# Generating basic molecules with Sequential VAE

## Making unique molecules out of CHONF with a Sequential VAE

Adam Xu, Chris Ho, Faz Zaidi

Subset of Louis

## Introduction

Our group is using Variational Autoencoders to generate novel molecules made up of carbon, hydrogen, oxygen, nitrogen, and fluorine atoms.

Our goal is twofold. The first is to generate chemically stable atoms that are actually possible. These molecules cannot violate fundamental laws of physics and chemistry. The second is to actually generate novel molecules. These molecules should not all be replicas of existing molecules. For example, we don't only want our molecule to generate carbon dioxide or hydrogen peroxide.

Our group members are particularly drawn to this problem because we have had experience working for pharmaceutical companies, AI startups, engineering companies, and more. We have real-world experience, particularly in the fields of AI, material science, and pharmaceuticals and would like to apply the knowledge we have gained in this DL Course to the real world to engage in real life problem solving that benefits society.

## DATA

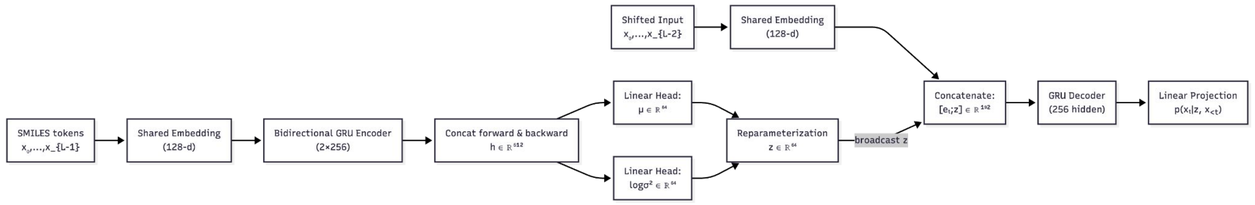
Dataset: We are using the QM9 dataset. It's widely used for basic molecular machine learning tasks. It contains 135k small stable molecules made up of C, H, O, N, F atoms. Each molecule in the dataset has two parts: a molecular structure (3D coordinates) and a set of properties (dipole moment, enthalpy, etc.).

Pre-Processing: We will be converting each set of coordinates into a SMILES string (which represents the molecule as a set of valid chemical operations). We will be tokenizing at the atom level. We will use RDKit to process and check for uniqueness. We'll build a vocabulary of the SMILES symbols, [PAD], [BOS], [EOS], and [UNK]. Then, we'll convert each sequence to token IDs and pad to a consistent length.

## METHODOLOGY

The Model: For our model, we are going to use a sequential VAE given the operation-based nature of SMILES strings. The goal is to create a latent space for SMILES strings that we can then decode off of to generate novel molecules. After tokenizing and converting to embedding vectors, we are going to use a bidirectional GRU (gated recurrent unit) to read through the sequence in order. We'll then just append their two hidden states together to create our z.

| Name | Elemental formula | SMILES String |
|---|---|---|
| Methane | $CH_4$ | C |
| Acetone | $C_3H_6O$ | CC(=O)C |
| Urea | $CH_4N_2O$ | O=C(N)N |
| Nicotine | $C_{10}H_{14}N_2$ | CN1CCCC1C2=CN=CC=C2 |
| Caffeine | $C_8H_{10}N_4O_2$ | Cn1cnc2n(C)c(=O)n(C)c(=O)c12 |




Decoder: Our GRU autoregressive decoder will start with the [BOS] token and our sampled z. Then it'll generate the next token in the SMILES sequence. Then it'll take that token and z and generate the next. It'll do this until it predicts [EOS].

Loss/Opt: We will use a combined cross-entropy and KL loss to ensure we have a smoother latent space and an Adam Optimizer. We'll also use teacher forcing during training so it doesn't spiral into a very incorrect molecule and not learn.

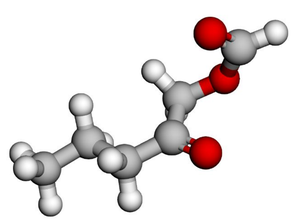
$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}}$$

$$\mathcal{L}_{\text{recon}} = -\frac{1}{N_{\text{tok}}} \sum_{\text{tokens } l} \log p_\theta\left(x_l \mid x_{<l}, z\right)$$

$$\mathcal{L}_{\text{KL}} = \frac{1}{2D} \sum_{j=1}^{D} \left(\mu_j^2 + \sigma_j^2 - 1 - \log \sigma_j^2\right).$$

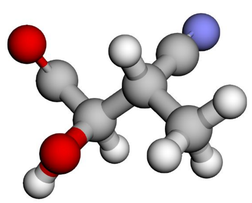
Generating Molecules: We'll finally generate new molecules just by sampling a random z, feeding into the decoder, and then using RDkit to check its uniqueness and validity.

Some valid molecules we generated that were novel to our dataset:



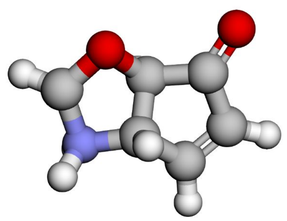
$C_6H_{10}O_3$ or

[H]C(=O)OC([H])([H])C(=O)C([H])([H])C([H])([H])C([H])([H])[H]



$C_5H_7NO_2$ or

[H]O[C@@]([H])(C([H])=O)[C@]([H])(C#N)C([H])([H])[H]

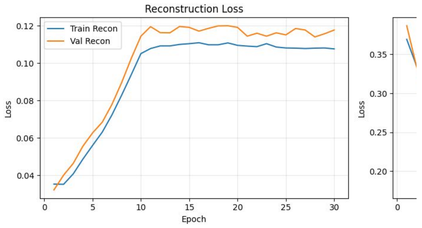


$C_6H_9NO_2^{2+}$ or

[H]C1=C([H])[C@]2([H])[NH2+][C+][H])O[C@@]2([H])C1=O

Gray: Carbon (C), White: Hydrogen (H), Red: Oxygen (O), Blue: Nitrogen (N)

## RESULTS




At this point, our model is generating molecules that are valid 98% of the time and novel 65% of the time. A valid molecule is a molecule that is chemically possible, meaning that it follows the fundamental laws of chemistry and physics. A novel molecule means that it is not found on the QM9 dataset we used to train our VAE. These results are fantastic, as we have already far exceeded our stretch goal.

## DISCUSSION

Lessons Learned: In terms of lessons learned, we learned to really experiment with model hyperparameters. Certain hyperparameters that we dismissed as being counterintuitive to change ended up being extremely helpful for our model's accuracy. Another lesson we learned is to spend more time exploring cutting edge AI research and datasets, as there is a lot of opportunity for interesting projects that come out of these projects.

Limitations: A core limitation of our project is that it is limited to 5 elements on the periodic table when there are 100+. More work particularly should be done on more understudied elements that could have a transformative potential.

Future Work: We hope that scientists and engineers that have more advanced knowledge in their domains of expertise (ie. medicine and materials) take the models we have created, make them applicable to more elements, and apply them to their research. We ultimately hope that this will help pioneer transformative innovations in these sectors that will help society and ultimately humanity.