**Preprocessing**

**ChemBERTa**

Machine learning models typically receive discrete or continuous quantitative data as inputs. The dataset previously mentioned is formatted as the SMILES representation followed by the other information about its molecular properties. Unfortunately, we cannot directly use the SMILES representation as an input to our model as they are a string of numbers and letters that represent its structural information. I had needed to find vector representations of these SMILES without losing critical information about their structures.

To solve this problem I decided to use a ChemBERTa natural language processing model to do this. Specifically, I decided to use a model trained on 100k SMILES strings from a benchmark dataset, ZINC. For the earlier stages of testing, this model will have suitable accuracy. In later stages of testing I will experiment with using different fine-tuned RoBERTa based models to assess performance and the effect they have on model performance.

A colorful dots on a white background

Description automatically generatedTo visualise the performance of the model for converting SMILES to Vectors I have used t-SNE. T-SNE is a statistical method for visualising high-dimensional data by reducing each datapoint to a location in two or three dimensions. The result of passing a SMILES representation through the ChemBERTa model is a Vector of size 768. Therefore, it is critical to reduce these dimensions for visualising. I first tested a smaller dataset of ~550 related molecules. To enable us to see patterns in the data clearer and before using t-SNE to visualise them, I used the k-means algorithm with the elbow method of cluster analysis to approximate the best number of clusters and provide every molecule with an appropriate cluster label. After doing this I then passed these vectors into our t-SNE algorithm and displayed the data using the previously generated labels to enable us to better understand the produced graph.

A colorful dots on a white background

Description automatically generated with medium confidence4 Clusters 3 Clusters

A screen shot of a computer screen

Description automatically generated

A graph with a line

Description automatically generatedA graph with a blue line

Description automatically generated44 Clusters – Elbow Method was used for this to calculate the optimal number of clusters

The above graphs shows the smaller dataset clustered multiple times with varying cluster numbers and visualised using t-SNE and two example graphs of the convergence of the elbow method.

When you hover over individual data points, you are provided with the labels and other information about the molecule.

A yellow background with black lines

Description automatically generated

A yellow background with black lines

Description automatically generated

Respective Molecules Skeleton for Arachidonoyl Serotonin and AM-404

A chemical structure of a molecule

Description automatically generatedA structure of a chemical formula

Description automatically generated

Aswell as clear patterns in the locations and distributions of clusters, As noted by the molecules shown above, cluster items have similar structural properties. I am confident that the ChemBERTa model provides more than adequate conversions of SMILES to Vectors, without losing the most important structural information.

Limitations – With larger datasets that require more clusters, without having the ability to hover over each data point and check their specific cluster numbers, it may be hard to see specific clusters because of similar shades of colours, I am working on a fix for this.

I have also experimented with the use of PCA for the visualisation of the data, but I have not seen any additional advantage over t-SNE.

A black and white arrow pointing to a black rectangle

Description automatically generated

The above image displays the simplified process of this component of the project.

Below display and example of the initial dataset provided.

A screenshot of a computer

Description automatically generated

After processing all the data from the original dataset above, an inputs csv is created for use with the Variational Auto-Encoder during training.

A screenshot of a computer

Description automatically generated

**Target/Ground Truth generation – Images of Chemical Skeletons**

In order to train the Self-Supervised Machine learning model from an initially unstructured dataset for the generation of molecules, we must make the program automatically generate data labels which are to be further used in the training process as ground truth. To do this I have decided to use the Python RDKit Library for the generation of images of chemical skeletons from their SMILES representation.

Firstly, to generate the molecules we pass the SMILES representation into the RDKit program which will then generate unscaled and unnormalized images of the skeletons. This means that the images are of different sizes and have different scales for bonds/lengths. We have to then make the program automatically standardise the size of these bonds between all the molecules in the dataset and then ensure that a large percentage of the molecules still fit in the resulting images. The following image shows some of the unprocessed molecules generated by rdkit.

A screenshot of a computer screen

Description automatically generated

As seen above all the images have different scales and sizes. To standardize this, the program gets the range of sizes of the images generated by RDKit, it then finds a suitable scaling bound to fit the images. Standardizes the sizes of the bonds and centers the molecules. The program ensures that >98% of the molecules fit into the frame. Unfortunately, there is a small fraction of molecules in the dataset that are far too large to fit into the images without scaling the rest of the images substantially down and losing data to compression.

For the <2% of molecules that do not fit into the frames, I have a working heuristic that covers a large percentage of the remaining large molecules. For the images where the width or height of the molecule is larger than the rescaled frame, we rotate it through 45 Degrees. After this there is a tiny fraction (Approximately 0.2%\*) of the dataset that still do not fit, even after this process. In later stages of development I will work on a solution for these molecules that compresses their representation, however, for the purpose of this project, it is not necessary to include these molecules in the dataset.

Here is are some examples of an image that does not fit into the frames. Note – if I increase the scaling bounds too substantially, smaller molecules lose data and images would have to be larger, therefore increase the computation required.

A black and red lines on a white background

Description automatically generatedA diagram of a molecule

Description automatically generated

In later experimentation I will adjust the bounds for the preprocessing of these images when I have access to the actual dataset of molecules and smiles.

Additionally, there are some issues with specific molecules that RDKit Struggles to generate because of their form in the dataset.

A black and white drawing of a crystal

Description automatically generatedA drawing of a circle

Description automatically generated

At a later stage in development I will develop an automatic detection system for finding these images and removing them, however, as they make up such a small portion of the current dataset, I can just manually remove them from the initial dataset.

**A screenshot of a computer screen

Description automatically generated**

Displayed above is the results of the preprocessing for the molecules.

All the images consist of 400x400 pixel images with standardized and proportional bond lengths to limit the model having to learn its own scaling approximation which would not provide us with any additional value, it would only serve to increases the processing time of the model.

Note : This program is designed to be entirely automated and only requires the user to provide a dataset of SMILES

A close-up of a sign

Description automatically generated

The above image displays the simplified process of this component of the project.

Summary

* Generates vector representations of SMILES for input to training Model
* Generates processed ground truth labels/images for model