Predicting and Analyzing Pre-Term Birth from Sleep and Physical Activity Proteomics Pregnancy Data

Perla Molina

Summer 2023

Stanford | MEDICINE

Introduction

It's me!

Perla Molina

- Incoming First Year PhD in BMI
- Bachelor's in Data Science at USF
  - DaVita Internship
  - AWM President
- Why Stanford?
  - Easy move
  - Meaningful research
  - Data science realm
- Research interests
  - Cancer and disease
  - gynecology/women's health
- Obsessed with kpop and horror movies

UNIVERSITY OF SAN FRANCISCO

Background Info + Material

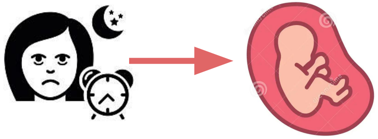
Previous PTB + Pregnancy Research feat. NALab

PTB → leading cause of mortality/morbidity for children under 5 *(1)*

Discovery of prevention measures based on maternal info (i.e previous PTB, socioeconomic background, quality of care visits, environment, etc) *(1)*

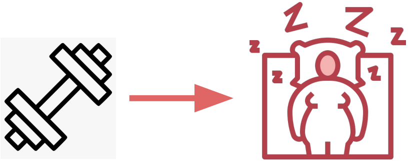
Multi-omics + ML techniques → precision medicine/healthcare to assess PTB associations and risk *(1)*; and predict neonatal outcomes *(2)*

Sleep + Physical Activity During Pregnancy



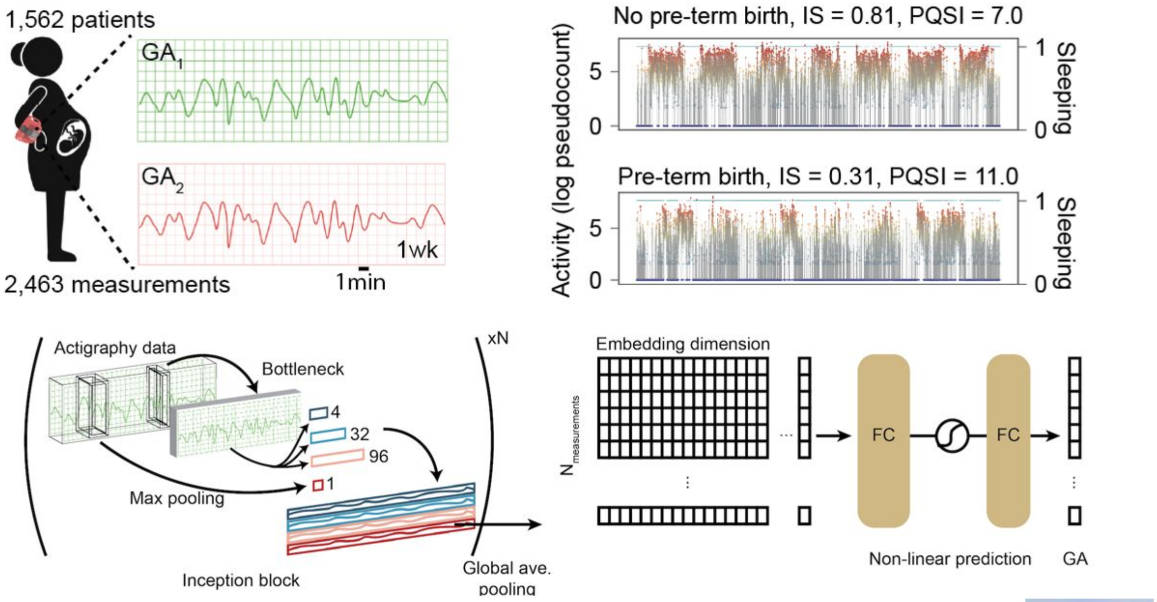
Sleep problems affect growth and fetal development due to transfer of melatonin from mother to fetus *(3)* → increase risk of PTB *(4)*

Regular exercise of moderate intensity increases placental blood perfusion → prevent placental abnormalities *(3)*



Relative sleep quality decreases in T3, but moderate PA improves sleep in T1 and T3 *(4)* → correlations of sleep + PA during pregnancy

NALab Research of Sleep + PA on PTB

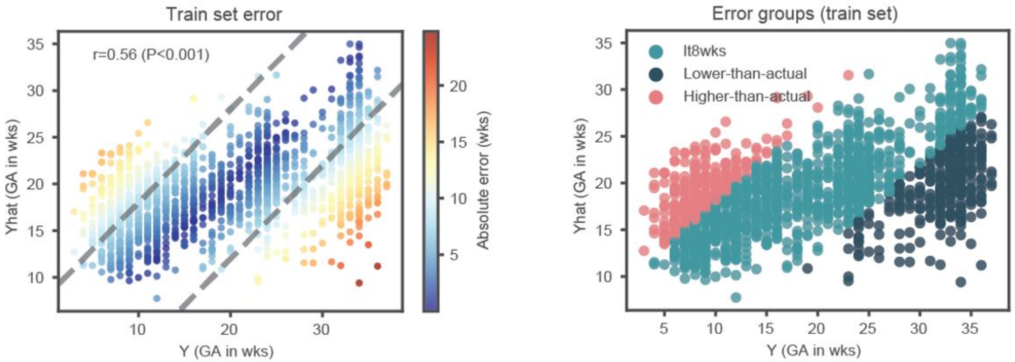


Neal Ravindra

- Smartwatch data to track sleep + PA
- Previous ML/NN models demonstrate link between PTB and sleep + PA
  - Computationally develop preventative measures w/o use of medications
  - Assess what is normal/good/bad

NALab Research of Sleep + PA on PTB Pt.2

Model has 3 distinct error modes



Real GA lower than actual GA: OR 0.49; pvalue 10e-152

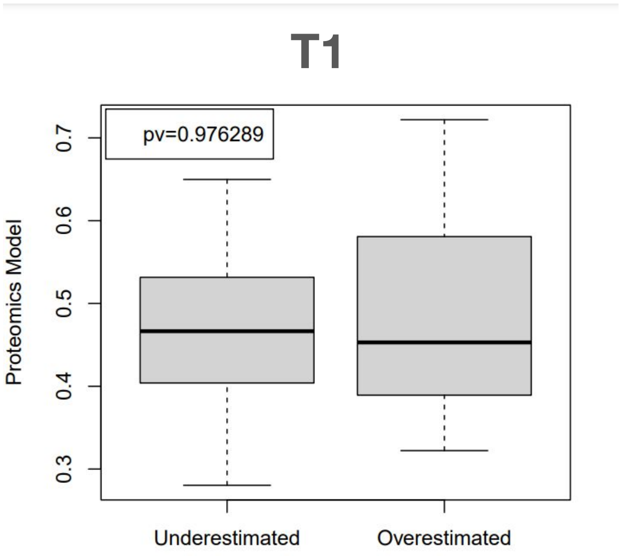
Real GA higher than actual GA: OR 1.33; pvalue 10e-27

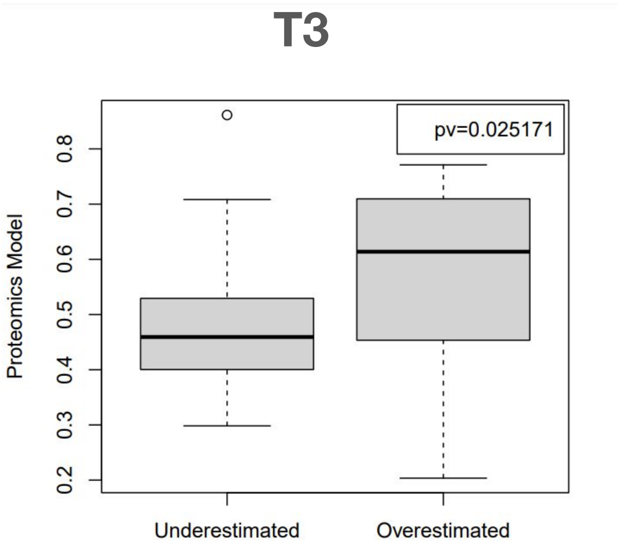
Neal Ravindra

- Sleep + PA have **major** impacts on PTB (look at difference of OR)
- Sleep + PA is dynamic throughout pregnancy

NALab Research So Far

- Predict PTB from sleep + PA proteomics (feat. outcome variable from EHR data)

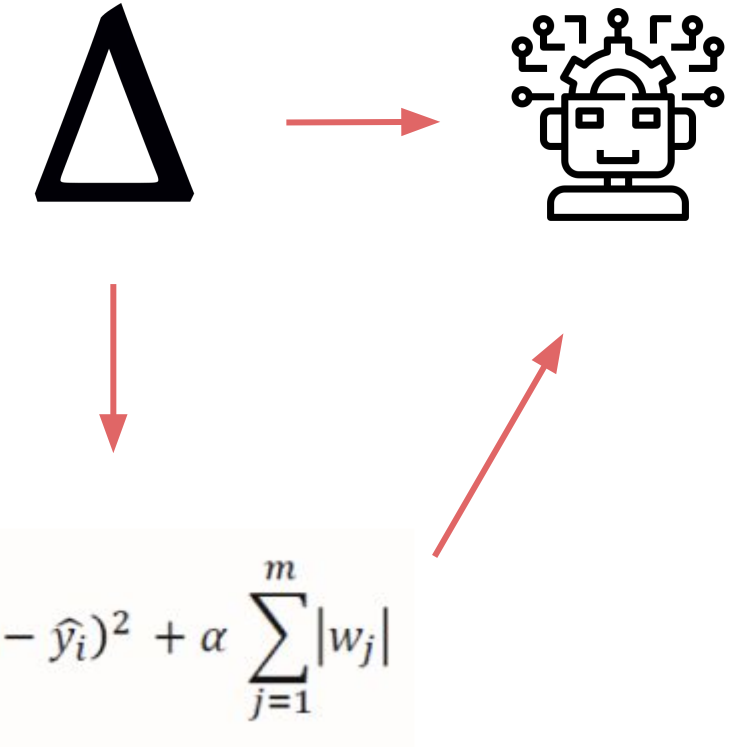




Objectives

My Task

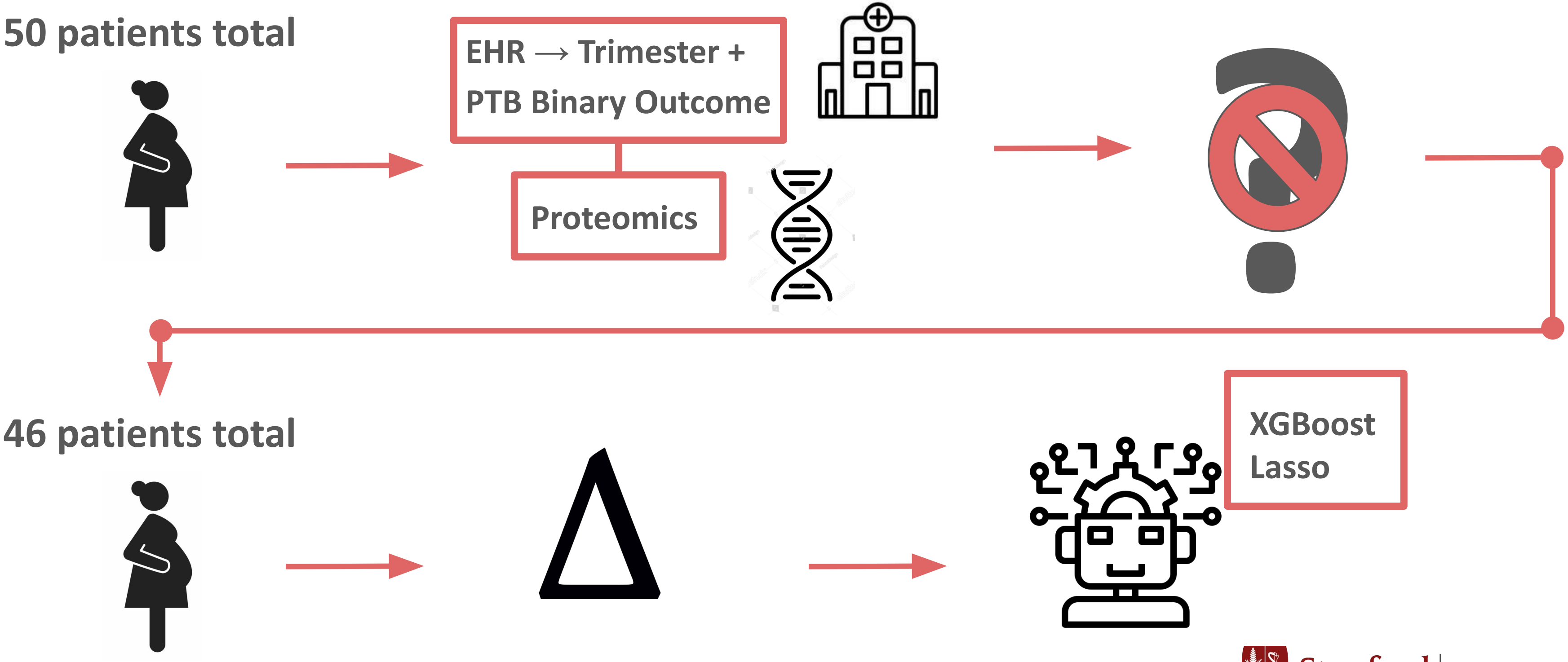
- Solve Delta(T1,T3) (Reproduce results from previous slide)
  - Analyze it
  - Assess significance + what it means in biological context
  - Compare to Prior PTB counts
- Run a Lasso Regression
  - Look at results
  - Make an assessment



$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{m}|w_j|$$

Methodology

Study Design



50 patients total

EHR → Trimester + PTB Binary Outcome

Proteomics

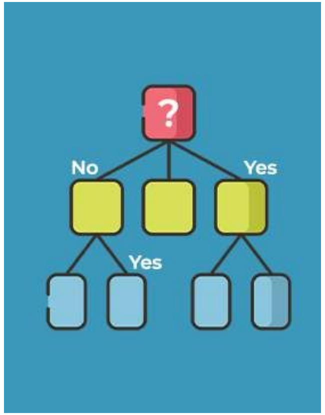
46 patients total

Δ

XGBoost
Lasso

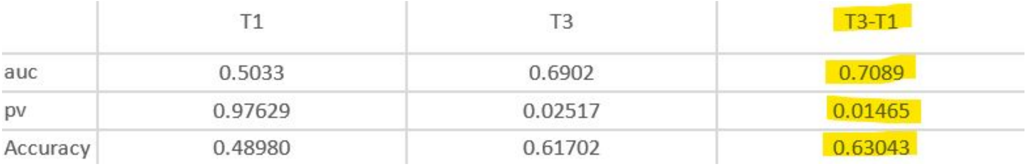
What I Used

- R: XGBoost + Lasso
    - Wilcoxon Test (p value)
    - AUC
    - Augmented accuracy
    - MSPE (for Lasso)
- Training-Test Split = 50/50 (for both)
- For Lasso, tested 200 alpha values between 0.005 and 1

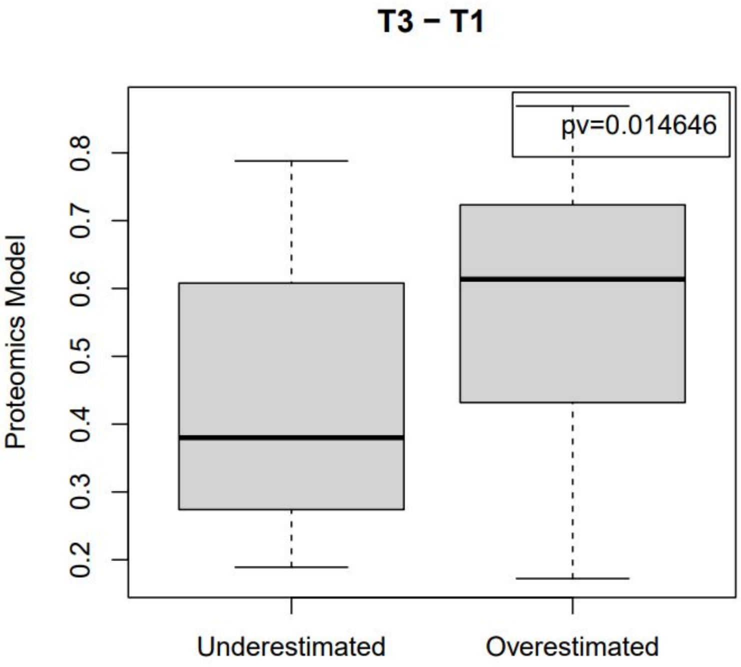


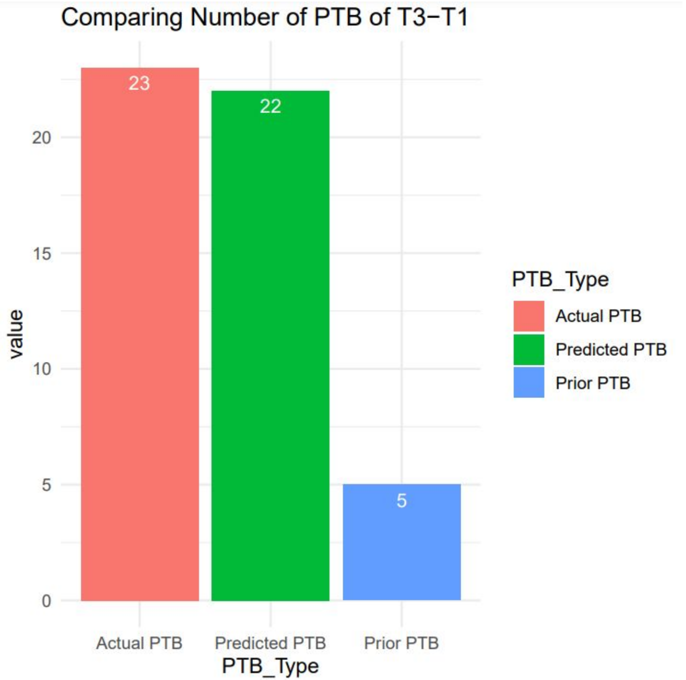
Results

Delta Results

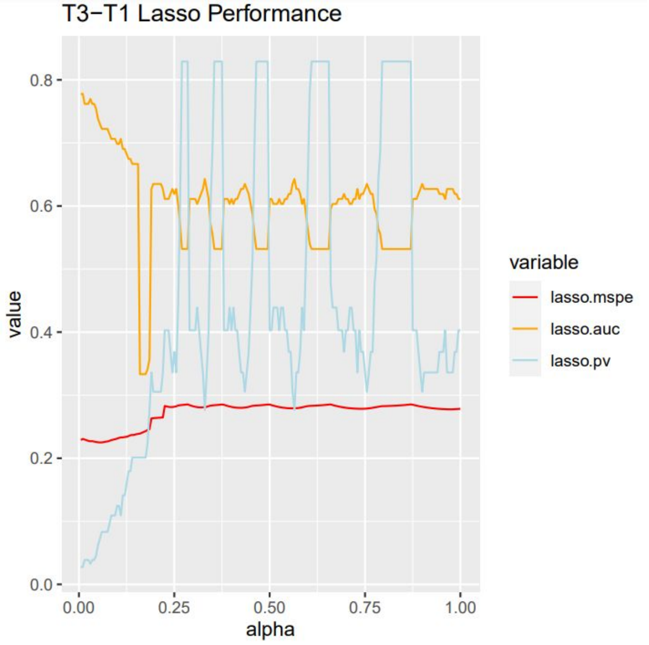


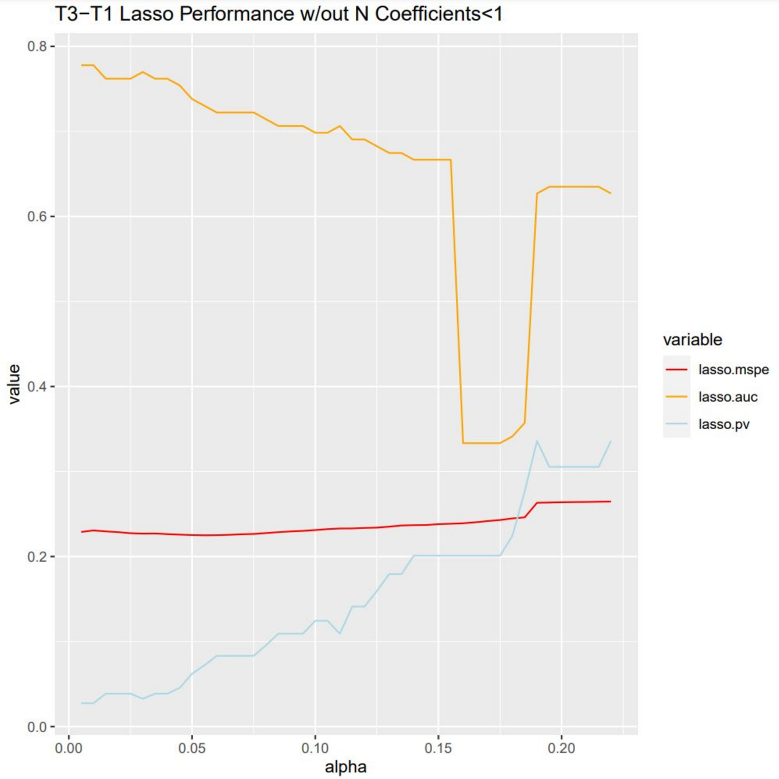
	T1	T3	T3-T1
auc	0.5033	0.6902	0.7089
pv	0.97629	0.02517	0.01465
Accuracy	0.48980	0.61702	0.63043



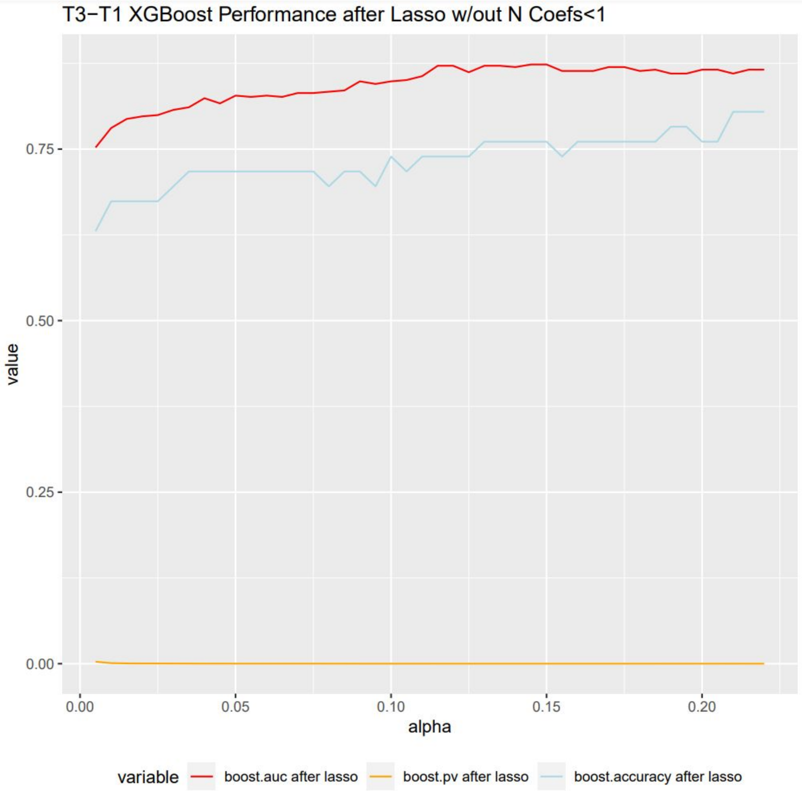


Lasso Results





Lasso Results Pt.2



Conclusions + Implications

Conclusions + Implications

What (was found)	So what (does it mean)	Now what (do we do)
Very difficult to predict anything in T1 compared to T3 looking at the difference in p values and accuracies.	Illustrates the importance of moderating pregnancy throughout. Conducting computational research requires holistic and comprehensive data/info of entire pregnancy. Corroborates literature + research on how dynamic sleep + PA is.	Longitudinal data is the way to go. Always acquire pregnancy data beyond a single trimester, a single doctor's visit, or even a single pregnancy (if applicable).
Significant value in delta (T3-T1) from its p value with improved accuracy and AUC. Better than either trimester individually.	May not be able to predict much in T1, but delta tells us there is attainable insights after T1 or between any or all trimesters.	Can possibly replicate this between or across doctor visits and not just between trimesters. Also apply process to other types of data (i.e. metabolomics, other pregnancy outcomes, placenta imaging, etc).

Conclusions + Implications

What (was found)	So what (does it mean)	Now what (do we do)
There were 23 actual PTBs versus 22 predicted with 5 having had prior PTB.	1) Previous research has demonstrated that prior PTB is the top increasing risk factor for PTB. These data results showed a lot more PTBs. 2) Illustrates and corroborates how much sleep + PA can truly effect likelihood of PTB aside from other risk factors.	1) Should conduct separate tests of predicting PTB on just those with prior PTB vs no prior PTB to see cohesive results than what I did. 2) Collect more proteomic data of sleep + PA. Perhaps even collect data at T2.
Lasso is tricky and can easily cause overfitting. Lower alpha values (less harsh of a penalty/smoother alphas) are better with lower p values and high AUCs.	1) Nearly all variables of proteomic data are necessary/important for predicting PTB. Probably do not need to consider or focus on dimension reduction. OR 2) Possibly need more data.	1) Collect and integrate more data (only 46 patients) to improve Lasso if high dimensionality is an issue when adding more patient data. 2) Test other algorithms or methods.

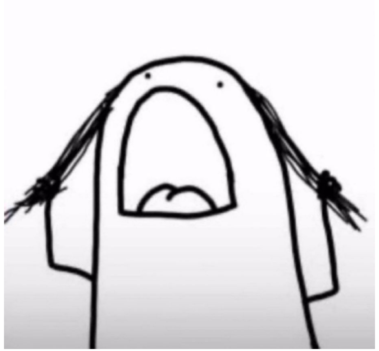
Conclusions + Implications

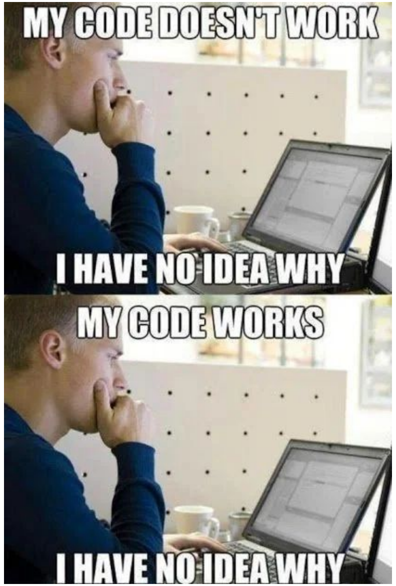
What (was found)	So what (does it mean)	Now what (do we do)
Overall, computational research corroborates non-computational research and literature of sleep + PA on PTB/pregnancy outcomes.	There is significant and translational value in what this lab is doing (I believe everyone knows that by now lol).	Keep doing what you’re doing!

Challenges & Opportunities

Challenges

- Figuring out how to do delta
    - It’s just basic matrix subtraction BUT different sizes between T1 and T3
- Coding Lasso and running XGBoost after
    - Ran into overfitting issues altogether with one alpha value ⟶ decided to test multiple alpha values
    - Had a lot of errors due to some alpha values resulting in only 1 coefficient (the intercept) ⟶ undefined matrix
- Basic coding errors
- Adjusting to grad school life right after undergrad

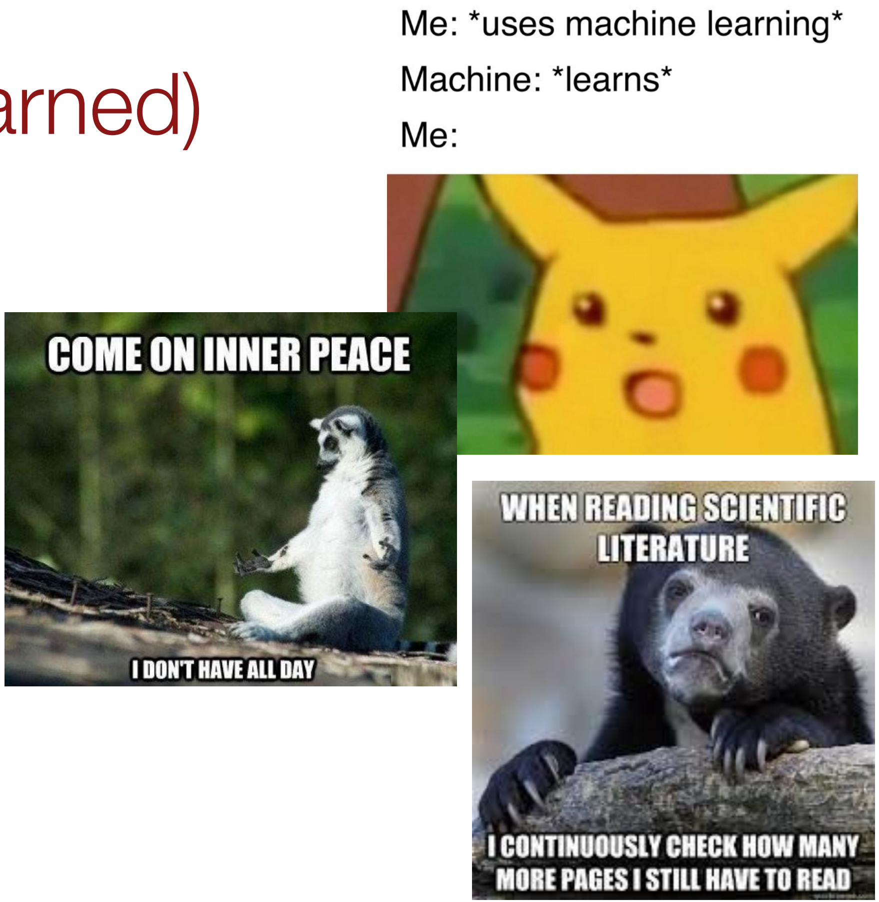




When you start grad school right after undergrad
Ah shit, here we go again.

Opportunities (What I Learned)

- How far along ML has gotten in biology/medical setting. Cool + impactful stuff happening
- Figuring out how to implement code you want can take A LONG time → taught me patience & it's okay to take time coding
- Reading science papers takes a while (help from ADVANCE)
- NALab is a cool lab to work at


Me: *uses machine learning*
Machine: *learns*
Me:
COME ON INNER PEACE
I DON'T HAVE ALL DAY
WHEN READING SCIENTIFIC LITERATURE
I CONTINUOUSLY CHECK HOW MANY MORE PAGES I STILL HAVE TO READ

If I Had More Time

- Test other ML techniques/algorithms (NN, Elastic Net, Logistic Regression, RF, etc)
- Rerun algorithm w/ more collected data
    - Rerun on different type of data (on genomics, metabolomics, etc)
- Perhaps test algorithm w/ different pregnancy outcome
- Try to see if i could implement PINNACLE *(5)*

References

References

(1) Espinosa, Camilo A et al. "Multiomic signals associated with maternal epidemiological factors contributing to preterm birth in low- and middle-income countries." Science advances vol. 9,21 (2023): eade7692. doi:10.1126/sciadv.ade7692

(2) De Francesco, Davide et al. "Data-driven longitudinal characterization of neonatal health and morbidity." Science translational medicine vol. 15,683 (2023): eadc9854. doi:10.1126/scitranslmed.adc9854

(3) Moreno-Fernandez, Jorge et al. "Impact of Early Nutrition, Physical Activity and Sleep on the Fetal Programming of Disease in the Pregnancy: A Narrative Review." Nutrients vol. 12,12 3900. 20 Dec. 2020, doi:10.3390/nu12123900

(4) Liwei Tan, Jiaojiao Zou, Yunhui Zhang, Qing Yang & Huijing Shi (2020) A Longitudinal Study of Physical Activity to Improve Sleep Quality During Pregnancy, Nature and Science of Sleep, , 431-442, DOI: 10.2147/NSS.S253213

(5) Li, Michelle M et al. "Contextualizing protein representations using deep learning on protein networks and single-cell data." bioRxiv : the preprint server for biology 2023.07.18.549602. 19 Jul. 2023, doi:10.1101/2023.07.18.549602. Preprint.

Q&A