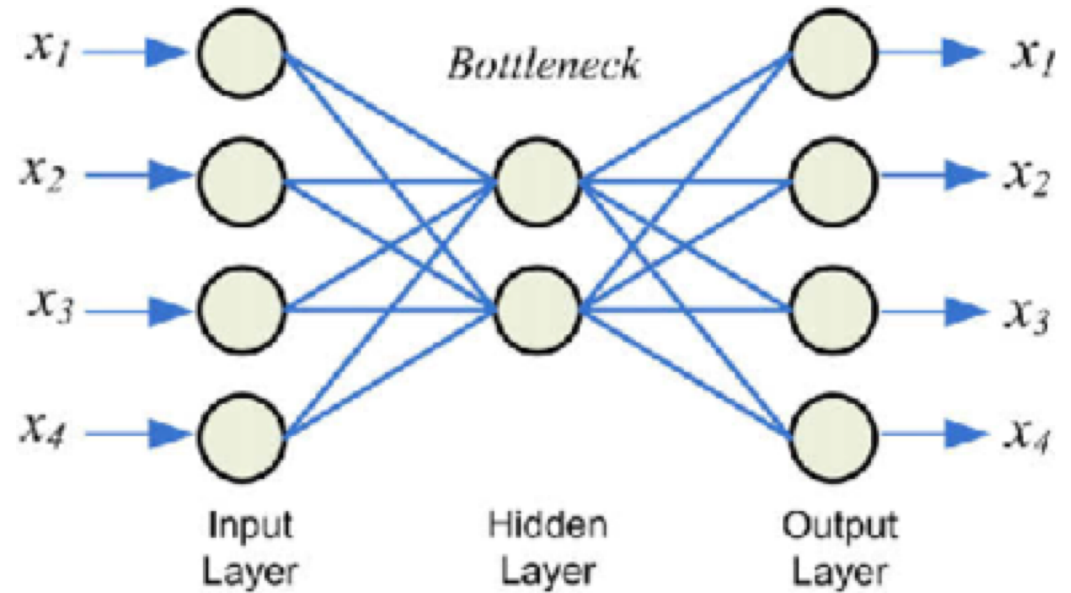


$$\mathbf{h}^{(1)} = g^{(1)} \left( \mathbf{W}^{(1)\top} \mathbf{x} + \mathbf{b}^{(1)} \right);$$
$$\mathbf{h}^{(2)} = g^{(2)} \left( \mathbf{W}^{(2)\top} \mathbf{h}^{(1)} + \mathbf{b}^{(2)} \right);$$

(Venelin Valkov, [Medium.com](#))

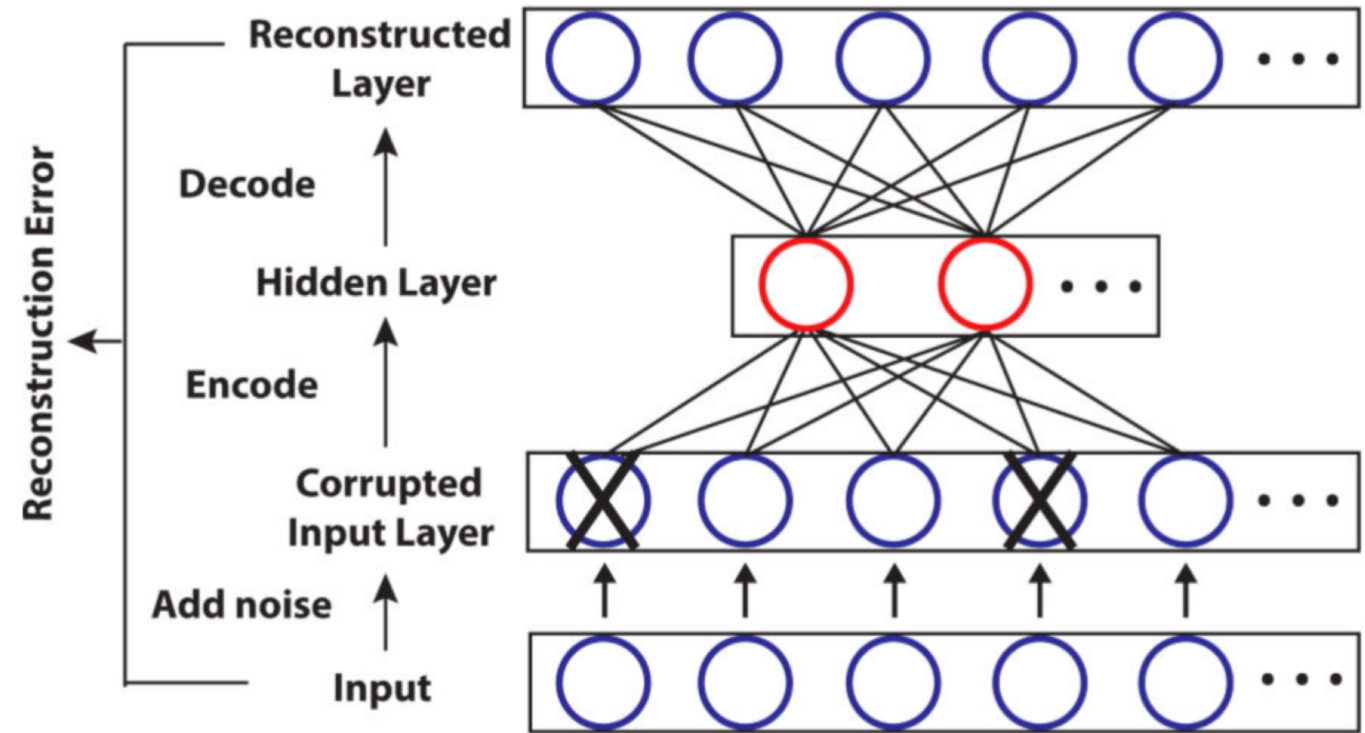
# Autoencoder: From input to input

- Autoencoder is a neural network that copies the input to its output.
- The hidden layer has a small number of nodes to force the input into a lower dimension. The loss calculates how well the input can be reconstructed from this low dimensional space back to its original space.
- The hidden layer is a “code” that efficiently represents the high dimensional input data in lower dimensions. For feature learning tasks, we take the hidden layer node as the “learned features.”



# Denoising Autoencoder (DAE)

- Autoencoder that maps the *corrupted input* to the *clean input*
- Avoids “perfect reconstruction” (decoder learning the inverse function of the encoder) and makes it able to denoise data



(Tan et al., 2015, Pac. Symp. Biocomput.)

In other words, DAE maps the input into a low dimensional space.

Evaluates the low dimensional space based on *how much information it contains about the original space* by trying to reconstruct the original space from it.



# 4 Network Modularity

# Newman and Girvan Modularity

PHYSICAL REVIEW E **69**, 026113 (2004)

## Finding and evaluating community structure in networks

M. E. J. Newman<sup>1,2</sup> and M. Girvan<sup>2,3</sup>

<sup>1</sup>*Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109-1120, USA*

<sup>2</sup>*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*

<sup>3</sup>*Department of Physics, Cornell University, Ithaca, New York 14853-2501, USA*

(Received 19 August 2003; published 26 February 2004)

We propose and study a set of algorithms for discovering community structure in networks—natural divisions of network nodes into densely connected subgroups. Our algorithms all share two definitive features: first, they involve iterative removal of edges from the network to split it into communities, the edges removed being identified using any one of a number of possible “betweenness” measures, and second, these measures are, crucially, recalculated after each removal. We also propose a measure for the strength of the community structure found by our algorithms, which gives us an objective metric for choosing the number of communities into which a network should be divided. We demonstrate that our algorithms are highly effective at discovering community structure in both computer-generated and real-world network data, and show how they can be used to shed light on the sometimes dauntingly complex structure of networked systems.

DOI: 10.1103/PhysRevE.69.026113

PACS number(s): 89.75.Hc, 87.23.Ge, 89.20.Hh, 05.10.–a

(Newman and Girvan, 2004, *Physical Review E*)



# Newman and Girvan Modularity

- A measure of the quality of ***modularization*** (clusterization) of a complex network
- Used in many clustering algorithms to assess the quality of clustering
- In the range  $[0, 1]$ , where 1 means the network has a strong community/clusterization structure

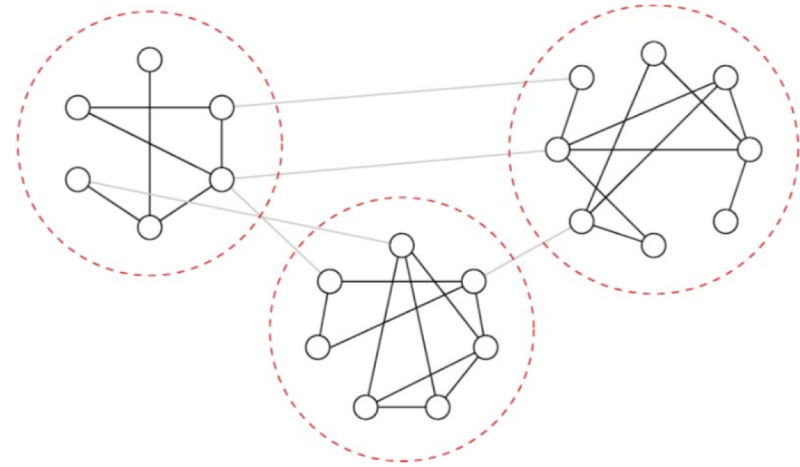


FIG. 1. A small network with community structure of the type considered in this paper. In this case there are three communities, denoted by the dashed circles, which have dense internal links but between which there is only a lower density of external links.

(Newman and Girvan, 2004, *Physical Review E*)

# Definition of Modularity

- **Actual fraction** of within-community edges (density of within clusters) minus the **expected value of the fraction** of within-community edges (density within clusters) in a **random graph**

$$Q = \sum_i (e_{ii} - a_i^2) = \text{Tr } \mathbf{e} - \|\mathbf{e}^2\|,$$

$$Q = \sum_i (e_{ii} - a_i^2) = \text{Tr } \mathbf{e} - \|\mathbf{e}^2\|,$$

Consider a network of  $k$  communities. Make a  $k$  by  $k$  matrix  $\mathbf{e}$  whose element  $e_{ij}$  is the fraction of edges in the network that link vertices in community  $i$  to vertices in community  $j$ .

$$\sum_i^k e_{ii} = \text{Tr}(\mathbf{e}) = \textit{fraction of edges that connect vertices in the same community}$$

Define row sums  $a_i = \sum_j^k e_{ij}$ , representing the fraction of edges that connect to vertices in community  $i$ .

$$* e_{ij} = a_i a_j,$$

*when network edges are randomly wired  
regardless of community structure*

First term = fraction of within-community edges

Second term = expected fraction of within-community edges for a random graph

# 5 Putting it together!

# A recap...

- Denoising Autoencoder: Maps data from the original high dimensional space to a low dimensional space
  - Assessment of the mapping: level of reconstruction
- Newman Girvan Modularity: Measure of the quality of a graph's clusterizations
  - Used in clustering algorithms to determine the best way to assign vertices to clusters in a graph
  - How is this related to microarray data?
  - Modularity is used in unsupervised learning (clustering) but we have labeled data!

## !!! 2 Key Ideas: What if...

1. What if we view microarray data not as a matrix, but as a *weighted sample graph* (a complex network)?
2. What if we do not use modularity to determine the quality of clusterization structures, but, the other way around, given the clusterization structures, we use modularity to determine the quality of the “space” (which determines the distance measures) that this clusterization structure is in?



# Idea 1: *Weighted Sample Graph (WSG)*

- Given an  $m$ -dimensional space where the samples are defined by  $m$  features. We create a *weighted sample graph* over all samples such that:
  - Each sample is a vertex.
  - It is a complete  $K_n$  graph where each vertex is connected to every other vertex, where  $n$  is the number of samples.
  - Weight of the edge connecting two vertices is determined by the  $m$ -dimensional Euclidean distance between the two samples in this given space.
- **!!** Clinical traits of the samples determine the clusterization structure of the graph
  - e.g. healthy patients in one cluster, and diseased patients in another cluster

# Idea 2: Modularity as an assessment of “space”

- Recall: You can make a WSG in the original, high-dimensional space or in the low-dimensional space mapped by DAE
  - Distance measures are different because the same sample is represented by different features
- A simple modification of modularity into *weighted modularity* allows us to calculate the modularity of WSGs

# Comparing the modularity of two spaces

- Assumption: In a “good” space that represents the samples, samples of different labels should be quite different.
  - In other words, the WSG with identical clusterization structures in a “good” space measured by its distance measures should have higher modularity measures
- DAE: the low dimensional space that contains similar amount of information as the original space is optimal
- Modularity: the low dimensional space that exhibits high quality clusterizations compared to the original space is optimal
- netDAE: I like both.

netDAE optimizes the **reconstruction error in its *decoded layer*** and the **modularity measure of its *encoded layer*** simultaneously

# Finale: The Loss Function

- Original data:  $\mathbf{X}$ ; labels of  $c$  classes (formulated as identity function):  $\delta_c(\mathbf{x}_i, \mathbf{x}_j)$
- Corrupted input:  $\tilde{\mathbf{X}}$
- Output:  $\mathbf{X}$
- Encoder:  $\mathbf{H} = f(\tilde{\mathbf{X}})$
- Decoder:  $\mathbf{Z} = g(\mathbf{H})$
- Modularity function:  $Q(\mathbf{H}, \delta_c)$
- Loss:

$$\begin{aligned} L(\tilde{\mathbf{X}}, \mathbf{X}, \delta_c) &= \lambda L_{\text{encod.}}(f(\tilde{\mathbf{X}}), \delta_c) + L_{\text{decod.}}(\mathbf{Z} = g(f(\tilde{\mathbf{X}})), \mathbf{X}) \\ &= -\lambda \log(Q(f(\tilde{\mathbf{X}}), \delta_c)) - \sum_{k=1}^d \{\mathbf{x}_k \log \mathbf{z}_k + (1 - \mathbf{x}_k) \log [1 - \mathbf{z}_k]\} \end{aligned}$$



# III Results

1. Datasets
2. (optional) Training strategies
3. Evaluation Method: node cutoffs
4. Comparison with ADAGE

# 1 Datasets

# METABRIC and TCGA compendia

- Identical to the 2015 ADAGE publication
- Two largest breast cancer gene expression data compendia: METABRIC (held by Cambridge Cancer Institute) and TCGA-BRCA (held by NIH)
- Preprocessing and normalization follows the ADAGE paper exactly
  - METABRIC:  $1992 + 144 = 2136$  samples
  - TCGA:  $525 + 22 = 547$  samples (recently updated to contain 1222 samples, but I use the old version for a fair comparison with the ADAGE paper)
  - Both contained information about breast cancer traits/subtypes
    - Basal, Her2 enriched, Luminal A, Luminal B, Normal-like
    - ER+/- signaling (whether patient benefits from endocrine therapy)



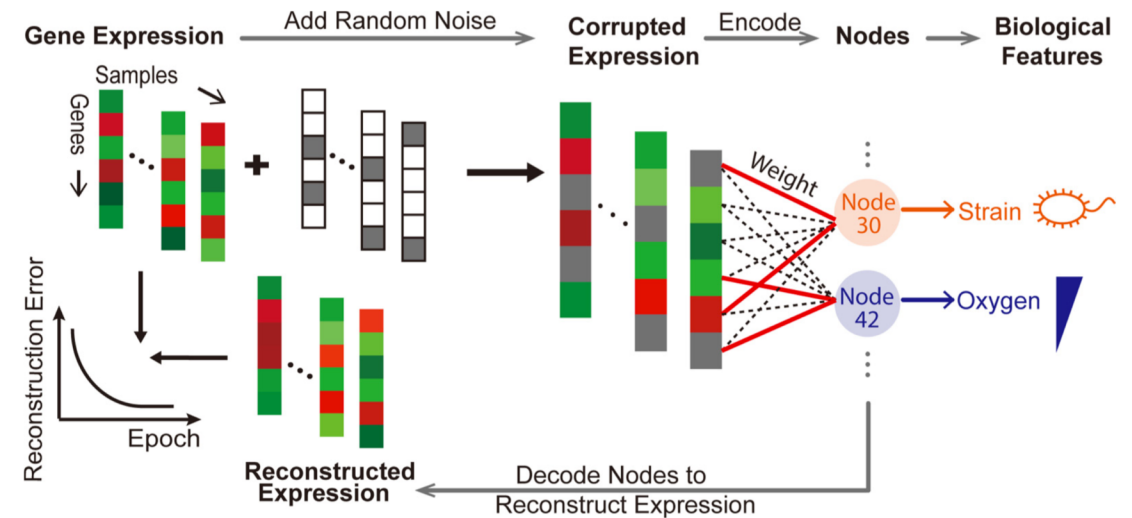
## 2 Evaluation Method

# Evaluation of Node Activations (computational)

1. Parameter tuning on the entire METABRIC dataset
2. Randomly split METABRIC into 2/3 discovery and 1/3 test sets
3. Given a label of interest, run through the discovery set:
  1. For each node in the hidden layer, find its range of activations and divide into 10 cutoffs
  2. Find the cutoff for this node that results in the best separation of all the discovery samples' labels
  3. Test the most accurate node from discovery set on the test set of METABRIC
  4. Test the most accurate node from the accuracy set on the TCGA set as validation (never seen TCGA set before)
4. Repeat 10 times and average accuracy to avoid split biases

# Evaluation of pathway discovery (biological)

- (Employed in ADAGE but not yet experimented with netDAE; will be a future task to do for netDAE)
- For given trait of interest (e.g. ER status), find the best node and retrieve the set of genes with highest weights to this node
- Run pathway enrichment analysis to see whether this gene set is biologically related to the clinical trait of interest



# 3 (Computational) Comparison with ADAGE

Table 1: Node cutoff accuracies reported in Tan *et al.*

	Tumor	ER+/-	Basal	Her2	LumA	LumB	Normal-Like
MB. discov.	0.970	0.848	0.929	0.761	0.780	0.755	0.750
MB. test	0.968	0.833	0.918	0.741	0.777	0.750	0.748
TCGA val.	0.996	0.749	0.992 (?)	0.712	0.800	0.717	0.733

Table 2: Node cutoff accuracies of netDAE (tuned on ER status only).

	Tumor	ER+/-	Basal	Her2	LumA	LumB	Normal-Like
MB. discov.	0.989	0.929	0.948	0.881	0.791	0.797	0.905
MB. test	0.987	0.880	0.932	0.880	0.774	0.780	0.905
TCGA val.	0.982	0.852	0.888	0.854	0.702	0.737	0.945

I did not have enough time to tune the hyperparameters for all the labels; therefore, I tuned with the ER status label and use the hyperparameters on all other datasets.

netDAE outperforms ADAGE for almost all labels except the one for Luminal A. The reason has not been investigated thoroughly, but I suspect that tuning the hyperparameters for Luminal A would resolve the issue. A more formal comparison would ensure ADAGE and netDAE use the exact same partitions each time and look into the t-test for statistical validity.



## IV Discussion and Future Works

1. Potential improvements of netDAE
2. Formal comparison to ADAGE using t-tests
3. Evaluate pathway enrichment of netDAE (biological evaluation)

# V Acknowledgements

- Mentor: Gil Alterovitz
- The PRIMES Program
- My parents
- All the researchers who made gene expression measurement possible and who provided inspirations to this work



Thank you for your time! Any questions?