
# ThermoDrift

Larissa Ault
Adam Chazin-Gray
Jacob Gershon
Nicole Panek
Jessie Zhao

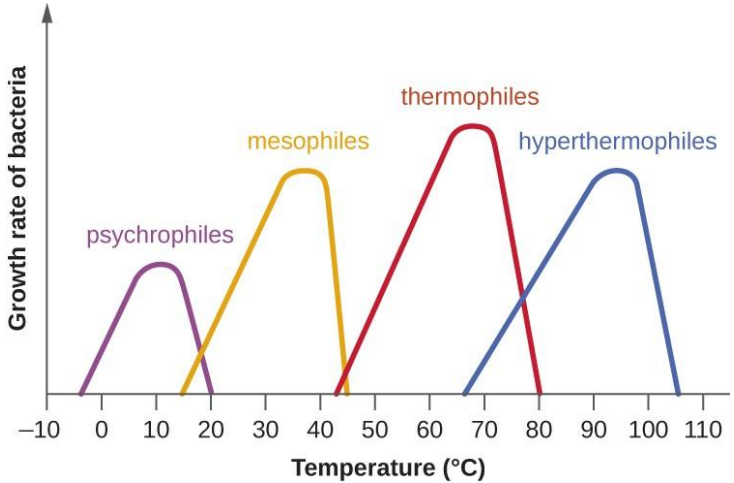
March 15, 2022

CHEME 545/546
Presentation Talk

# Project Overview

- **Scope:**
  - Develop a CNN classifier for automated prediction of any protein sequence as thermophilic, mesophilic, or psychrophilic.
  - Make this model user-facing through a published GUI.
  - Open source project codebase

- **Impact of our product:**
  - No existing computational classifiers for thermostable protein prediction
  - Accessible to a scientific user base with little machine learning experience
  - Various use cases: metagenomics, *de novo* protein design, crowdsourced protein prediction



# ThermoDrift Software Component Design

| GUI | Data Configuration | Classifier Model | Prediction & Metrics |
|---|---|---|---|
| **Dash GUI**<br><br>**Input:**<br>Protein Sequence FASTA File<br><br>**Output:**<br>CSV and Data Table with protein sequence, protein classification | **Filtering**<br>Removing proteins that do not start with M and are less than 75 AA; crop sequences longer than 500 AA<br><br>**One Hot Encoder**<br>Convert AA sequence to numerical position in a matrix | **PyTorch CNN**<br><br>Inspired by MNIST CNN architecture<br><br>Dataset assembled with sequences from Uniprot DB<br><br>60k total examples; 20k examples per class | **Phenotype Prediction:**<br>Thermophilic, mesophilic, or psychrophilic<br><br>**Confidence score:**<br>Probability that phenotype prediction is true<br><br>**Visualization:**<br>Training logs |

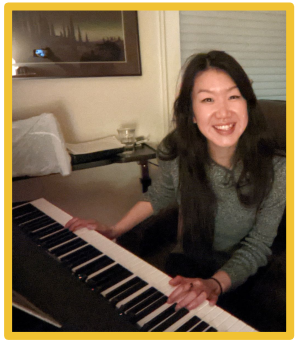
# Project Management

GUI

Data Configuration

Classifier Model

Prediction & Metrics



# Future Steps

| GUI | Data Configuration | Classifier Model | Prediction & Metrics |
|---|---|---|---|
| Report confidence of classification predictions in output CSV | Expand dataset to cover proteins beyond species *thermophilus, mesophilus, psychrophilia* | Apply transformer to CNN output to select for more specific features from the CNN | Output sequence motifs that correlate to protein classification |
| Ability to read in PDB to pull structural information | Improve time complexity of data processing method | Layer optimization | Add 3D atom coordinates to analyze structural motifs |
| | | | Generate continuous training logs after every epoch |

# GUI DEMO

# Acknowledgements

**Instructors:**
Dave Beck
Stephanie Valleau

**TAs:**
Nisarg Joshi
Evan Komp

**Technical consultation:**
Sanaa Mansoor (Baker Lab)

**Compute resources:**
UW-IT

# APPENDIX

# ThermoDrift Use Cases

USE CASE 1: A scientist's metagenomic samples from Yellowstone NP have become mixed up upon arrival back to their lab. These samples come from a cold high elevation tundra, a low-lying warm grassland, and a thermal spring environment. The researcher wants to use ThermoDrift to recover the identities of the various samples. The researcher first sequences various samples, and then uses the orfipy tool to predict ORFs for organisms in the sample. Then they compile these ORFs into DNA fasta files and then translate them into protein fasta files. Finally they upload their fasta files into ThermoDrift to classify the thermostability of the proteins encoded in the metagenomics samples and thus uncover the source environment of each metagenomic sample.

USE CASE 2: A graduate student at the UW IPD wants to design a highly thermostable protein nanoparticle, but the student is unsure whether the institute's current set of validated components meets their design specifications. The student wants to design a new set of de novo components that are highly thermostable. The student plans to take all current components and feed them into ThermoDrift to predict thermostability parameters. Based off of ThermoDrift predictions, the student plans to select those components predicted with the greatest thermostability and analyze the sequence similarity and apply these patterns to design additional components with similar sequence identity but variations in secondary structure.

USE CASE 3: Biochemist with a love for machine learning is interested in seeing if he can improve the accuracy of ThermoDrift's predictions by incorporating 3D structures as an input to the model. To do this the researcher takes the training set of sequences and folds them using Alphafold or Rosetta fold and selects all structures which the network is confident in folding. With this set of folded proteins from the three different classes, the researcher adds another CNN which will read in the coordinates of the structure and define a latent space. On the backend the researcher implements an attention module which reads in the latent space generated by the sequence and structure CNNs and then predicts an class for the paired sequence and structure. With the new code pushed to the model, people are now able to predict which environment the protein will classify into given a structure or sequence alone.