

MSSE 277B: Machine Learning Algorithms
Homework assignment #10:
Recurrent Neural Network, LSTM
Assigned Apr. 13 and Due Apr. 25

1. **LSTM applied to SMILES string generation.** (10 pt) Using the SMILES string from the ANI dataset with upto 6 heavy atoms, build a LSTM generative model that can generate new smiles string with given initial character.
 - (a) (3pt) Process the smiles strings from ANI dataset by adding a starting character at the beginning and an ending character at the end. Look over the dataset and define the vocabulary, use one hot encoding to encode your smiles strings.
 - (b) (7pt) Build a LSTM model with 1 recurrent layer. Starting with the starting character and grow a string character by character using model prediction until it reaches a ending character. Look at the string you grown, is it a valid SMILES string?