Assignment 1 - Project Proposal

Enhancing the Prediction of Chemical Reaction Activation Energies with 3D Information

Tobias Jechtl - 01535670

October 23, 2024

1 Introduction

Predicting the speed of a chemical reaction is a crucial challenge in fields like drug design, materials science, and environmental chemistry. One of the most important factors that determines reaction speed is the activation energy—the minimum energy required to initiate a reaction. Machine learning models, especially those based on neural networks, have become powerful tools for predicting such chemical properties from data.

Traditionally, machine learning models have used 2D representations of molecules, like SMILES (Simplified Molecular Input Line Entry System) strings [1], which encode the structure of molecules as text. However, molecules also have 3D structures, and these geometric details can have a significant impact on how they behave in reactions. This project proposes an improvement to existing methods by incorporating 3D molecular information into machine learning models to enhance their ability to predict activation energies.

2 Objectives

The goal of this project is to develop a machine learning model that can better predict the activation energies of chemical reactions. Activation energy is crucial because it tells how much energy is needed to make a reaction happen. The project builds on existing methods that use 2D molecular representations, but with a key improvement: integrating information about the 3D structure of molecules to make more accurate predictions. Therefore, this project resembles the "***Bring your own Method***" project type.

3 Background

3.1 Molecules in 2D and 3D

A molecule can be described in many ways. One of the simplest ways is through a SMILES string, which is a 2D representation that encodes the atoms and bonds in a molecule, see Fig. 2. However, a molecule also has a 3D structure in space, which affects how it reacts with other molecules. For example, two molecules might look similar in 2D but behave very differently in 3D because of their shape and orientation.

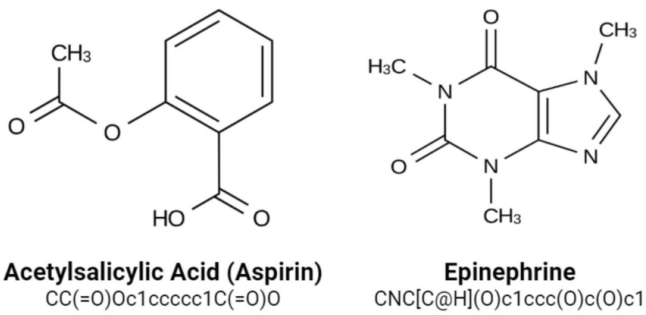

Figure 1: Example of two SMILES notation for Aspirin and Epinephrine, taken from [2].

Cartesian XYZ-coordinates are a common way to represent the 3D structure of a molecule. These coordinates specify the exact position of each atom in a molecule in 3D space.

3.2 Current Methods

Machine learning models called message passing neural networks (MPNN) [3] have been widely used to predict the properties of molecules. These models learn from the 2D structure of molecules and have been successful in many applications.

One approach, the Condensed Graph of Reaction (CGR) model [4], uses MPNNs to predict activation energies by analyzing the changes in the 2D structure of molecules during a reaction. However, this model does not account for the 3D structure of molecules, which can limit its accuracy.

In order to improve the prediction of activation energies, this project introduces 3D molecular information into the MPNN model. Specifically, the project will use a method called MACE [5], which computes 3D descriptors from XYZ coordinates. These 3D descriptors capture how molecules interact in 3D space, providing a more detailed picture of the molecules involved in a reaction.

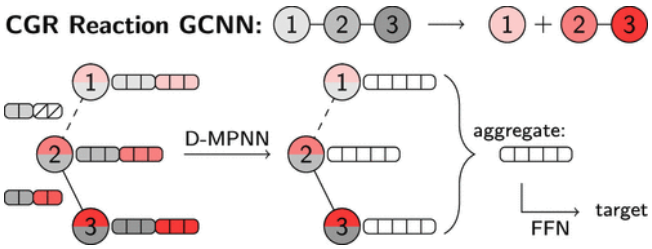

Figure 2: Schematic of the CGR Reaction GCNN (Graph Convolutional Neural Network) architecture. The D-MPNN processes the condensed graph of a reaction by passing messages between atoms, followed by aggregation and a feedforward neural network to predict target properties, taken from [4].

4 Dataset

The dataset for this project is called Transition1x (T1x) [6], which contains data from over 10,000 chemical reactions. For each reaction, the dataset provides:

- Reactants: The molecules before the reaction.
- Products: The molecules after the reaction.
- Transition states: Intermediate structures that form during the reaction.

The dataset includes both 2D SMILES strings and 3D XYZ coordinates for the reactants, products, and transition states. It also includes the activation energy for each reaction, which serves as the target value to be predicted by the model.

5 Proposed Approach

The project will proceed in two stages:

5.1 Baseline Model

The initial model will use the CGR-MPNN approach, which takes the 2D SMILES strings as input and predicts activation energies. This represents the current state-of-the-art in activation energy prediction.

5.2 Improved Model with 3D Information

In the second stage, the model will be improved by adding the 3D molecular descriptors calculated by MACE. These descriptors will be combined with the 2D information from the SMILES strings to give the model a more complete understanding of each reaction.

5.3 Expected Outcome

By comparing the performance of the baseline model with the improved model, it is aimed to demonstrate that including 3D molecular information leads to better predictions.

6 Methodology

The key steps in the project are:

Task 1 Preparation of the T1x Dataset: Extract the reactants, products, and transition states for each reaction, including both the SMILES strings and XYZ coordinates. (1 day)

Task 2 Train the Baseline Model: Use the CGR-MPNN model to predict activation energies from the 2D SMILES strings. (2 days)

Task 3 Compute MACE Descriptors: Install the MACE software and compute the 3D molecular descriptors from the XYZ coordinates. (3 days)

Task 4 Fine-Tune MACE: Adjust the MACE model using the training data to improve its accuracy in predicting energies and forces. (4 days)

Task 5 Train the Improved Model: Use the enhanced CGR-MPNN model, which incorporates both 2D SMILES strings and 3D MACE descriptors, to predict activation energies. (4 days)

Task 6 Evaluate Performance: Compare the accuracy of the baseline model and the improved model to assess the impact of incorporating 3D information. (2 days)

Task 7 Summarizing Results: Building a small application and shipping the project. Preparation of the final presentation. (2 days)

Fig. 3 illustrates the time schedule of the project by a Gantt chart.

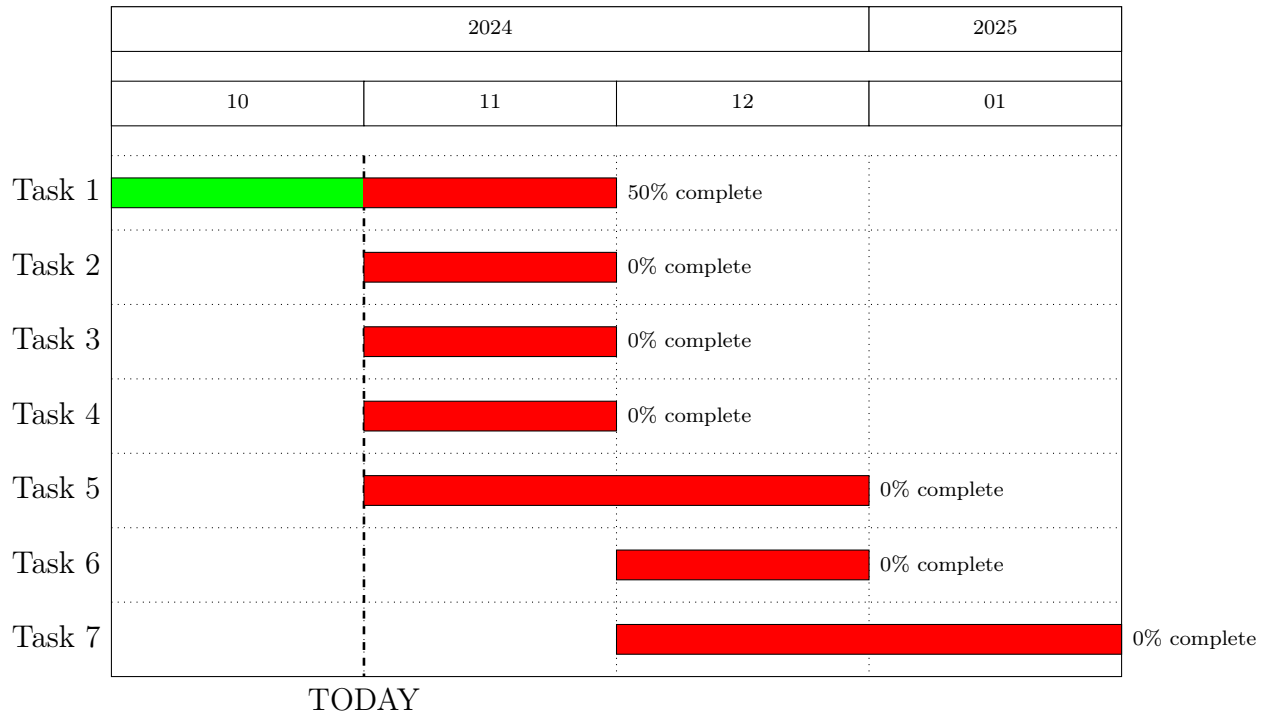

Figure 3: Time schedule of the planned tasks.

Bibliography

[1] A. P. Toropova, A. A. Toropov, E. Benfenati, G. Gini, *Chemical Biology & Drug Design* **2011**, *77*, 343–360.

[2] J. Yasonik, *Journal of Cheminformatics* **2020**, *12*, 14.

[3] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl in International conference on machine learning, PMLR, **2017**, pp. 1263–1272.

[4] E. Heid, W. H. Green, *Journal of Chemical Information and Modeling* **2021**, *62*, 2101–2110.

[5] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, G. Csányi, *Advances in Neural Information Processing Systems* **2022**, *35*, 11423–11436.

[6] M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, O. Winther, *Scientific Data* **2022**, *9*, 779.