1 Method

1.1 Generating Crops

The generation of crops was implemented as specified in the assignment and lectures. The feature matrix was created using the sequence and evolutionary data, and then split into 64×64 sized crops for the model. An IterableDataset to wrap the crop generation was implemented so that the pytorch dataloader can easily send batches of them to the model.

1.2 Model

The current model uses 4 set of 4 dilation blocks. More would be preferred, however memory issues were encountered when increasing the number of blocks beyond this. The learning rate was manually decreased from 0.025 to 0.005 throughout training. After some experimentation, it was found that a batch size of 20 was the largest possible given memory constraints.

1.3 Training

Training was fairly straightforward and was carried out on the GPU. Each of the 25 epochs took approximately 25 minutes for a total training time of about 10.5 hours. This was done in sessions throughout the past few days, with the model‘s state_dict being saved and loaded after and before the training. Contact prediction accuracy was not calculated during training.

1.4 Validation

A validation set was used both to tune the hyperparmeters and to determine when to stop training. As coded, training will continue until the validation accuracy is higher than the 5-epoch rolling accuracy sum. Unfortunately, this point was never reached; time was the limiting constraint in training.

2 Evaluation

Evaluation for each sequence was split into multiple steps as described below.

2.1 Crop Aggregation

Because of the nature of batching by number of crops, each batch may contain crops from several different sequences. This was handled by assigning each sequence its own unique ID, and checking each crop‘s associated sequence ID.

For all crops from a given sequence, softmax was applied to their features, and the crops were combined into a sequence_length-sized 2D tensor that held the probabilities of contact at each index pair. It is worth noting that crop generation for evaluation was done with a crop stride of 32 (half of the total crop size), thus there was a significant amount of overlap between crops. Once all crops from a sequence were processed into the aggregate tensor, the accuracy was calculated.

2.2 Calculation of Accuracy

For each sequence, accuracy was calculated for short, medium, and long contacts using the top $\frac{L}{k}$ predictions, where $k = \{1, 2, 5, 10, 20, 50, 100\}$. The assignment states that if the probability of contact between 2Å and 8Å is greater than 0.5, that there is a contact. The current model, however, does not predict enough contacts to allow for correct calculation of accuracy for the top L predictions. As such, contact accuracy with the 0.5 threshold is not entirely accurate. In the output below, there are tables for accuracy with and without the threshold.

3 Results

Accuracies for top $\frac{L}{k}$ predictions:

k	Short	Medium	Long
1	0.230443	0.221818	0.230140
2	0.347676	0.307315	0.296237
5	0.511692	0.431034	0.376984
10	0.615373	0.506860	0.438546
20	0.693295	0.586715	0.480660
50	0.762319	0.630435	0.526087
100	0.818116	0.660507	0.556522

Accuracies for top $\frac{L}{k}$ predictions with thresholding (>0.5):

k	Short	Medium	Long
1	0.736420	0.759364	0.784539
2	0.736420	0.759364	0.784539
5	0.736397	0.760673	0.785875
10	0.742424	0.762611	0.794528
20	0.763906	0.774161	0.829209
50	0.800072	0.797717	0.847826
100	0.835290	0.825797	0.849275

3.1 Conclusion

As shown above, the contact prediction accuracy is relatively good when thresholding, but the accuracy falls off significantly when the predictions are not thresholded and when greater than $\frac{L}{10}$ of the predictions are considered. From this, it can be concluded that the model is not predicting enough contacts with high enough confidence to effectively estimate the protein structure. However, it is worth noting that while training, the prediction accuracy has only increased. Given more time, and perhaps more dilation blocks, the model may be able to achieve much better accuracy.

3.2 Script Output

Note that not all of the training output was recorded, as it was done over several days.

```
Epoch 22, 70,880 crops:
        Train loss per batch = 0.049809
        Valid loss per batch = 0.052133
Epoch 23, 73,640 crops:
        Train loss per batch = 0.047115
        Valid loss per batch = 0.050202
Epoch 24, 69,980 crops:
        Train loss per batch = 0.046847
        Valid loss per batch = 0.050519

Test loss per crop: 0.022741

---Accuracies for L/k sequences--
        short       med      long
1    0.230443  0.221818  0.230140
2    0.347676  0.307315  0.296237
5    0.511692  0.431034  0.376984
10   0.615373  0.506860  0.438546
20   0.693295  0.586715  0.480660
50   0.762319  0.630435  0.526087
100  0.818116  0.660507  0.556522
```