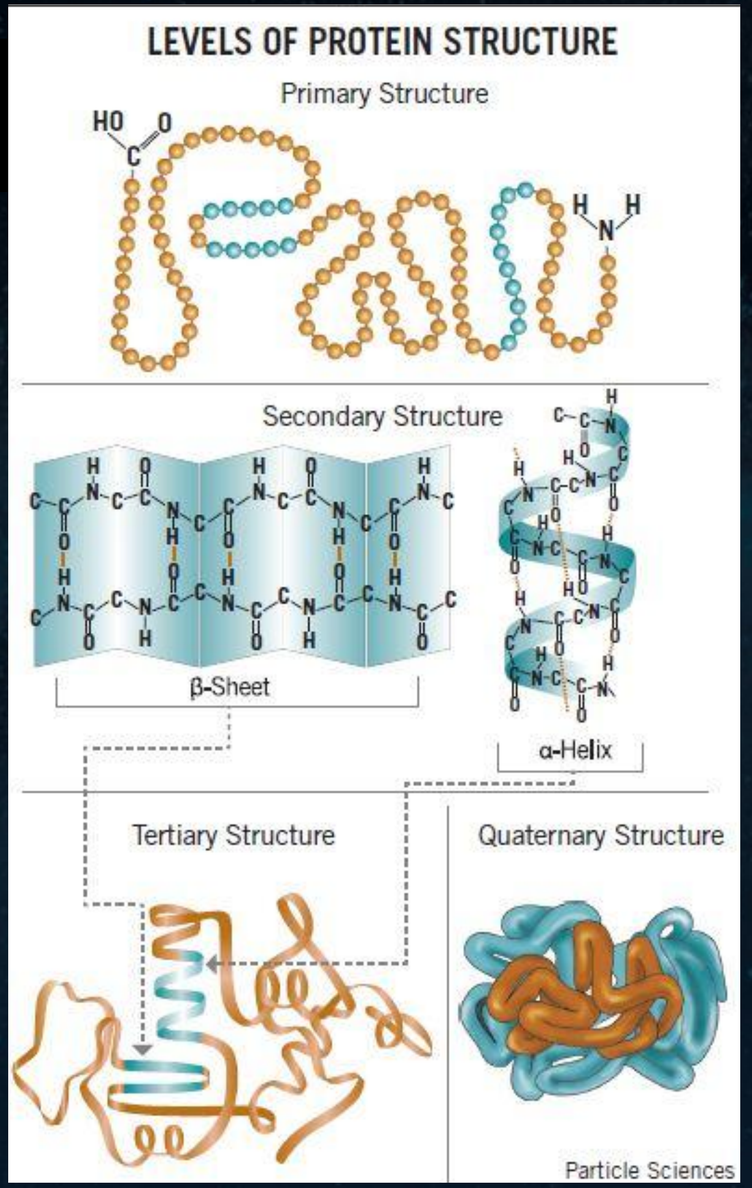
711th Human Performance Wing

Prediction of Protein Secondary Structure through Topological Analysis of Amino-Acid Subsequences

Aaron Hammer, Matt Piekenbrock, Kevin O'Neill,
Caroline Bablin, Kyle Hixenbaugh, Rory McIntosh

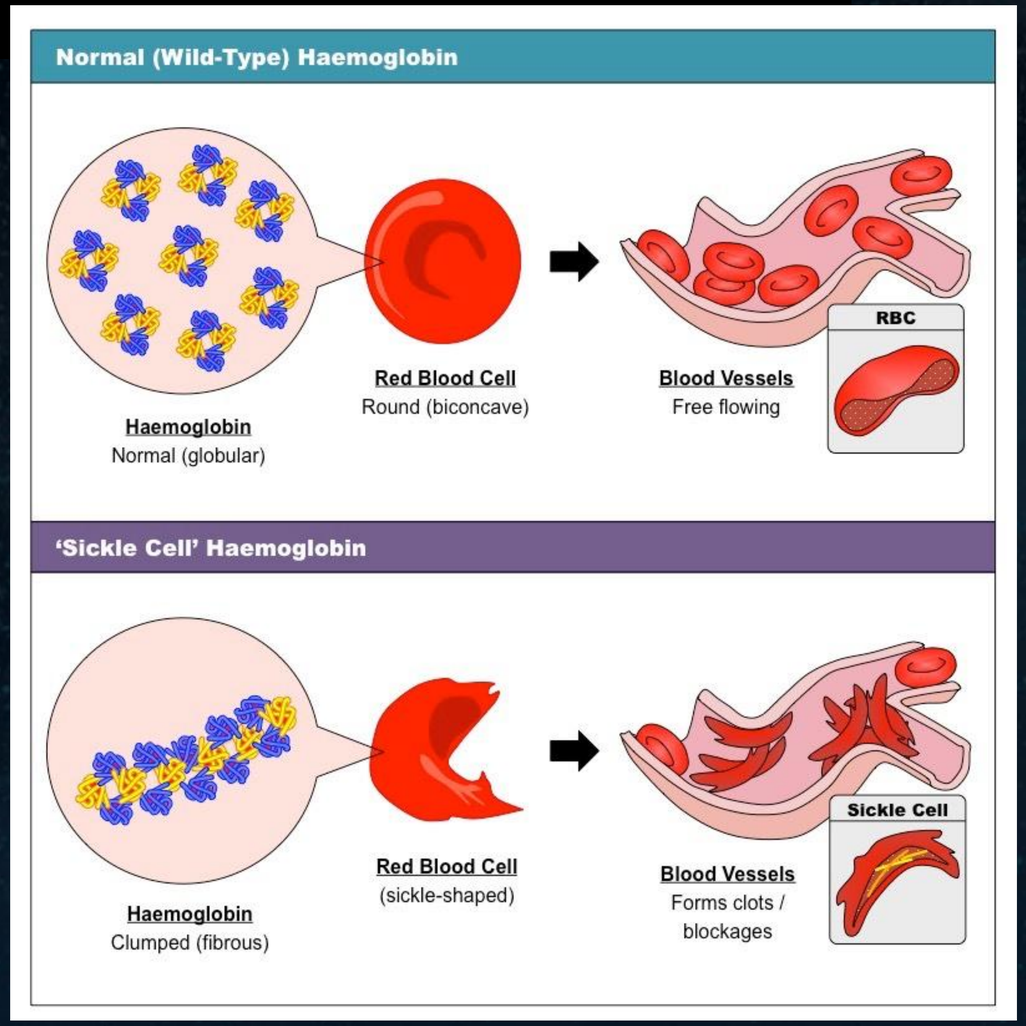
Protein Folding

- Proteins at the most basic level are chains of amino acids which creates primary structure
- Interactions between atoms within and between amino acids form the secondary structure
  - a local structure
    - helix, sheet, coil are three most prevalent structures
- Long-range interactions between amino acids determine how the protein will fold in 3-dimensional space; the result is the tertiary structure
- One of the most challenging machine learning problems



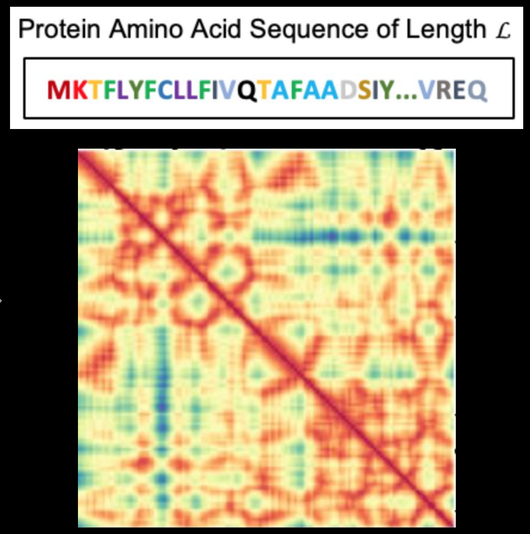
Impact

- Shape determines function
- Many diseases believed to be caused from improper folding of proteins
  - Parkinson’s, Huntington’s, Sickle Cell Anemia, Cystic Fibrosis, and more
  - Tertiary structure determination could allow for protein synthesis to aid in treatment of these diseases
- Multiple DOD applications
  - Protein capture elements for sensing



The Challenge

- IEEE ICMLA 2019 Challenge focused on Protein Inter Residue Distances Prediction
- Given Input Features:
  - 1-D features
    - PSIpred-helix, PSIpred-sheet, PSIpred-coil, solvent accessibility, Shannon Entropy
  - 2-D features
    - cmpred, pstat pots, free contact
  - Chosen features were not transparent at first
- Goal: Predict secondary structure



Given an amino acid sequence...

MKTFLYFCLL...

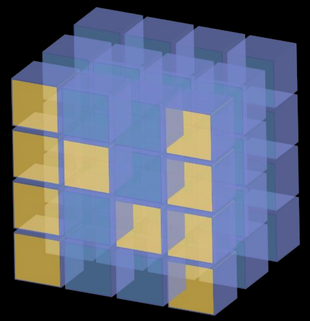
Predict secondary structure label

HHHCCEEEE...

Data Pre-Processing

1. Transform IEEE data into a standard format (numpy to CSV)
2. Augment IEEE data with extra features
   - Pull protein info from DSSP and PDB
   - Calculate torsion angles for each amino-acid pair in each protein
   - Add this to IEEE data
3. Segment distance matrix data using chipping strategy
   - Chip Sizes: 6, 11, 26
   - Strides: 3, 5, 12

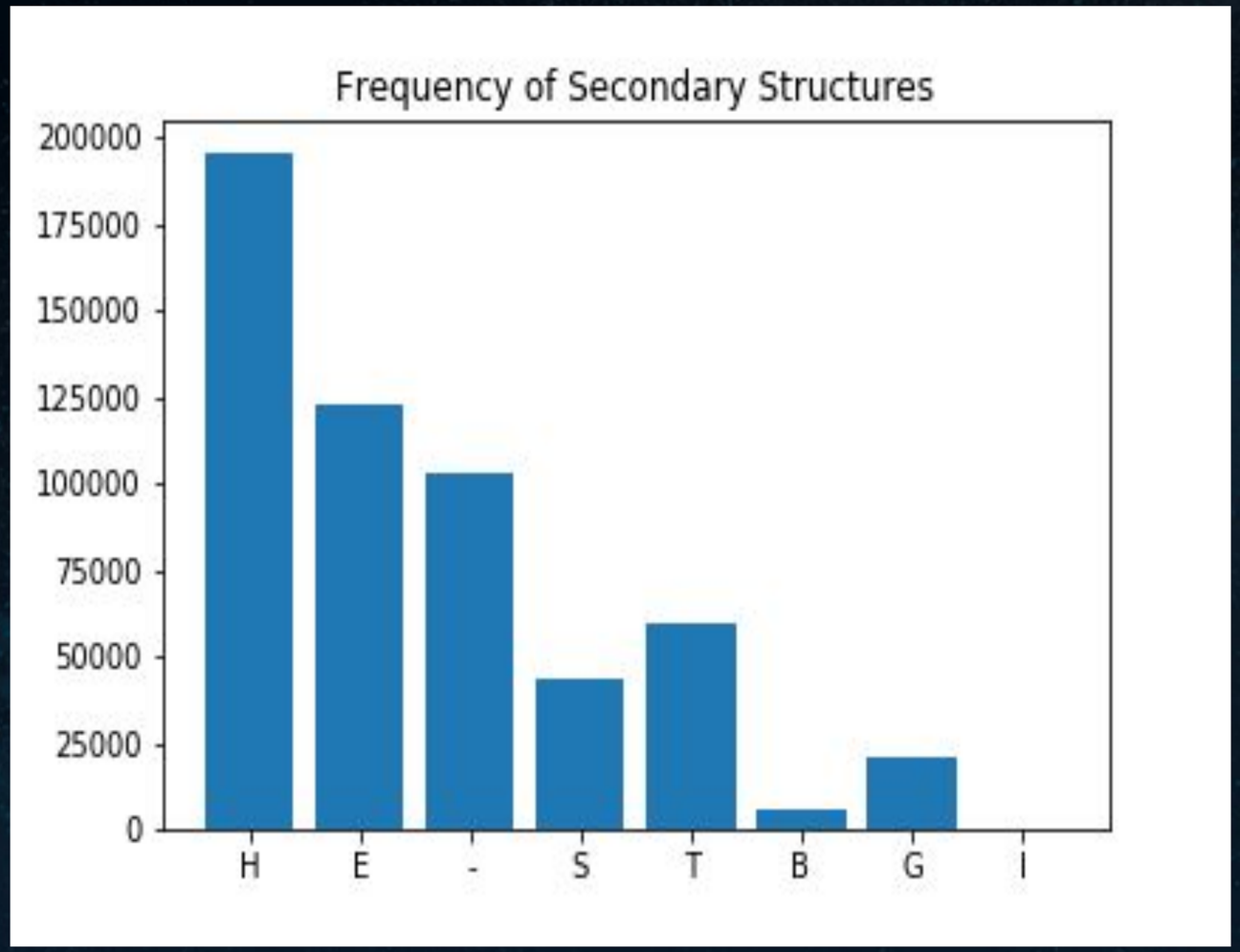
RCSB PDB
PROTEIN DATA BANK



NumPy

Secondary Structure Types and Frequencies

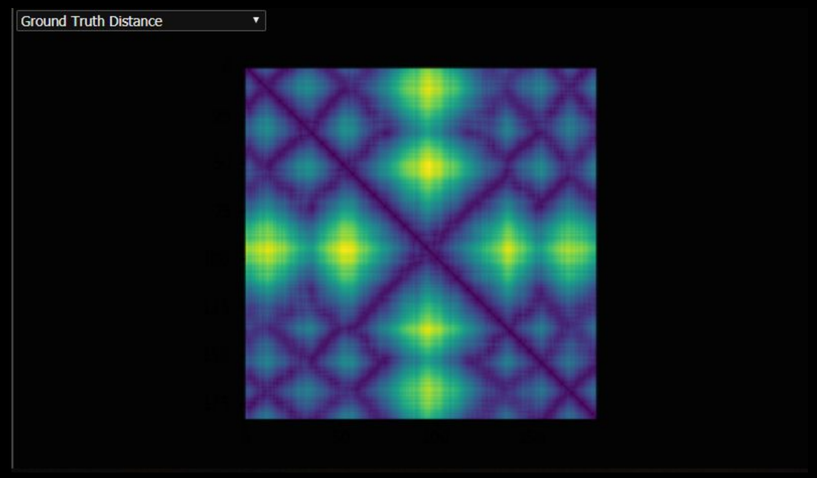
Key
H: Alpha Helix (Helix)
E: Extended Strand (Sheet)
-: Unknown (Coil)
S: Bend (Coil)
T: Hydrogen Bonded Turn (Coil)
B: Isolated Beta-Bridge (Sheet)
G: 3-Helix (3/10 Helix)
I: 5-Helix (Pi Helix)

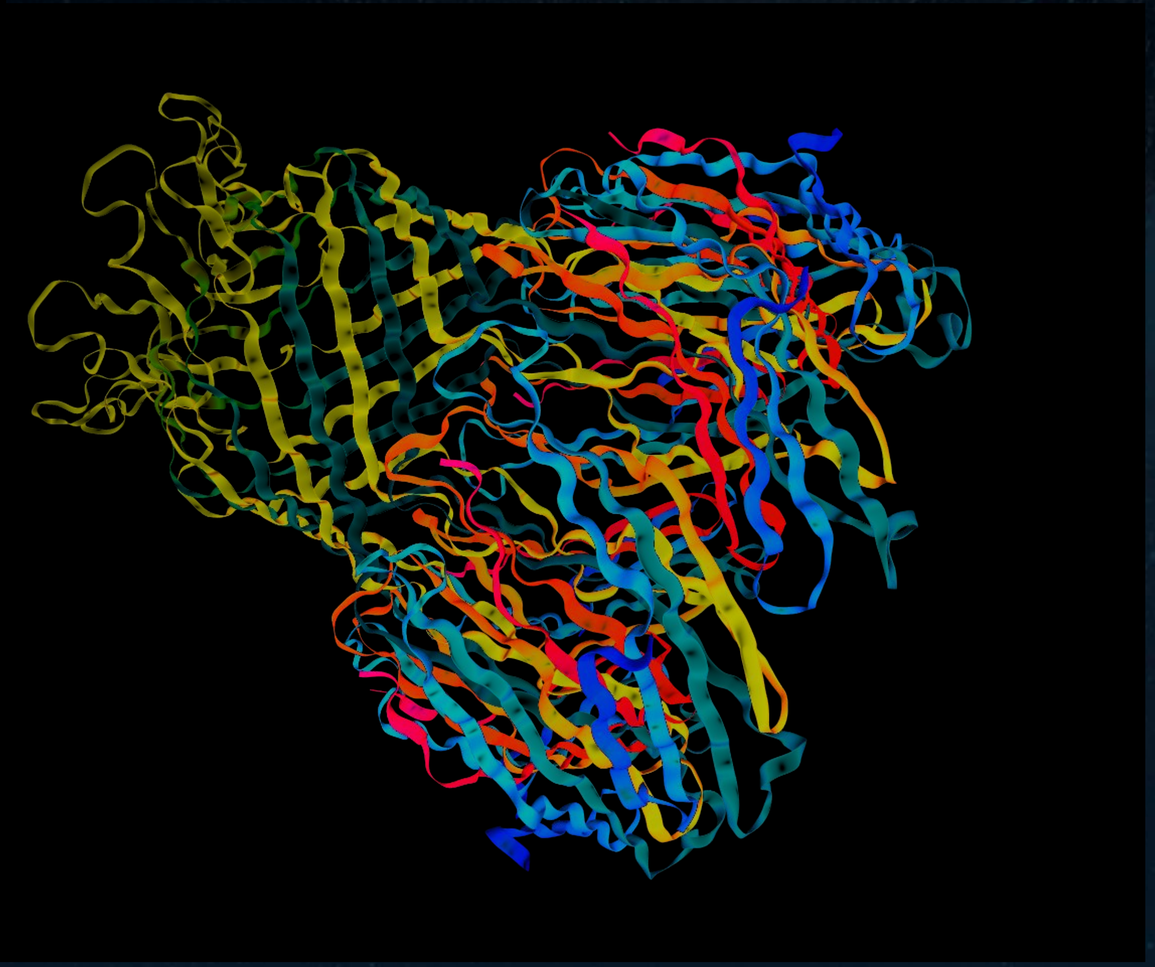


PDB Information

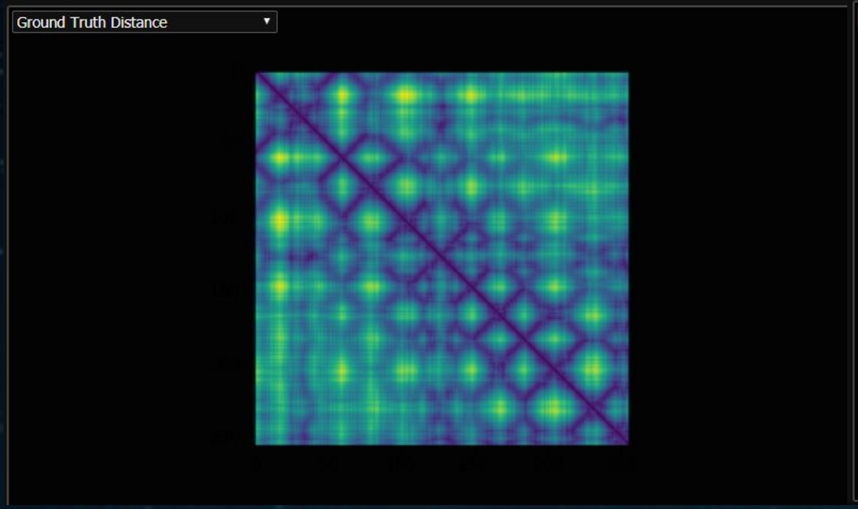
1UUN Information

- Classification: Porin
- Organism: Mycobacterium Smegmatis
- Expression System: E. Coli
- Mutations: 1
- Chains: A, B
- Sequence Length: 184
- X, Y, Z Positions

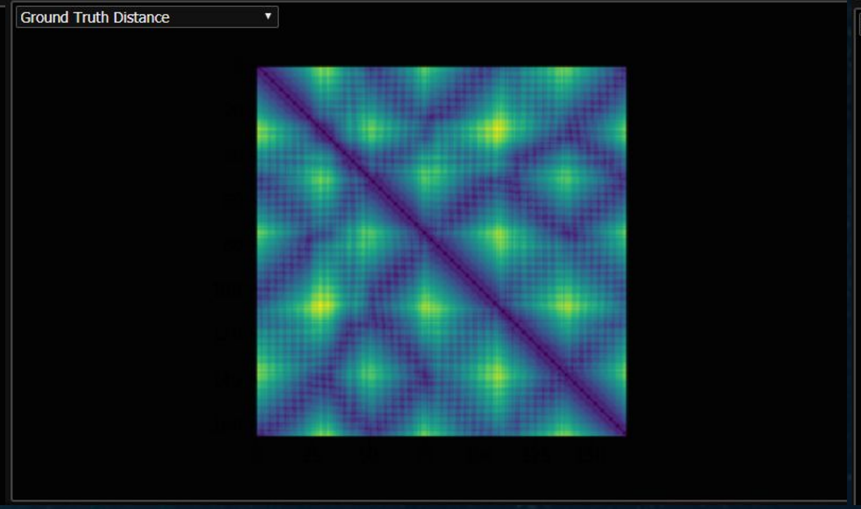




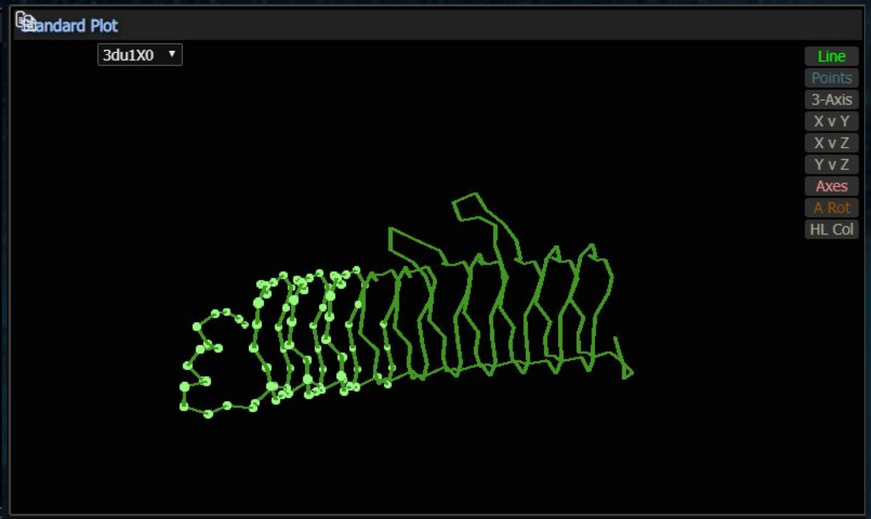
Protein of Sheets

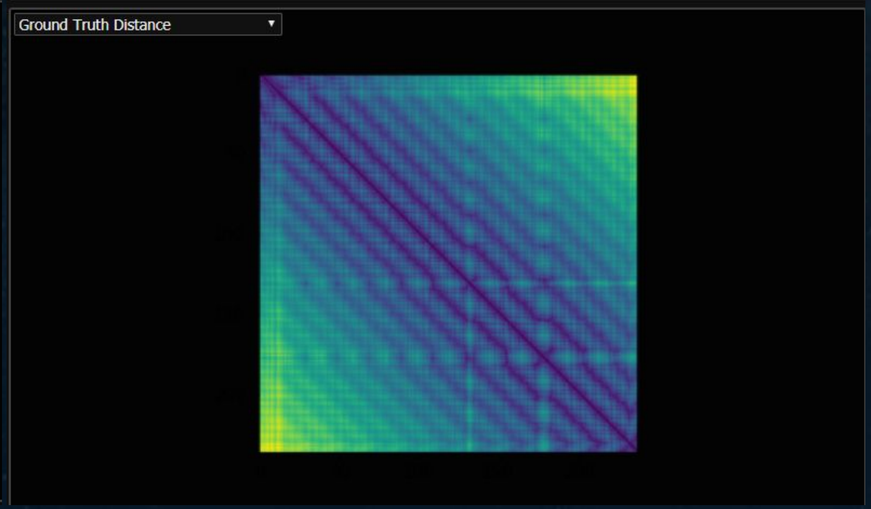


Protein of Helices

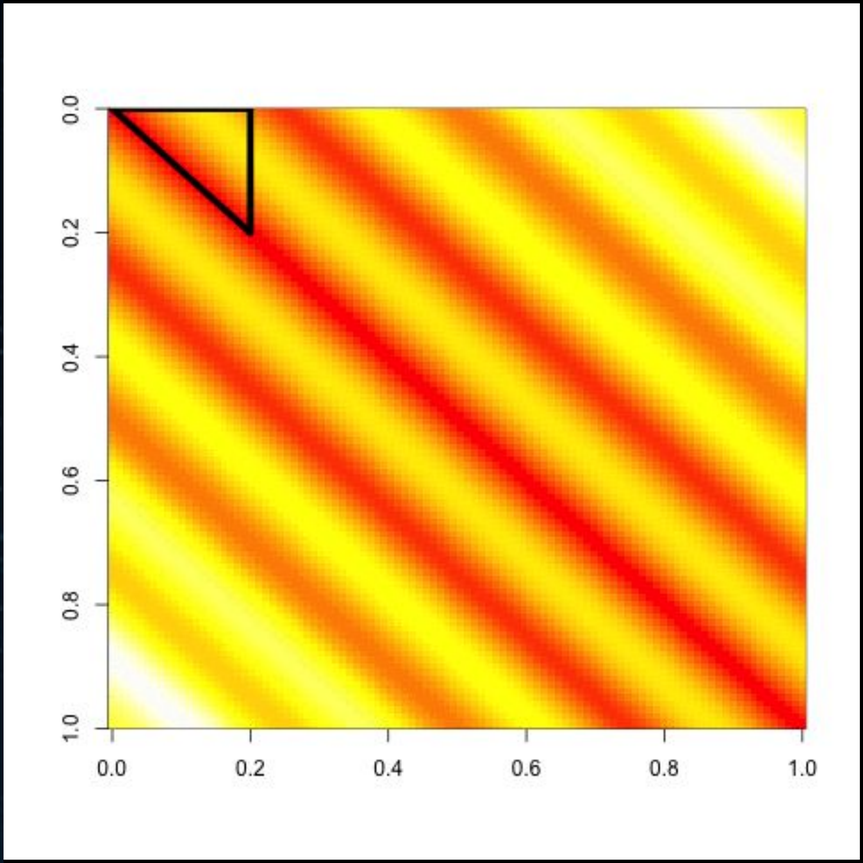


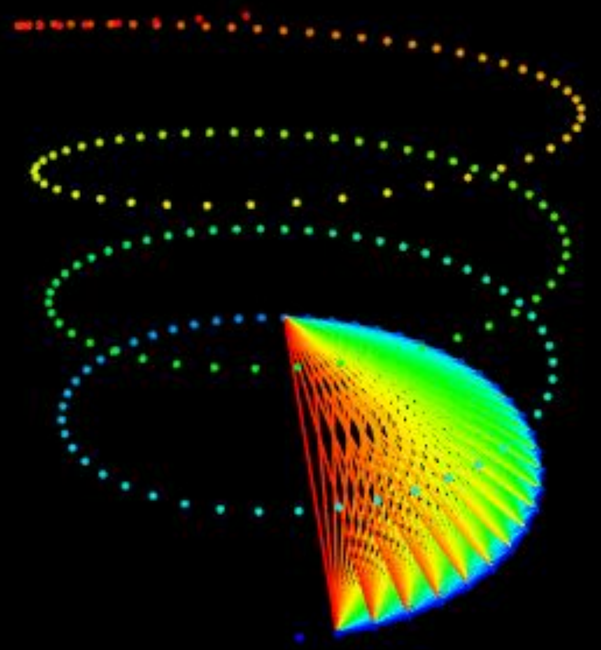
Protein of Coils





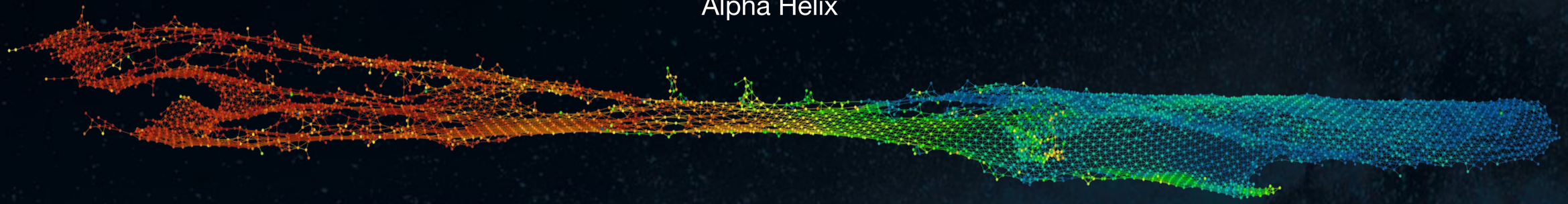
Chipping



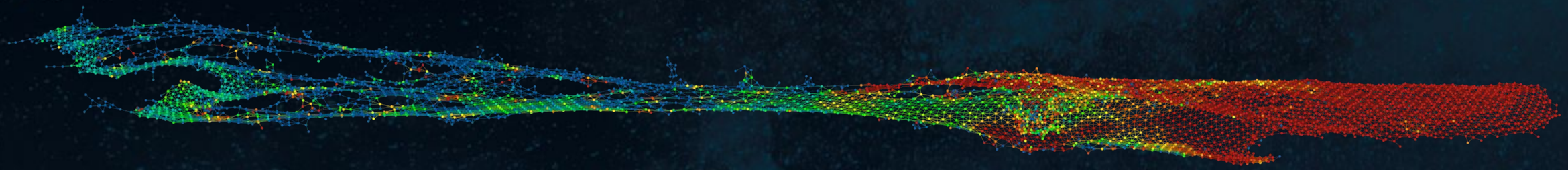


Protein Models

Alpha Helix

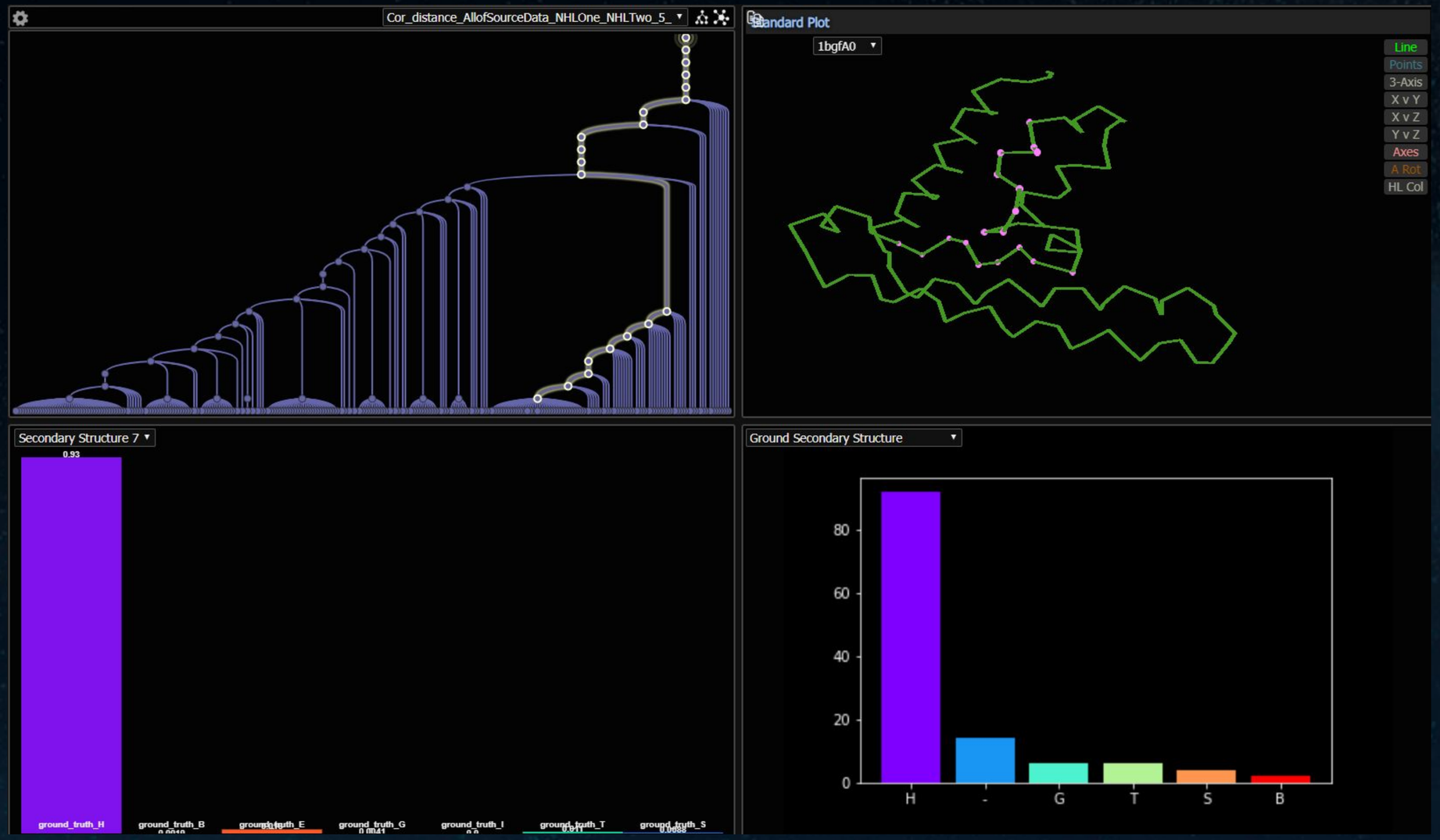


Beta Sheet

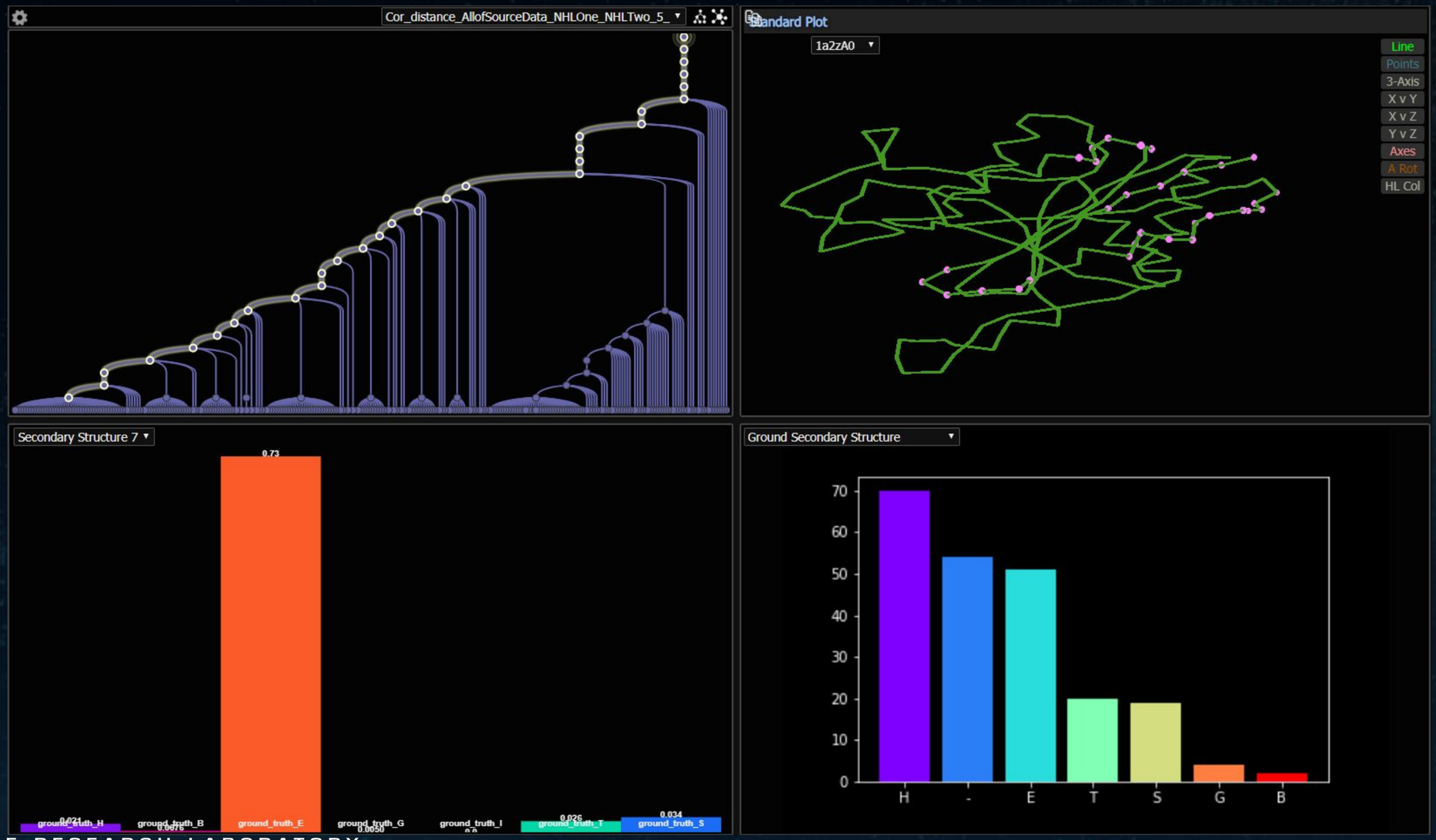


3/10 Helix

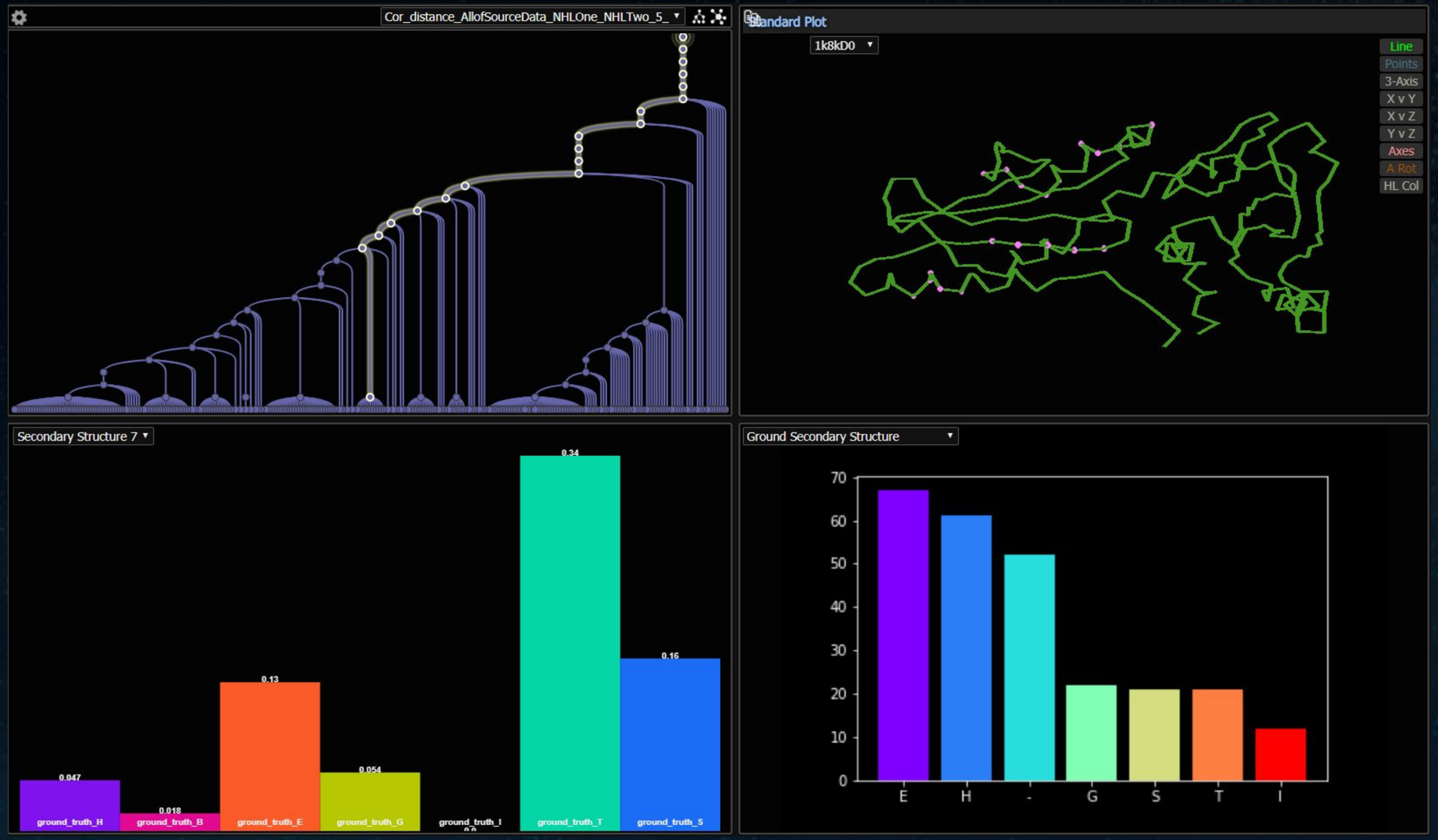
Interactive Dashboard for THDs - Alpha Helix



Interactive Dashboard for THDs - Beta Sheets

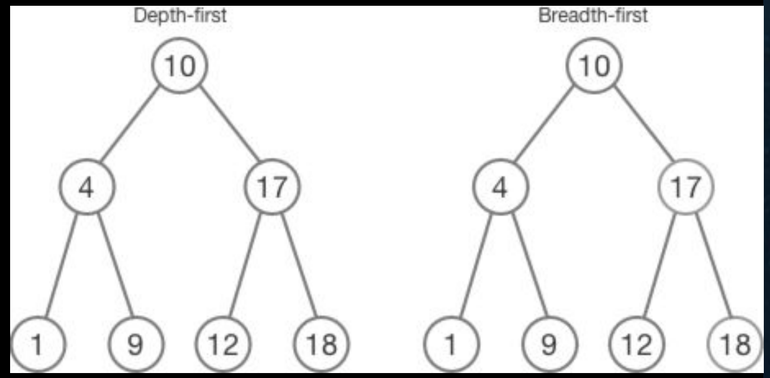


Interactive Dashboard for THDs - Coils



Prediction with Sequence Alignment

- Need means of predicting which *group* in the THD a given sequence is most similar to
- Solution: Compare sequences via biologically-motivated *edit distance*
  - i.e. minimum # additions, deletions, or edits, to transform one string to another
  - We use the mPAM*-matrix to weight the similarity measure
- Prediction strategy: Traverse THD in a bottom-up manner
  - Find node with minimum weighted-alignment distance
  - Use ℓ_2-loss to settle ties
  - Edit distance can be computed in $O(n^2)$
  - Finding the minimal node can be solved via dynamic programming



*Xu, Weijia, and Daniel P. Miranker. "A metric model of amino acid substitution." *Bioinformatics* 20.8 (2004): 1214-1221.

Sequence Alignment Results for 3cx5I0

	protein_id	chip_id	start	stop	node_ids	loss	file_id	aligned	subsequence
1	3cx5I0	chip_0	0	5	60	0	30.0.30	AEVTQL	AEVTQL
2	3cx5I0	chip_1	1	6	60	2	30.0.30	AEVTQL	EVTQLS
3	3cx5I0	chip_2	2	7	55	2	30.0.25	TQLSNG	VTQLSN
4	3cx5I0	chip_3	3	8	55	0	30.0.25	TQLSNG	TQLSNG
5	3cx5I0	chip_4	4	9	468	2	13.0.14	QLSNMI	QLSNGI
6	3cx5I0	chip_5	5	10	673	2	12.0.8	LSNEIV	LSNGIV
7	3cx5I0	chip_6	6	11	27	0	27.0.0	SNGIVV	SNGIVV
8	3cx5I0	chip_7	7	12	33	2	30.0.3	NIVYVA	NGIVVA
9	3cx5I0	chip_8	8	13	30	2	30.0.0	GIVATT	GIVVAT
10	3cx5I0	chip_9	9	14	485	0	14.29.16	IVVATE	IVVATE
11	3cx5I0	chip_10	10	15	486	2	15.16.31	AVVATH	VVATEH
12	3cx5I0	chip_11	11	16	483	2	12.0.1	VAEVDN	VATEHN
13	3cx5I0	chip_12	12	17	383	0	15.0.1	ATEHNP	ATEHNP
14	3cx5I0	chip_13	13	18	383	2	15.0.1	ATEHNP	TEHNPS

Prediction with Sequence Alignment

Prediction with Sequence Alignment

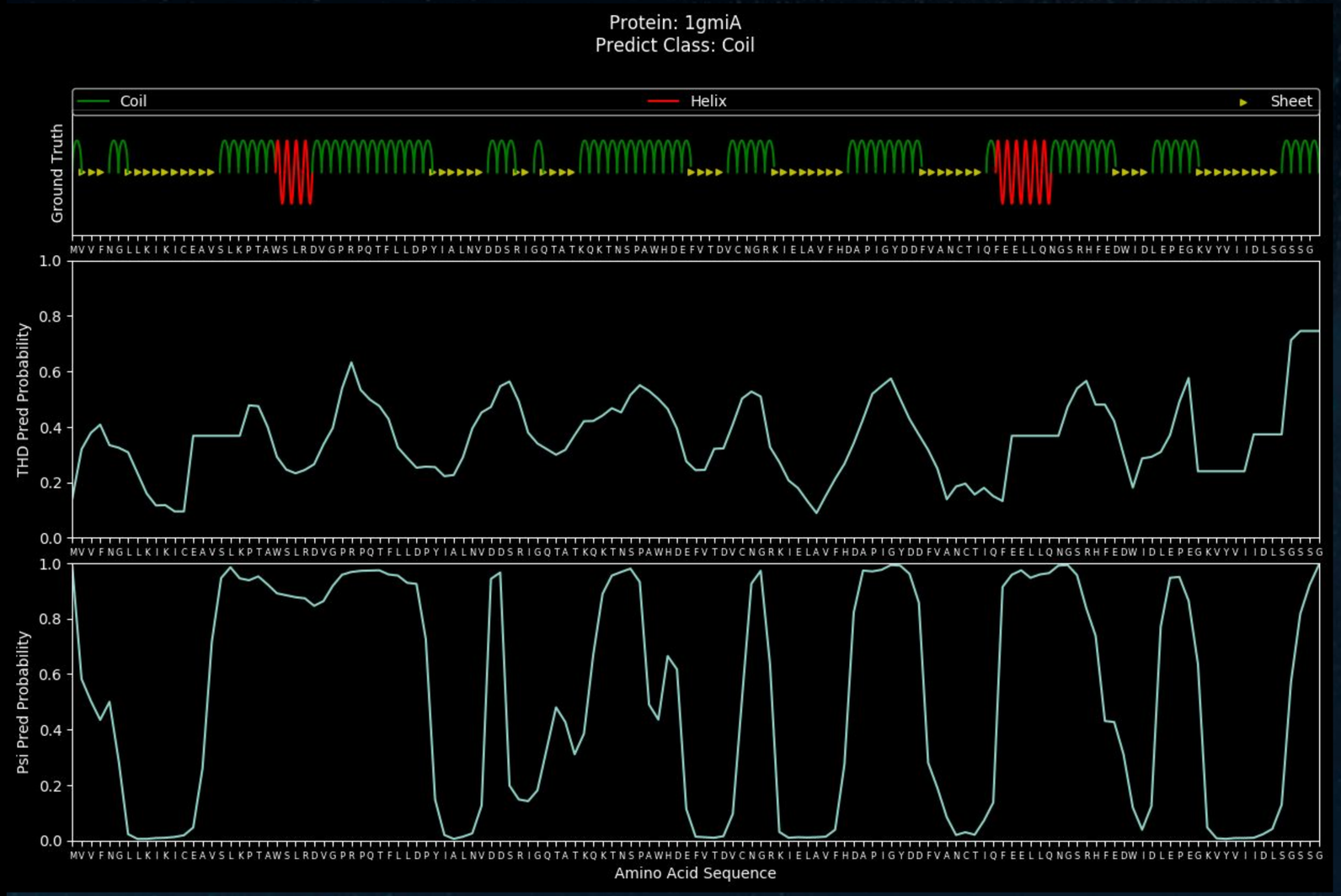
- Chip test protein with stride of 1
  - Find most similar chips within nodes of THD
  - For each node which is matched to a given subsequence, find the average secondary structure proportion across that node
- For each amino acid in the test protein, set the predicted secondary structure to the average secondary structure of the chip with the lowest edit distance
  - average between multiple chips in the case of ties

THD Pred vs PSI Pred for Coil

Ground Truth:

Our Predictions:

PSI Predictions:

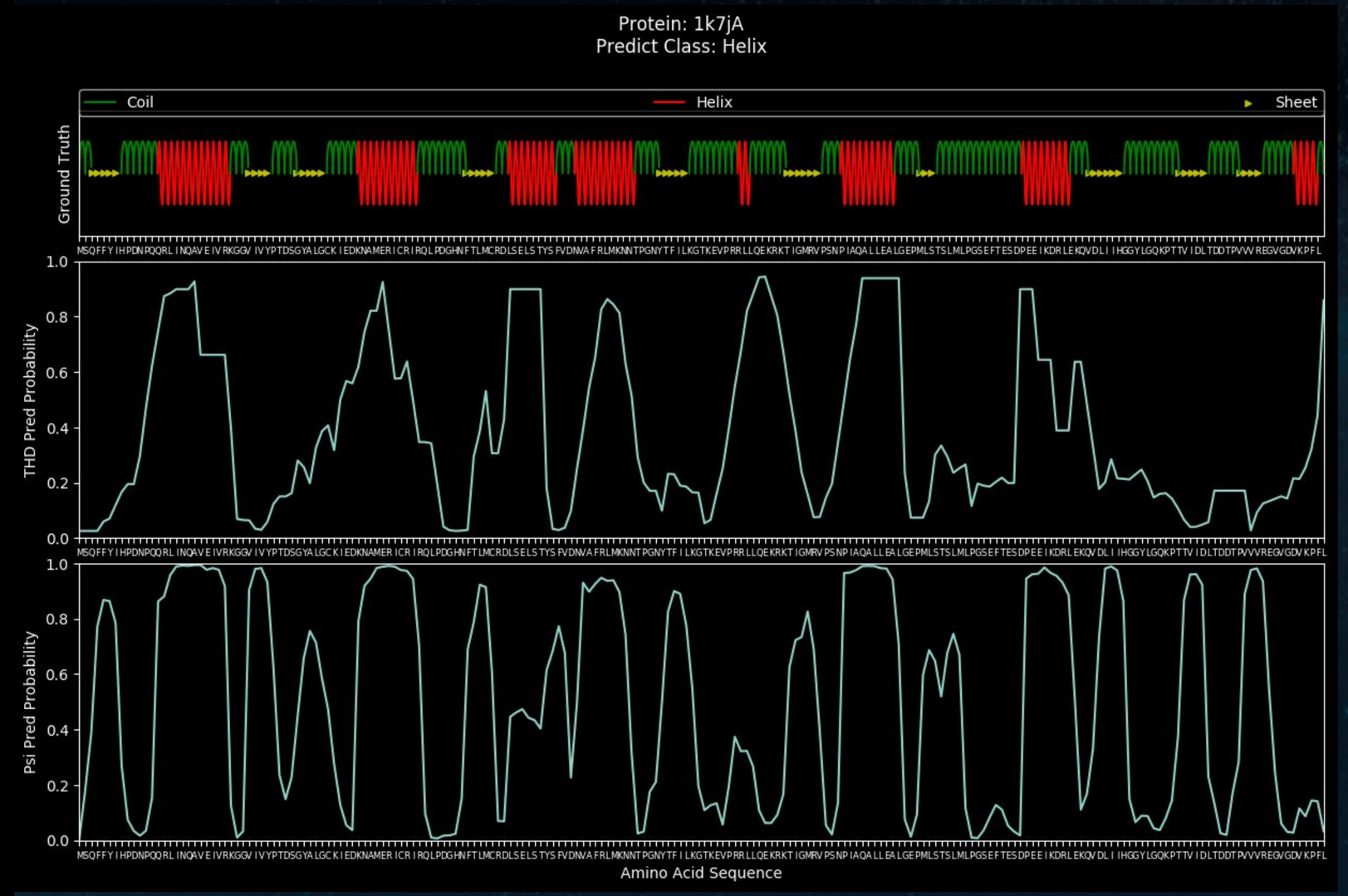


THD Pred vs PSI Pred for Helix

Ground Truth:

Our Predictions:

PSI Predictions:

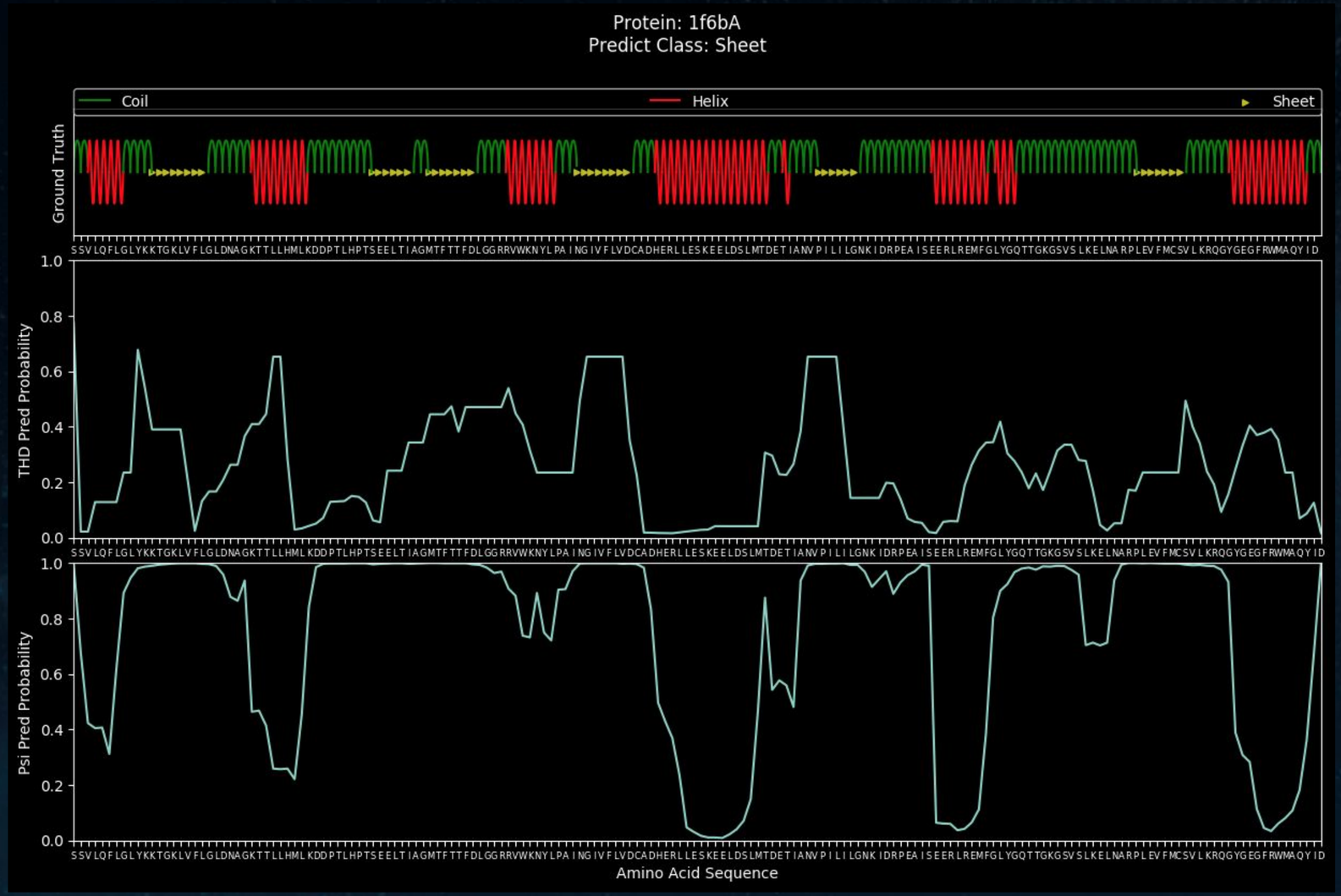


THD Pred vs PSI Pred for Sheet

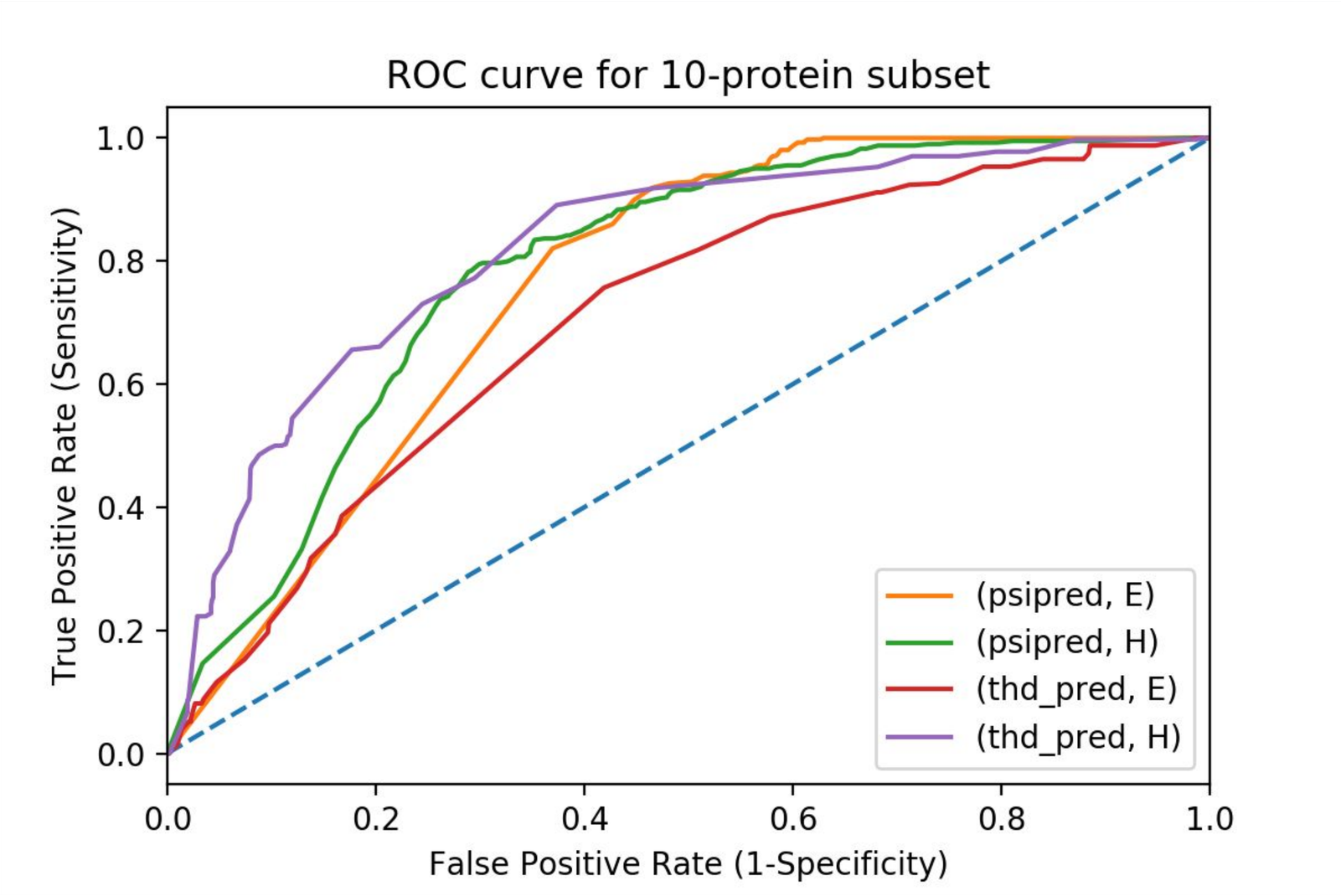
Ground Truth:

Our Predictions:

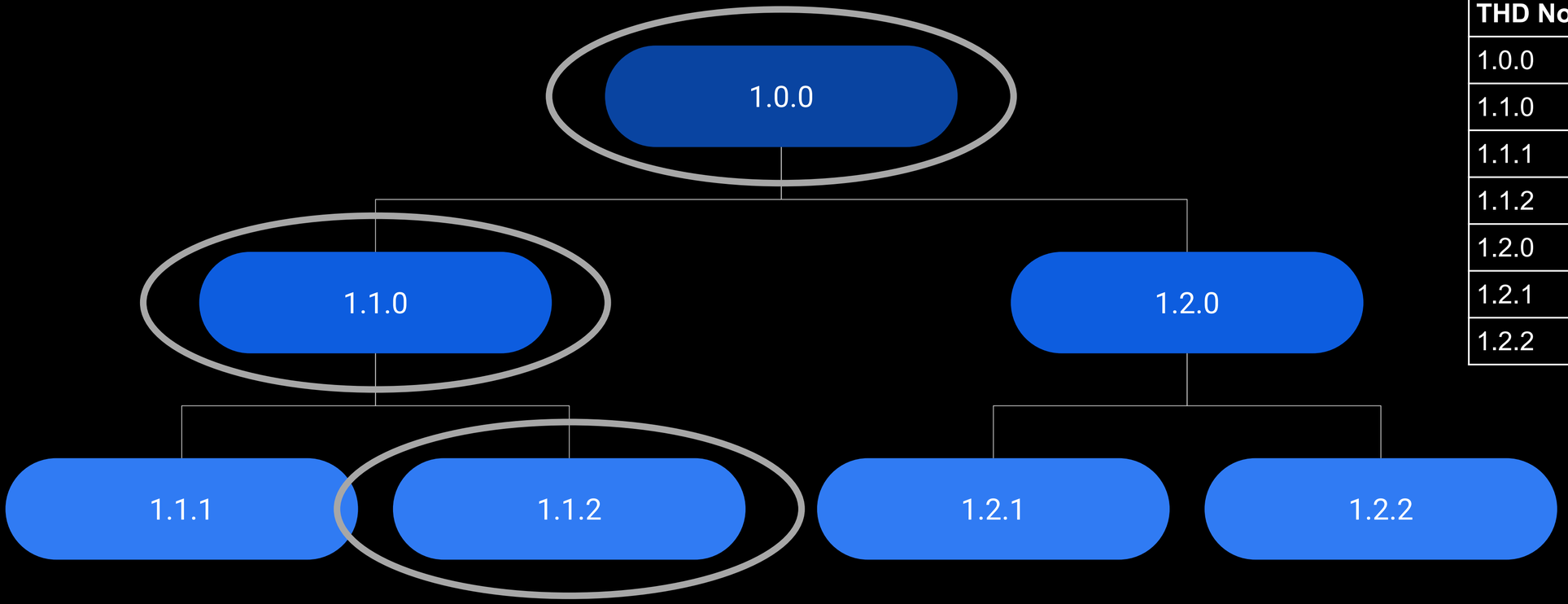
PSI Predictions:



THD Pred vs PSI Pred ROC



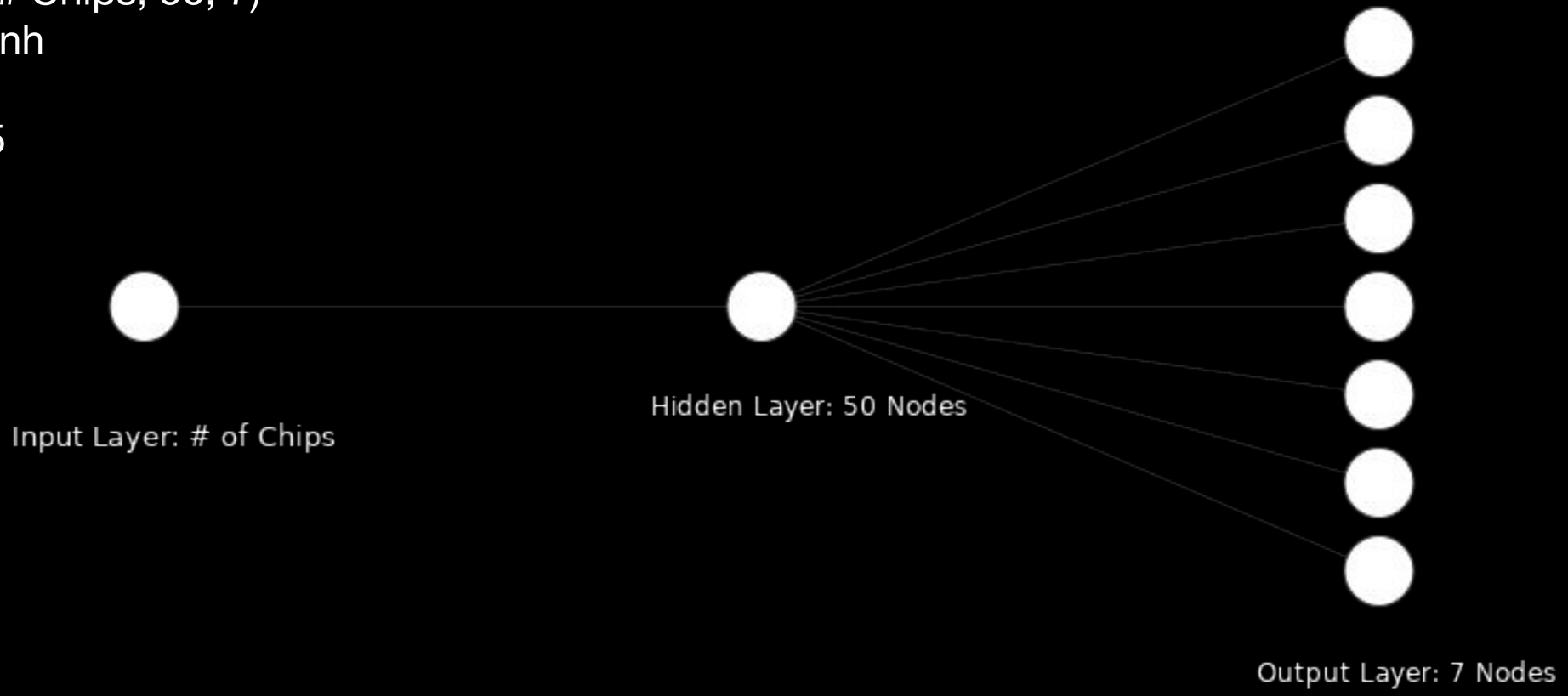
Secondary Structure THD Assessment



THD Node ID	Membership of Chip
1.0.0	1
1.1.0	1
1.1.1	0
1.1.2	1
1.2.0	0
1.2.1	0
1.2.2	0

Secondary Structure THD Assessment

1. Layer Sizes: (# Chips, 50, 7)
2. Activation: Tanh
3. Solver: Adam
4. Alpha: 0.0005



Secondary Structure THD Assessment

MAE for NN (hidden_layer_sizes=50, activation='tanh', solver='adam', alpha=0.0005) Predictions on Chips								
Data	ground_truth_H	ground_truth_B	ground_truth_E	ground_truth_G	ground_truth_I	ground_truth_T	ground_truth_S	All
Distance_10_5	0.126	0.028	0.128	0.063	0.007	0.088	0.076	0.074
Distance_All	0.14	0.018	0.144	0.059	0.002	0.096	0.081	0.078
Distance_25_12	0.173	0.035	0.135	0.061	0.025	0.074	0.065	0.081
Distance_5_3	0.144	0.022	0.155	0.063	0.004	0.108	0.088	0.084
Torsion_25_12	0.236	0.034	0.173	0.066	0.027	0.076	0.067	0.097
Torsion_10_5	0.323	0.02	0.227	0.065	0.01	0.108	0.088	0.12
Torsion_All	0.34	0.018	0.249	0.066	0.003	0.127	0.1	0.129
Torsion_5_3	0.375	0.019	0.278	0.069	0.005	0.155	0.116	0.145

THD vs Original Data | Times and Error

	Original		Group Membership	
Name	MAE	Time (s)	MAE	Time (s)
Distance_5_3	0.075	7.95	0.083	28.31
Distance_10_5	0.069	5.05	0.072	21.6
Distance_25_12	0.057	4.4	0.074	2.54
Torsion_5_3	0.073	8.08	0.144	33.19
Torsion_10_5	0.067	5.59	0.12	28.59
Torsion_25_12	0.056	4.94	0.093	2.6

Future Work

- Refine prediction strategy for secondary structure
  - Combine information from several THD's in one prediction
  - Aggregate multi-sized chip information
- Incorporate torsion angle information to predict secondary and tertiary structure
- Inspect other metrics
- Refine neural net optimization
- Predict long-range distances between amino acid residues
  - Competitive at predicting secondary structure
  - Could we incorporate tertiary structure into prediction strategy?