APPLICATIONS OF MACHINE LEARNING IN HEALTHCARE

Yves Greatti

AGENDA

- Viome: Health Related Biological Pathway Analysis
- VAEN: Anti Cancer Drug Response Prediction
- SARD: End-Of-life Patient Prediction
- Time Series Forecasting

ABOUT ME

- Masters
  - Master's Degree in Computer Sciences
  - Master's Degree in Mathematics
  - Enrolled in Bioengineering Master at JHU
- In my last position
  - Analyzed multi-omics data sets, including next-generation sequencing data
  - Investigated biological pathways using pathway analysis methods
- Data Scientist
  - Analyzed different data modalities
  - Worked with large-scale datasets using cloud computing
  - Contributed to development of ML models
  - Implemented highly scalable and fast software applications
  - Brought models from conception to production
  - Worked with statistical models for complex datasets including RNA-Seq data, effectively measuring the goodness of fit
- https://github.com/ygrepo

VIOME

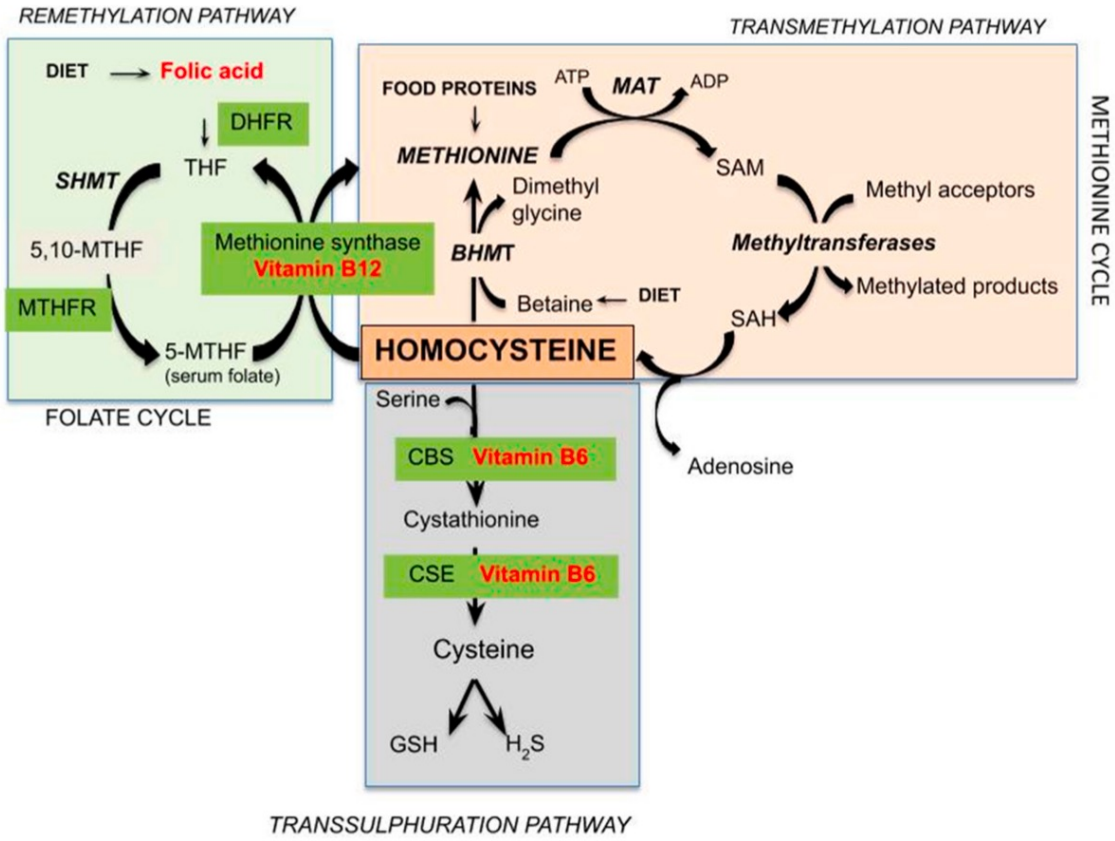
BACKGROUND

SCORING HEALTH RELATED BIOLOGICAL PATHWAYS

BACKGROUND

- A pathway is activated when a sufficient number of genes are activated
- Pathways are a conceptualization of the understanding of cell biology and their boundaries are arbitrary

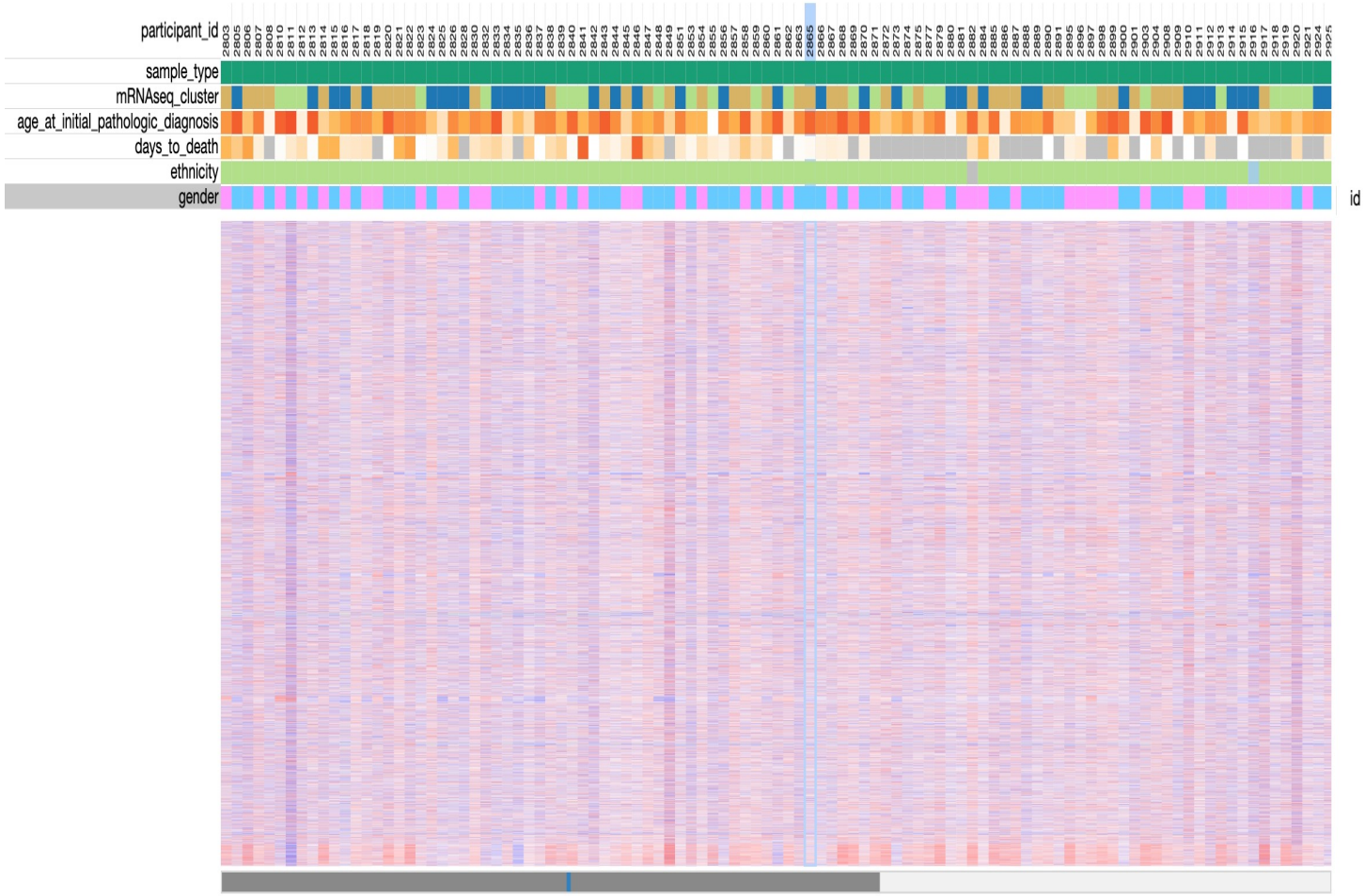
Example: Homocysteine Pathway



Azzini, et al. IJMS 2019

BACKGROUND

- High-dimensional data
- Data is sparse



Source: The Broad Institute, Morpheus, Glioma Cohort

WORKFLOW FOR RNA-SEQ COMPUTATIONAL ANALYSES

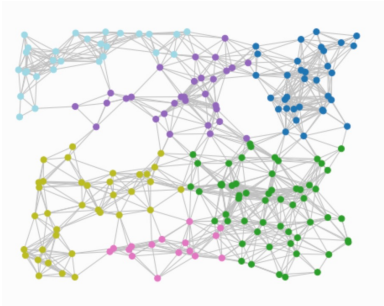
1. Preprocessing

- ✓ Adjust for batch effect
- ✓ Low raw count removal
- ✓ Low variance genes filtering
- ✓ Outlier Detection
- ✓ Normalization within sample

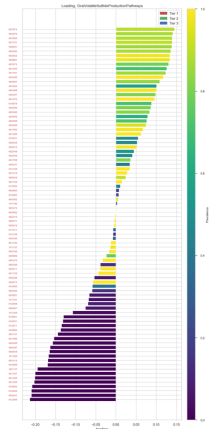
2. Gene Expression Data

Sample Gene	S1	S2	...	S200
ENSGID1				
ENSGID2				
...				
ENSGID100				

3. Network construction

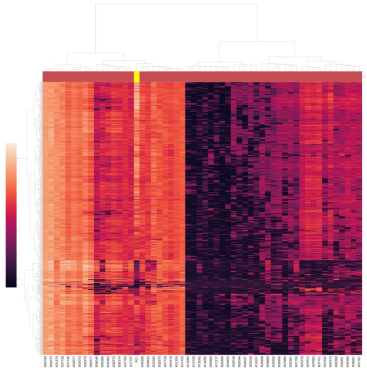


Compute a ranking of highly differentially expressed genes in each cluster (MWW test)



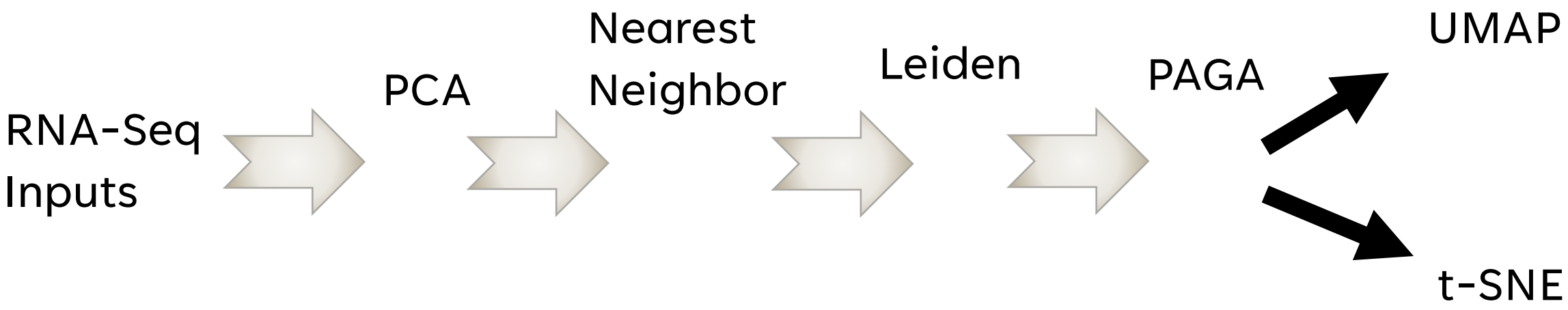
A correlation plot between genes on Y-axis and first principal component, the PCA loadings are on the X-axis

4. Validation



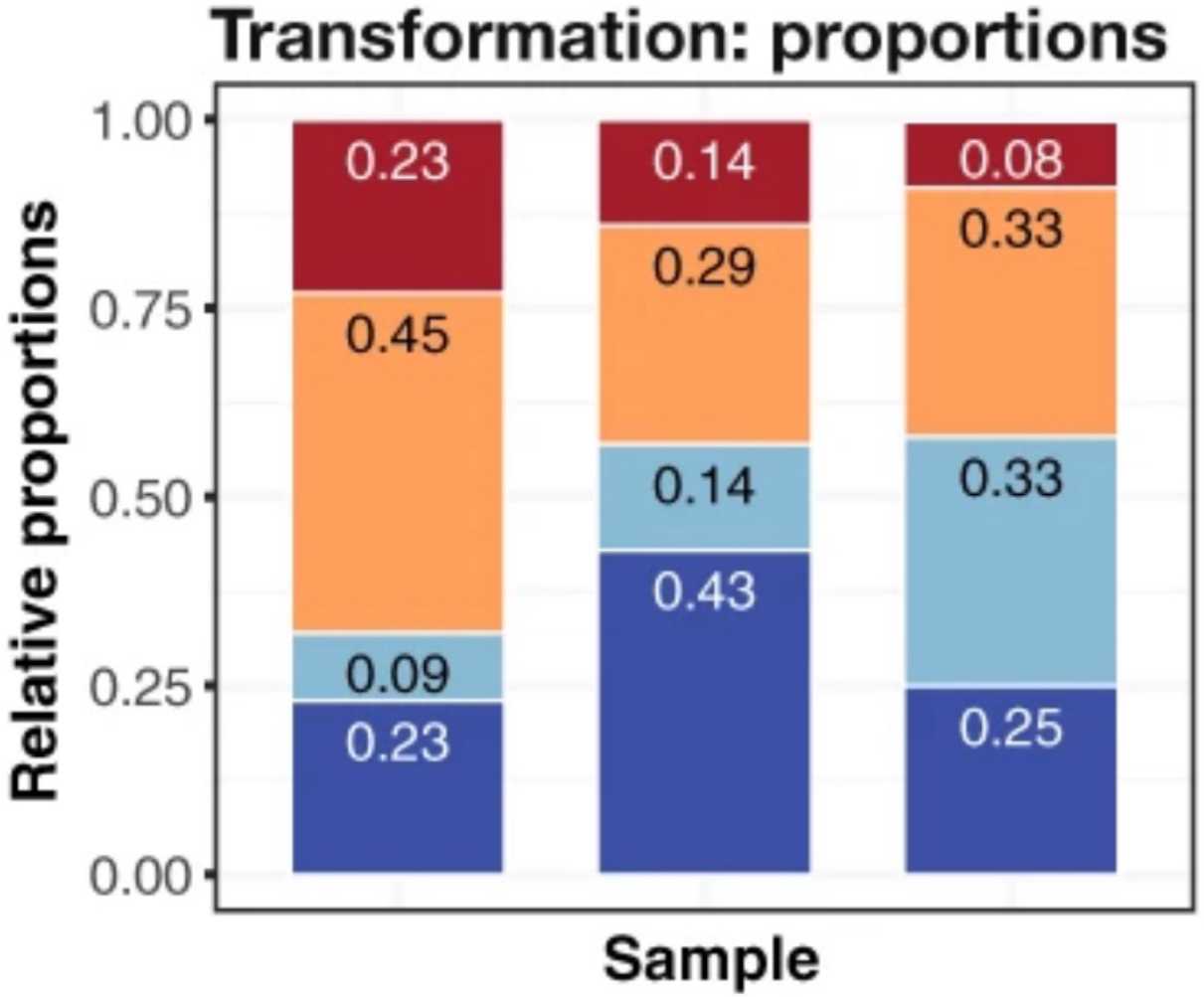
A heatmap display with each gene represented by a column, and each row giving the data from one consumer

WORKFLOW FOR RNA-SEQ COMPUTATIONAL ANALYSES



ANALYSIS DETAILS

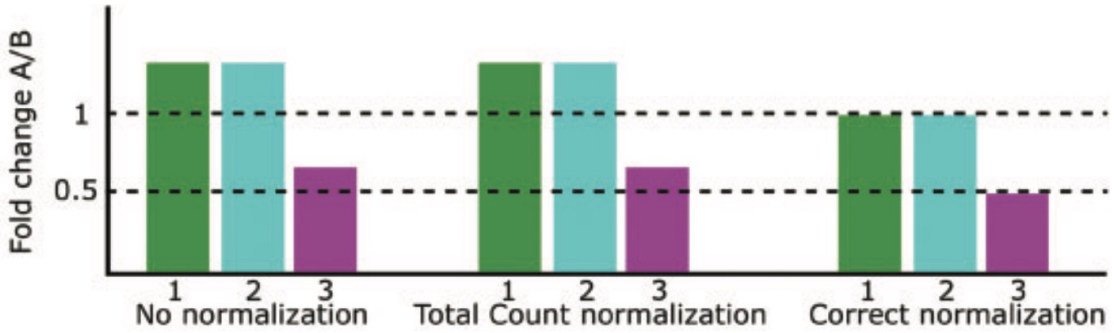
MICROBIOME DATA TRANSFORMATION



- Taxa and bacterial functions were filtered so that only those with more than 10 reads in at least five samples were selected
- Raw counts transformed using centered log-ratio transformation

RNA-SEQ DATA TRANSFORMATION

- Only genes with more than 10 reads in at least 2 samples were selected
- Genes with less than 25% variances were also filtered
- RNA-Seq data was normalized using trimmed mean of M-value and converted to log2(CPM)

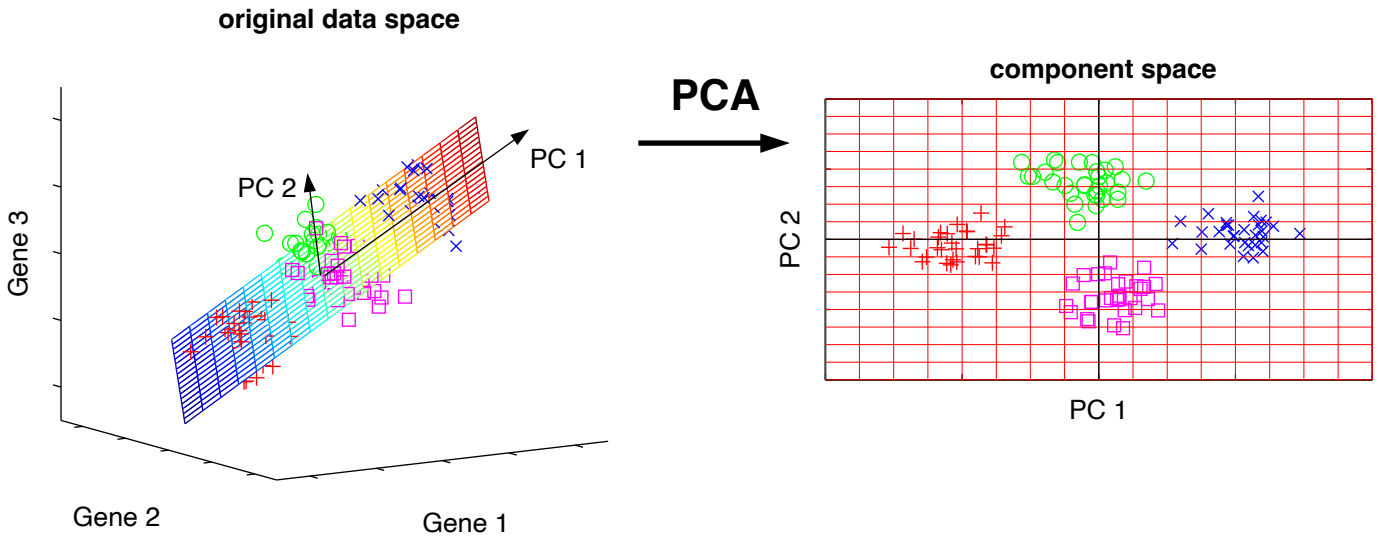


Difference in read counts from technical variability

SCORING DETAILS

PCA

- The effects of a pathway can be captured by a small number of ”super genes”
- The super genes may correspond to the linear combinations of genes that explain the variations of gene expression



VALIDATION

- Mixed-effect models were created
  to validate molecular scores with health related scores

- Additional feature extraction models were used
  to relate saliva-based genes to gut molecular scores (domain transfer)

MODEL RESULTS

- About 60 molecular scores were created within two months
- Significant results for some of the scores (significant reduction in depression following recommendation, p-values < 0.001)
- New biological scores were scheduled to be delivered
- Scoring methodologies were to be described in research papers

"I've been implementing my specific recommendations for months now and have noticed a tremendous difference in my energy level and overall gut function. I would recommend Viome to everyone as it provides tools and insight for all of us to come into balance and vibrancy to a healthy lifestyle."
Alice, Viome Customer

"I was blown away by my results & have started eating the suggested foods! I feel that my inflammation is going down - I’m happy!!!"
Rhonda, Viome Customer

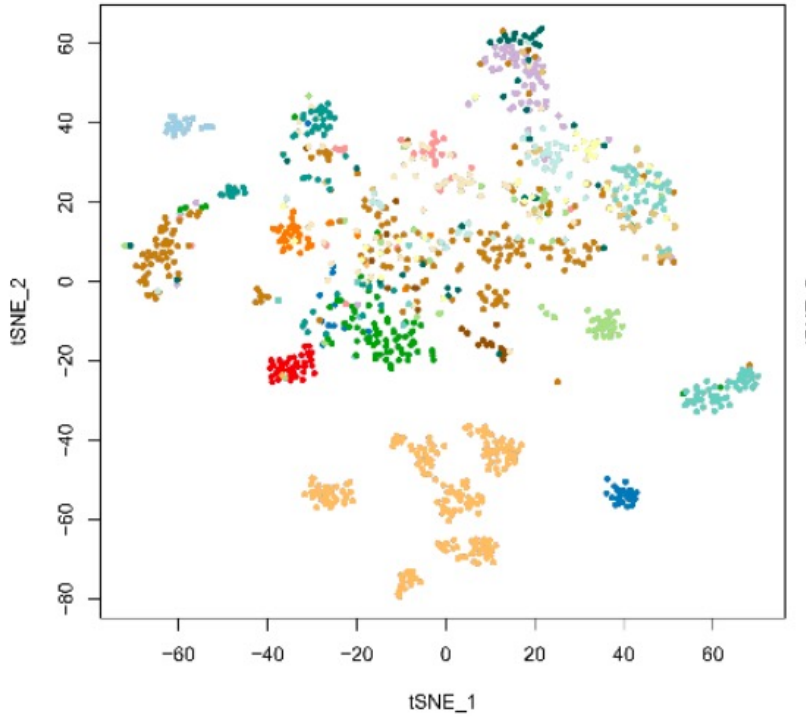
"...Can’t put a price on good health."
Margareta, Viome Customer

UTHealth Houston School of Biomedical Informatics

BACKGROUND

IMPROVE THE UNDERSTANDING OF HOW GENETIC ALTERATIONS ARE ASSOCIATED WITH DRUG RESPONSE

BACKGROUND

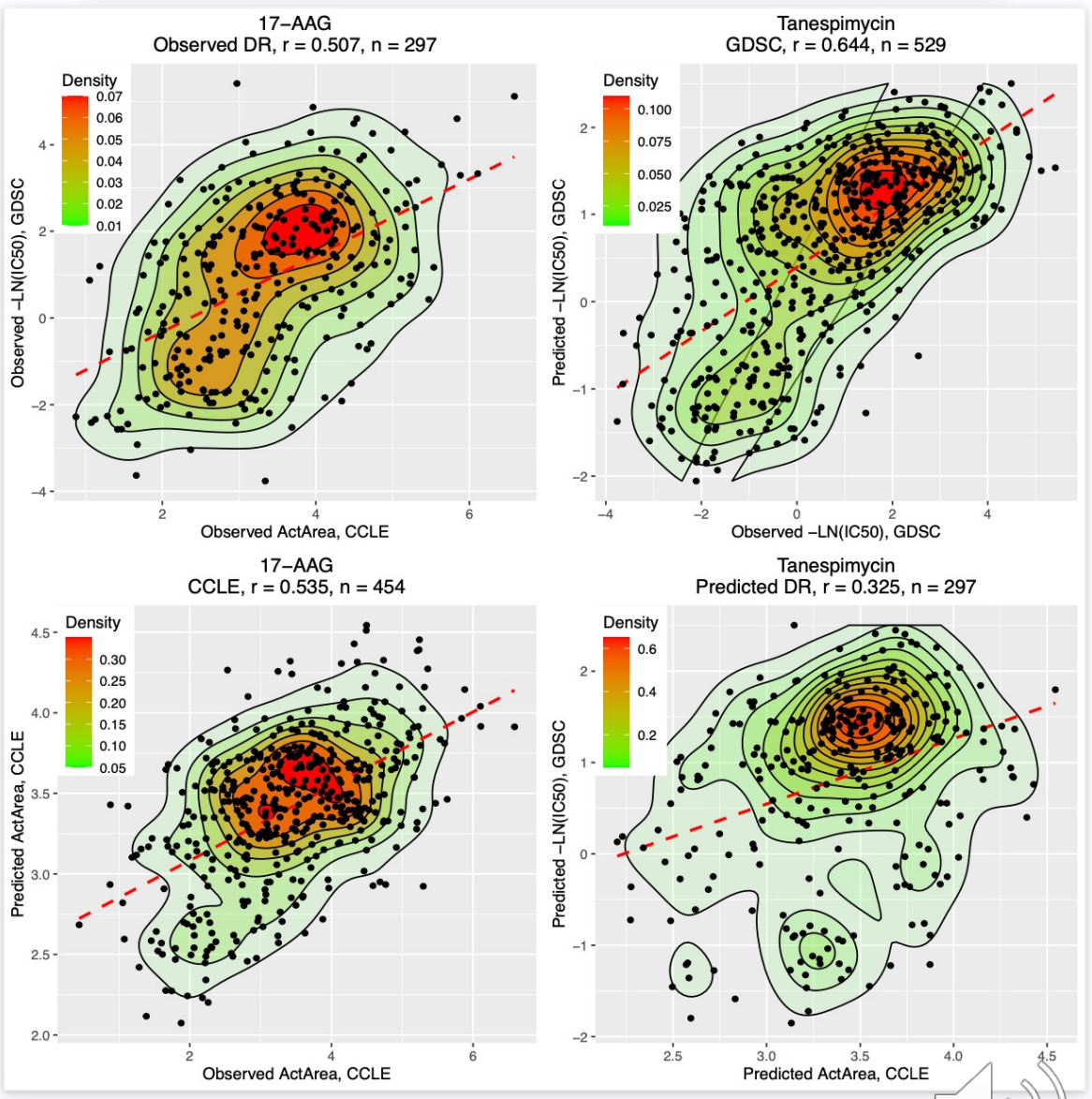


t-SNE plot showing the distribution of cell lines with their tissue origins – Jia et al, Nature 2021

- 1100 cell lines from 19 tissues
- Each tissue required to have at least 20 cell lines
- Filter genes with low expression variability
- RNA-Seq data was transformed using log2(RPKM + 1)

BACKGROUND

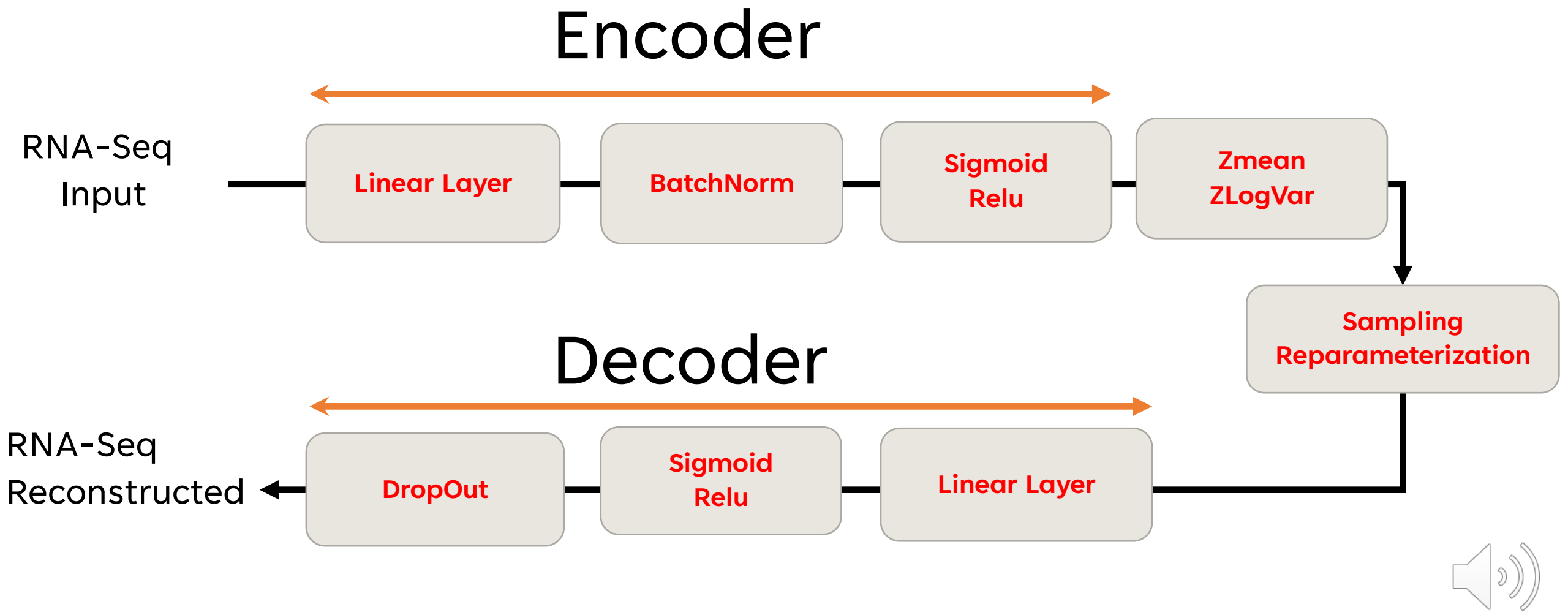
- Variational Autoencoder models followed by Elastic Net regression to predict drug response from CCLE, GDSC and TCGA
- For each drug A-model (all cell lines) and S-model (solid tumor)
- Model efficiency evaluated using R^2 on holdout samples



Predicted vs observed drug response – Jia et al, Natu[illegible] 2021

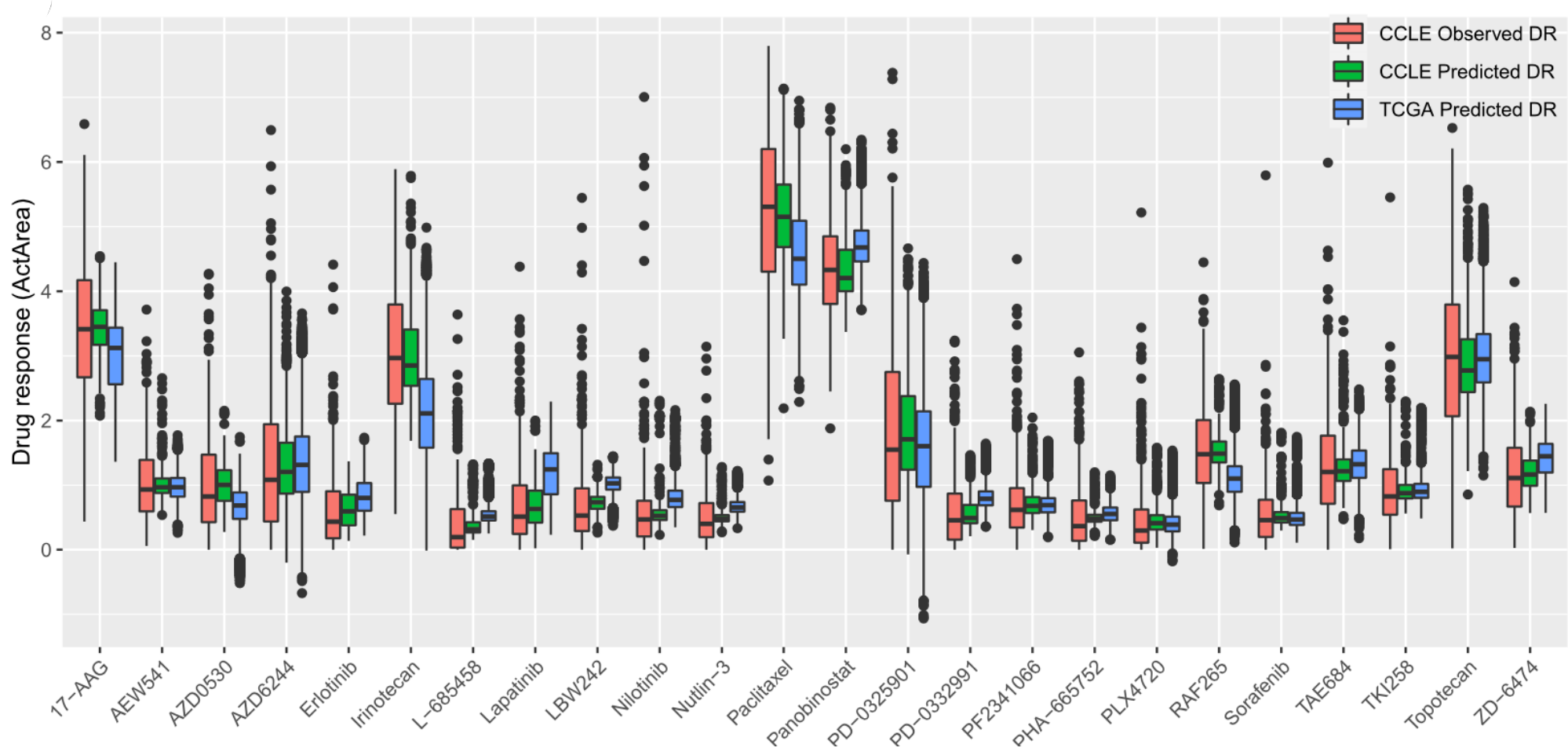
MODEL DETAILS

VAE MODEL


Encoder
RNA-Seq
Input
Linear Layer
BatchNorm
Sigmoid
Relu
Zmean
ZLogVar
Sampling
Reparameterization
Decoder
RNA-Seq
Reconstructed
DropOut
Sigmoid
Relu
Linear Layer

MODEL VALIDATION

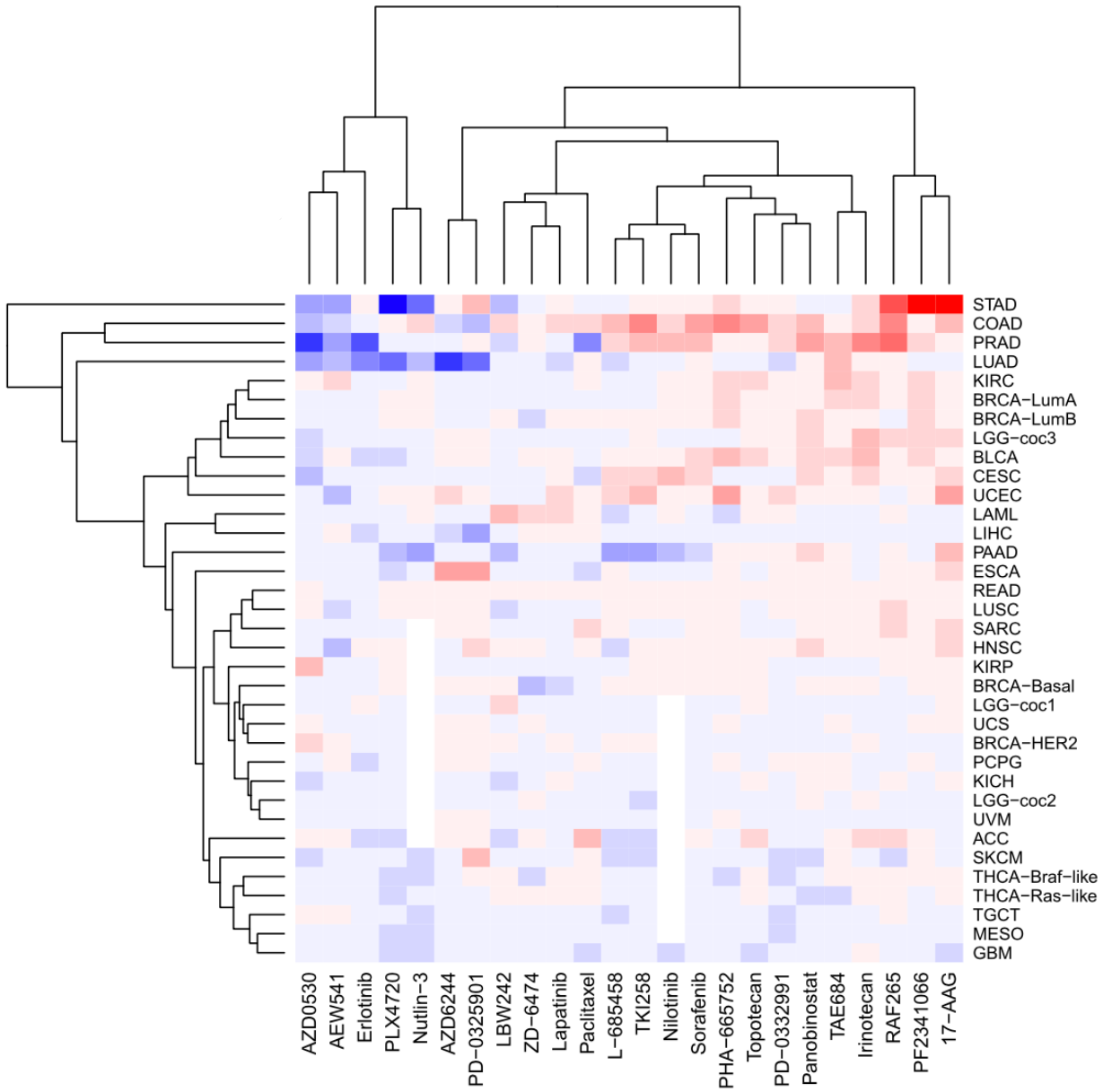
PREDICTED RESPONSE SIMILAR TO OBSERVED DATA



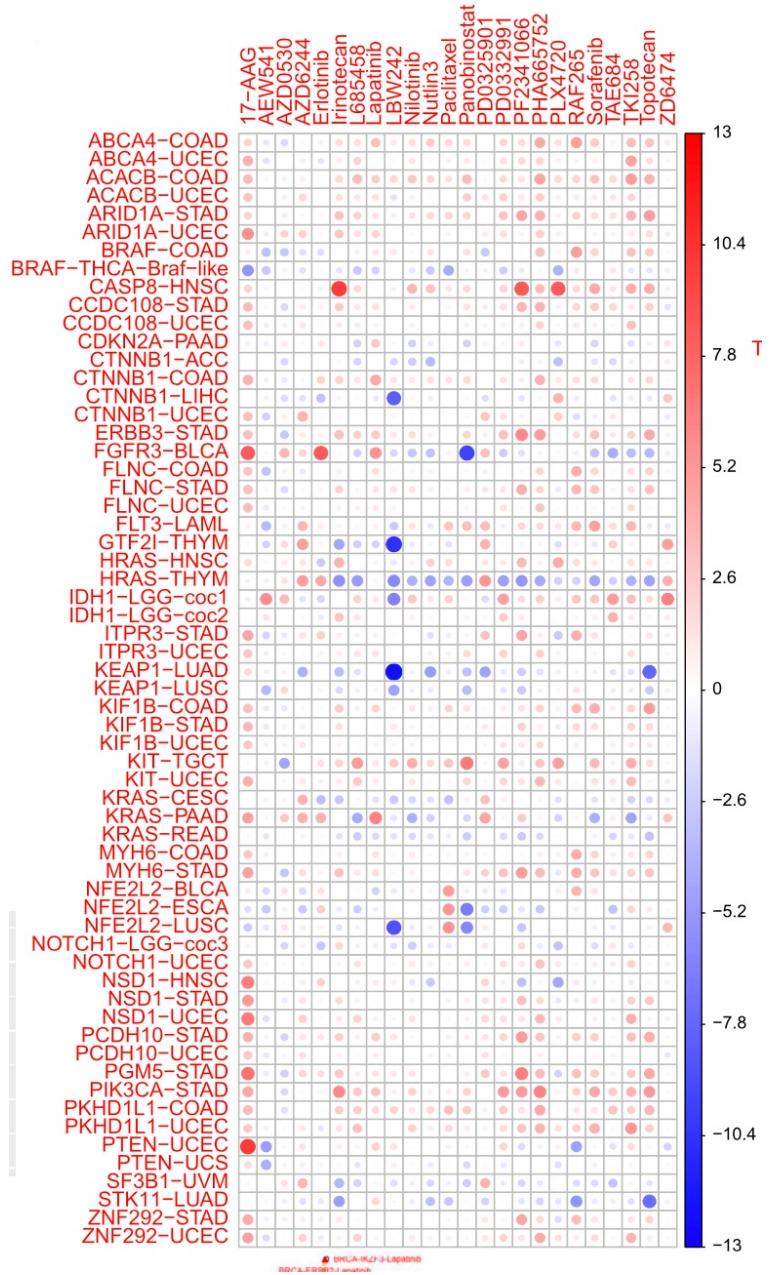
Distribution of the observed and predicted drug response – Jia et al, Nature 2021

MODEL RESULTS

Significant drug-TMB or drug-gene association



Drug-TMB association– Jia et al, Nature 2021



Drug-gene association – Jia et al, Nature 2021

Association of somatic mutation with drug response

Mutation	Drugs
ARID1A (SNF/SWI complex)	Topotecan
FLT3 (mutation cluster at 594-605)	Sorafenib
HRAS, KRAS, NRAS (G12/G13 mutations)	AZD6244, PD-0325901
Amplification on chromosome 7 (EGFR)	17-AAG, RAF265
Amplification on chromosome 8 (MYC)	17-AAG, AZD0530, TAE684
WWOX deletion	Azd6244, Paclitaxel, PD-0325901, TAE684

DRUG RESPONSE ASSOCIATED WITH TME

- Negative correlation between increase CAF/ECM gene expression and sensitivity to lapatinib (HER2 inhibitor)
- Strong association between TIS genes and irinotecan, nilotinib, PHA665752, PLX4720, and RAF265 indicating that patients with a high-level of T-cell inflammation might be more sensitive to these drugs
- EGFR signaling pathways confounding factor of the association between BRAF mutations and response to BRAF inhibitors

DISCUSSION

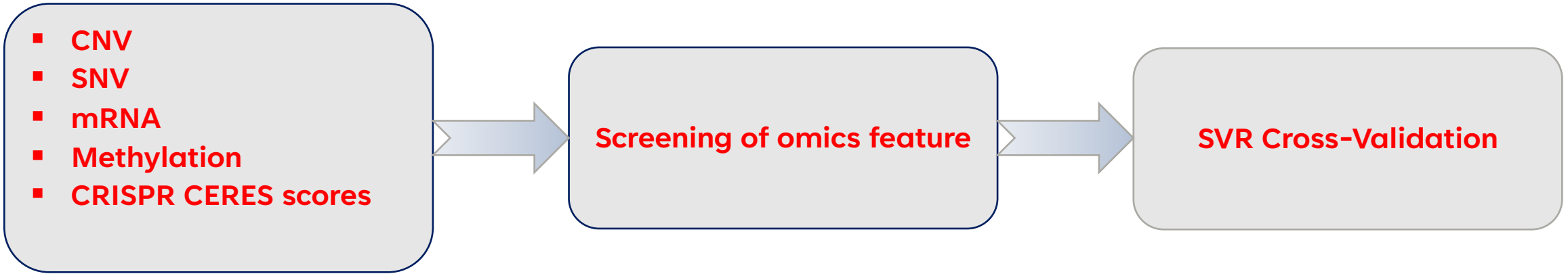
DROEG

A method for cancer drug response prediction based on omics and essential genes integration

BACKGROUND

- Use standard ML methodology: Support Vector Machine for Regression (SVR)
- Drug response prediction model integrating genomic, transcriptomic, methylation data and CRISPR essential genes(CERES score)
- CCLE, GDC, and TCGA Pan-GI patient data
- Filter cell lines with more than 20% missing drug sensitivity values

BACKGROUND


CNV
SNV
mRNA
Methylation
CRISPR CERES scores
Screening of omics feature
SVR Cross-Validation

MODEL PERFORMANCE COMPARISON

Method	Input	Datasets	PCC
DROEG	**CNV, mRNA, mutation, Methylation, CRISPR**	**GDSC**	87.85% > 0.5
		CCLE	[0.55, 0.73]
VAEN	**CNV, mRNA, mutation**	**GDSC**	80.88% > 0.5
		CCLE	[0.33, 0.77]

Wu et al. Briefings in Bioinformatics 2023

DRIVER GENES AND MUTATIONS ARE "ACTIONABLE"

A genomic alteration can be considered "actionable" if it:

1. Affects the function of a cancer-related gene and can be targeted directly or indirectly with approved or investigational therapies
2. Predicts therapy response (sensitivity or resistance)
3. Has demonstrated the ability to establish diagnosis or influence prognosis

Meric-Bernstam, et al. JNCI 2015

OUTCOME DETERMINANTS

- Need to include clinical and epidemiological data (age, gender, smoker status, BMI)
- Disease characteristics (histology, mutation status, TNM status, tumor extrinsic and intrinsic features like HLA aberrations)
- Treatment data (treatment line, toxicity)
- Blood counts: neutrophil, lymphocyte counts
- MRI radiomics
- Immune System Pathway

NETBIO

Identify biological pathways located to immunotherapy

target in a PPI network to predict immunotherapy response

Kong et al., Nature 2022

NETWORK-BASED BIOMARKERS (NETBIO)

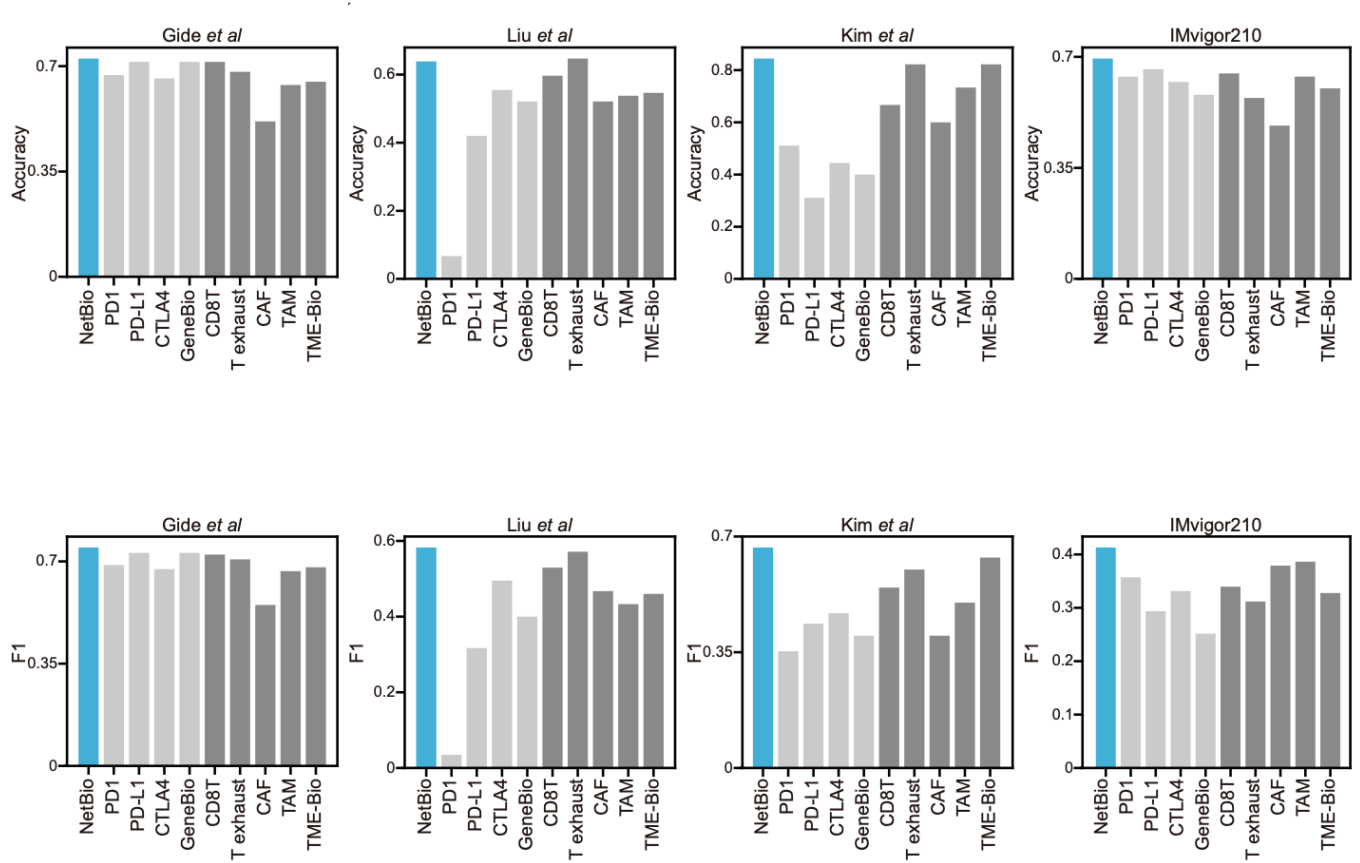
- Use STRING protein-protein interaction (PPI) network (STRING score > 700)
- Apply network propagation (page rank) using PD1 and PD-L1 targets (CD274, PDCD1)
- Select gene with high influence scores (top 200) and identify biological pathways (Reactome) enriched with these genes
- Use the expression values from these biological pathways to predict immunotherapy response

NETWORK-BASED ML APPROACH TO PREDICT IMMUNOTHERAPY RESPONSE IN CANCER PATIENTS

- 8 melanoma-treated cohorts (nivolumab, pembrolizumab, atezolizumab)
- Higher accuracy on identified genes than drug response predictions based on PD1, PD-L1, cytotoxic T-lymphocyte antigen 3 (CTLA4) and markers associated with TME (CD8 T cell, T-cell exhaustion, cancer tumor-associated fibroblast (CAF), and tumor-associated macrophage (TAM))
- Combining identified gene expression levels and TMB improve the prediction of the overall survival in immune checkpoint inhibitor (ICI) treated bladder cancer patients

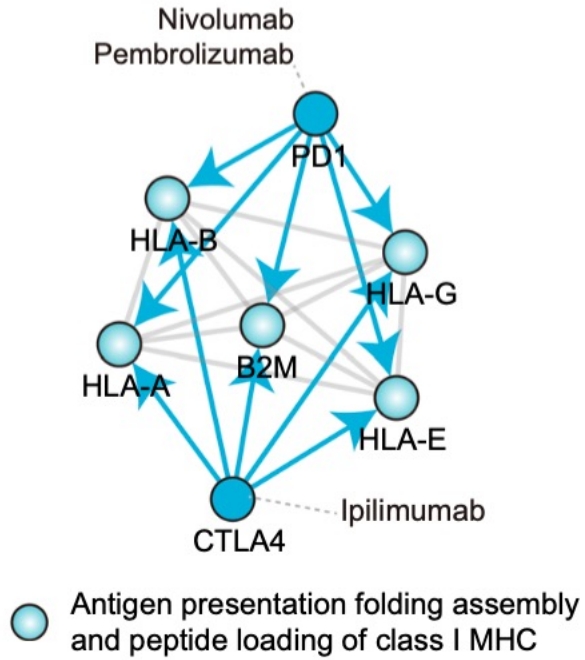
Kong et al., Nature 2022

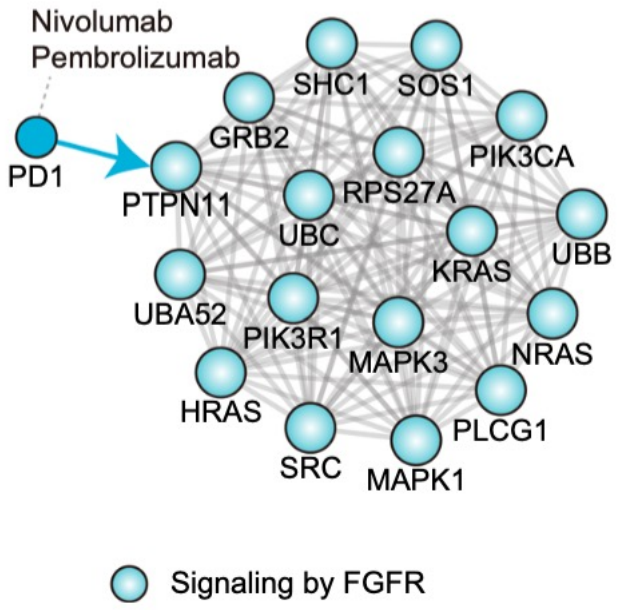
IMPROVED PERFORMANCE WITH NETWORK-BASED GENES



Kong et al., Nature 2022

BIOLOGICAL PATHWAYS ASSOCIATED WITH IMMUNOTHERAPY RESPONSE





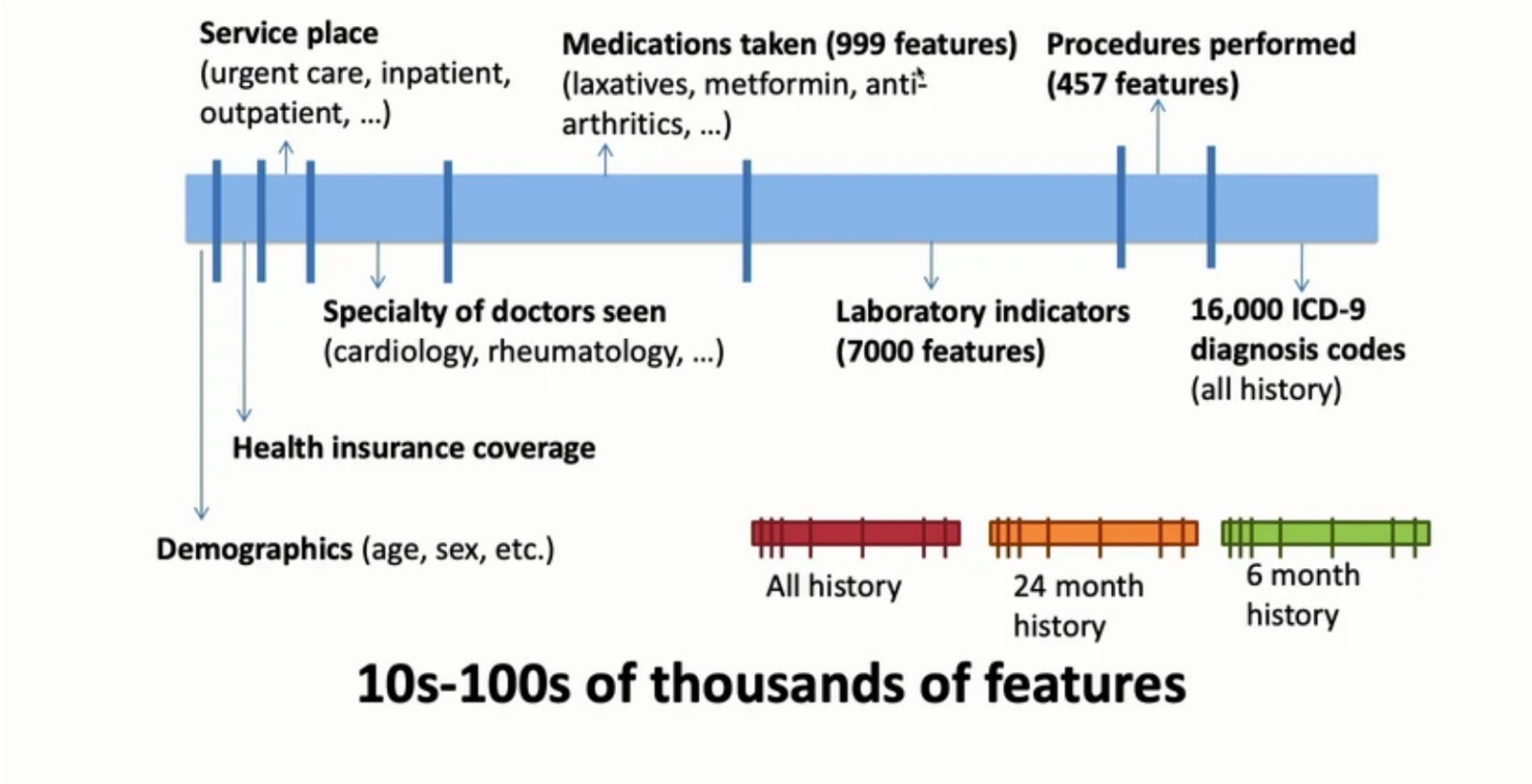
"Antigen presentation folding assembly and peptide loading of class I MHC"
and FGFR signaling pathways have the highest positive correlation with CD8 T-cell proportions

Kong et al., Nature 2022

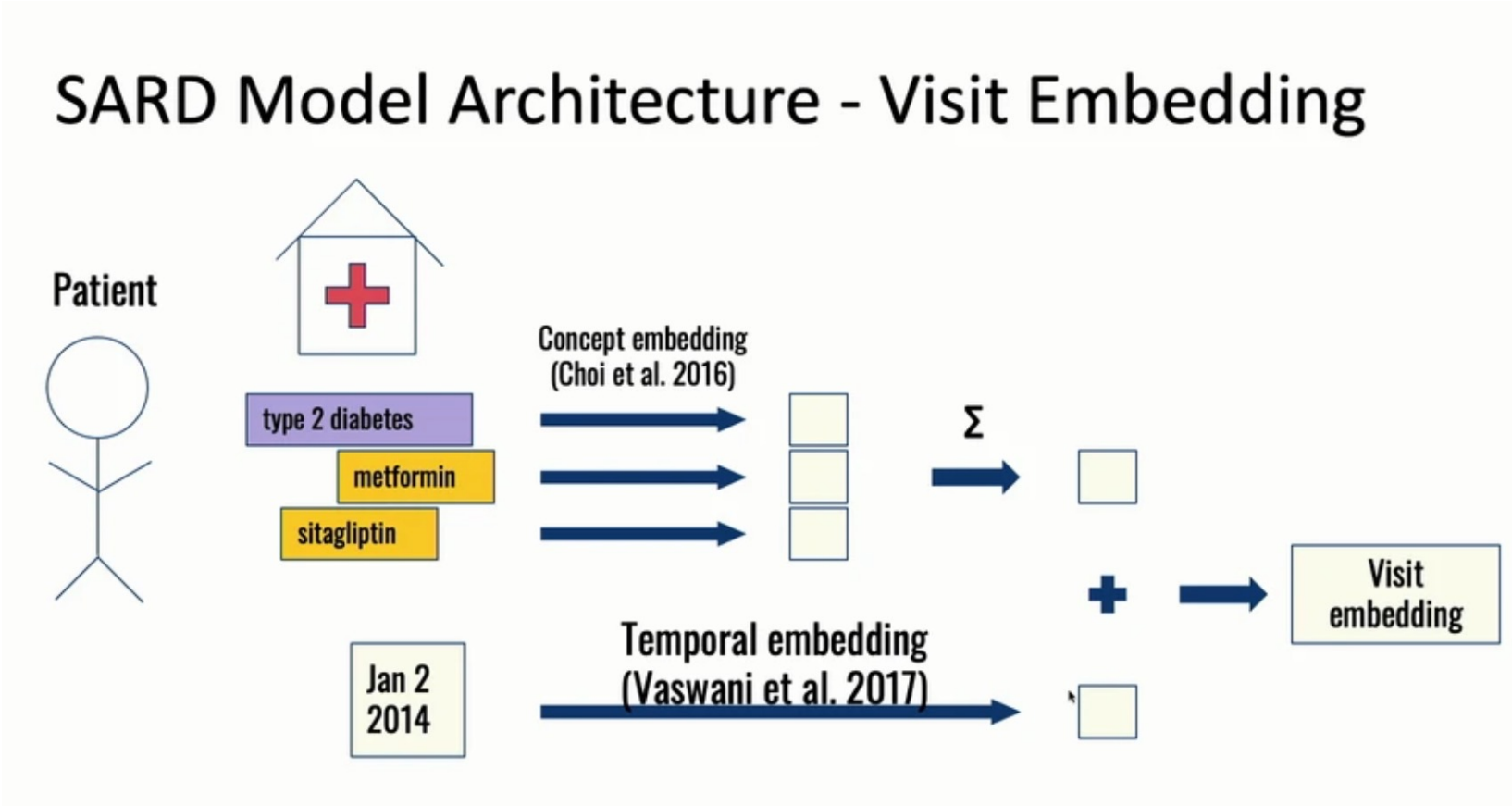
BACKGROUND

Predicts end-of-life for a patient using longitudinal clinical datasets

BACKGROUND

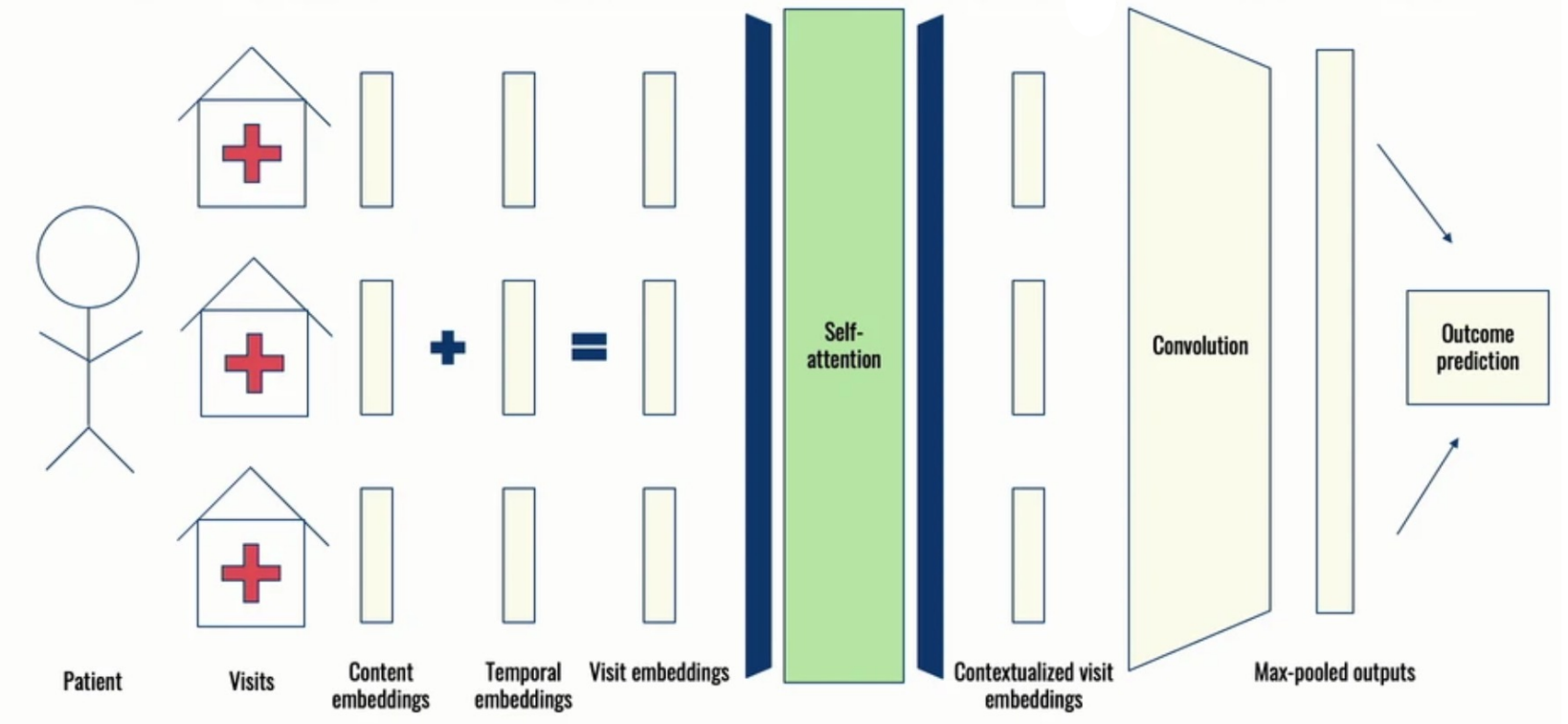

Service place
(urgent care, inpatient, outpatient, ...)
Medications taken (999 features)
(laxatives, metformin, anti-arthritics, ...)
Procedures performed
(457 features)
Specialty of doctors seen
(cardiology, rheumatology, ...)
Laboratory indicators
(7000 features)
16,000 ICD-9 diagnosis codes
(all history)
Health insurance coverage
Demographics (age, sex, etc.)
All history
24 month history
6 month history
10s-100s of thousands of features

BACKGROUND


SARD Model Architecture - Visit Embedding
Patient
type 2 diabetes
metformin
sitagliptin
Concept embedding
(Choi et al. 2016)
Σ
Jan 2
2014
Temporal embedding
(Vaswani et al. 2017)
Visit
embedding

MODEL DETAILS

SARD MODEL


Patient
Visits
Content embeddings
Temporal embeddings
Visit embeddings
Self-attention
Contextualized visit embeddings
Convolution
Max-pooled outputs
Outcome prediction

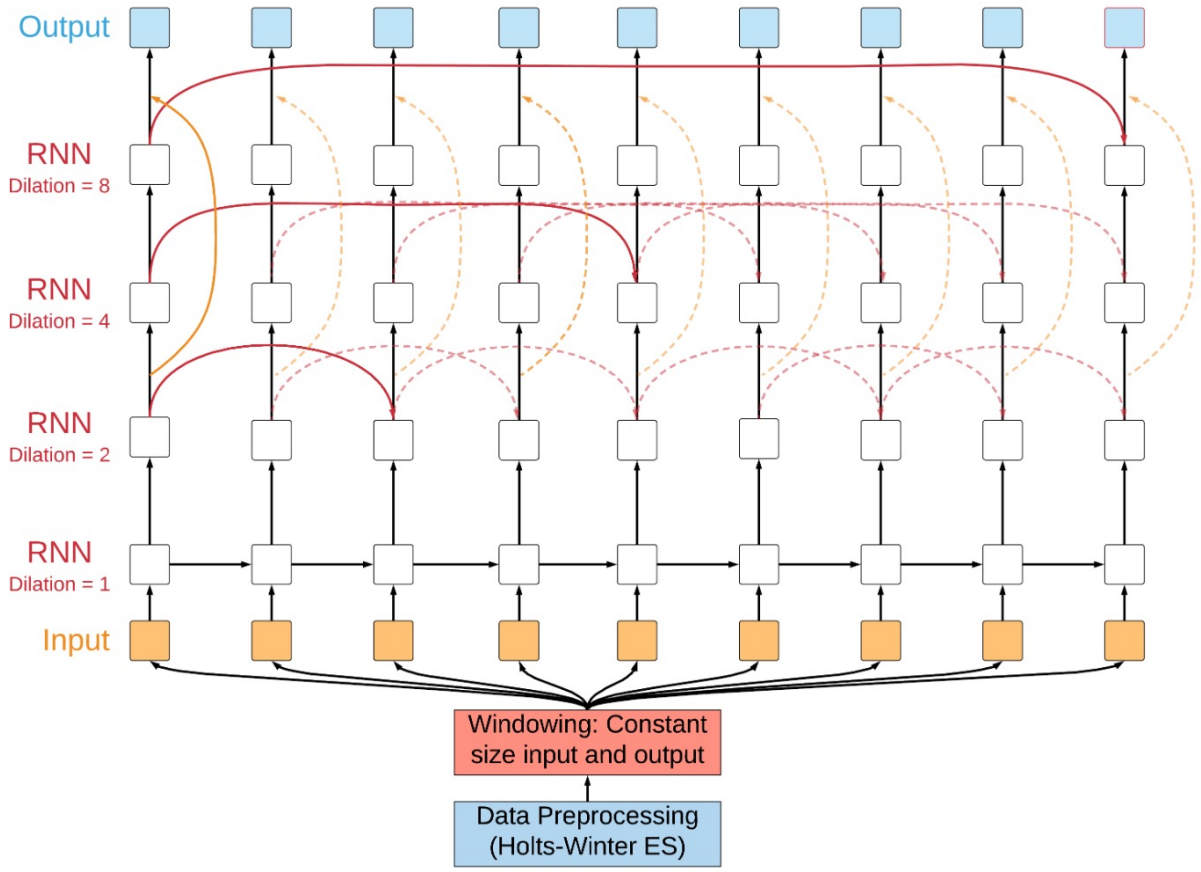
MODEL VALIDATION

- Validation based on AUC (~0.81)
- Examination which visits were the most influential by looking at the attention weights for specific patients

BACKGROUND

Forecasting of audience tuning during a campaign

DEEP LEARNING TIME SERIES FORECASTING



ALGORITHM

- Given an history $[y_1, \cdots, y_T]$
- H: horizon, T: length of observations, m: periodicity of the data
- TASK: predict $[y_{T+1}, \cdots, y_{T+H}]$
- Standard scale-free metrics in the practice of forecasting:
  - sMAPE: symmetric Mean Absolute Percentage Error
  - MASE: Mean Absolute Scaled Error

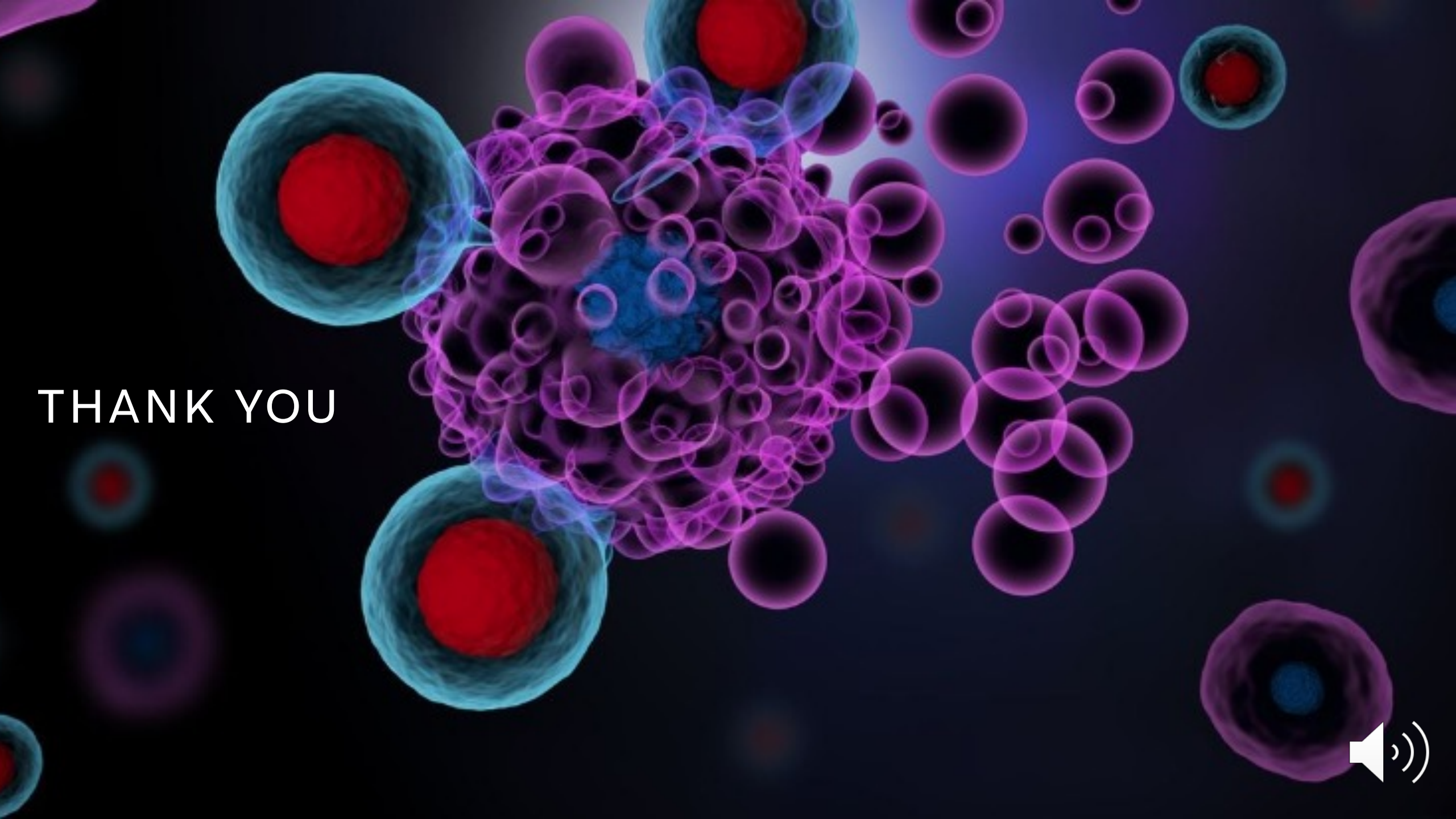
$$\text{sMAPE} = \frac{2}{H} \sum_{i=1}^{H} \frac{|y_{T+i} - \hat{y}_{T+i}|}{|y_{T+i}| + |\hat{y}_{T+i}|} \text{ , } \text{MASE} = \frac{1}{H} \sum_{i=1}^{H} \frac{|y_{T+i} - \hat{y}_{T+i}|}{\frac{1}{T+H-m} \sum_{j=m+1}^{T+H} |y_j - y_{j-m}|}$$

MODEL RESULTS

- Back testing using historical data was successful
- Patented and cited during all-hands meeting

REFERENCES

- Nguyen, Tuan-Minh, et al. “Identifying Significantly Impacted Pathways: A Comprehensive Review and Assessment.”
- Ma, Shuangge, and Michael R. Kosorok. “Identification of Differential Gene Pathways with Principal Component Analysis.”
- Koklesova, Lenka, et al. “Homocysteine Metabolism as the Target for Predictive Medical Approach, Disease Prevention, Prognosis, and Treatments Tailored to the Person.”
- Nehme, Ali, et al. “Atlas of Tissue Renin-Angiotensin-Aldosterone System in Human: A Transcriptomic Meta-Analysis.”
- Peilin Jia, Ruifeng Hu, Guangsheng Pei, Yulin Dai, Yinying Wang, Zhongming Zhao(2021), Deep generative neural network for accurate drug response imputation. *Nature Communications*, 12:1740.
- Wu, Peike, et al. “DROEG: A Method for Cancer Drug Response Prediction Based on Omics and Essential Genes Integration.” *Briefings in Bioinformatics*, Jan. 2023, p. bbad003
- Kong, JungHo, et al. “Network-Based Machine Learning Approach to Predict Immunotherapy Response in Cancer Patients.” *Nature Communications*, vol. 13, no. 1, June 2022, p. 3703
- David Blei et al, Variational Inference: A review for statisticians
- David Sontag et al., Deep Contextual Clinical Prediction with Reverse Distillations


THANK YOU

EXTRAS

UNSUPERVISED FEATURES SELECTION ALGORITHMS

Algorithm	Type	Reference
PCA	Largest Variance	Principal Component Analysis
MCFS	Sparse Leaning	Unsupervised Feature Selection for multi-cluster data
NDFS	Sparse learning	Unsupervised Feature Selection Using Nonnegative Spectral Analysis
Deep Learning Based	Neural Network	Discover Latent structure of the feature space

OTHER ML APPLICATIONS

BACKGROUND

Predicting who is watching media content cross devices and what the audience is watching and when

BACKGROUND

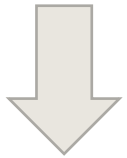
- **DMAS** were designed when TV local stations dominated the video market.
- They are outdated geographical areas which no longer apply to today’s streaming and mobile media landscape
- Some DMAS have very small panel coverage
- Meters are not always accurate

CHALLENGES

- Multimodal data: demographic, temporal, unstructured (text)
- Petabytes of data
- Different languages

AGGREGATION AND ENCODING

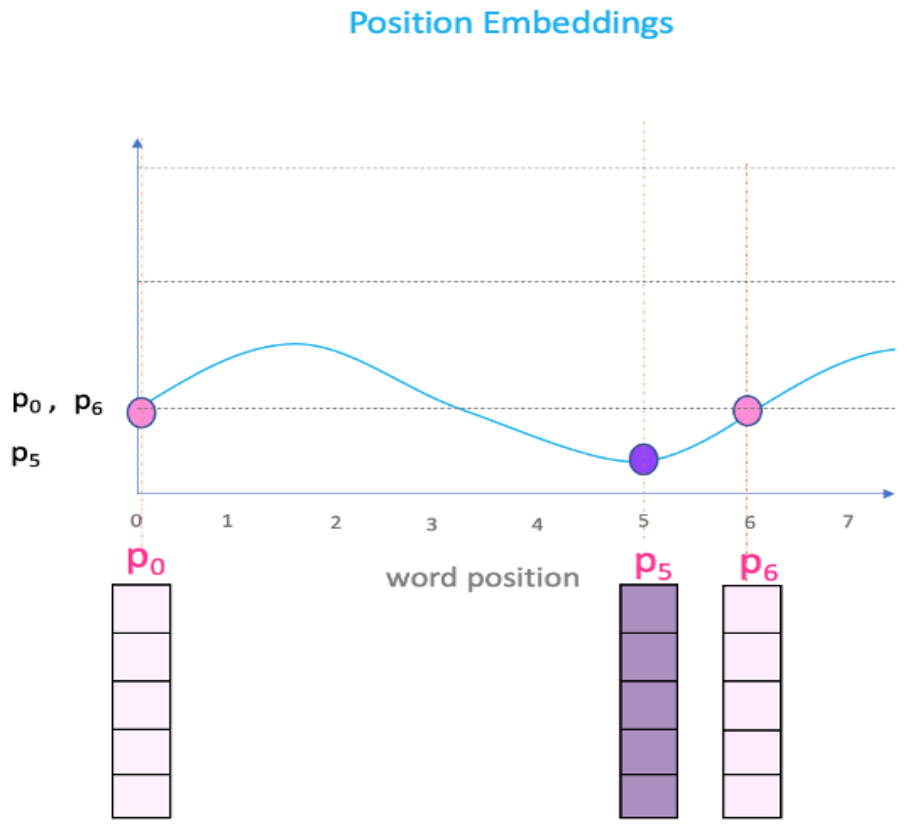
	Age	Gender	Start	End
Subject 1	30	M	10:01	10:50
Subject 2	30	M	9:30	10:30



	Demos	Start	End	Count
Event 1	[30, M]	9:30	10:01	1
Event 2	[30, M]	10:01	10:30	2
Event 3	[30, M]	10:30	10:50	1

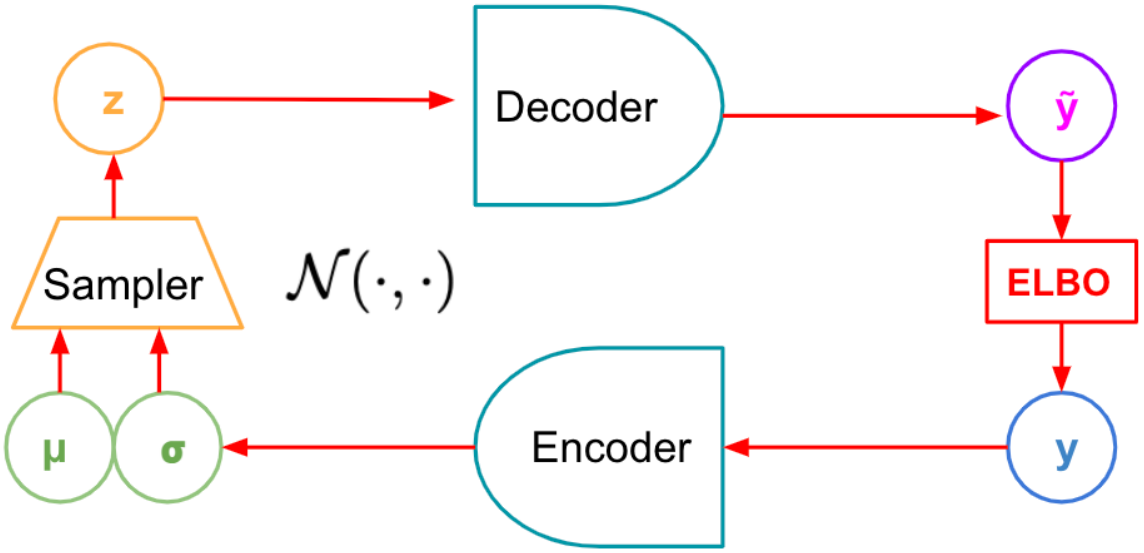
TIME ENCODING

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$



RATING VARIATIONAL AUTOENCODER

	Encoder	Decoder
# Layers	3	3
Activation	LeakyRelu	LeakyRelu Sigmoid



MODEL RESULTS

MODEL RESULTS

- Model scaled efficiently, only a few hours to process the data and train a model
- Back testing in past major media events was successful
- Became a legitimate product for Nielsen to predict audiences for future shows based on their description

MATHEMATICAL DETAILS

VAE MATHEMATICAL FORMULATION

$$\mathrm{P}(\theta|\mathcal{D}) = \frac{\mathrm{P}(\theta)\mathrm{P}(\mathcal{D}|\theta)}{\mathrm{P}(\mathcal{D})}$$

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p(\mathbf{x};\theta)$$

$$\mathrm{P}(x_i|z_i = k) \approx \mathcal{N}(\mu_k,\, \sigma_k)$$

Intoducint latent variables $\mathbf{z}$, the inference problem is to compute the conditional distribution of the latent variables given the observations: $\mathrm{P}(\mathbf{z}|\mathbf{x})$

$$\mathrm{P}(\mathbf{z}|\mathbf{x}) = \frac{\mathrm{P}(\mathbf{z},\mathbf{x})}{\mathrm{P}(\mathbf{x})}$$

$$\mathrm{P}(\mathbf{x}) = \int \mathrm{P}(\mathbf{z},\mathbf{x})\, d\mathbf{z}$$

The last integral is usually intractable so in varational inference, it is replaced with finding a surrogate distribution $\mathbf{q}^*$ the closest in KL and consists in solving this optimization problem:

$$\mathrm{q}^* = \arg\min \mathbf{KL}(\mathrm{q}((\mathbf{z})||\mathrm{P}(\mathbf{z}|\mathbf{x}))$$

VAE MATHEMATICAL FORMULATION

Reparameterization

$$z = \mathbb{E}(z) + \epsilon \odot \sqrt{\mathbb{V}(z)}$$

VAE Loss

$$l(x, \hat{x}) = l_{\text{reconst}} + \beta\, l_{\text{KL}}(z, \mathcal{N}(0, I_d))$$

$$l_{\text{reconst}} = -\sum_{i=1}^{n} [x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i)]$$

$$\beta\, l_{\text{KL}}(z, \mathcal{N}(0, I_d)) = \frac{\beta}{2} \sum_{i=1}^{d} \left(\mathbb{V}(z_i) - \log[\mathbb{V}(z_i)] - 1 + \mathbb{E}(z_i)^2\right)$$

LOSSES FORMULATION

Sard Loss

$$\mathrm{l}_{RD}(x) = -p_c g_w(x) log f_\theta(x) - (1 - g_w(x))\ log(1 - f_\theta(x))$$
$$\mathrm{l}_{CE} = -p_c y(x) log f_\theta(x) - (1 - y(x))\ log(1 - f_\theta(x))$$
$$l_{\text{tune}} = \mathrm{l}_{CE}(x) + \alpha \mathrm{l}_{RD}(x)$$

where $g_w(x)$ is a logistic regression model and p_c a weight class.

RNN Loss

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha) l_{t-1}$$
$$s_t = \beta \frac{y_t}{l_{t-1}} + (1 - \beta) s_{t-m}$$
$$\hat{y}_{\text{win}} = \textbf{ES-RNN}(\frac{y_{ti}}{s_{ti} l_{ti}})$$
$$y_{\text{truth}} = (\frac{y_{to}}{s_{to} l_{to}})$$

where l is a state variable, s is a seasonality coefficient, α and β are network parameters, m is the periodicity of the data.