No Time (To Die): Machine Learning for Predicting Time of Death

Tillie Slosser[1], Olivia Liau[2], Ivan Betancourt[3] ; MS: Qiaochu Liu[4], Deep Suchak[4] ; Advisors: Drs.: Chun-Kit Ngan[4], Chen Fu[5]
[1]Smith College, [2]University of Southern California, [3]Amherst College, [4]Worcester Polytechnic Institute, [5]UMass Chan Medical School

Motivation

- Gene expression reveals differences between healthy and diseased patients
- Sample time (typically **Time of Death / TOD**) explains variance for **circadian** (rhythmically expressed) genes

Expression of Gene NR1D2[1]

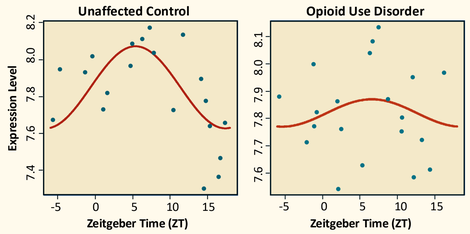


Fig 1. Example of a gene changing circadian expression under diseased conditions

The Problem: Gene expression datasets **lack crucial TOD information.**

Our Goal: Demonstrate **superiority** of **NEW method** to process **genetic time-series data** **via a comprehensive** development of pipelines.

Data Provenance

The Source: *Chen et al. 2016[2].* A paper examining circadian gene expression in younger vs. older adults

The Sample:

Subjects:	Mean Age:	% Male:	% Caucasian:
146	50.7	75	85

The Final Dataset:

146 patients * 2 brain areas = **292 total samples**

After Feature Selection: **20,000 original genes** -> **235 circadian genes**

Example Observation for One Patient

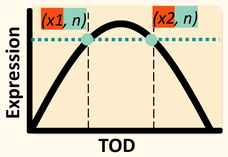
TOD	Age	Sex	Brain Area	PER3	LYPLA2	. . .
0.035	52	1	BA11	6.95	6.14	. . .
0.035	52	1	BA47	6.85	6.07	. . .

1 = Male; 0 = Female

235 columns of circadian gene expression levels

Problem with Current Methodology

At any given expression level y = n:



Circadian genes follow a sinusoidal curve → **The Problem:** 2 possible TOD values per *y* value

Previous Methods (Baseline 1)

- Uses **raw** gene expression values

Problem: Ignores sequential nature of data

Solution: Sequentiality in Pre-Training

Convolutional Neural Networks (CNN) (Baseline 2)

- Learns sequence through CNN layers

Problem: Low explainability (black box), computationally expensive

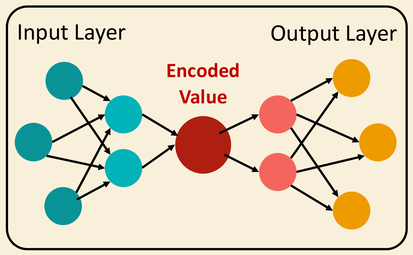
Autoencoders (Our Method)

- Encodes **sequentiality** via **sliding windows** before training models

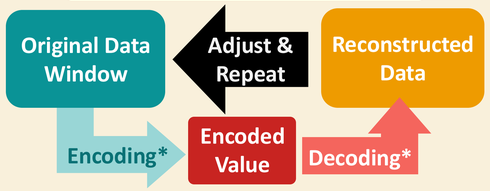
Proposed Method: Sequentiality via AutoEncoders

How AutoEncoders Work:

Autoencoder Structure



Training Process



***Encoding:** Compresses window to single **encoded** value
***Decoding:** Reconstructs original values to evaluate autoencoder performance

AutoEncoder Application

A. Create Windows

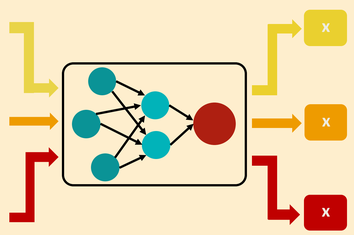
TOD	GeneX
1	3.28
2	8.10
3	1.33
4	6.74
5	9.02

N - 1	N=2	N + 1
3.28	8.10	1.33

N - 1	N=3	N + 1
8.10	1.33	6.74

N - 1	N=4	N + 1
1.33	6.74	9.02

B. Apply AutoEncoder



C. Append Encoded Values to Data

Age	Sex	TOD	GeneX
45	0	2	x
50	1	3	x
31	1	4	x

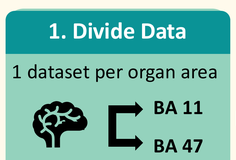
Pseudocode for creating windows, example size = 1:

```
for gene in gene_list:      #235 gene columns
   for TOD in TOD_column:   #each row
      window = [            #list of values:
        gene[TOD-1],        #previous value
        gene[TOD],          #current value
        gene[TOD+1]]        #next value
```

Model Development Pipeline

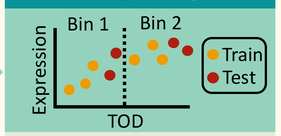
1. Divide Data

1 dataset per organ area



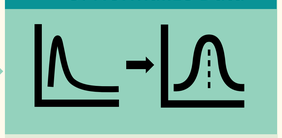
We use **2** brain areas

2. Train/Test Split



We test **3** splits

3. Normalize Data



We test **2** techniques

4. Reduce Dimensions

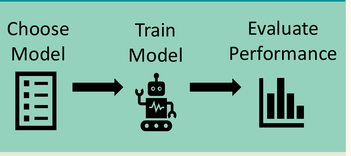
4a. Encode windows

We test **AutoEncoders** with **3** window sizes

4b. Reduce Dimensionality

We test **4** reduction methods

6. Train Regressors



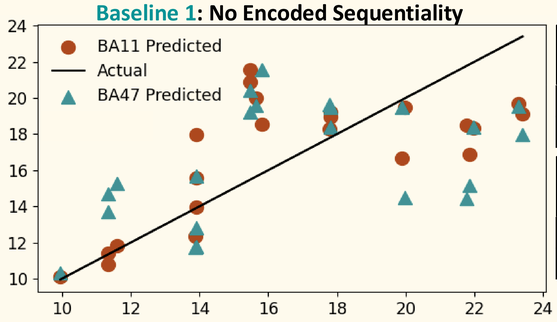
We train **16** regressors

(**5** Single, **8** Ensemble, **3** Deep Learning)

Results

Regression graphs for best model per method/baseline and brain area

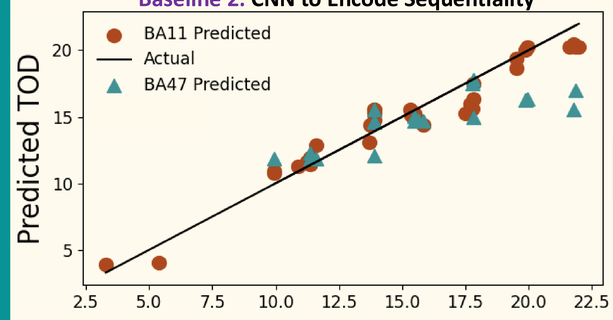
Baseline 1: No Encoded Sequentiality



Stand. Dev. of Error (Hours):	
BA11	3.077
BA47	3.823

Mean Absolute Error (Hours):	
BA11	2.424
BA47	3.274

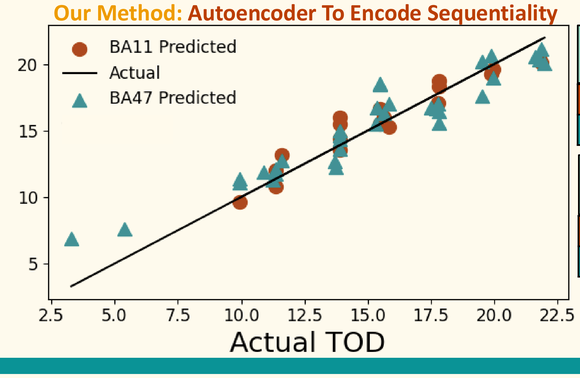
Baseline 2: CNN to Encode Sequentiality



Stand. Dev. of Error (Hours):	
BA11	1.107
BA47	2.201

Mean Absolute Error (Hours):	
BA11	0.945
BA47	1.757

Our Method: Autoencoder To Encode Sequentiality



Stand. Dev. of Error (Hours):	
BA11	0.996
BA47	1.451

Mean Absolute Error (Hours):	
BA11	0.839
BA47	1.227

Conclusions & Future Work

- We demonstrated that including sequentiality in input data **greatly improves** model performance.
- Autoencoders **perform better** than CNN, **more explainable**, and typically **less computationally expensive.**

In the future, we will:

- Explore larger window sizes for encoding
- Thoroughly evaluate performance time
- Generate gene expression profiles for out-of-sample timestamping

Acknowledgments & References

Thank you to NSF for REU Site Grant: 2349370 | Applied Artificial Intelligence for Advanced Applications (2024-2026).

Chen CY, Logan RW, Ma T, Lewis DA, Tseng GC, Sibille E, et al. Effects of aging on circadian patterns of gene expression in the human prefrontal cortex. Proceedings of the National Academy of Sciences. 2016 Jan 5;113(1):206–11.
Xue X, Zong W, Glausier JR, Kim SM, Shelton MA, Phan BN, et al. Molecular rhythm alterations in prefrontal cortex and nucleus accumbens associated with opioid use disorder. Transl Psychiatry. 2022 Mar 26;12(1):1–13.