Team 4 – An Army of One

Enzyme Stability Prediction Model

Post Mortem

- Successes
  - Developed a general-purpose predictor with >50% Spearman correlation for ranks in a short timeframe
  - Utilized lightweight processes
  - Able to incorporate protein structure information
- Challenges
  - Training data wasn't clean, highly skew, varying precision across sources
  - Kaggle data not similar to training data
- Future Updates
  - Data
    - Over/under sample to reduce pH bias, or consider modeling differently by pH (e.g. neutrals vs. acids vs. bases)
    - Look into standardizing stability scores by source/lab
    - Consider supplemental data sources
  - Model
    - General
      - Look into expanded use of biopython and similar packages
      - Examine skip-gram sequences/longer amino acid sequences
      - Try other models (e.g. neural networks)
      - Model pruning
      - Tune hyperparameters
    - Kaggle results
      - Examine features in the baseline/wildtype sample, to ensure inclusion in feature set
      - Focus model on most similar sequences
- Lessons learned
  - Business use case must align closely to training data
  - Use of external models/data can improve performance