## Generative Modeling of Small Molecules Using VAE

### Data Preparation

During the data preparation phase, the qm9.csv dataset was accessed directly from the project requirements. Once the dataset was loaded, I split the dataset into training and validation subsets, using an 80:20 ratio(>50%), which is a common practice to ensure both robust training and effective model validation on unseen data. To facilitate efficient processing and consistency in input data format, each SMILES string in the dataset was padded to a uniform length of 29 characters, the maximum length observed in the dataset. This padding ensures that all input data fed into the model maintains a standardized shape, crucial for the neural network to process the data effectively without introducing bias toward shorter sequences. Thus, it might experience a decrease in computing performance.

### Model Selection

In the final project, I chose to implement a Variational Autoencoder (VAE) for synthesizing small molecules, due to its effectiveness in learning meaningful latent spaces and my familiarity with the model. The VAE architecture features an encoder that compresses input SMILES strings into a latent representation, and a decoder with LSTM units to reconstruct these strings, maintaining long-term dependencies essential for accurate molecular modeling. The training process involves minimizing reconstruction errors using binary cross-entropy loss and optimizing the latent space via Kullback-Leibler divergence. This configuration helps capture the complexity and diversity of molecular structures effectively.

### Training the Model

To train the Variational Autoencoder (VAE) model for SMILES representations, a tokenizer was used to convert chemical structures into numerical indices based on a constructed vocabulary from training data. These indices were padded and turned into batches using PyTorch Data Loaders. The VAE with an LSTM-based encoder and decoder was trained to minimize a loss function accounting for reconstruction accuracy and latent space continuity. Training involved forward and backward passes with the Adam optimizer, and the model's generalizability was validated after each epoch. Finally, training and validation losses were plotted to ensure convergence, and the model parameters were saved post-training.

### Evaluation

* Validity: 0.17, Novelty: 0.94, QED: 0.47
* Valid SMILES: ['CC(C)CCC(OC)C([O-])=O', 'CCCCCCCCCCC', 'CCCCCCCCCCCC=O', 'CCCCCC=O', 'CCCCCCCCCCC', '[NH3+]CCCC([O-])=O', 'CCCCCCCCCCCC', 'O1C2C3C1C1CCCC3C21', 'CCCCCCCCCCCN', 'CCCCCO', 'CCCCCCCCCCCCO', 'CCCCCCCCCCCCO', 'CC1cnnn1', 'CC1CCCCCCCC(C)C=1', 'CC1(CCCCCCCCC)CC1', 'CC1CCCCCCCCCCC1', 'CC(CCCCCC#C)C#N']

### Conclusion

In summary, the VAE model for SMILES representation faces challenges, including potential information loss or redundancy due to fixed-length sequences and padding methods. The model's generative capabilities could also be limited by the diversity and size of the training dataset, potentially leading to a lack of novelty in the generated molecules.

To address these limitations, exploring more sophisticated representations like Graph Neural Networks could capture the topological intricacies of molecules better than sequence-based methods. Enhancing the decoding architecture with advanced techniques such as RNNs or Transformers could also improve the model's ability to manage longer sequences with more accuracy.

For increasing molecular diversity, incorporating regularization strategies that encourage variety and leveraging enhanced VAE architectures can promote a richer latent space exploration. Additionally, extending the training dataset with more diverse molecular structures will likely improve the model's generality and creativity in molecule generation. Enhancing the model with these improvements aims to yield higher quality and more diverse molecular predictions, broadening the potential applications of the model in computational chemistry and related fields.

A graph with blue and orange dots

Description automatically generated

A screenshot of a cell phone

Description automatically generated