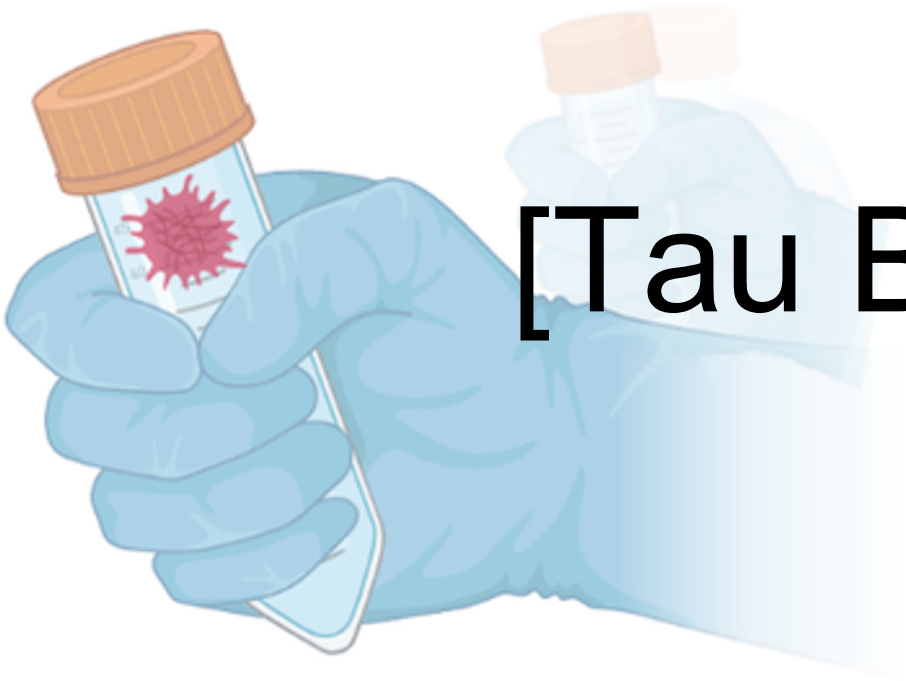


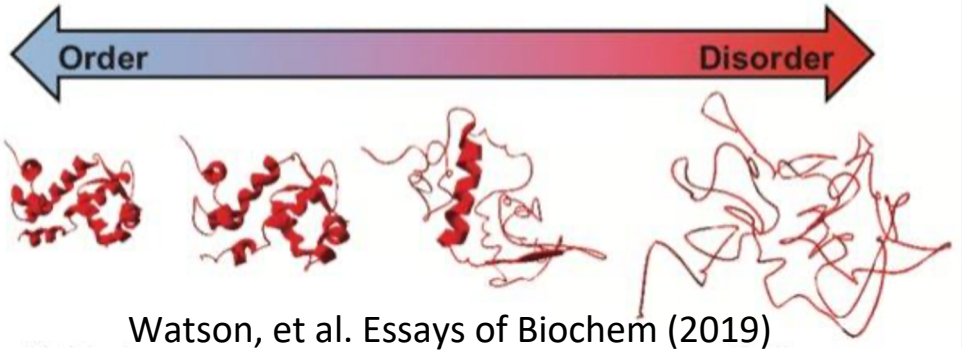
TBD
[Tau Be Damned]

December 15, 2021

Jiawei He
Elena Holland
Robert Pecoraro

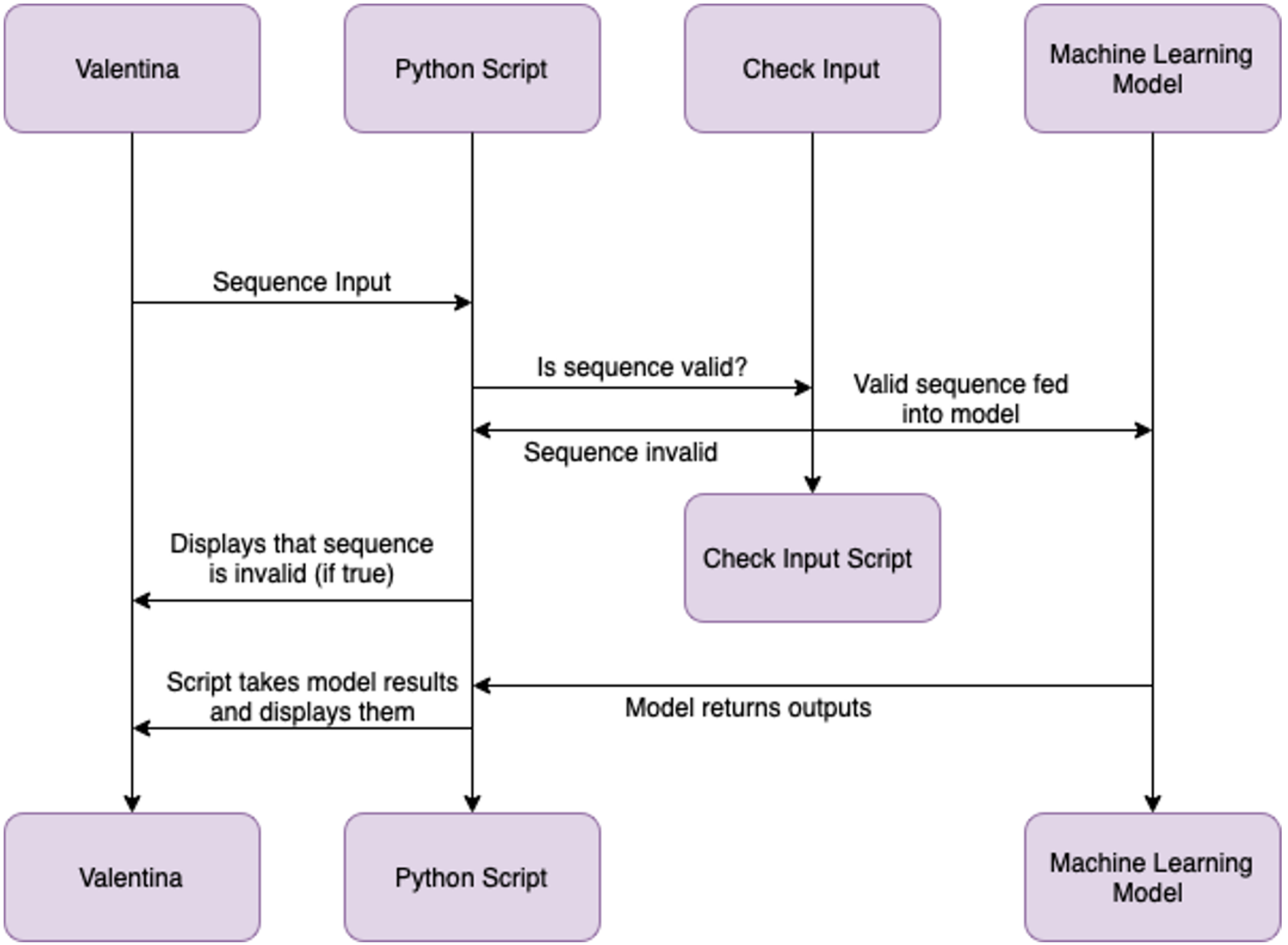
Introduction/Background

- Goal: build a tool to identify whether a protein is disordered based on its amino acid sequence
- While many proteins fold into regular conformations which can be easily analyzed on a structural basis, intrinsically disordered proteins (IDPs) do not
- IDPs are implicated in diseases such as Alzheimer's (tangles formed by the disordered tau protein)
- We collected amino acid sequences for ordered and disordered proteins from publicly available datasets to train a machine learning model to perform the classification task

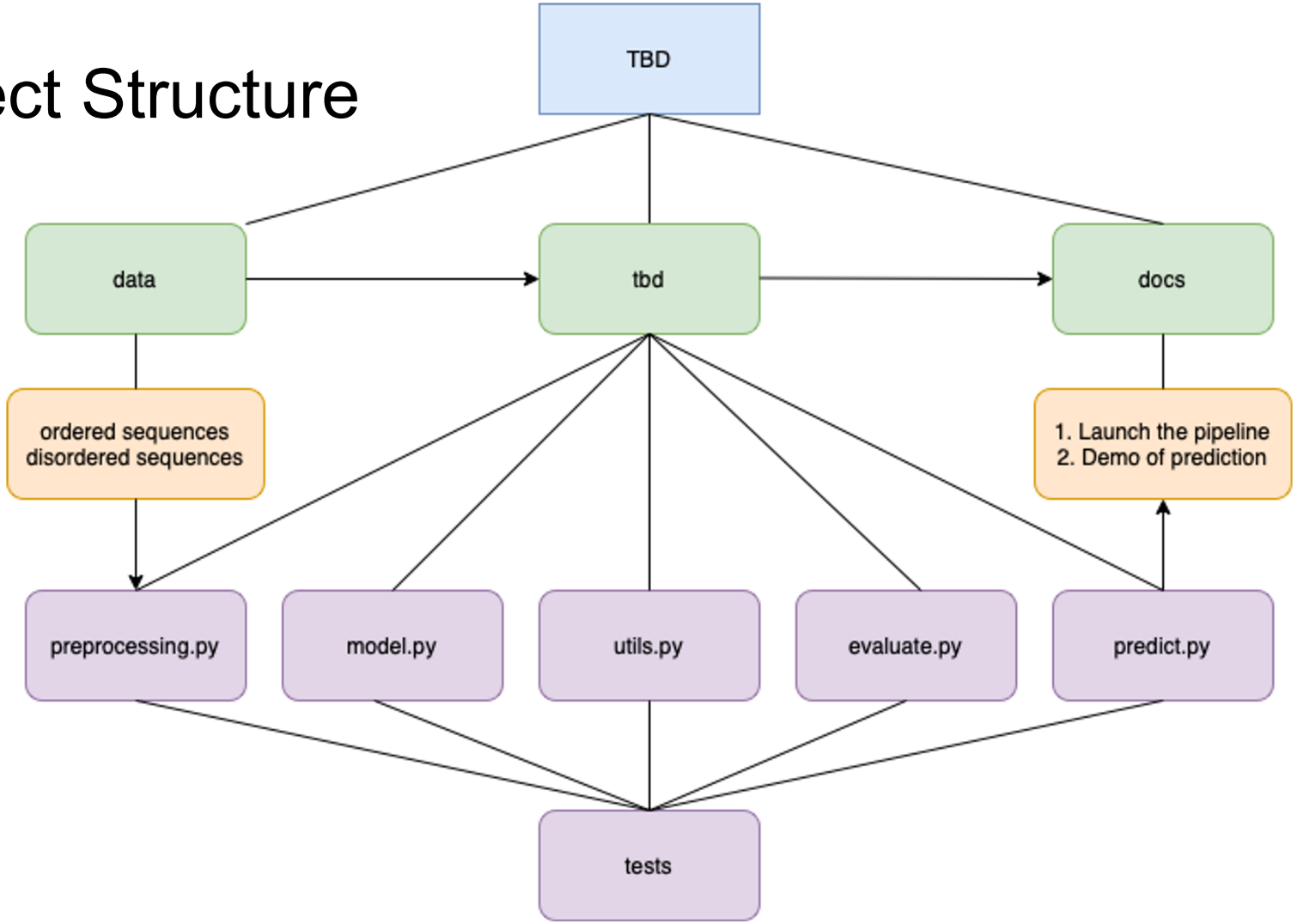


Watson, et al. Essays of Biochem (2019)

Design


Valentina
Python Script
Check Input
Machine Learning Model
Sequence Input
Is sequence valid?
Valid sequence fed into model
Sequence invalid
Check Input Script
Displays that sequence is invalid (if true)
Script takes model results and displays them
Model returns outputs
Valentina
Python Script
Machine Learning Model

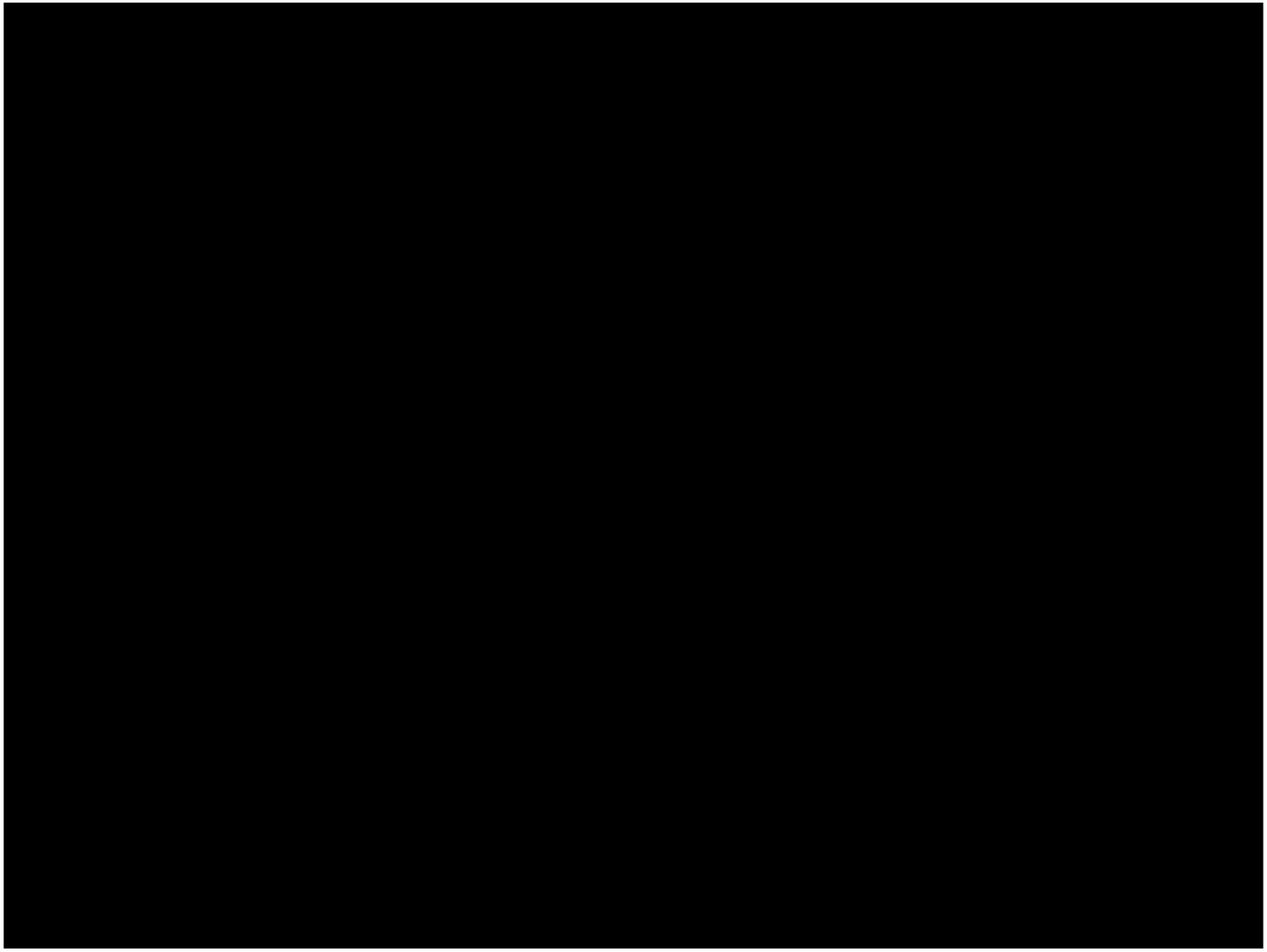
Project Structure



Technologies Employed

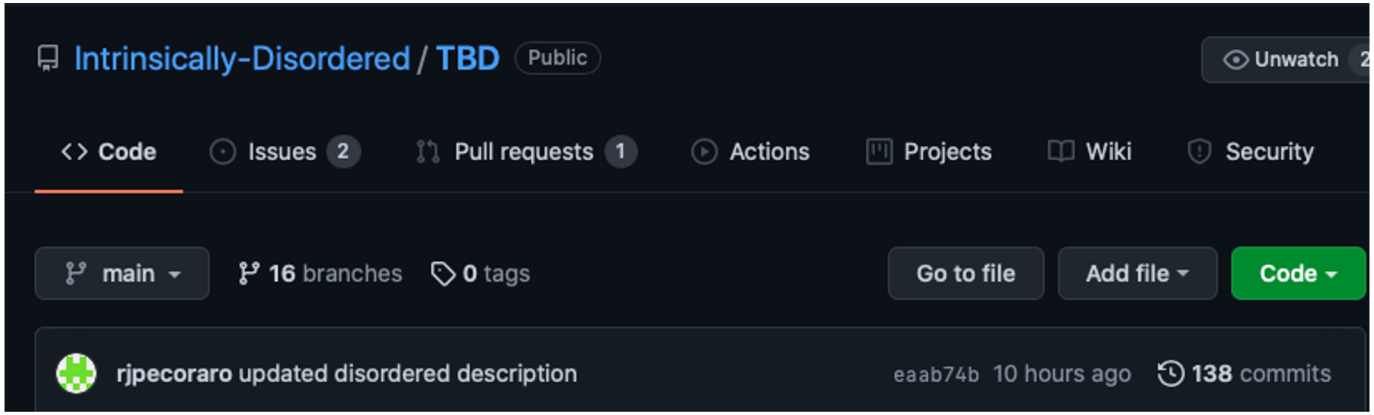
- **Algorithm**
  - Convolutional Neural Network (CNN)
- **Machine learning package**
  - TensorFlow
  - Scikit-learn
- **Continuous Integration service**
  - Travis CI
  - Coveralls
  - GitHub Actions
- **Testing tool**
  - pytest
- **Package management**
  - setup.py
  - conda
  - pip
- **Code style**
  - flake8
  - Git hooks

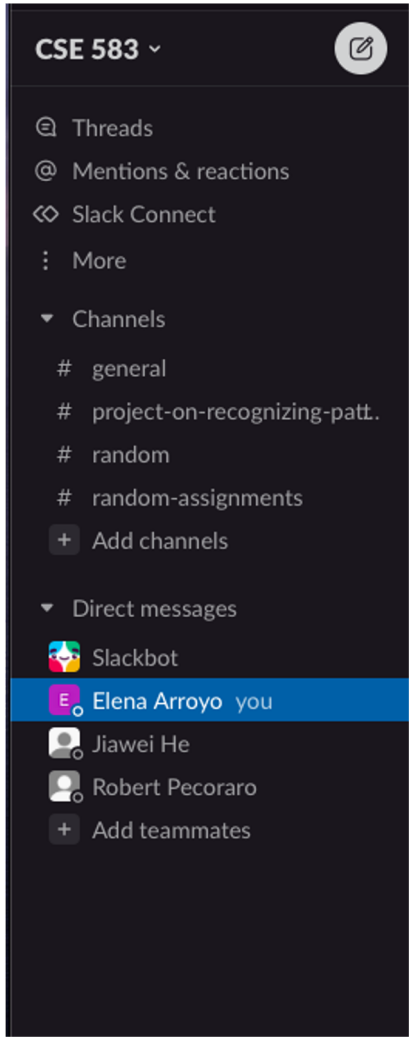
Demo



Collaboration Strategies

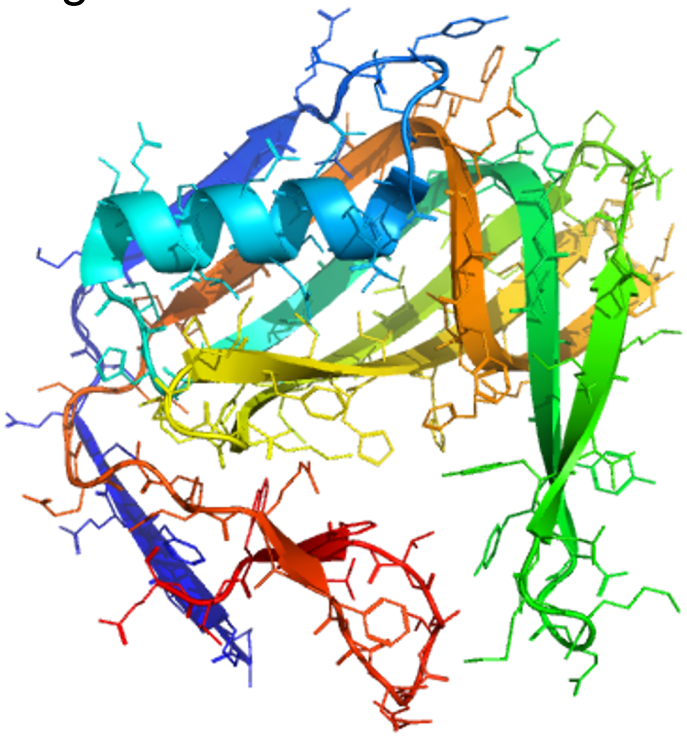
- Meetings
  - In person
  - On zoom
- On GitHub
  - Branches
  - Pull requests
  - Issues
- In Slack
  - Channels
  - Direct messages





Challenges

- Had to reduce very broad technical goals for our model/package
- Figuring out balance of work given diverse backgrounds
- Coordinating on GitHub is not easy!!
- Availability of data
- Setting up continuous integration



Lessons Learned

- Continuous integration
  - Flake while writing scripts
  - Test before triggering
  - Automatic code checking, testing, deployment
- Setting up package
  - Uninstall first
  - Data can be included too
- Writing more tests
  - Sample data for test
  - Cannot be enough
- Publishing package
  - Not allowed to submit package?
  - Change version number or skip existing
  - Repository secrets
- CNN & Protein sequence
  - Encoding
  - Sample weights
  - It's software class

Thanks Dave, Anant, and the rest of the class for listening!