AI BIO INNOVATE CHALLENGE

Team Name: Vageesh.l

Team Members:
Vageesh Jangra
Sweety Khut
Shreyansh Lodha

Table of contents

01

Problem Statement

PROBLEM STATEMENT

- Our project focuses on improving the field of organic chemistry by addressing a crucial aspect often overlooked—predicting the conditions necessary for chemical reactions during retrosynthetic analysis.
- Traditional methods rely on manual selection by experts, but we aim to automate this process using machine learning.
- Our approach involves building a seq2seq-based model using multitask learning that learns the relationship between reaction components and conditions.
- This innovation aims to accelerate synthesis planning, offering chemists a predictive tool to enhance efficiency and speed up chemical discovery.

02

Data Processing

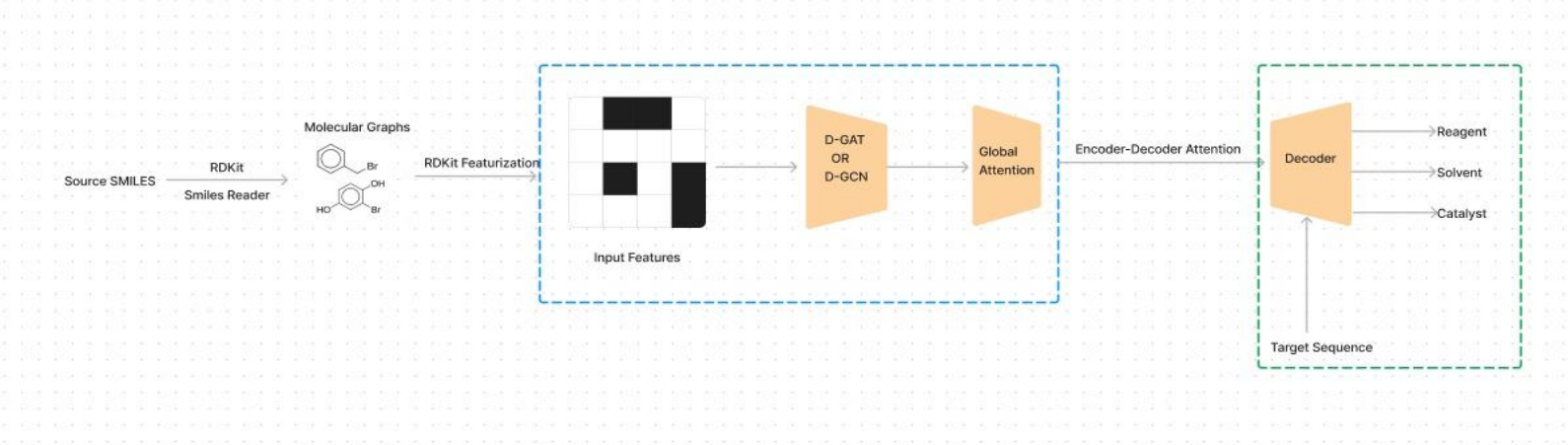
DATA PROCESSING

- Converting **SMILES** into molecular graphs for Graph Neural Networks involves using RDKit for featurization, converting SMILES to molecular objects, adding hydrogen atoms, and creating a graph with atom features and connectivity information.
- The conversion of **SMILES** to graph extends this by embedding the structure in 3D space, extracting features, and constructing an adjacency matrix.
- A mapping is generated to efficiently represent categorical data, assigning unique tokens random integers on first encounter and converted them into PyTorch tensor for compatibility with models.
- The PyTorch "Dataloader" class manages chemical reaction data using a DataFrame and a set sequence length.

03

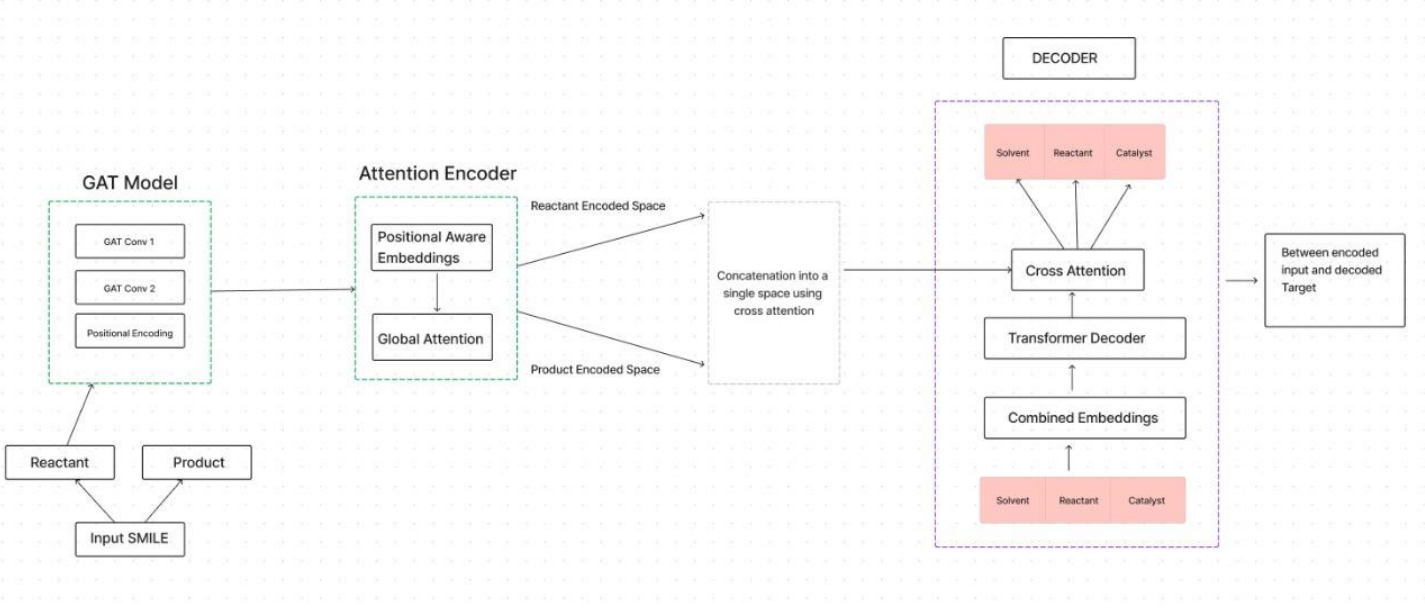
Methodology

Flow Of The Variables



- The source smiles are made into molecular graphs using RDKit after that we convert them to graph features which are parsed into our Graph Network followed by the Attention Encoder.
- The Decoder takes input of target sequence and concatenated with the encoder output by cross attention gives the probability output for each task.

Model Architecture



- The input SMILES are split into reactant and product, each fed separately into a GAT model with two layers. The model output is processed through an attention encoder, generating position-aware embeddings and applying global attention.
- In the decoder, the tasks are processed through a transformer decoder, cross-concatenated with the input sequence, and outputs are generated for three tasks using separate heads.

Inference

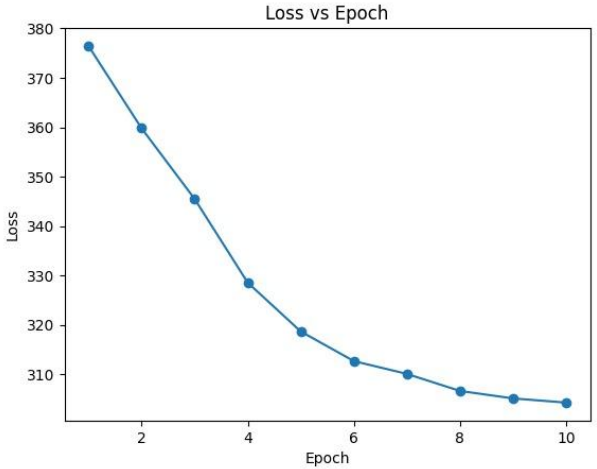
```
[32] o1,o2,o3=inference(model,'CC(C)(C)OC(=O)N1CC2CC(CN(Cc3ccccc3)C2)C1>>CC(C)(C)OC(=O)N1CC2CNCC(C2)C1')

     This is the original Reaction CC(C)(C)OC(=O)N1CC2CC(CN(Cc3ccccc3)C2)C1>>CC(C)(C)OC(=O)N1CC2CNCC(C2)C1 The reactant is  CC(C)(C)OC(=O)N1CC2CC(CN(C
     The Reagant of the original Reaction is phosphorus iodide ; 0 methanol|water water|platinum palladium iv water|platinum ammonium ethanol|nickel p
     The Solvent of the original Reaction is azide|cerium iodide|diisopropylamine
     The Catalyst of the original Reaction is iodide|diisopropylamine sodium peroxide|trifluoroacetic
```

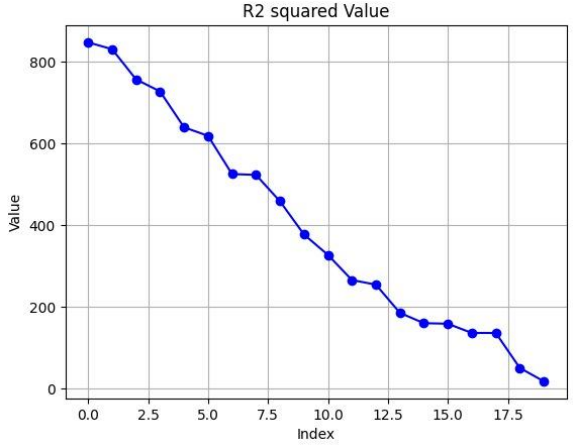
- Here is the inference result, the input is a reaction in form of SMILES and below are the predicted reagent, solvent and catalyst. The output from the model is decoded to string using our mapping dictionary that we created

04

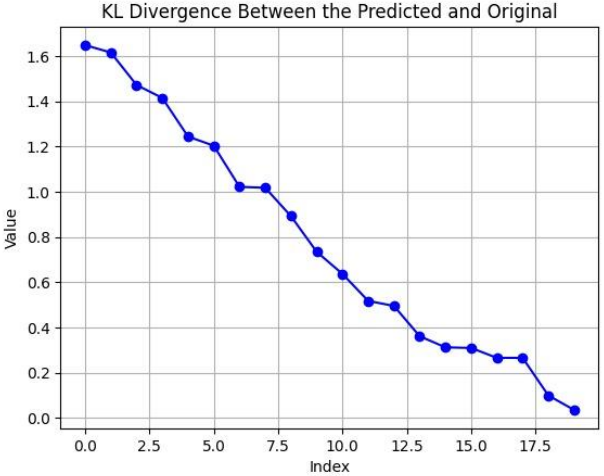
Evaluation Metrics



MSE LOSS



R2 Square



KL Divergence

05

Future Scope

Future Scope

- A User Friendly Interface
- More Advanced mechanism for attentions
- Making a more Deeper Network
- Improving the Model Performance

Thank You