PREDICTING THE EFFECTS OF MUTATIONS ON PROTEIN STABILITY

UPDATES

AUM KHATLAWALA (2020113008)
SARTHAK AGGARWAL (2020101008)

MOTIVATION BEHIND USING 3D STRUCTURES AS INPUT

3D CNN (ThermoNet):

1. The ΔΔG of a point mutation can be sufficiently captured by modeling the 3D biophysical environment around the mutation site. Thus, the 3D input is given to understand the spatial arrangement of the atoms around the mutation site.
2. It thus extracts predictive features by treating protein structures as if they were 3D images and voxelizing the space around the mutation site of both the wild-type structure and the corresponding mutant structural model.

MOTIVATION BEHIND USING 3D STRUCTURES AS INPUT

ProS-GNN:

1. The idea is the same as the 3D CNN paper, but instead of voxelizing the space around the mutation site, the paper proposes that we get the molecular information (in terms of feature vectors) and the adjacency matrix for the atoms in the vicinity of the mutation site.

BASELINE MODEL

We were able to compile and run the baseline model of ProS-GNN on our local system and observed that the model outperforms the previous work (ThermoNet) by a considerable margin.

BASELINE MODEL – RESULTS

epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
0	3.105	1.740	0.134	0.402	1.330	37.569
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
1	2.387	1.545	0.212	0.396	1.245	35.505
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
2	2.144	1.615	0.277	0.404	1.286	35.019
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
3	2.046	1.525	0.327	0.427	1.225	34.696
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
4	2.050	1.546	0.319	0.424	1.234	34.643
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
5	1.913	1.662	0.375	0.429	1.244	34.412
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
6	1.870	1.508	0.403	0.438	1.247	34.819
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
7	1.880	1.416	0.399	0.470	1.196	35.010
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
8	1.827	1.396	0.421	0.480	1.193	34.900
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
9	1.769	1.459	0.450	0.451	1.224	34.317
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
10	1.819	1.357	0.426	0.484	1.179	33.913
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
11	1.768	1.490	0.447	0.426	1.230	34.119
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
12	1.778	1.443	0.446	0.471	1.191	34.034
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
13	1.745	1.418	0.457	0.463	1.201	34.090
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
14	1.722	1.449	0.469	0.463	1.203	33.930
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
15	1.748	1.458	0.457	0.474	1.191	34.075
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
16	1.742	1.417	0.460	0.476	1.198	34.197

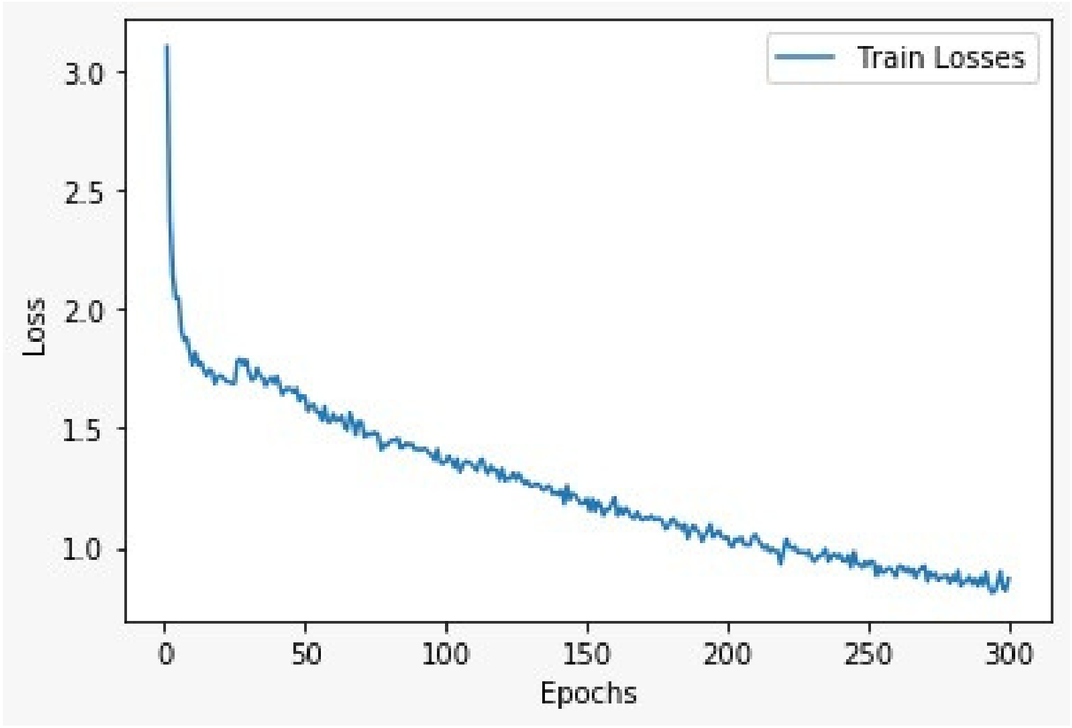
BASELINE MODEL – RESULTS

epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
283	0.854	1.298	0.783	0.562	1.141	33.321
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
284	0.859	1.354	0.782	0.569	1.162	33.408
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
285	0.877	1.279	0.778	0.596	1.134	34.004
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
286	0.847	1.262	0.787	0.587	1.136	33.486
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
287	0.868	1.235	0.780	0.592	1.120	33.563
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
288	0.838	1.471	0.791	0.546	1.193	33.455
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
289	0.874	1.458	0.779	0.506	1.197	33.635
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
290	0.843	1.185	0.786	0.624	1.098	33.231
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
291	0.896	1.355	0.771	0.577	1.145	33.459
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
292	0.839	1.290	0.788	0.577	1.144	34.246
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
293	0.810	1.389	0.796	0.544	1.177	33.194
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
294	0.814	1.679	0.795	0.521	1.207	33.445
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
295	0.854	1.198	0.785	0.598	1.104	33.775
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
296	0.899	1.363	0.771	0.584	1.141	33.652
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
297	0.827	1.200	0.790	0.607	1.102	33.862
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
298	0.818	1.384	0.794	0.568	1.167	33.479
epochs	train_losses	test_losses	pcc_train	pcc_test	rmse	time
299	0.870	1.208	0.778	0.597	1.104	33.277

COMPARISON

1. ProS-GNN (claimed) (400 epochs):
   - Root Mean Square Error: 1.23
   - Pearson Correlation Coefficient: 0.61
2. ProS-GNN (running locally):
   - Root Mean Square Error: approx. 1.15
   - Pearson Correlation Coefficient: 0.597
3. ThermoNet:
   - Root Mean Square Error: 1.56
   - Pearson Correlation Coefficient: 0.47

BASELINE MODEL – TRAINING LOSS CURVE



BASELINE MODEL – TEST PCC CURVE