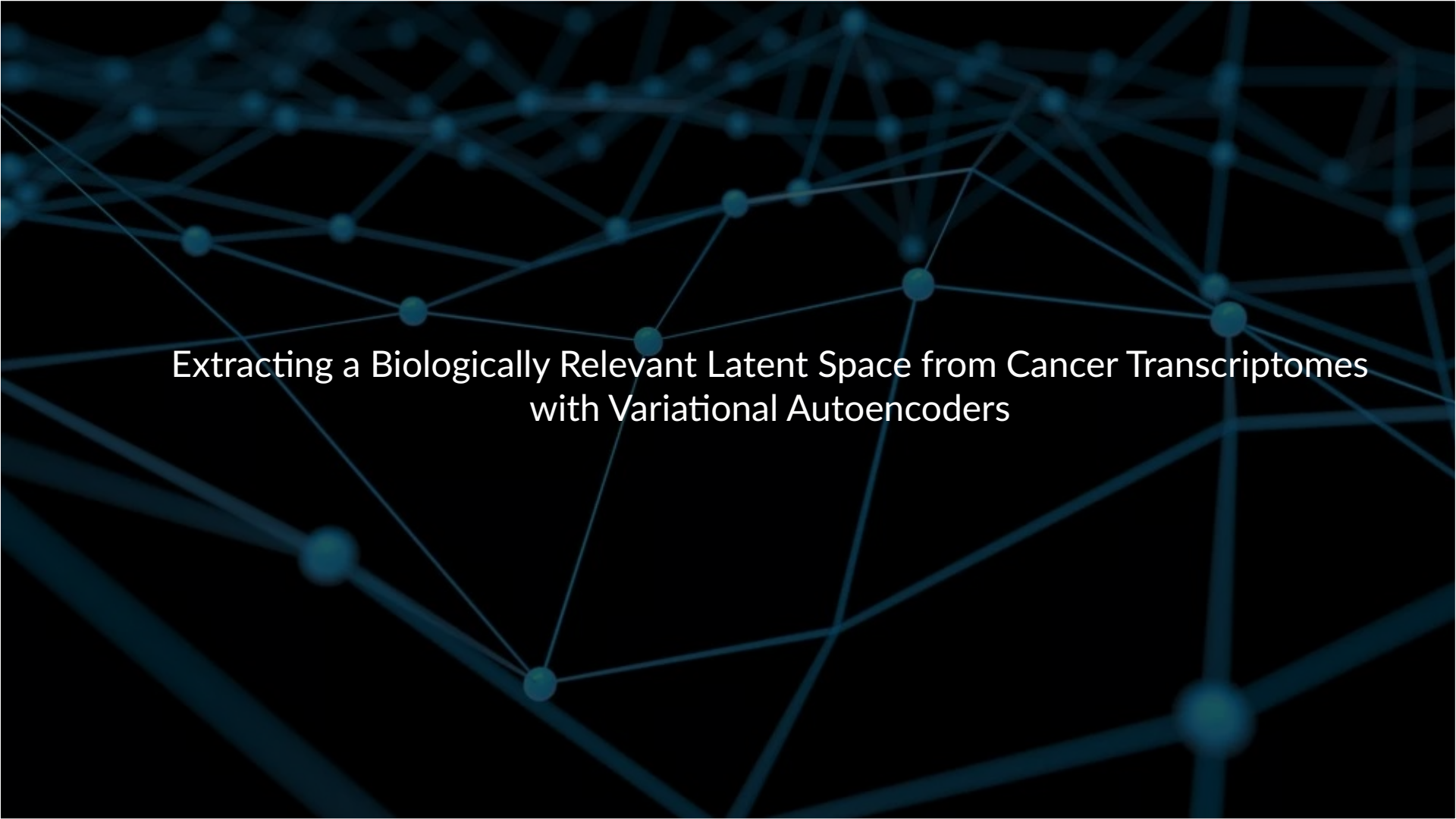
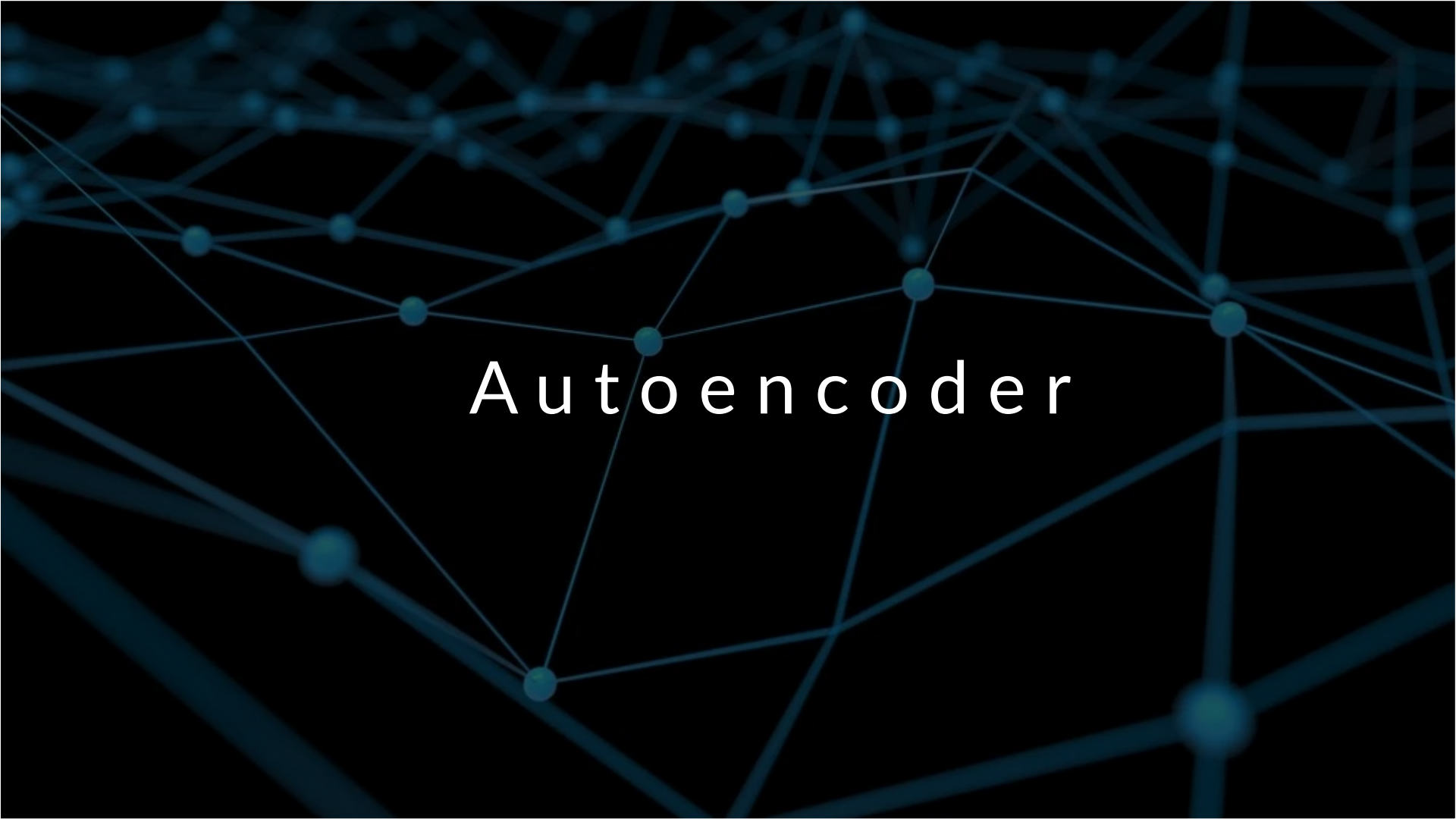
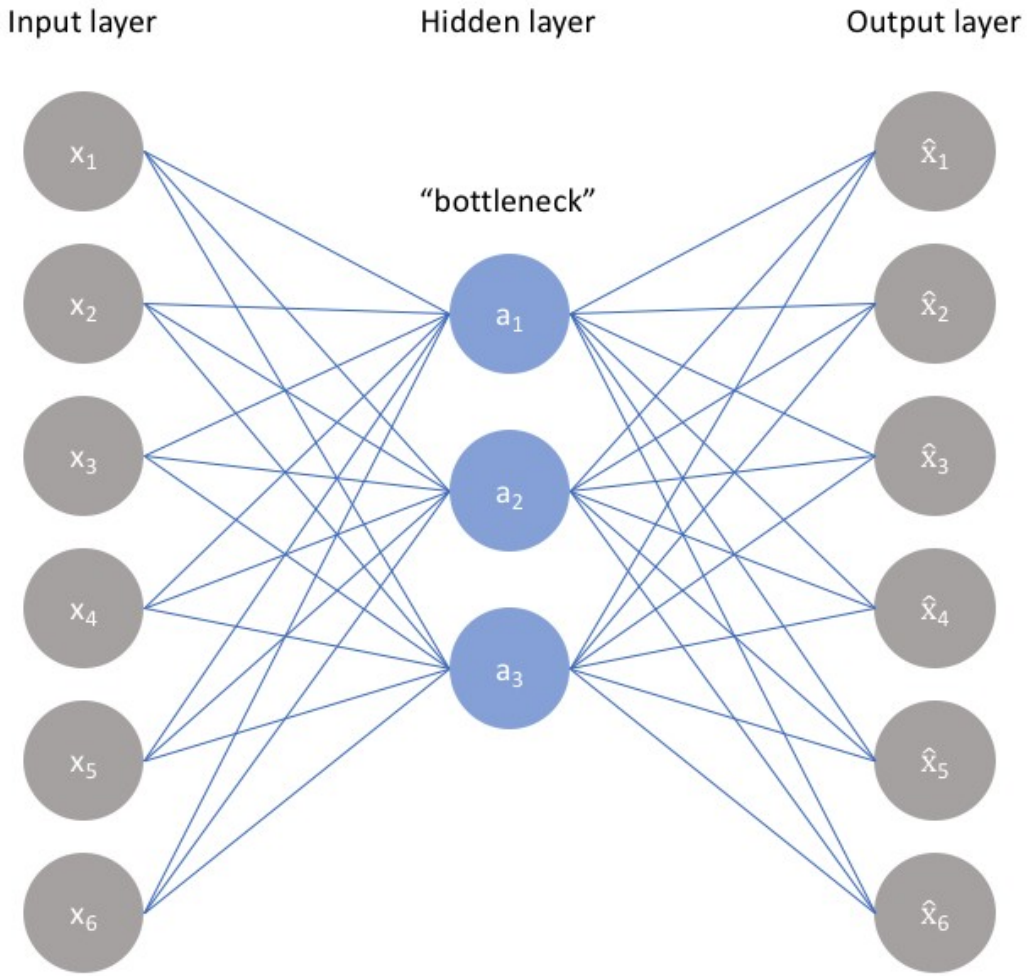
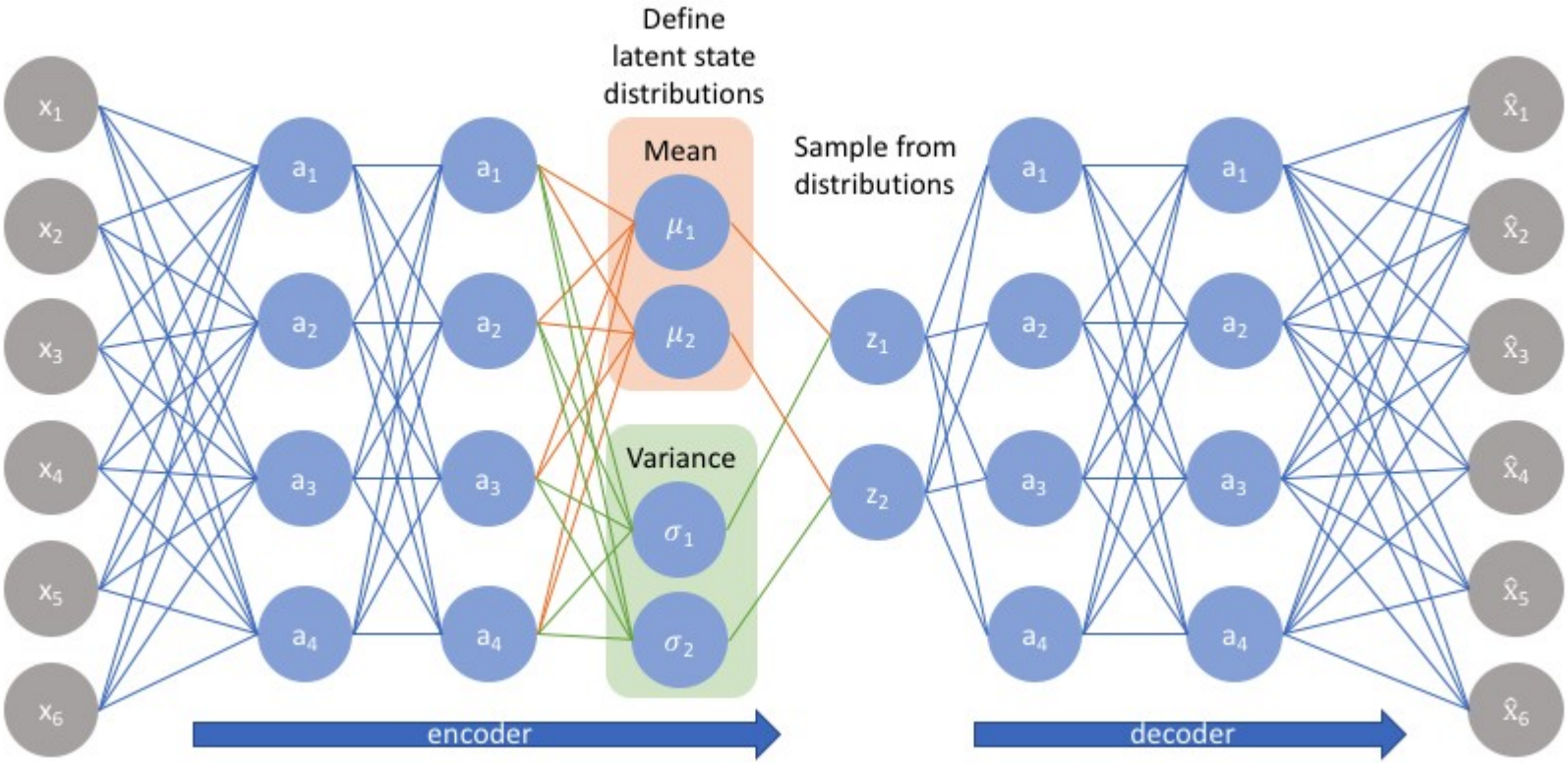

Extracting a Biologically Relevant Latent Space from Cancer Transcriptomes
with Variational Autoencoders


Autoencoder


Input layer
Hidden layer
Output layer
"bottleneck"
x_1
x_2
x_3
x_4
x_5
x_6
a_1
a_2
a_3
$\hat{x}_1$
$\hat{x}_2$
$\hat{x}_3$
$\hat{x}_4$
$\hat{x}_5$
$\hat{x}_6$

Variational Autoencoder


Define
latent state
distributions
Mean
Variance
Sample from
distributions
encoder
decoder

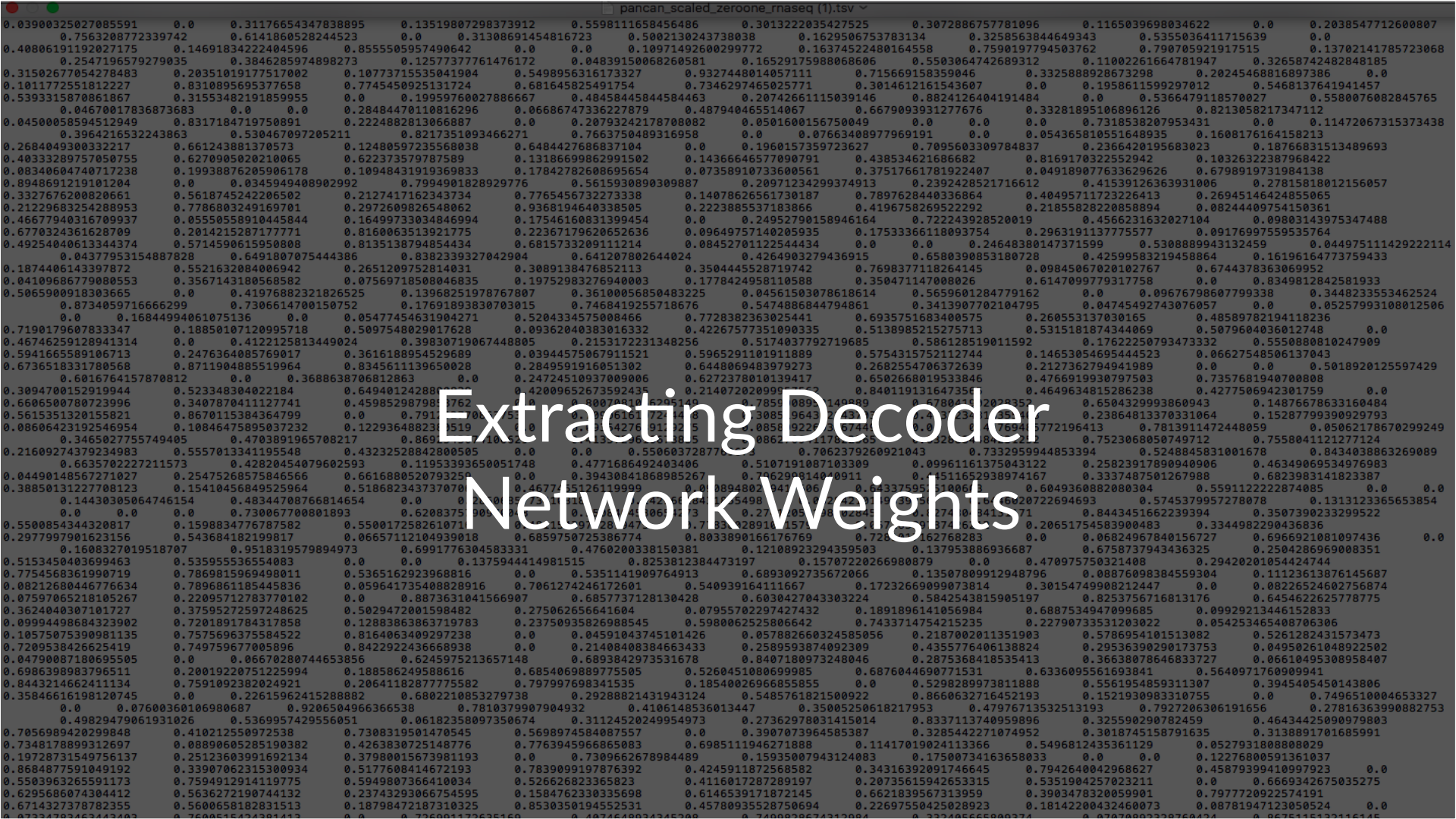
Loss Function

- Generative Loss
- Latent Loss

$$\mathcal{L}\left(x, \hat{x}\right) + \sum_{j} KL\left(q_j\left(z|x\right)||p\left(z\right)\right)$$

pancan_scaled_zeroone_rnaseq (1).tsv
Genes

pancan_scaled_zeroone_rnaseq (1).tsv
Gene Expression Level


pancan_scaled_zeroone_rnaseq (1).tsv
Extracting Decoder
Network Weights

Feature 82 → Patient Sex

In [15]:
```
# There are 17 genes with high activation in node 82
# All genes are located on sex chromosomes
sex_node_plot.head(17)
```

Out[15]:

	encoding 82	encoding 85
EIF1AY	-0.543605	0.013644
UTY	-0.514367	0.019198
CYorf15A	-0.514013	0.016753
KDM5D	-0.499488	0.019061
DDX3Y	-0.497709	0.022053
TMSB4Y	-0.495302	-0.003714
USP9Y	-0.494874	0.022164
TTTY15	-0.487078	0.014047
NCRNA00185	-0.473645	-0.019878
CYorf15B	-0.472546	0.019967
ZFY	-0.453348	0.018549
RPS4Y1	-0.452888	0.016435
NLGN4Y	-0.371427	-0.013003
DAZ1	-0.369805	-0.008085
XIST	0.316281	-0.023227
TSIX	0.298300	-0.030515
PRKY	-0.294522	0.017516

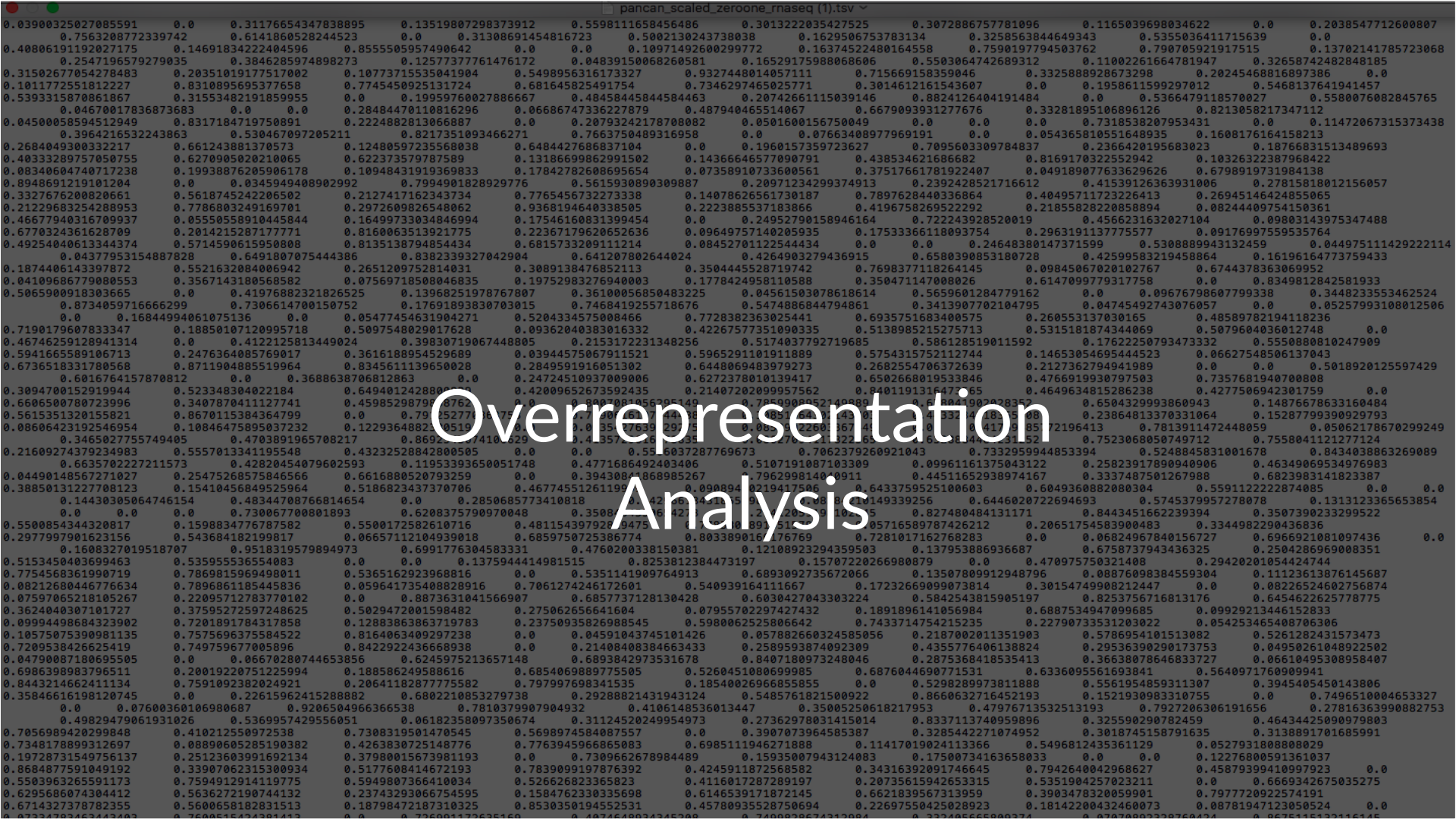
Feature 66 and 53 → Primary and Metastatic SKCM tumours

Out[18]:

genes	encoding 66	direction
ROPN1	0.115224	positive
ROPN1B	0.108729	positive
ADIPOQ	0.107172	positive
SOX10	0.105733	positive
CIDEA	0.103107	positive

Out[19]:

genes	encoding 53	direction
HCG22	0.084451	positive
RXRG	0.081329	positive
MGAT4C	0.076966	positive
PLP1	0.076440	positive
C10orf90	0.076005	positive


pancan_scaled_zeroone_rnaseq (1).tsv
Overrepresentation
Analysis

GO terms

geneset	description	link	C	O	E	R	PValue	FDR	overlapGene	OverlapGene_UserID
GO:0030300	regulation of intestinal cholesterol absorption	http://amigo.geneontology.org/amigo/term/GO:0030300								5
GO:1904729	regulation of intestinal lipid absorption	http://amigo.geneontology.org/amigo/term/GO:1904729								5
GO:0006069	ethanol oxidation	http://amigo.geneontology.org/amigo/term/GO:0006069	7	3	0.0731343283582					
GO:1904478	regulation of intestinal absorption	http://amigo.geneontology.org/amigo/term/GO:1904478	7	3						
GO:0032371	regulation of sterol transport	http://amigo.geneontology.org/amigo/term/GO:0032371	20	4	0.20895					
GO:0032374	regulation of cholesterol transport	http://amigo.geneontology.org/amigo/term/GO:0032374	20	4						
GO:0016042	lipid catabolic process	http://amigo.geneontology.org/amigo/term/GO:0016042	96	7	1.0029850746268					
GO:0006067	ethanol metabolic process	http://amigo.geneontology.org/amigo/term/GO:0006067	8	3	0.08358					
GO:0030299	intestinal cholesterol absorption	http://amigo.geneontology.org/amigo/term/GO:0030299	9	3						
GO:0098856	intestinal lipid absorption	http://amigo.geneontology.org/amigo/term/GO:0098856	9	3	0.09402					

HO
OH
NH2
OCH3
OCH3
HO
HN
H3CO
Cb =pH[H+]
[OH-]

5 Key Takeaways

- VAE's are generative models that allows for reconstruction of original input data as well as identification and recapturement of variations/patterns within the data.

- The weights in the decoder network are able to learn which genes contributed to a particular feature.

- Through the decoder network, we are able to understand that the VAE was able to construct features representing patient sex and another capturing SKCM tumours.

- The genes with the highest weight are the most expressed meaning they have a higher chance of being the cause of the state of a cancerous tumour.

- We are able to identify the functional processes of the highly expressed genes within a feature, using ORA. This can help researchers identify which particular gene products or pathways to target when curing for a disease.

Discussion Questions

Pro:

- The papers presented in #NIPS2017
- Potential for discovering genetic/biological patterns for certain conditions (strong enough conclusion?)
- The manifolds of the VAE provide the potential for representing differing pathway activations, transitions between cancer states, or indicating particular tumors vulnerability to specific drugs; versatile applications? Transfer learning potential?
- This is a open source project and for code and dataset of this paper search /Greenelab/Tybalt
- Very simple VAE structure
- The model has two direct benefits of modeling cancer gene expression data.

  Automatically engineer non-linear features

  Learning the reduced dimension manifold of cancer expression space
- Side note: VAEs are different from deterministic autoencoders because of the added constraint of normally distributed feature activations per sample. This constraint not only regularizes the model, but also provides the interpretable manifold.

Cons:

- Is the genetic expression of gender found in this paper enough to validate the hypothesis on genetic expression of disorders?
- What are the evaluation/validation techniques are used by the authors (sounds like pattern matching from select examples)?
- How would validation need to take place for more extensive research?
- Why and how normalization happened and how we can measure it effects
- Predefined type of cancers in latent space!!! Very selective choices
- Measuring the relation of the depth of VAE to latent space size and interpretability of extracted features