Building Aging Clocks Using Transcriptomics and EEG Data

Davyd Sadovskyy

11/15/2024

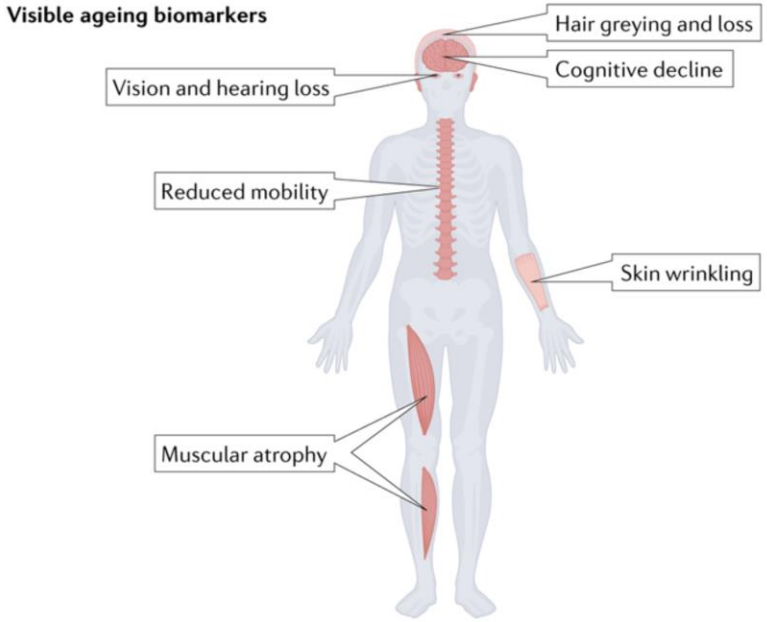
KrishnanLab

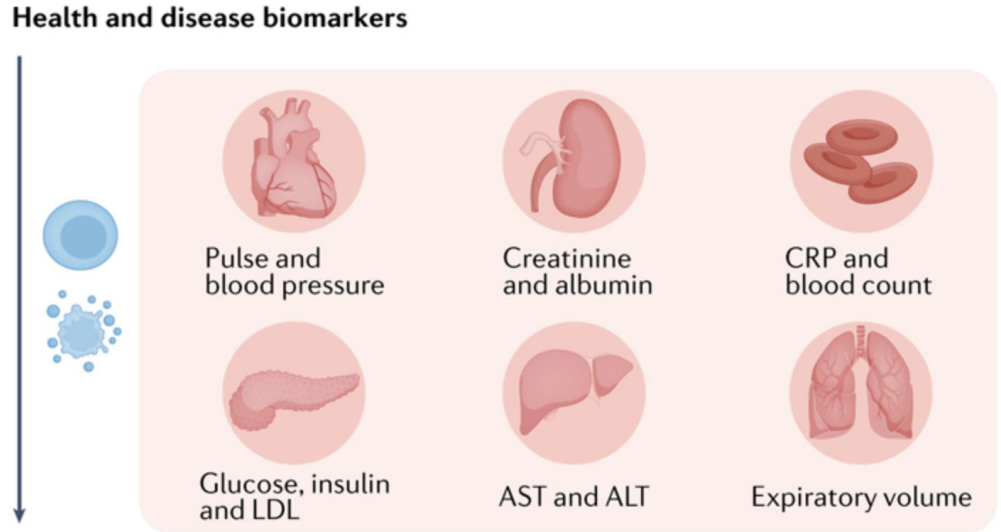
University of Colorado
Anschutz Medical Campus

Motivation

(Rutledge et al., 2022)

- It has been shown that aging of an organism can be manipulated via caloric restriction, heterochronic parabiosis, partial epigenetic reprogramming, and certain drug interventions.
- We need a way to quantify biological age to be able to tell if an intervention is working.
- Chronological vs Biological age. Someone's biological age can be very informative.

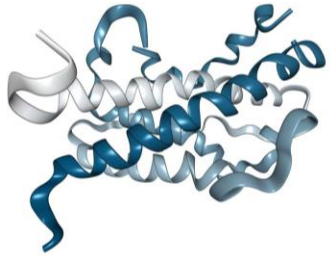




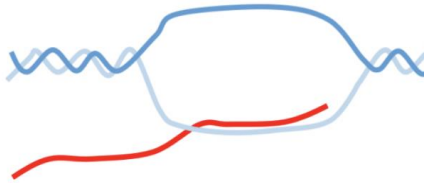
Statistical Learning Model
Y = Chronological Age

(Rutledge et al., 2022)

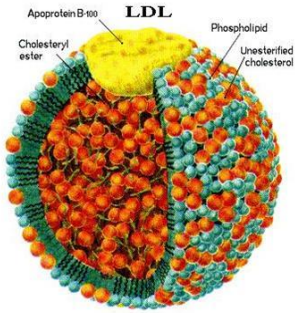
DNA methylation
- X = CpG methylation sites
- Hannum, Horvath clocks



Proteomics
- X = protein levels in plasma or CF fluid
- SomaLogic platform can quantify over 7,000 proteins, but its currently not yet possible to quantify entire proteome

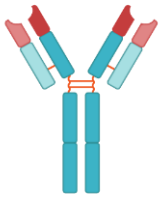


Transcriptomics
- X = RNA expression levels measured with microarray or RNAseq
- Issue of combining data from different sources

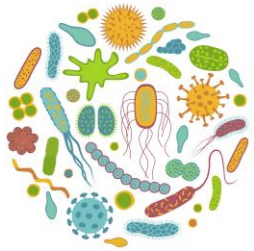


Metabolomics
- X = metabolites in plasma like albumin, LDL, amino acids
- PRO: data is widely available in biobanks
- CONS: low accuracy despite large sample sizes

<u>Less Commonly Used:</u>



Glycomics



Microbiome Composition

<u>Not Used Yet:</u>

EEG Waves

My Transcriptomics Dataset

- RNAseq is a newer technology that can detect a broader range of expression levels

RNAseq

- 14232 observations x 25570 transcripts
- 768 columns with mean and standard dev equal to 0 --> 14232 x 24802
- Data was normalized with arcsin

```
refine_normed = np.arcsinh(refine)
```

Microarray

- 16107 observations x 18478 transcripts
- No columns dropped
- I did not need to do normalization because that was already done earlier in the pipeline using log transformation.

- Both datasets were merged with their respective, separate labels data set to obtain the age label, dropping some rows that had NA age
- Biologically informed age bins were used:

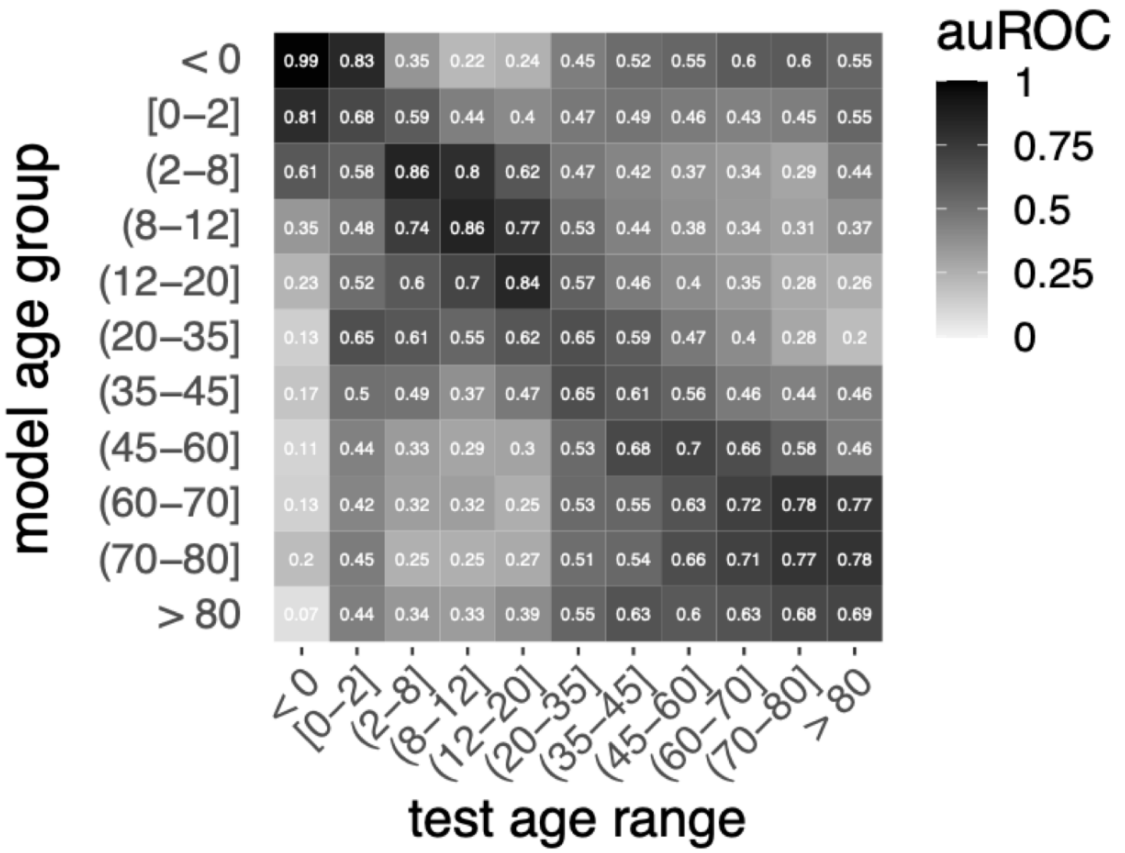
age
12wk
12wk
16wk

```
age_mapping = {
    'Carnegie Stage 23': 0.16,  # ~ 8 weeks (~0.16 years)
    'Carnegie Stage 22': 0.15,  # ~ 7.5 weeks
    'Carnegie Stage 21': 0.14,  # ~ 7 weeks
    'Carnegie Stage21': 0.14,  # account for the typo
    'Carnegie Stage 20': 0.13,  # ~ 6.5 weeks
    'Carnegie Stage 19': 0.12,  # ~ 6 weeks
    'Carnegie Stage 18': 0.11,  # ~ 5.5 weeks
    'Carnegie Stage 17': 0.10,  # ~ 5 weeks
```

Old Approach

- One vs All Classifier
- Elastic Net Logistic Regression
- A model trained on a certain age group predicts adjacent age categories better than distant categories

$$\text{Loss}_{\text{Elastic Net}} = \text{SSE} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

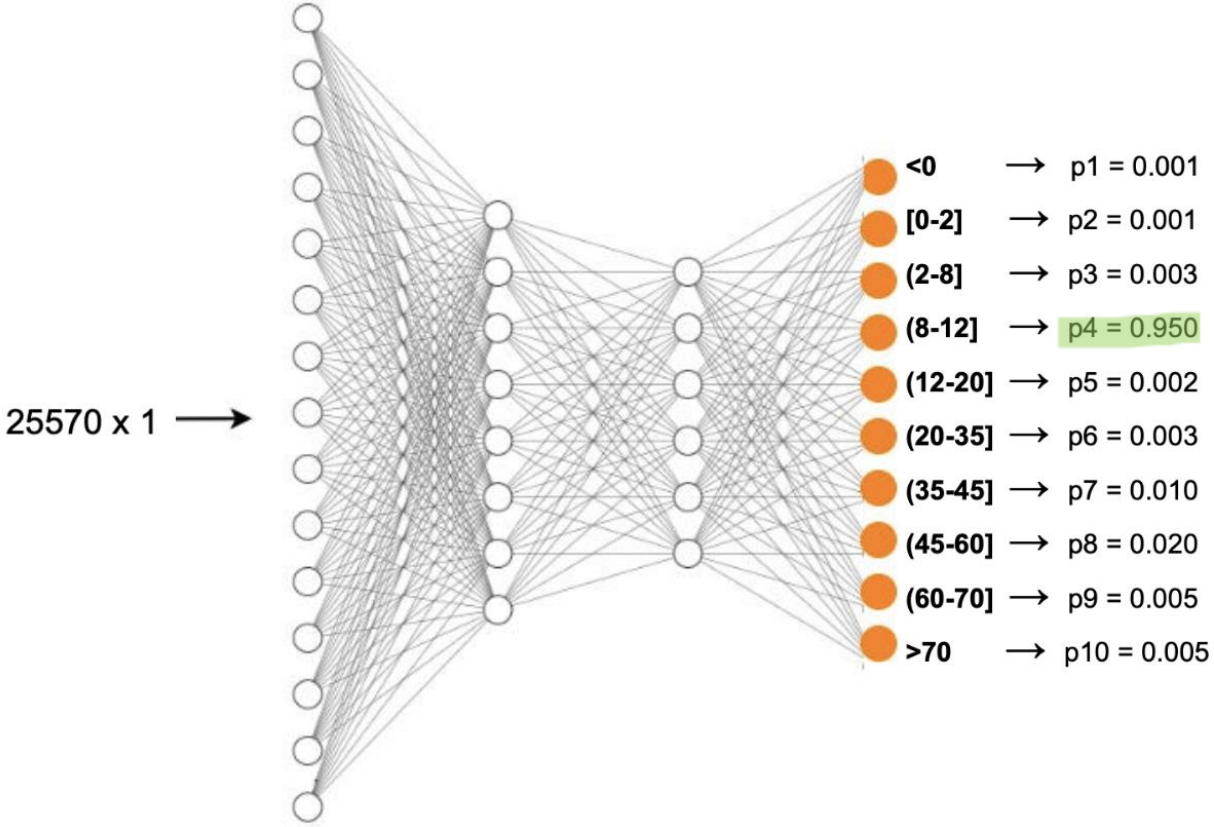


Source: *(Johnson & Krishnan, 2023)*

- What's the problem with treating age as a continuous variable?

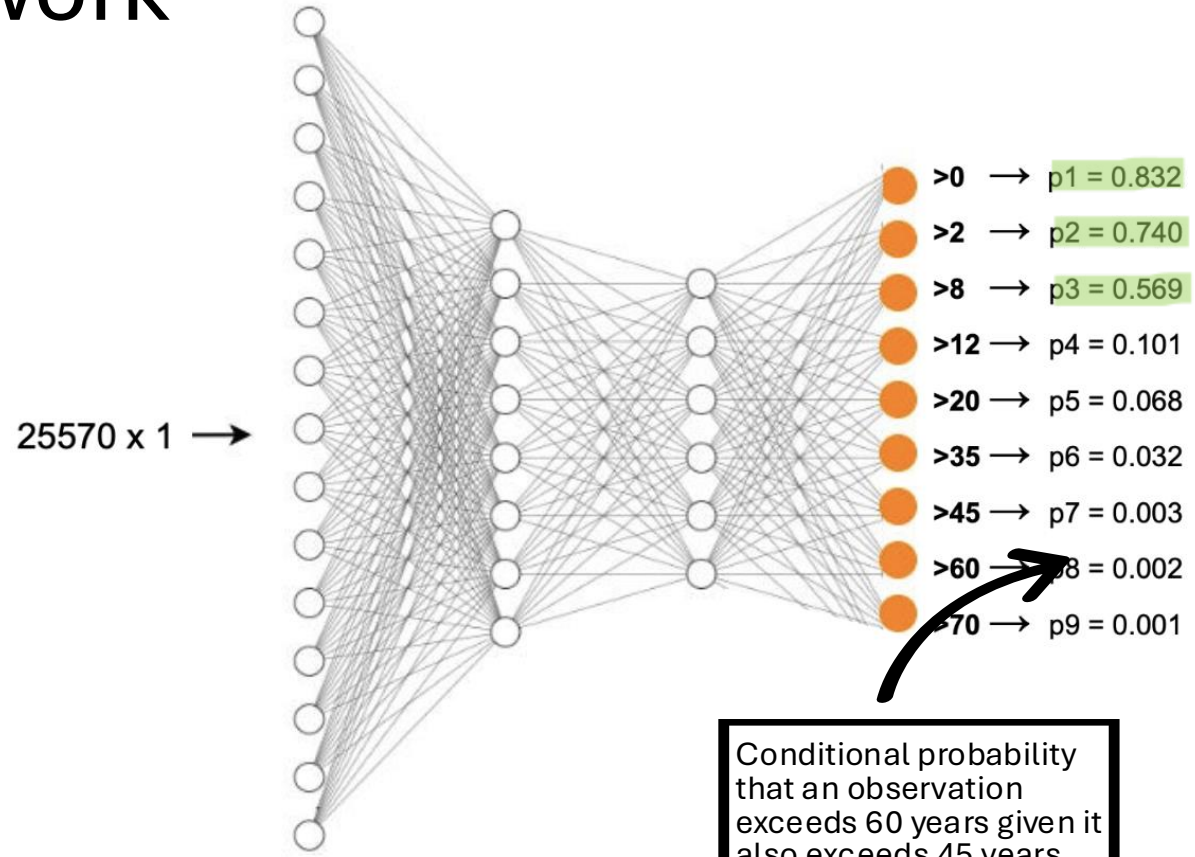
Rank-Consistent Ordinal Regression Neural Network

(Shi et al., 2021)



Regular MLP model

- No concept of ranking.
- Loss function doesn't respect "close predictions"
- Ranking concept can be incorporated by having output nodes >0, >2, >8, etc. However, this has issue of rank inconsistency



MLP model with CORN

- One less output node
- Loss function respects ordinal relationships
- Uses chain rule for probabilities
- Learned using conditional training subsets.

CORN MLP vs Regular MLP on RNAseq Dataset

5 Fold CV – Ordinal Neural Network (CORN) vs MLP

	Age Bin Label	Count	CORN MAE	MLP MAE
0	[0–2]	1243	1.0829	1.237329
1	(2–8]	760	0.7539	0.689474
2	(8–12]	1098	0.6630	0.652095
3	(12–20]	2157	0.5614	0.607789
4	(20–35]	1867	0.7568	0.857525
5	(35–45]	1404	0.6282	1.037037
6	(45–60]	2711	0.4921	0.838436
7	(60–70]	1554	0.7117	0.958816
8	(70–80]	929	1.0032	1.257266
9	> 80	483	1.0414	1.455487

Overall MAE Comparison:

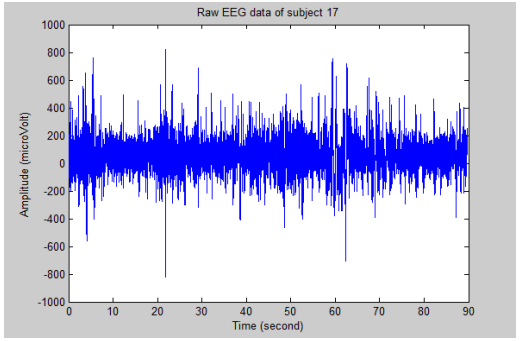
	Model	Overall MAE
0	CORN	0.711673
1	MLP	0.899621

- On average, the CORN model's predicted age bin was 0.712 bins away from the actual age bin on the test data.
- MAE is more informative than measures of classification accuracy in the one vs all approach

** Performance was notably worse on microarray dataset

EEG Sleep Data

(McConnell et al., 2023)



Single channel recording 726 Participants

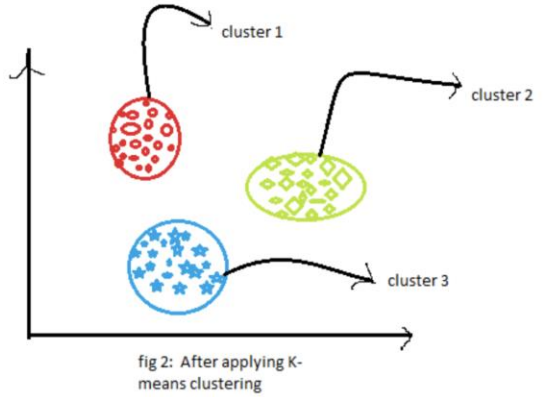
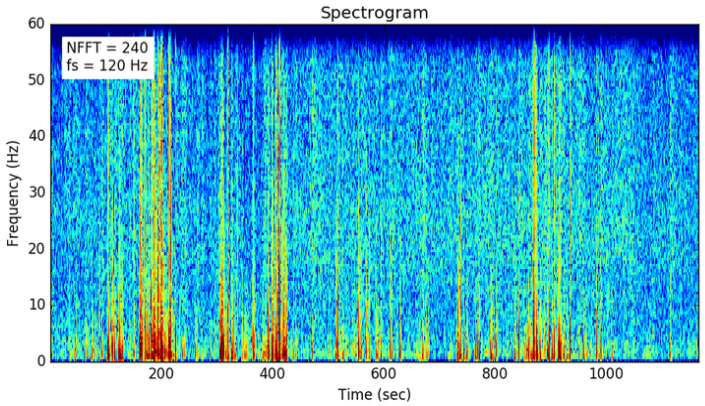


fig 2: After applying K-means clustering

Identify and Cluster Slow Waves



Spectrogram

- roundness, average pixel intensity of events, etc, etc
- 36 features * 2 clusters = **72 features**

Slow Waves

- waveform metrics were things like distances between zero crossings, amplitudes at peaks troughs, etc.
- 15 waveform metrics * 3 frequency groups * 13 summary statistics * 2 clusters = **1170 features**

Short-time Fourier transform (STFT)

- mean, sd, skewness,kurtosis
- 7 frequency ranges * 6 time regions * 4 summary statistics * 2 clusters = **336 features**

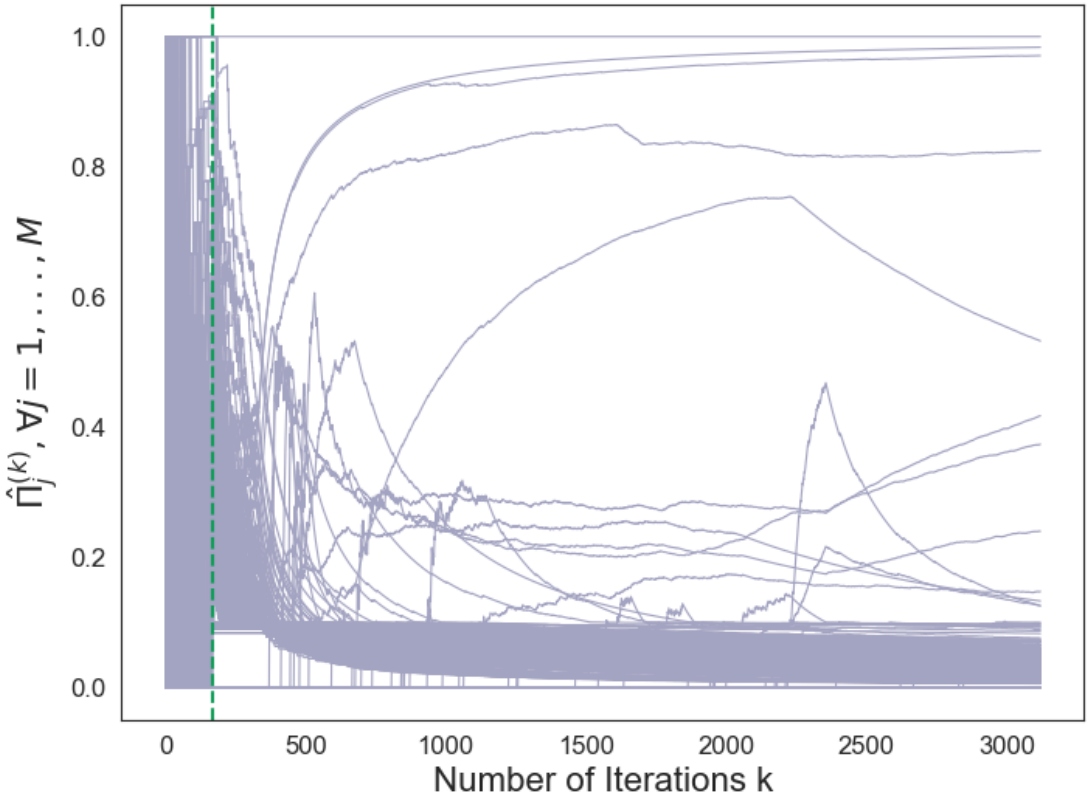
My Goal: Feature Selection

(Yao & Allen, 2020)

- Basics were already implemented on the dataset. Recursive feature elimination. SHAP (SHapley Additive exPlanations) values.

- I used "minipatch feature selection" technique from 2020.
  - A minipatch is a subsample of n observations and m features without replacement.
  - A base selector is applied on the minipatch (decision tree or OLS)
  - Unique algorithm for exploring the feature space

- Ran into bugs that I couldn't work out completely in time.



Selection frequency of a feature

Acknowledgments

- Dr. Arjun Krishnan
- Dr. Brice McConnell
- Parker Hicks and Krishnan Lab Members

References

Rutledge, J. C., Oh, H., & Wyss-Coray, T. (2022). Measuring biological age using omics data. *Nature Reviews Genetics, 23*(12), 715–727. https://doi.org/10.1038/s41576-022-00511-7

Johnson, K. A., & Krishnan, A. (2023). Human pan-body age- and sex-specific molecular phenomena inferred from public transcriptome data using machine learning. *bioRxiv*. https://doi.org/10.1101/2023.01.12.523796

Shi, X., Cao, W., & Raschka, S. (2021). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *arXiv*. https://arxiv.org/abs/2111.08851

McConnell, B. V., Liu, Y., Biswas, A. K., Bettcher, B. M., Medenblik, L. M., Broussard, J. L., Lucey, B. P., Ramos, A. R., & Kheyfets, V. O. (2023). On monitoring brain health from the depths of sleep: Feature engineering and machine learning insights for digital biomarker development. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2024.02.27.581950v1

Yao, T., & Allen, G. I. (2020). Feature selection for huge data via minipatch learning. *ArXiv*. https://arxiv.org/abs/2010.08529