CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Evaluating and optimizing Transformer models for predicting chemical reactions

An analysis for assessing and improving the effectiveness of the Chemformer model in retrosynthesis

Master's thesis in Applied Data Science

Siva Manohar Koki
Supriya Kancharla

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023

Master's thesis 2023

Evaluating and optimizing Transformer models for predicting chemical reactions

An analysis for assessing and improving the effectiveness of the Chemformer model in retrosynthesis

Siva Manohar Koki
Supriya Kancharla

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
University of Gothenburg
Chalmers University of Technology
Gothenburg, Sweden 2023

Evaluating and optimizing Transformer models for predicting chemical reactions
An analysis for assessing and improving the effectiveness of the Chemformer model in retrosynthesis
SIVA MANOHAR KOKI
SUPRIYA KANCHARLA

Supervisor from division of Data Science and AI: Rocío Mercado
Examiner from division of Data Science and AI: Ola Engkvist
Supervisor at AstraZeneca: Samuel Genheden
Co-Supervisor at AstraZeneca: Annie Westerlund

Master's Thesis 2023
Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2023

Evaluating and optimizing Transformer models for predicting chemical reactions
An analysis for assessing and improving the effectiveness of the Chemformer model in retrosynthesis

Siva Manohar Koki & Supriya Kancharla
Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology

Abstract

In this thesis, we assess the effectiveness of a transformer model specifically trained to predict chemical reactions. The model, named Chemformer, is a sequence-to-sequence model that uses the transformer's encoder and decoder stacks. Here, we employ a pre-trained Chemformer model to predict single-step retrosynthesis and evaluate its performance for diverse chemical reaction categories using various metrics such as Top-k accuracies, and Tanimoto similarity. We compare and analyse the results of the evaluations to those of the present template-based model. Based on the findings of the analysis, we fine-tuned the Chemformer model for specific chemical reactions, such as Ugi, Suzuki-Coupling, Rearrangement, Diels-Alder and Ring-Forming. In this project, we address five research questions, including whether the Chemformer model has higher accuracy than template-based model, which reactions it performs better and worse on top-k accuracies, the level of diversity in the results, and what fine-tuning strategies should be employed to enhance its performance. Using attention-based explainable AI, we scrutinize the input features that impact the transformation in the produced molecule. The results presented here may be used in the future to design fine-tuning strategies.

The evaluation results of pre-trained Chemformer model yields average Top-k accuracies across most of the reaction classes suggesting that the model struggles to accurately predict the reactions on in-house test data. When evaluating the model's performance on USPTO data, we found similar results. While the results demonstrate that the pre-trained model outperforms the template-based model, there is still potential for enhancing its performance. This potential for further improvement paves the way for the fine-tuning process. By applying fine-tuning to specific sub-tasks such as Ugi, Suzuki-Coupling etc., we managed to significantly enhance the model's performance. The fine-tuned model consistently outperforms the both pre-trained and template-based models, exhibiting a notable 50% improvement in accuracy over the pre-trained model. This substantial progression reinforces the effectiveness of transfer learning as a powerful approach for enhancing Chemformer models.

Keywords: Chemformer, transformer, evaluation, explainable AI, fine-tuning, machine learning.

Acknowledgements

We would like to express our gratitude to our supervisors at AstraZeneca, Samuel Genheden, and Annie Westerlund, for creating the groundwork for our thesis and providing us with exceptional support throughout the project. Their guidance and feedback have been instrumental in the project's success so far. Also, we extend our appreciation to our supervisor, Rocio Mercado, and examiner, Ola Engkvist, from the Department of Computer Science and Engineering at Chalmers and GU, for their valuable guidance, advice, and feedback throughout the process. Thank you all for your assistance in helping us achieve our goals.

Siva Manohar Koki and Supriya Kancharla, Gothenburg, April 2023

Contents

List of Figures

List of Tables

List of Tables

1

Introduction

In this chapter, we begin by providing background information on retrosynthesis, computer-aided retrosynthesis, SMILES, Chemformer, and both template-based and template-free approaches, which are all essential topics for our thesis. Following that, we will outline the thesis objectives and present our research questions. Lastly, we will address the limitations of the thesis and provide an overview of the remaining sections of the report.

1.1 Retrosynthesis

Synthesis planning is a process to determine how to synthesize a chemical compound from available starting materials [1]. A reverse method called retrosynthetic analysis starts with the desired product and works backward to identify the sequence of simpler molecules, called reactants, that can be combined to make the product. The concept of retrosynthesis was pioneered by E.J. Corey, who received the Nobel Prize in Chemistry in 1990 for his contributions to the theory and methodology of organic synthesis [2]. This process is used to plan the synthesis of chemical compounds from available starting materials. Retrosynthesis is a critical aspect of organic chemistry, involving the strategic dissection of complex target molecules into simpler precursors (simpler building blocks generated by breaking complex molecules) to facilitate their synthesis [3] through a series of reverse synthetic steps as shown in Figure 1.1, guided by established disconnection rules and strategic bond disconnections [4], [5]. As a fundamental tool in the design and planning of synthetic routes, retrosynthetic analysis has significantly impacted various fields, including drug discovery, materials science, and natural product synthesis [4], [6]. Ultimately, the goal is to identify a practical and efficient synthetic pathway that yields the target molecule from readily available starting materials.

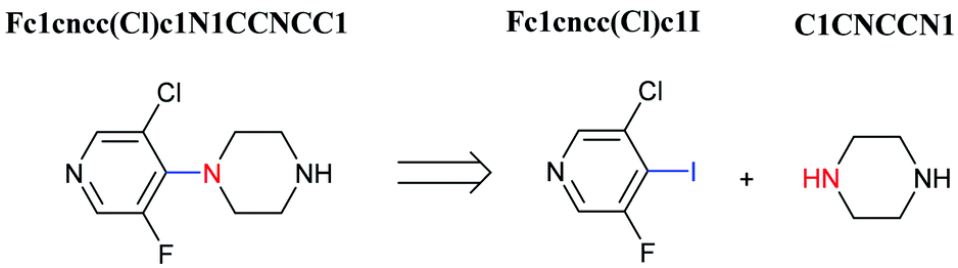

Figure 1.1: *An example of a retrosynthesis reaction. The target compound is shown on the left side of the arrow, while possible reactants can be found on the right. The corresponding SMILES notations for each of these compounds are also displayed. Figure taken from [7].*

1.2 Computer-aided Restrosynthesis prediction

A Computer-aided retrosynthesis prediction is a computational approach for designing efficient synthetic routes for complex organic molecules [8]. The introduction of machine learning techniques has significantly improved the capabilities, scalability, and generalizability of computer-aided retrosynthesis systems, leading to better predictions, more efficient route planning, and accelerated discovery in chemistry. It involves using machine learning algorithms to learn from a large database of known chemical reactions and then using this knowledge to propose retro-synthetic steps for a given target molecule.

Computer-aided retrosynthesis prediction has the advantage of significantly speeding up the drug discovery process by quickly identifying viable synthetic routes for a target molecule. This is especially important because wet-lab experiments to identify viable synthetic routes can be time-consuming and expensive [9], and using computational tools can help to prioritize which routes are worth pursuing in the lab. It also has the potential to reduce the cost of drug development by minimizing the number of synthetic steps required to produce the desired compound. Several companies and research groups are currently developing and refining computer-aided retrosynthesis prediction tools [10]. These tools are expected to become increasingly important in the drug discovery process as computational methods continue to play an ever larger role in the field of chemistry.

Computer-aided retrosynthesis prediction involves both template-based and template-free approaches [11]. These two approaches will be detailed further below.

1.2.1 Template based approach

The template-based approach [12] is a widely used method in several gold standard models [13] for retrosynthesis prediction. This approach was one of the initial approaches used in retrosynthesis prediction and it involves using a library of known reactions and retrosynthetic transformations [14].

While the template-based approach is a relatively simple and straightforward method, it has limitations, such as the availability and accuracy of the reaction database used to generate the templates. However, it still remains a useful and effective approach for generating potential synthetic routes. Several gold standard models for retrosynthesis prediction, such as Chemetica [15], have incorporated the template-based approach into their algorithms. These models have been trained on large datasets of known reactions and have been shown to achieve high accuracy in predicting synthetic routes for a given target molecule.

1.2.2 Template free approach

The template-free approach [11] does not rely on pre-defined templates. Instead, these methods utilize machine learning algorithms and other computational techniques to learn the underlying chemical patterns and transformations directly from the data. One application of template-free approaches is the generation of new reactant SMILES (Simplified Molecular Input Line Entry System) for a given target molecule. SMILES Figure 1.2 are compact and standardized textual representations of chemical structures and are widely used in chemical informatics and drug discovery. These models can automatically learn the transformations without relying on predefined templates.

There are several types of template-free approaches used in retrosynthesis prediction, including, graph-based models [16] and machine learning-based models [12]. For e.g., graph-based models use graph theory to represent chemical structures and reactions as graphs and these models suggest reactants typically using graph-convolutional neural networks which performs convolution-like operation on the graph (e.g. LocalRetro [17], Variational Graph Neural Networks [18]) Machine learning-based models such as Molecular Transformer [19], DeepSMILES [20], and Chemformer [21] use supervised or unsupervised algorithms to learn patterns.

Template-free approaches have several advantages over template-based approaches, including their ability to predict complementary and more complex reactions suggested by template-based approaches. However, they also have limitations, such as the quality and quantity of the training data is much more limited. Despite these limitations, template-free approaches are a promising area of research in retrosynthesis prediction and are being actively developed and improved [21].

1.3 SMILES

The chemical notation system SMILES (Simplified Molecular Input Line Entry System) was created for contemporary chemical information processing. SMILES, which uses molecular graph theory as its foundation, enables rigorous structural specification through the use of an extremely compact and intuitive grammar. It is a linear notation that uses symbols and characters to denote atoms, bonds, and other structural features. SMILES notation is widely used in chemical databases and computational chemistry applications because of its concise and unambiguous representation of molecular structures [22]. In SMILES notation, each atom is represented by its

atomic symbol, while bonds between atoms are denoted by various characters such as hyphens and equals signs. Other characters are used to describe ring structures, branching, and other molecular features. SMILES representation for molecules is as shown in Figure 1.2. High-speed machine processing works nicely with the SMILES notation scheme. Numerous highly effective chemical computer applications, such as the creation of a distinctive notation, constant-speed (zeroth order) database retrieval, flexible substructure searching, and property prediction models, can be designed which results to the ease of use by the chemist and machine compatibility.

The SMILES notation is easily accessible to chemists, yet flexible enough to allow interpretation and generation of chemical notation independent of the specific computer system in use [23]. Similar to conventional chemical notation, it improves on conventional software methods by greater speed and better use of computer capacity. Molecular structures are uniquely and accurately specified and can be used with chemical databases.

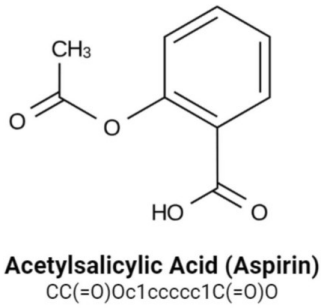

Figure 1.2: *SMILES notation for Aspirin. The Aspirin SMILES string represents a molecule known as phenylacetic acid. It consists of a phenyl ring (c1ccccc1) bonded to an acetic acid moiety (CC(=O)O). Figure taken from [24].*

1.4 Chemformer

Chemformer [21] is built on the BART (Bidirectional and Auto-Regressive Transformer [25]) language paradigm, which makes use of the Transformer's encoder and decoder stacks. Since the input sequences to the encoder and decoder are treated individually, which lowers the amount of computing required in comparison to a model that processes both sequences together. Therefore it is extremely ideal for seq2seq tasks like reaction prediction and molecular optimization. By utilizing merely the encoder stack, the BART paradigm can also be easily applied to exclusionary tasks. Examples of such tasks include classification [26] and similarity search [27].

1.5 Aim

The primary objective of this thesis is to assess and enhance the efficacy of the Chemformer model for predicting chemical reactions. To accomplish this, we employ the Chemformer model fine-tuned for the retrosynthesis task [19], and evaluate

its performance for diverse chemical reaction categories using various metrics. The results of the evaluations are then compared to those of the present template-based model. By employing attention-based explainable AI, we scrutinize the input features that impact the transformation in the produced molecule by producing heat maps. Based on the findings of the analysis, we fine-tune the Chemformer model for specific chemical reactions, such as Ugi, Suzuki-Coupling, Diels Alder, ring-forming and rearrangement.

1.5.1 Research questions

1. How does the performance of Chemformer compare to existing template-based models?
2. For which reactions does Chemformer perform better, and for which reactions does it perform poorly and why?
3. What is the level of diversity in the results generated by Chemformer?
4. Can Explainable AI help us understand and validate the model predictions?
5. Can we enhance the Chemformer's performance on specific reaction classes by employing fine-tuning?

1.6 Limitations

In the case of the Transformer model, transfer learning [26] is the most commonly used strategy, wherein the Chemformer model is pre-trained on a comprehensive and varied collection of chemical reactions and subsequently fine-tuned on a specific, smaller set of reactions. This approach can enhance the model's ability to generalize and improve its performance on particular reaction categories. However, other strategies, such as Multi-Task Fine-tuning [28], Data Augmentation [29], and Ensemble Fine-tuning [30], have also proven effective in boosting model performance. However, due to time constraints, we will only employ transfer learning to address our research queries in this thesis paper. For more investigation, these other approaches will be described under future work in the conclusion chapter.

1.7 Thesis Outline

The first chapter provides a concise summary of retrosynthesis, Chemformer, and the template-free approach. Our research scope and the constraints of the thesis have also been outlined.

- In the theory section, we will provide an overview of seq2seq models, AiZynthFinder, and SMILES. We will also provide in-depth background into the Transformer model, Retrosynthesis with transformers, various evaluation metrics, different chemical reactions, attention as an explainability method, and seq2seq fine-tuning in greater detail.

- In the methods section, we will outline our methodology for data preparation, discuss the specific chemical reactions that are the focus of this thesis, and detail the implementation of our model.
- In the results section, the findings of our study will be presented and their implications for the research questions will be discussed.
- In the conclusion section, the main conclusions and potential approaches for future work are proposed.

2

Theory

In this chapter, we have provided a background for all the topics this thesis will address. We begin with highlighting specific chemical reactions of interest followed by Sequence-to-sequence models. Then, we introduce the currently most common transformer architecture for sequence-to-sequence problems. We first describe the architecture of the Transformer model in detail and then explain how the Transformer model is used for retrosynthesis prediction and how the model was pre-trained. Subsequently, We will provide a concise overview of various evaluation metrics and in the final sections, we will discuss explainable AI for natural language processing and the fine-tuning of sequence-to-sequence models.

2.1 Chemical Reactions

Chemical reactions involve the transformation of chemical substances into new substances with different chemical and physical properties and can be classified into several categories based on the changes that occur in the reactants and products. Ring-Forming reactions cleave cyclic compounds to form open-chain or linear structures[31], while rearrangement reactions involve the rearrangement of atoms or functional groups to form a different isomer or structural variant [32]. The Suzuki coupling reaction is a versatile cross-coupling method for the formation of carbon-carbon bonds between aryl or vinyl halides and organoboron compounds. It has broad substrate scope, mild reaction conditions, and is widely used in organic synthesis. [33], while Diels-Alder reactions involve the addition of a conjugated diene to an alkene to form a cyclic compound with a six-membered ring [34]. Finally, Ugi reactions synthesize peptidomimetic compounds from an aldehyde, an amine, an isocyanide, and a carboxylic acid via a four-component condensation process, and are widely used in drug discovery and development due to their efficiency and versatility [35].

2.2 Sequence-to-sequence models

Sequence-to-sequence (seq2seq) models are a type of deep learning model that can learn to map an input sequence to an output sequence, typically of a different length or format. seq2seq models are widely used in natural language processing (NLP) tasks such as machine translation, text summarization, and chatbot generation.

The basic architecture of a seq2seq model consists of two recurrent neural networks (RNNs) - an encoder RNN and a decoder RNN. The encoder RNN reads the input sequence and produces a fixed-length representation of the input called the context vector, which is used by the decoder RNN to generate the output sequence [36].

At each time step t, the encoder RNN takes in the input sequence x_t and the previous hidden state h_{t-1} and produces the current hidden state h_t using the following equation:

$$h_t = f_{enc}(x_t, h_{t-1}) \text{ where } f_{enc} \text{ is the encoder RNN function.}$$

The final hidden state of the encoder RNN, h_T, is used as the context vector c to initialize the decoder RNN. At each time step t' in the output sequence, the decoder RNN takes in the previous output $y_{t'-1}$ and the previous hidden state $h_{t'-1}$ and produces the current output $y_{t'}$ and the current hidden state $h_{t'}$ using the following equations:

$$h_{t'} = f_{dec}(y_{t'-1}, h_{t'-1}, c)$$
$$y_{t'} = g(h_{t'})$$

where f_{dec} is the decoder RNN function,

g is a function that maps from $h_{t'}$ to $y_{t'}$,

and c is the context vector produced by the encoder RNN.

One of the key advantages of seq2seq models is that they can handle input and output sequences of variable lengths [37]. This makes them particularly useful for tasks such as machine translation, where the length and structure of the input and output sentences can vary widely. seq2seq models can also learn to capture the context and meaning of a sentence, which can be challenging for traditional rule-based or statistical NLP models.

Seq2seq models can be further enhanced with various techniques, such as attention mechanisms that allow the decoder to selectively focus on different parts of the input sequence when generating the output sequence. Other improvements include using bidirectional RNNs to capture both past and future context or using convolutional neural networks (CNNs) as an alternative to RNNs for sequence modeling [38].

2.3 Transformer

The Transformer employs the encoder-decoder architecture and leverages the attention mechanism [38], which gives more weight to certain components of the input. The attention mechanism operates in three ways: self-attention in the encoder, self-attention in the decoder, and cross-attention between the encoder and decoder[26].

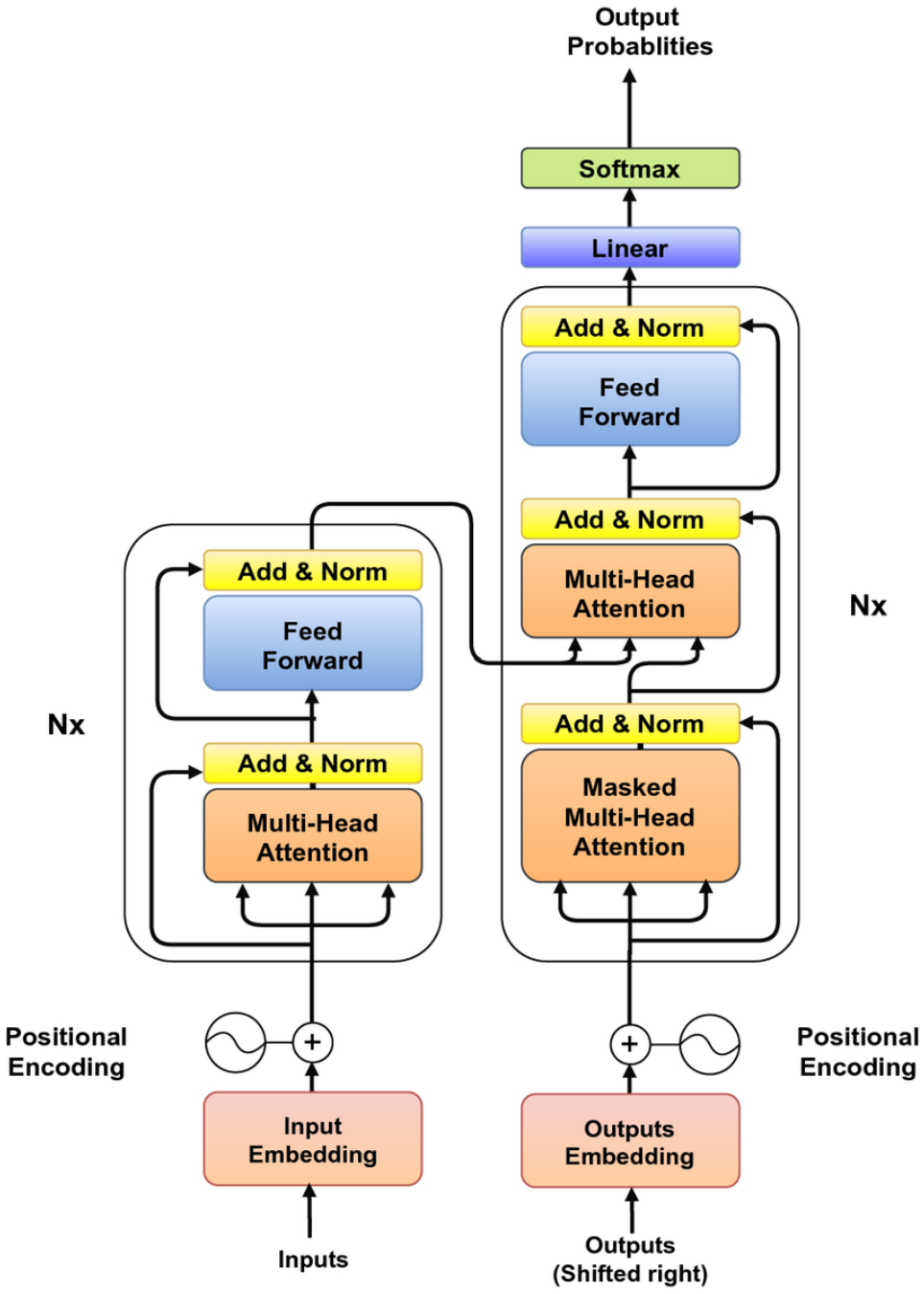

Figure 2.1: *The Transformer model architecture, where the left and right halves illustrate how the encoder and decoder of the Transformer, respectively, work using point-wise, fully-connected layers with stacked self-attention. Figure taken from [39].*

These attention mechanisms are represented by the dark yellow attention blocks in Figure 2.1.

Self-attention pertains to the connections between different positions within a single sequence to create a representation of that sequence. For example, consider the sentence "The animal did not eat the food because it was too full". It may not be

clear to an algorithm whether "it" refers to the food or the animal, but self-attention enables the model to associate "it" with "animal" by looking at other positions in the input sequence for context. On the other hand, cross-attention involves linking two different sequences and providing information from the input sequence to the decoding layers to predict the next token in the sequence, which is then added to the output sequence. This project focuses only on cross-attention since it has a more direct relationship with the output sequence.

2.3.1 Input embedding

The input embedding process in the Transformer model involves converting words to word embeddings using an embedding layer, followed by combining positional encoding information. Word embeddings are obtained by multiplying one-hot encoding vectors with a learned weight matrix. These real-valued vectors capture semantic meaning. Positional encodings are added to incorporate relative or absolute positions of tokens in the sequence. Various methods, such as sine and cosine functions, are used for encoding positional information.

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{2.1}$$

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{2.2}$$

The model learns to attend to the linear properties of these functions, where "pos" refers to the position and "i" refers to the dimension [26].

2.3.2 Scaled dot product attention

When processing a sequence of vectors $x = [x_1, \ldots, x_n]$ with an attention mechanism, each vector x_i undergoes separate linear transformations to generate query, key, and value vectors qi, ki, and vi. These vectors represent what information is being searched for, its relevance to the query, and the actual content of the input, respectively. The attention head then calculates attention weights α for all pairs of words using a dot product between the query and key vectors, which are then normalized using softmax to a value between 0 and 1. The dimension of keys, dk, is used in the normalization process. Finally, the attention head generates an output o as a weighted sum of the value vectors.

$$\alpha_{ij} = \frac{\exp(q_i^T k_j/\sqrt{d_k})}{\sum_{l=1}^{n} \exp(q_i^T k_l/\sqrt{d_k})} \tag{2.3}$$

$$o_i = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{2.4}$$

Here, α_{ij} is the attention weight between query vector q_i and key vector k_j, and o_i is the output vector for query q_i. d_k represents the dimension of the key vectors, v_j represents the value vector for the j-th input token, and n is the number of input tokens in the sequence.

In practice, the attention function is computed on a set of queries, keys, and values simultaneously, packed together into a matrix Q, K, and V

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The dot product of queries with all keys is normalized with d_k, as described in reference [26]. In self-attention, keys, values, and queries are generated from the same sequence, while in cross-attention, the queries come from a different sequence than the key-value pairs. Each attention function corresponds to computing one head, and to perform the attention function in parallel, multiple heads are used in the model[26].

2.3.3 Multi-head attention

The multi-head attention mechanism computes attention not just once, but multiple times to capture different aspects of the input. This approach enables the model to gather information from various sub-spaces, as illustrated in Figure 2.2. Each head is a distinct linear projection of the input representation, serving as the query, key, and value. During the learning process, the scaled dot product attention is computed h times in parallel, and the resulting outputs are concatenated. A single linear projection is applied using:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{2.5}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.6}$$

The projections are parameter matrices W_{Qi}, W_{Ki}, W_{Vi}, and W_{Oi} [26].

2.3.4 Overall model architecture

The Transformer architecture follows an encoder-decoder structure, similar to seq2seq models. Both the encoder and decoder are composed of N identical layers stacked on top of each other. Figure 2.1 illustrates the architecture of one encoder-decoder layer. The encoder consists of two sub-layers: the first is a multi-head self-attention mechanism, and the second is a position-wise fully connected feed-forward network. Each sub-layer is followed by normalization and a residual connection. The output of each sub-layer is given by *LayerNorm(x + Sub-layer(x))*, where *Sub-layer(x)* is the function implemented by the sub-layer itself. In the original Transformer model, the dimension of all sub-layers and embedding layers is set to d_{model}= 512.

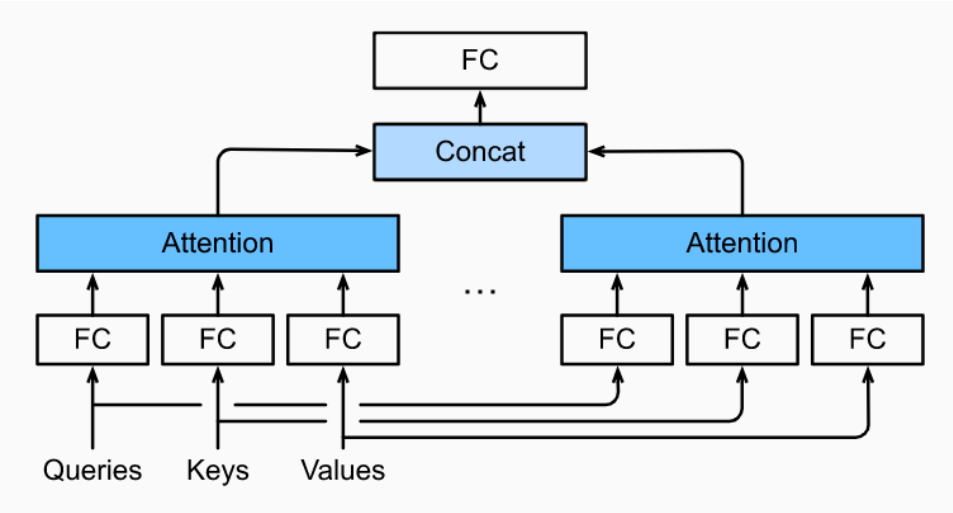

Figure 2.2: *The multi-head attention scheme, where multiple heads are concatenated and then linearly transformed.*

In addition to the two sub-layers in each encoder layer, the decoder has a third sub-layer that performs multi-head attention over the output of the encoder stack. To avoid the decoder positions attending to future positions, the self-attention layer masks the future positions, ensuring that the prediction for position i only depends on the known outputs at positions less than i [26].

2.4 Retrosynthesis with Transformers

As mentioned earlier, retrosynthesis is the process of planning the synthesis of a target molecule by breaking it down into simpler starting materials as shown in Figure 1.1. Transformer models have recently shown promise for retrosynthesis planning, it is a type of neural network architecture that has shown impressive performance in natural language processing tasks.

Transformer models are seq2seq models. As such, a transformer model for retrosynthesis [40] takes the product SMILES string as input. The target molecule uses SMILES representation, where nodes represent atoms and edges represent chemical bonds. The SMILES are then fed into a Transformer model, which directly predicts the reactant SMILES. The model takes into account both the target molecule and the available starting materials, as well as the reactions that are likely to occur.

One advantage of using Transformer models for retrosynthesis [40] planning is that they can capture long-range dependencies between atoms and chemical bonds, which can be difficult to model using traditional rule-based approaches. Additionally, the models can learn from large amounts of data, allowing them to generalize to new molecules and reaction types. However, there are also some challenges to using Transformer models for retrosynthesis planning. One issue is that the models require large amounts of training data, which can be difficult to obtain for rare or novel reactions. Additionally, the models may generate sub-optimal reaction pathways or

overlook important reaction conditions.

2.5 Chemformer-based Pre-trained Model

In this thesis, we use the Chemformer model, which is a pre-normalized transformer model trained to carry out computer-aided synthesis planning. In a pre-normalized transformer model, normalization is applied before each sub-layer, including the self-attention and feedforward layers. This pre-normalization technique is intended to address the problem of vanishing gradients that can occur in deep neural networks with many layers.

The Chemformer model was first pre-trained in a self-supervised manner on AstraZenecas in-house dataset comprising 18 million records.We evaluate this base model on both the in-house test set and the USPTO dataset. To further improve the models performance, we utilize attention-based explainable AI to debug the evaluation results and devise a strategy for fine-tuning the model.

We conduct in-depth analysis on specific hard reaction classes of interest, such as Ring-forming reactions, Rearrangement reactions, Diels-Alder, Ugi and Suzuki-Coupling, and evaluate the models performance on these classes. To improve the model's performance on specific reaction classes, we utilize a seq2seq fine-tuning method. This involves loading pre-trained weights for each reaction class and then applying a task-specific fine-tuning process. Figure 2.3 illustrates how the pre-training and downstream fine-tuning tasks are applied to the Chemformer model.

Our study aims to not only improve the Chemformer model's performance but also provide insights into the feasibility of using attention-based explainable AI for debugging model evaluation results.

2.6 Sequence-to-sequence fine-tuning

Sequence-to-sequence (seq2seq) models [41] are widely used in Natural Language Processing (NLP) tasks, such as machine translation and text summarization. One of the challenges with seq2seq models is that they need to be trained on large amounts of data to generate accurate predictions. Training seq2seq models from scratch can be computationally expensive and time-consuming.

Fine-tuning pre-trained seq2seq models [42] can be a useful strategy to overcome this challenge. Fine-tuning is the process of taking a pre-trained model and adapting it to a new task or dataset by further training it on that specific task or dataset. In the context of seq2seq models, fine-tuning involves taking a pre-trained encoder-decoder model and adapting it to a new translation task [41], for example, by training it on a new parallel corpus. Fine-tuning a seq2seq model involves updating the model's weights using a smaller dataset specific to the task of interest. The pre-trained weights are used to initialize the model, which is then fine-tuned using the new dataset. The fine-tuning process typically involves training the model for

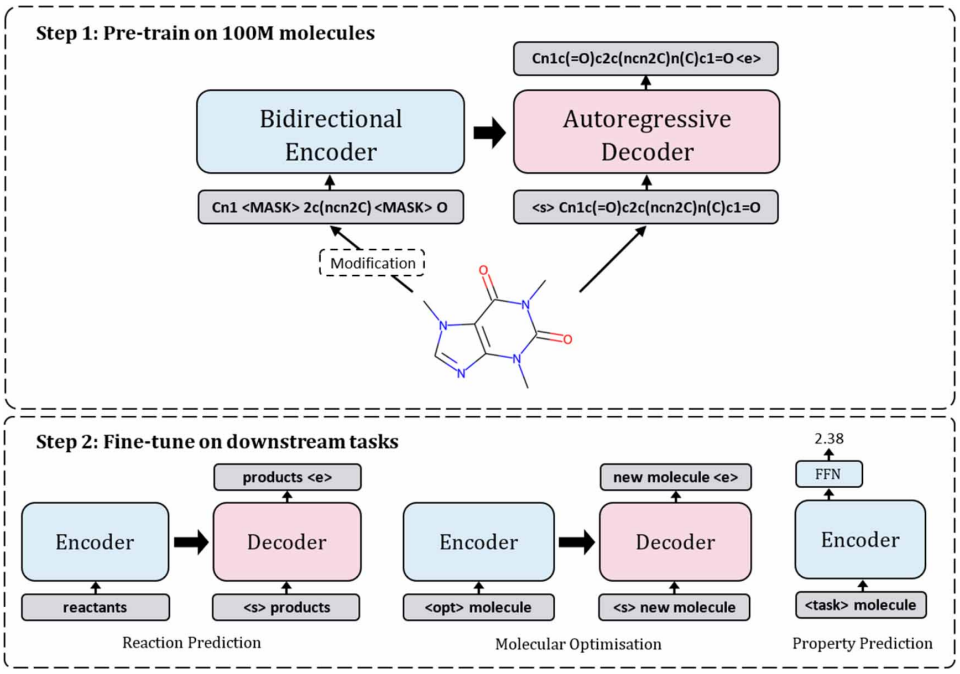

Figure 2.3: *Illustration of the pre-training and fine-tuning processes for downstream tasks, where pre-training refers to training the model on AstraZeneca in-house data prior to fine-tuning its weights for specific chemical reactions. Figure taken from [21].*

a few epochs, which allows the model to adapt to the new data and improve its performance on the target task.

The general steps involved in sequence-to-sequence fine-tuning are:

1. Load a pre-trained seq2seq model
2. Prepare a new dataset specific to the task of interest
3. Fine-tune the model on the new dataset
4. Evaluate the performance of the fine-tuned model on a validation set and make any necessary adjustments to the training process.

Fine-tuning can be an effective way to leverage pre-trained models and reduce the amount of data required for training. It also enables the transfer of knowledge from pre-trained models to new tasks, which can be especially useful in scenarios where there is limited data available for a specific task.

2.7 Evaluation Metrics

The effectiveness of the statistical or machine learning model is assessed using evaluation metrics. For each project, evaluating machine learning models or algorithms is crucial. The Chemformer model is tested using a wide variety of evaluation metrics. These evaluation metrics are first applied to the pre-trained Chemformer model us-

ing the in-house test data. After fine-tuning, the model is finally evaluated on the USPTO dataset. We also assess how well the model generalizes to new reaction classes.

2.7.1 Top-k accuracy

Top-k accuracy is a widely used evaluation metric in machine learning and deep learning [43]. It is a measure of how well a model can correctly predict the class or label of a given input.

In classification tasks, a model takes an input and outputs a probability distribution over a set of possible classes. The top-1 accuracy measures the proportion of examples for which the predicted label matches the true or target label exactly. For example, a model may assign a probability to each possible label or class, and the predicted label is the one with the highest probability [44]. In the case of chemformer, an assignment is performed for each token present in the SMILES notation. The computation of top-k accuracy hinges upon this token-wise assignment process. Notably, the accuracy measure considers not just a generic match but insists on exact correspondence with the SMILES notations. Thus, the precise matching of these SMILES tokens is fundamental to our evaluation process and plays a pivotal role in the computation of top-k accuracy.

Top-k accuracy is a simple and intuitive evaluation metric that measures how often a model makes the correct prediction among all possible classes. However, it can be limited in case of retrosynthesis where there are multiple correct answers or when the predicted class is close to the true class but not exactly the same. In such cases, other evaluation metrics such as top-k accuracy or mean average precision may be more appropriate.

2.7.2 Tanimoto Similarity

The Tanimoto similarity is a measure of similarity between two sets. It is commonly used in cheminformatics [45] and information retrieval to compare the similarity of molecules. The Tanimoto similarity coefficient is a number between 0 and 1, with 1 indicating perfect similarity and 0 indicating no similarity.

Tanimoto similarity coefficient is defined as the ratio of the intersection of two sets to the union of the sets. In this context, sets A and B represent the sets of substructures (such as atoms or bonds) present in two molecules. Mathematically, the Tanimoto similarity coefficient between sets A and B is defined as:

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.7}$$

where |A| and |B| denote the cardinality (number of elements) of sets A and B, respectively, and |A B| denotes the number of elements that are common to both sets.

In the context of our study, the predicted molecules are subjected to the comparison against the target molecules using the fingerprints [46]. This comparison is carried out by passing the fingerprints of the predicted and target molecules to the 'rdkit.DataStructs.TanimotoSimilarity' module. The Tanimoto similarity calculation performed by this module serves as a fundamental metric that quantifies the extent of overlap between the structural features encapsulated within the fingerprints. By producing a numerical representation of their similarity, this calculation offers valuable insights into the degree of resemblance between the predicted and target molecules. Such a quantitative assessment of molecular similarity is crucial in evaluating the effectiveness and accuracy of our prediction model.

2.7.3 Fraction Invalid

Fraction Invalid is a metric used to evaluate the chemical validity of the SMILES generated by the Chemformer model. It measures how many invalid SMILES (i.e. that do not correspond to the physical molecule) were proposed by the Chemformer model [47]. SMILES may be considered invalid for a variety of reasons, such as invalid syntax, infeasible bonding, unstable structures, incorrect valence or incomplete or mismatched ring closures. A lower fraction invalid suggests that the model has successfully produced a substantial number of valid SMILES. Consequently, this increases the likelihood of achieving higher accuracy.

2.7.4 Diversity score

Diversity plays a crucial role in assessing reaction class diversity among the top-k predicted SMILES. Different approaches exist for defining diversity, one of which is entropy. The diversity score with entropy is a quantitative measure used to assess the diversity or variety within a dataset [48]. It utilizes the concept of entropy from information theory to quantify the uncertainty or randomness associated with the distribution over reaction classes. To calculate the diversity score using entropy, we consider the probabilities of each reaction class within the dataset. A flat distribution, where probabilities are evenly spread across different reaction classes, results in high uncertainty within the dataset. This high uncertainty leads to an increase in entropy. Consequently, a high entropy value indicates a higher diversity among the predicted reaction classes. On the other hand, if one reaction class is dominant among the predictions, the entropy and the diversity will be low. In such cases, the dominance of a single reaction class reduces the overall uncertainty and limits the diversity of the predicted reactions.

Given a set of reaction classes denoted by X, where each reaction class x is associated with a probability or frequency $P(x)$, entropy can be defined as follows:

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

The resulting entropy value is normalized to the interval [0, 1]. The normalization ensures that the entropy measure falls within a standardized range, allowing for

meaningful comparisons and interpretations of diversity among reaction classes.

where Σ denotes the sum over the variable's possible values. The choice of base for log, the logarithm, varies for different applications, base value 2 is used to calculate the entropy.

2.8 Explainable AI for NLP

Complex domains such as natural language processing have seen high accuracy achieved by deep learning algorithms. However, models can be compared to black boxes where the input and output can be observed but the reasons behind the model's predictions remain unknown. To address this issue [49], researchers have developed various techniques to make NLP models more interpretable. One popular approach is to use attention mechanisms, which allow the model to focus on certain parts of the input text that are most relevant to the prediction. Explainable AI comprises a set of tools and frameworks to aid in the comprehension and interpretation of deep learning models. Heat maps are created in relation to attention scores, providing an interpretable visual representation of how the model's focus on input tokens influences the resulting output tokens. When a model is considered explainable, it denotes the ability of humans to comprehend the outcomes of an AI-generated solution [50]. This understanding can encompass the role played by various parts of the model in learning specific tasks and the significance of different input features in making predictions.

3

Methods

In this section, we will discuss the essential steps for evaluating data. Initially, we will outline the process of selecting data for the dataset. Later on, we will elucidate how the dataset was partitioned and categorized according to the reaction types. Lastly, we will provide a comprehensive explanation of how the evaluation is carried out.

3.1 Data

The project utilizes chemical reaction data comprising various chemical classes (approximately 70), such as Ugi reactions, Diels-Alder reactions, Suzuki-Coupling rearrangement, Ring-Forming, and Rearrangement reactions. These reactions can be unimolecular, bimolecular, or trimolecular. Each record includes the reactants and product of the chemical reaction, encoded as a reaction SMILES.

3.1.1 Datasets Used

There are two datasets to work with:

- **In-house dataset**: A CSV file comprises 18 million reactions sourced from AstraZeneca's proprietary databases such as Reaxys [51], Pistachio [52] and in-house electronic lab notebooks (ELNs).
- **USPTO dataset**: A CSV file comprises 1 million reactions obtained from publicly accessible data from the United States Patent and Trademark Office (USPTO [53], [54]) is provided.

The former is used to train and evaluate the models, and the latter only evaluates the models. The data is ready for modeling after being pre-processed [55].

3.1.2 Pre-processing Datasets

Pre-processing steps performed by AstraZeneca researchers are:

- removing all SMILES that cannot be sanitized with RDKit [56],
- removing any product that also is identical to a reactant,

- performing role assignment based on the atom-mapping (all reactants should share at least one atom-mapping number with a product),
- removing duplicate reactions based on reactants and products.

3.1.3 Datasets Preparation

For evaluation purposes, data splitting and categorization were performed. The in-house data was randomly divided into training, validation, and test sets at a 98:1:1 ratio, resulting in 18 million records for training and 186k records each for validation and testing.

We applied the below rules to categorize specific reaction classes.

- Ring-Forming reactions: Is indicated with the Ring-Forming column in the dataset. The process for identifying reactions suitable for the RingBreaker class is based on atom-mapping [55].
- Rearrangement reactions: These have a reaction class name that ends with "rearrangement"
- Diels-Alder cycloaddition: These have a reaction class name starting with 3.11.3
- Ugi reaction: These have a reaction class name starting with 2.1.28
- Suzuki-Coupling: These have a reaction class name starting with 3.1

As the dataset contained numerous reactions, we concentrated on specific reactions, as detailed in Section 3.1. The NextMove software was used to provide the reaction classification. In order to target specific reaction classes for evaluation, we segmented the data into two distinct parts. The first part encompassed only the category name (e.g., Ugi reaction), while the second part was solely devoted to the category index or identifier (e.g., 2.1.28). Separate data files were then created for each reaction, to assess them individually.

Conversely, in the USPTO dataset, the reactants and products were extracted from the dataset and atom-mapping is removed. There was also no designated category for segregating data into training or testing groups. As a result, we incorporated a new classification that would distinctly designate the data for testing. Some reactions in this dataset were not categorized, so we classified them according to the aforementioned rules. Subsequently, we divided the datasets into distinct CSV files based on reaction classes for individual evaluation.

3.1.4 Proportion of reaction classes in datasets

The proportion of chemical reactions in datasets can vary depending on the focus and scope of the dataset. Datasets may also be biased towards certain types of reactions, it's important to keep in mind that the proportion of chemical reactions in a dataset is not necessarily representative of all chemical reactions that occur in nature or are of scientific interest. Here in this study, our main focus is on particular

reaction classes as explained in (section 3.1.3). The following figures will illustrate more clearly how the proportions of reaction classes varied.

3.1.4.1 Proportion of reaction classes in the in-house dataset

The in-house dataset that was created in-house was divided into training and testing sets. During the training and testing process, the model was exposed to 18 million records. However, it was observed that the Ugi had a lower proportion in comparison to other reactions, which had a higher proportion. This distribution for training dataset can be visualized in the Figure 3.1 and distribution for test dataset can be visualized in the 3.2

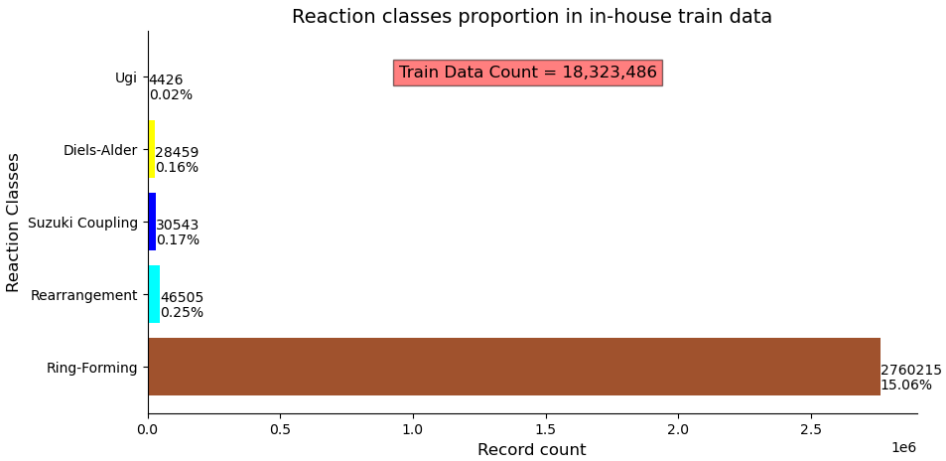

Figure 3.1: *The figure displays the percentage distribution of specific reaction classes within the in-house dataset. The proportions are determined in relation to the total count of approximately 18 million data points in the in-house dataset.*

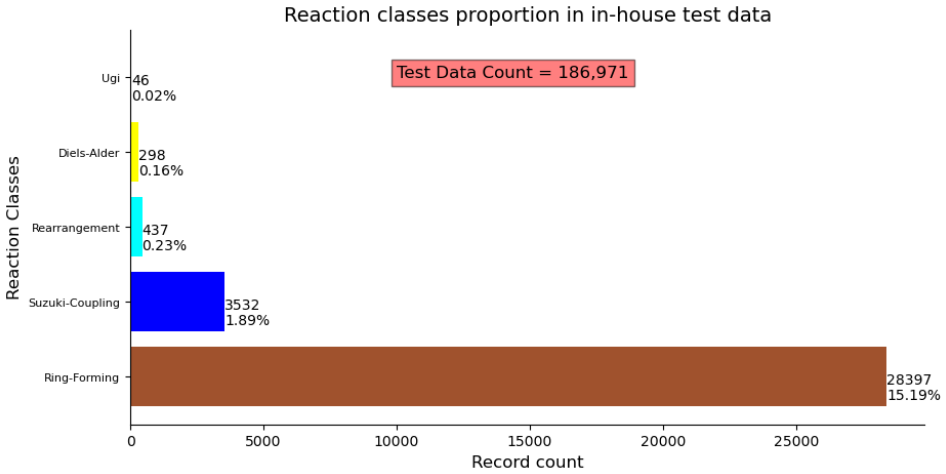

Figure 3.2: *The figure displays the percentage distribution of specific reaction classes within the in-house dataset. The proportions are determined in relation to the total count of approximately 186k data points in the in-house dataset.*

3.1.4.2 Proportion of reaction classes in the USPTO dataset

In the USPTO dataset, the Ugi class is underrepresented as it is only present in small portion. Conversely, Ring-Forming, Suzuki-Coupling and Diels reactions are more prevalent in the USPTO dataset, with a considerable proportion of samples belonging to these classes. This distribution has also been depicted in the Figure 3.3.

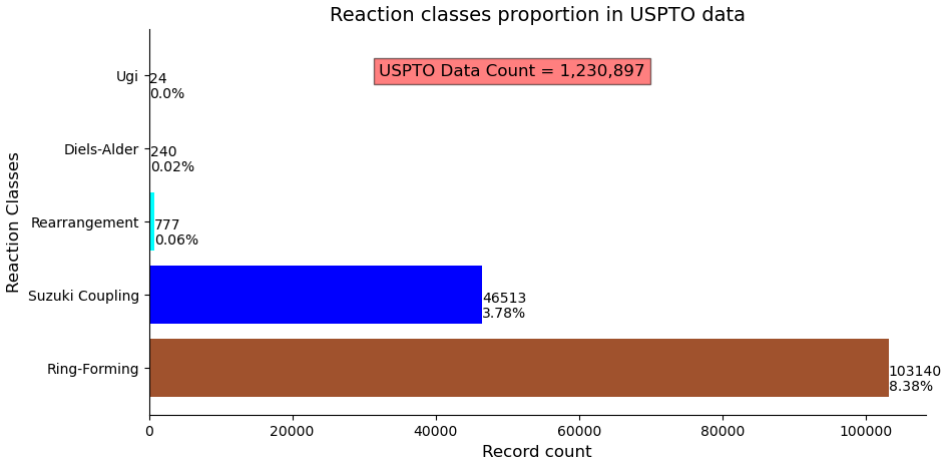

Figure 3.3: *The figure displays the percentage distribution of specific reaction classes within the USPTO dataset. The proportions are determined in relation to the total count of approximately 1.2 million data points in the USPTO dataset.*

3.2 Evaluation

1. The pre-trained Chemformer model undergoes evaluation at multiple stages, including:
    - In-house test set: consisting of 186,000 records
    - USPTO dataset: containing 1.1 million records
    - Individual reaction classes: evaluating the model's performance within specific challenging reaction classes of interest, such as Ring-Forming reactions, Rearrangement reactions, Diels-Alder, Ugi, and Suzuki-Coupling.
2. To carry out the evaluation, the following steps are taken:
    - The Chemformer model is loaded, and data is processed in batches of 128 records.
    - The model takes input products and generates sample SMILES and their unique values in output files
    - These files are further processed to canonicalize both target and sampled SMILES, which aids in generating the necessary evaluation metrics
    - A variety of evaluation metrics are employed to perform diverse analyses on the Chemformer model's effectiveness. These evaluation metrics are

detailed in Section 2.8.

3.3 Transfer learning

We explore the application of transfer learning on pre-trained model with the objective of fine-tuning specific chemical reaction classes - Ugi, Suzuki-Coupling, Rearrangement, Diels-Alder and Ring-Forming reactions. Our approach commenced with extracting distinct subsets of data from in-house training set for each of these reaction classes. Each subset of data is then used to fine-tune the pre-trained model using a specified learning rate of 0.01, a cross-entropy loss function, a batch size of 32, and iterating for 70 epochs. The process of fine-tuning is executed individually for each reaction class, resulting in five uniquely fine-tuned models. During the fine-tuning procedure, checkpoints are established at regular intervals to record progress and monitor performance. The final fine-tuned model is selected based on the checkpoint where accuracy is observed to converge. This final model is then subjected to evaluation on both our in-house testing set as well as the USPTO datasets. The purpose of this approach is to assess the efficacy of transfer learning in predicting specific reaction classes and understand the impact of the fine-tuning process on model accuracy and performance.

3.4 Explainable AI

Our process uses the BART model, called Chemformer, to provide explanations for its predictions. The explanations take the form of feature importance values and attention heat maps for given molecular transformations. These explanations help in understanding which parts of the input molecules are important for a given predicted transformation.

4

Results

This chapter aims to detail the outcomes of our research, providing an analytical perspective on the gathered data and addressing the potential implications. Our discussion commences with an overview of the distribution of various chemical reactions, followed by an evaluation of these reactions based on distinct metrics and a review of their fine-tuning processes. Finally, we explore the application of Explainable AI as a tool for interpreting the outcomes produced by the model.

4.1 Evaluation

4.1.1 Evaluation on overall and specific reaction classes for in-house test data

The model's performance on our proprietary test data, as depicted in Figure 4.1, reveals consistent results across various metrics. The Top-1, Top-3, Top-5, and Top-10 metrics demonstrate performance of 34%, 49%, 55%, and 61%, respectively across the overall in-house test set. This indicates that the model's accurate predictions increase in relation to the increase in Top-k. In other words, the Top-10 metrics contain more accurate predictions compared to the Top-5, Top-3 or Top-1. A Tanimoto similarity coefficient of 0.7 suggests that the reactions predicted by the model exhibit a 70% correspondence with ground truth reactions. Furthermore, the significant Fraction Unique score of 0.77 indicates a high level of diversity within the model's predictions and with a Fraction Invalid score of 0.01, demonstrates that only 1% of predicted reactions are invalid, emphasizing the overall validity of model's output.

Upon evaluating individual reaction classes, the Ugi, Diels-Alder and Suzuki-Coupling reaction classes consistently showcases high performance, ranging from 80% for Top-1 to 89% for Top-10 for Ugi, 70% for Top-1 to 88% for Top-10 for Diels-Alder and 37% for Top-1 to 79% for Top-10 for Suzuki-Coupling. In contrast, the Ring-Forming and Rearrangement reaction class demonstrates performance below the overall test set metrics, ranging from 32% for Top-1 to 58% for Top-10 for Ring-Forming and 24% for Top-1 to 49% for Top-10 for Rearrangement. This indicates that the prediction of Ring-Forming and Rearrangement reactions presents a significant challenge, due to their complex mechanisms and intramolecular interactions.

Additionally the Tanimoto similarity for Ugi, Diels-Alder is 0.98, for Rearrangement,

Ring-Forming and Suzuki-Coupling is 0.58, 0.70 and 0.78. While Fraction Invalid is 0.01 for Ring-Forming, Diels-Alder and Suzuki-Coupling reaction classes. Whereas, Rearrangement and Ugi exhibits 0.0 for Fraction Invalid. The Fraction Unique for Ugi, Diels-Alder, Ring-Forming, Rearrangement, Suzuki-Coupling are 0.43,0.53, 0.72, 0.81 and 0.86 respectively. High Tanimoto similarity and low Fraction Unique for both Ugi and Diels-Alder proves that they are quite consistent and predictable reaction mechanisms.

These metrics indicate that the pre-trained model demonstrated good performance on the in-house test set, particularly in the case of Ugi and Diels-Alder reactions which exceeded performance metrics across all other reaction categories. The substantial Tanimoto similarity signifies a high degree of accuracy in the model's predictive capabilities. Combination of low Fraction Invalid and high Fraction Unique provide a good evaluation of models quality and diversity.

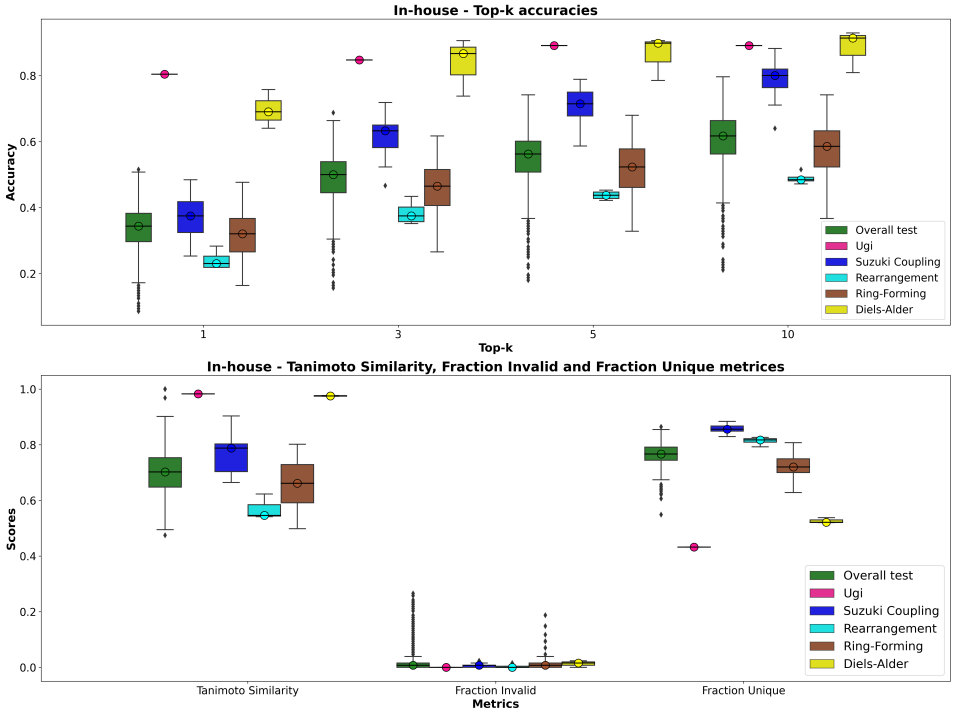

Figure 4.1: *Box plots illustrating the various evaluation scores for the pre-trained Chemformer model when assessed on an in-house dataset. The upper plot illustrates the Top-k accuracy metrics and the lower graph presents the Tanimoto Similarity, followed by the Fraction Invalid and Fraction Unique values.*

4.1.2 Evaluation on overall and specific reaction classes for USPTO data

The evaluation of the model's performance on USPTO data, as depicted in Figure 4.2, reveals insightful results. Across all reaction classes, the Top-1, Top-3, Top-5, and Top-10 metrics achieve accuracies of 42%, 61%, 67%, and 73%, respectively, surpassing the evaluation on the overall In-house dataset. Conversely, the

Tanimoto similarity, Fraction Invalid, and Fraction Unique metrics exhibit remarkable accuracy levels of 0.81, 0.0, and 0.81, respectively, across all reaction classes in the USPTO dataset.

Upon evaluating individual reaction classes, the Ugi, Diels-Alder and Suzuki-Coupling reaction classes consistently showcases high performance, ranging from 58% for Top-1 to 75% for Top-10 for Ugi, 69% for Top-1 to 98% for Top-10 for Diels-Alder and 45% for Top-1 to 86% for Top-10 . In contrast, the Ring-Forming and Rearrangement reaction class demonstrates performance below the overall USPTO test set, ranging from 32% for Top-1 to 63% for Top-10. for Ring-Forming and 36% for Top-1 to 58% for Top-10 for Rearrangement. This mirrors the same pattern observed in the in-house test set.

Additionally the Tanimoto similarity for Ugi and Diels-Alder is 0.98, 0.60 for Rearrangement and Ring-Forming, 0.83 for Suzuki-Coupling. While Fraction Invalid is 0.0 for all reaction classes. The Fraction Unique for Ugi, Diels-Alder, Ring-Forming, Rearrangement and Suzuki-Coupling are 0.57, 0.52, 0.78, 0.79 and 0.86 respectively. Despite the lower top-k accuracy metrics for Ring-Forming and Rearrangement reactions, the high Fraction Unique value indicates a considerable diversity in these reactions. This suggests that even though precise prediction of these reactions may pose challenges, the model can generate a broad spectrum of unique results within these categories.

The metrics observed are largely comparable to those from the in-house test set, demonstrating consistency in the model's performance across different test sets. An exception to this similarity is the presence of more outliers in the USPTO set. The reason behind these outliers could be attributed to a higher number of data batches from the USPTO, as well as the fact that the model has been pre-trained on the in-house test set, thereby influencing its familiarity with the chemical space, which could be different when compared to the USPTO set.

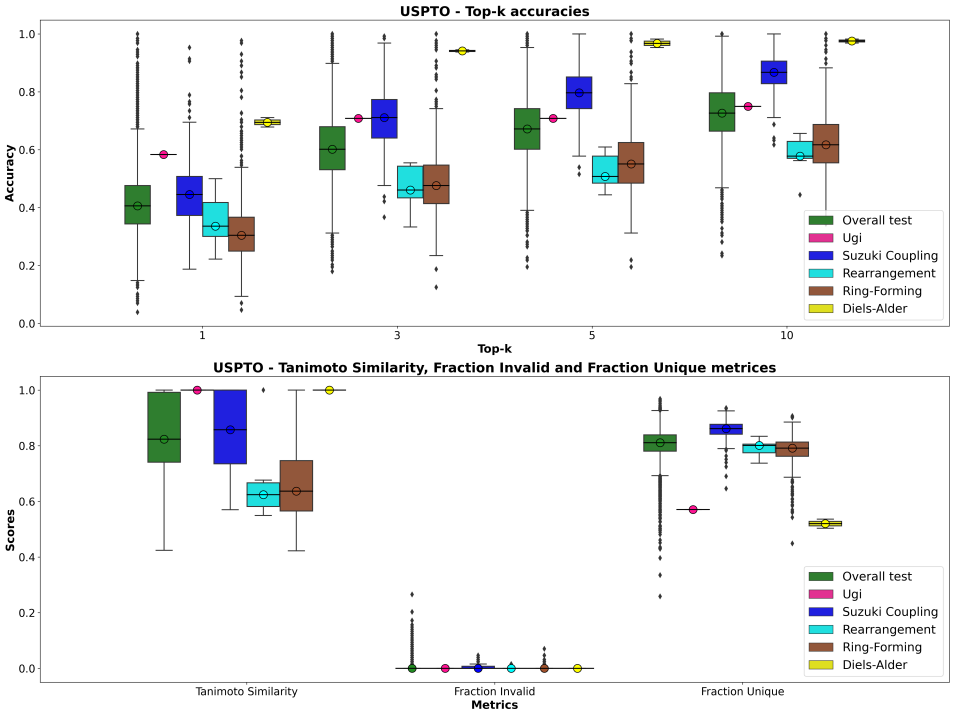

Figure 4.2: *Box plots illustrating the various evaluation scores for the pre-trained Chemformer model when assessed on overall USPTO data. The upper plot illustrates the Top-k accuracy metrics and the lower graph presents the Tanimoto Similarity, followed by the Fraction Invalid and Fraction Unique values.*

When compared with all the reaction classes in USPTO dataset, Ugi and Diels-Alder performed well. These reaction classes are multi-centered, often yielding products that are less diverse than those derived from other reaction types, such as Suzuki-Coupling. Both the Ugi and Diels-Alder reactions involve the creation of distinct bond types and structural arrangements, which leads to a reduced variability in the resultant products, thereby simplifying prediction tasks. In contrast, the compounds produced from Suzuki-Coupling can exhibit substantial structural diversity, potentially adding a layer of complexity to the prediction process. For instance, Figure 4.3 represents the true reactions, and Figure 4.4 represents the predicted reactions. Although the predicted reaction does not appear to be a typical Suzuki reaction at first glance, it can be categorized as a cross-coupling Suzuki reaction owing to the presence of the bromine chemical element. These exceptional predictions reduce the overall accuracy to 45% on the USPTO dataset. However, upon more detailed examination, it can be concluded that 67% of the predicted reactions can indeed be classified as Suzuki reactions.

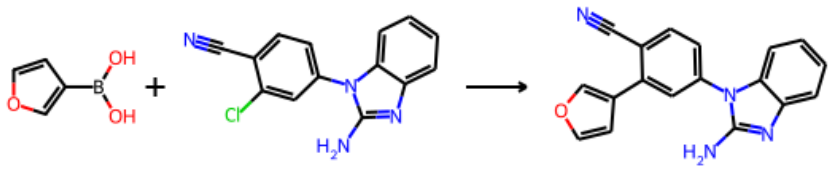

Figure 4.3: *This is an example reaction for Suzuki-Coupling from USPTO dataset, which is a true reaction.*

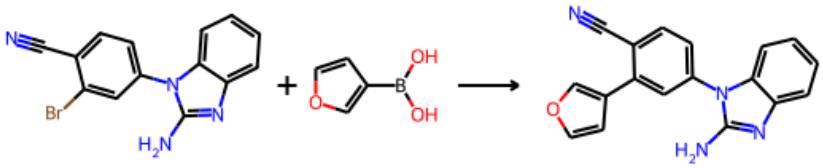

Figure 4.4: *This is an example reaction for Suzuki-Coupling from USPTO dataset, which is a predicted reaction.*

4.1.3 Fraction Invalid

The "Fraction Invalid" is a measurement employed to determine the quantity of incorrect predictions within a given set. It is essentially the proportion of inaccurate SMILES strings compared to the complete number of SMILES strings that have been produced. Evident in Figure 4.1 and Figure 4.2, there is a substantial Fraction Invalid detected across the overall test sets in specific batches. This observation is largely attributed to certain reactions, which are inclined to produce a considerable fraction of invalid outcomes, as demonstrated in Figure 4.5. Therefore, if data batches, that are assembled randomly, have these specific reactions, it leads to an overall higher fraction of invalid data within these particular batches.

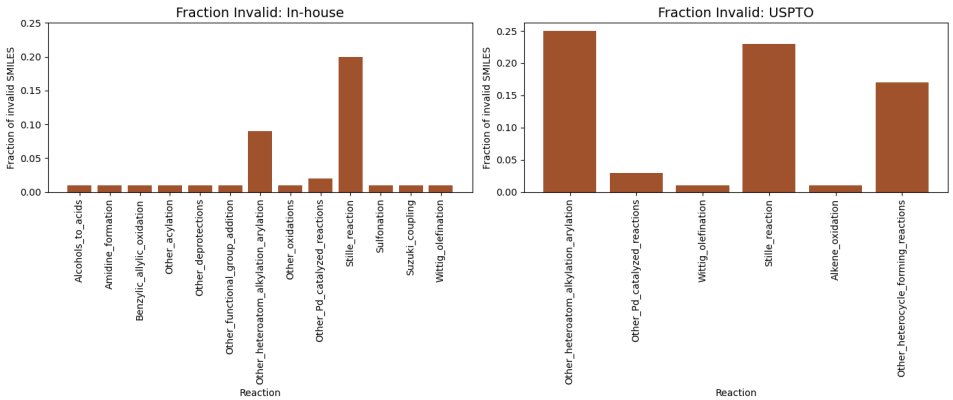

Figure 4.5: *Bar plots illustrating the fraction of invalid SMILES by pre-trained Chemformer model on in-house and USPTO datasets respectively. Other reactions that are not part of this plot have zero Fraction Invalid.*

4.1.4 Chemformer vs template-based performance on an in-house data

The below Figure 4.6 presents performance comparison between the Chemformer and template-based methods for predicting the Top-1 and Top-10 reaction outcomes across five distinct reaction classes (Ring-Forming, Rearrangement, Diels-Alder, Ugi, and Suzuki-Coupling) using our proprietary in-house data.

The term "Chemformer (Top-1)" refers to the proportion of reactions for which the Chemformer model has accurately predicted the most likely outcome, while the corresponding proportion for the template-based model is also indicated. Similarly, "Chemformer (Top-10)" signifies the proportion of reactions for which the Chemformer model has accurately predicted one of the ten most likely outcomes, with a corresponding measure for the template-based model.

For instance, in the Ring-Forming class, the Chemformer model achieved a Top-1 prediction accuracy of 32% and a Top-10 prediction accuracy of 58%, while the template-based model yielded a Top-1 prediction accuracy of 29% and a Top-10 prediction accuracy of 43%. Similarly, in the Rearrangement class, the Chemformer model attained a Top-1 prediction accuracy of 24% and a Top-10 prediction accuracy of 49%, whereas the template-based model achieved a Top-1 prediction accuracy of 15% and a Top-10 prediction accuracy of 25%.

Notably, the Chemformer model outperformed the template-based model across all reaction classes. For example, in the Diels-Alder class, the Chemformer model achieved a Top-1 prediction accuracy of 70% and a Top-10 prediction accuracy of 80%, whereas the template-based model achieved a Top-1 prediction accuracy of 35% and a Top-10 prediction accuracy of 49%. Similarly, in the Ugi and Suzuki-Coupling reaction classes, the Chemformer model demonstrated a Top-1 prediction accuracy of 80%, 69% and a Top-10 prediction accuracy of 89%, 80% surpassing the template-based model with its respective Top-1 prediction accuracy of 52%, 29% and Top-10 prediction accuracy of 67% and 60% respectively.

These findings indicate that the Chemformer model consistently outperforms the template-based model in predicting both the Top-1 and Top-10 outcomes across all reaction classes in an in-house test set.

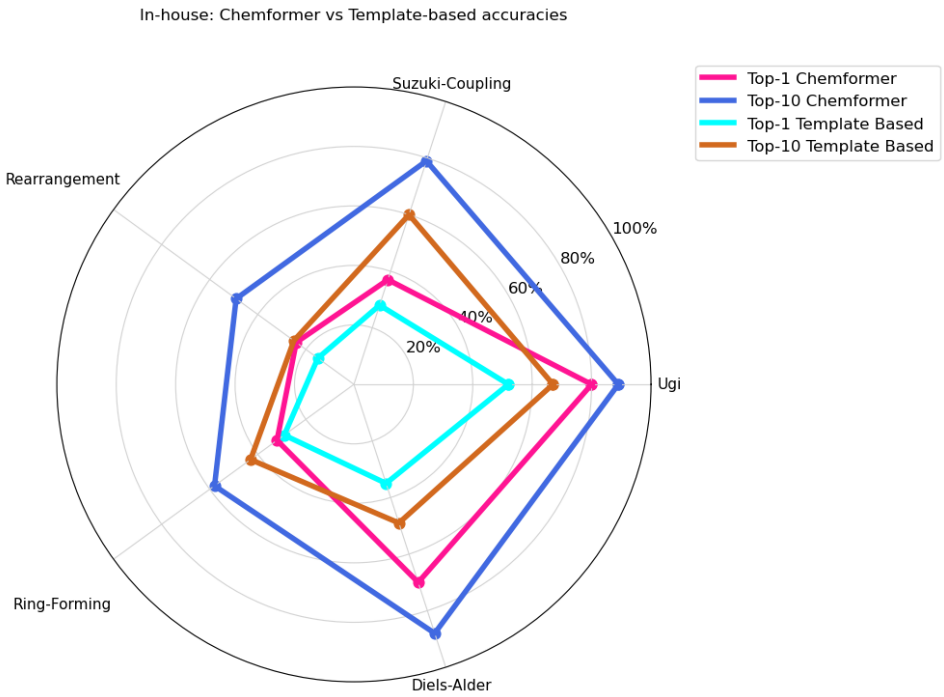

Figure 4.6: *Radar plot illustrating the evaluation scores for the pre-trained Chemformer model and template-based model when assessed on an in-house dataset.*

4.1.5 Chemformer vs template-based performance on USPTO data

The below Figure 4.7 presents performance comparison between the Chemformer and template-based models for predicting the Top-1 and Top-10 reaction outcomes across five distinct reaction classes (Ring-Forming, Rearrangement, Diels-Alder, Ugi, and Suzuki-Coupling) using our proprietary in-house data.

The "Chemformer (Top-1)" represents the fraction of reactions for which the Chemformer model accurately predicted the Top-1 outcome, while the corresponding fraction for the template-based model is indicated in the "Production template-based (Top-1)" column. Similarly, the "Chemformer (Top-10)" column denotes the fraction of reactions for which the Chemformer model correctly predicted the Top-10 outcome, while the "Production template-based (Top-10)" column displays the corresponding fraction for the template-based model.

For instance, in the Ring-Forming class, the Chemformer model achieved a Top-1 prediction accuracy of 32% and a Top-10 prediction accuracy of 63%, while the template-based model yielded a Top-1 prediction accuracy of 28% and a Top-10 prediction accuracy of 46%. Similarly, in the Rearrangement class, the Chemformer model attained a Top-1 prediction accuracy of 36% and a Top-10 prediction accuracy of 58%, whereas the template-based model achieved a Top-1 prediction accuracy of 30% and a Top-10 prediction accuracy of 46%.

Notably, the Chemformer model outperformed the template-based model across all

reaction classes. For example, in the Diels-Alder class, the Chemformer model achieved a Top-1 prediction accuracy of 69% and a Top-10 prediction accuracy of 98%, whereas the template-based model achieved a Top-1 prediction accuracy of 42% and a Top-10 prediction accuracy of 69%. Similarly, in the Ugi and Suzuki-Coupling reaction classes, the Chemformer model demonstrated a Top-1 prediction accuracy of 58%, 49% and a Top-10 prediction accuracy of 75%, 89% surpassing the template-based model with its respective Top-1 prediction accuracy of 29%, 30% and Top-10 prediction accuracy of 42% and 67% respectively .

These findings indicate that the Chemformer model consistently outperforms the template-based model in predicting both the Top-1 and Top-10 outcomes across all reaction classes in an USPTO dataset.

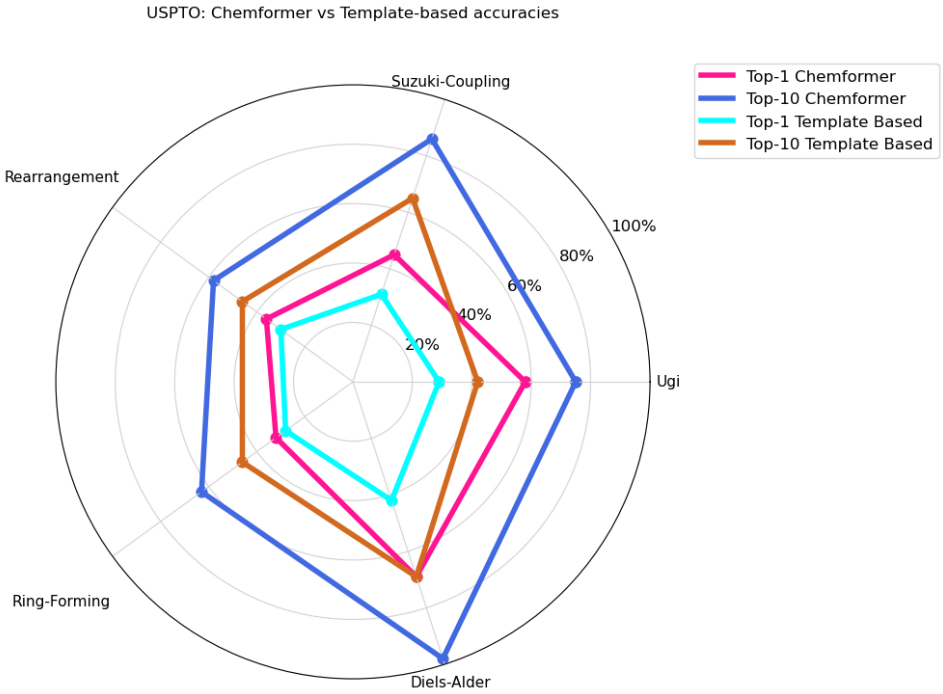

Figure 4.7: *Radar plot illustrating the evaluation scores for the pre-trained Chemformer model and template-based model when assessed on USPTO dataset.*

4.1.6 Evaluation on other reaction classes

In addition to the particular reaction classes explored in previous sections, the model's overall performance across various reaction classes is also assessed by evaluating on additional reaction classes found in both in-house and USPTO test sets as shown in Figure A.1 and Figure A.2. The evaluation outcomes from this analysis confirm that the pre-trained Chemformer has surpassed the performance of template-based models, barring the Stille reaction, which is not considered a crucial reaction class, primarily due to its higher toxicity. Moreover, there are alternative cross-coupling reactions that yield similar outcomes to the Stille reaction. One such example is the Suzuki-Coupling reaction, which is already a central focus of our

study. This underscores the general efficacy of Chemformer across a broad range of reaction classes.

4.1.6.1 In-house: Best and worst performing five reactions

Beyond the specific reactions addressed in Section 4.1.4, the Figure 4.8 presents a comparison of the five best and least effective reactions when evaluated on the Chemformer model in comparison to the template-based model for the in-house data.

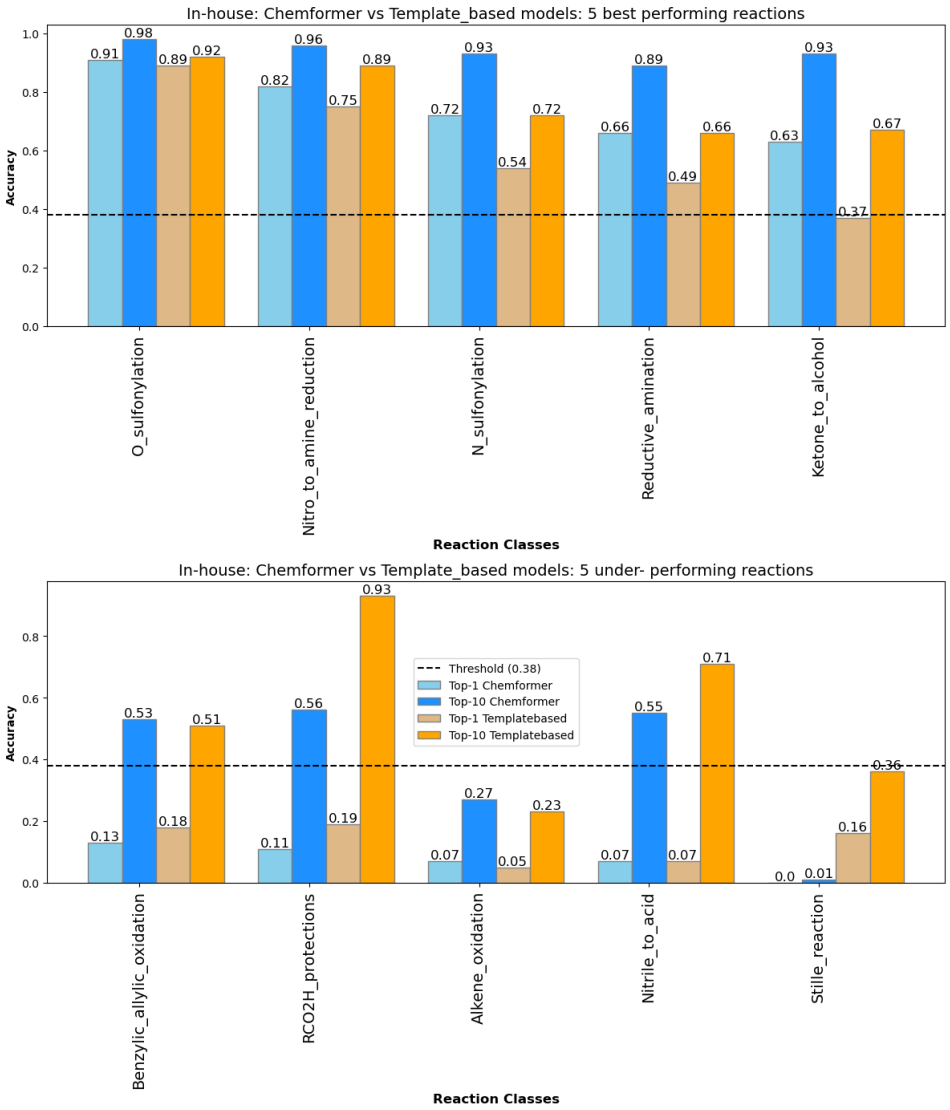

Figure 4.8: *Bar plots illustrating the evaluation scores for the pre-trained Chemformer model compared with template-based model on an in-house dataset. First sub plot illustrates best 5 reactions and second sub plot illustrates least 5 reactions.*

4.1.6.2 USPTO: Best and worst performing five reactions

Beyond the specific reactions addressed in Section 4.1.5, the Figure 4.9 presents a comparison of the five best and least effective reactions when evaluated on the

Chemformer model in comparison to the template-based model for USPTO data.

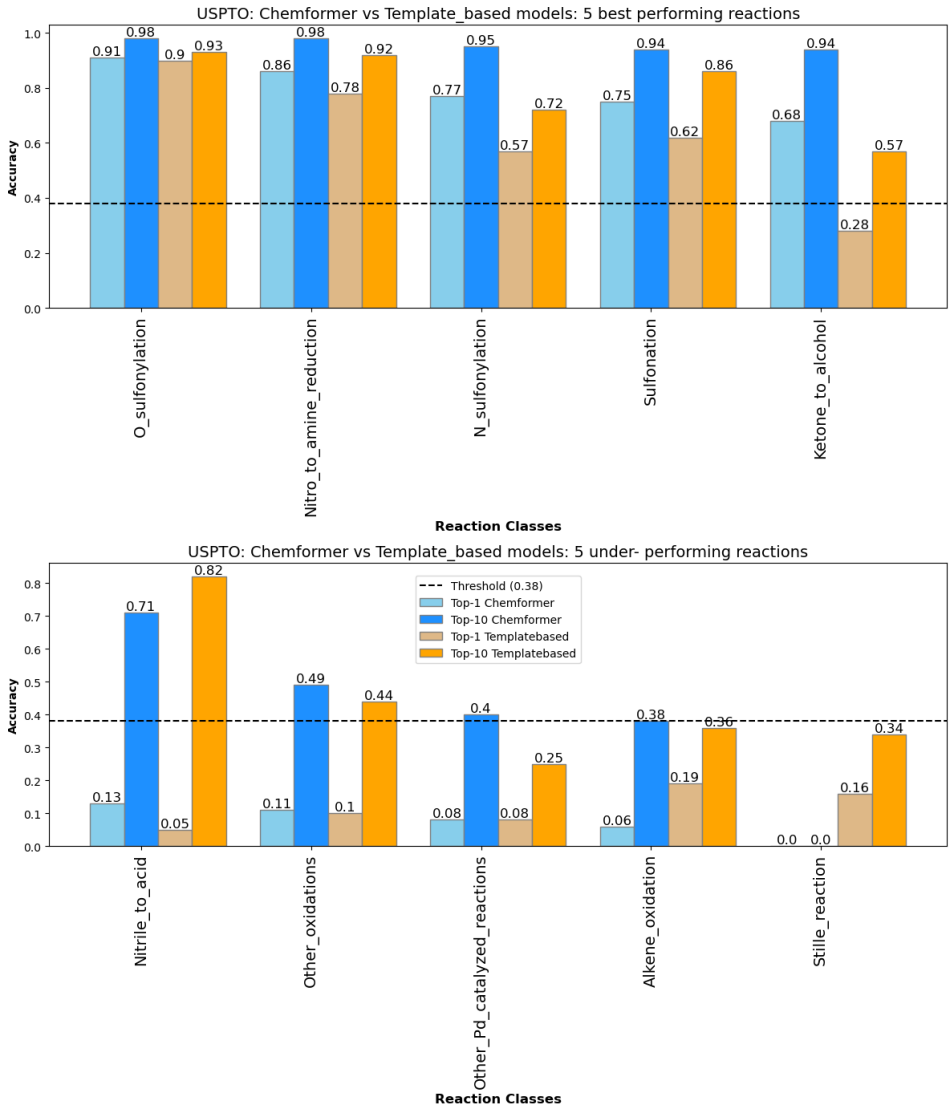

Figure 4.9: *Bar plots illustrating the evaluation scores for the pre-trained Chemformer model compared with template-based model on USPTO dataset. First sub plot illustrates best 5 reactions and second sub plot illustrates least 5 reactions.*

4.1.7 Diversity score on overall in-house test set

The distribution plot in the Figure 4.10 visualizes the diversity scores within in-house test set. Which ranges from 0 to 1 and serve as a measure of the diversity within a dataset. A diversity score of 0 indicates low diversity, while a score of 1 represents high diversity. The overall diversity score for the in-house test set is calculated to be 0.49618955. This score suggests that a significant portion of the samples or categories within the dataset have unique reactions, concentrated towards the extremes of 1 and 0, which greatly impacts the overall dataset diversity. The scatter plot, accompanied by a best fit line, showcases the relationship between the relative entropy and number of unique reaction classes. It is important to note that

even if the unique reactions may be the same for different products, the entropy values differ due to variations in the distributions of reaction classes.

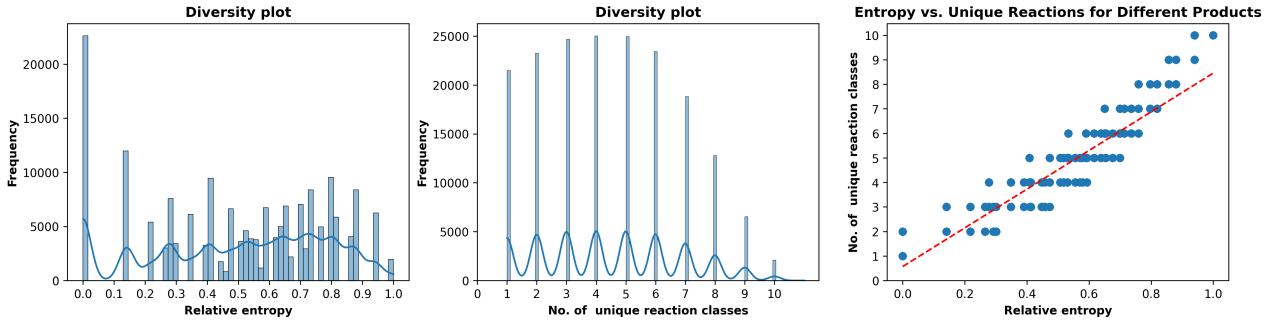

Figure 4.10: *The diversity score distribution plot visualizes the distribution of diversity scores, illustrating the spread and concentration of diversity across the In-house test set.*

4.2 Fine-tuning

After pre-training, the Chemformer model is fine-tuned on downstream datasets by applying transfer learning. For this we investigated five reaction classes as shown in Figure 4.11. In line with evaluation, the in-house and USPTO test datasets are used to benchmark Chemformer against template-based model on retrosynthesis prediction task. A cross-entropy loss function and a learning-rate of 0.01 are used to train the model.

4.2.1 Fine-tuning on in-house and USPTO dataset

The subplot for in-house dataset in Figure 4.11 shows delta values indicating performance improvements achieved by fine-tuning the Chemformer model. Notably, the suzuki-coupling reaction demonstrates a significant increase with a delta of 32% for top-1 accuracy. Similarly, Diels-Alder and ugi reactions exhibit noticeable enhancements with deltas of 18% and 2%, respectively. The Rearrangement reaction shows the highest improvement with a delta of 51% for top-1 accuracy. The same trend is observed for top-10 accuracies, with Suzuki-Coupling, Diels-Alder, Ugi, and Rearrangement reactions having deltas of 15%, 5%, 11%, and 43%, respectively.

Additionally, the subplot for the USPTO dataset in Figure 4.11 displays delta values reflecting performance gains achieved through fine-tuning the Chemformer model. The Suzuki-Coupling reaction shows a substantial improvement with a delta of 37% for top-1 accuracy, while Diels-Alder and Ugi reactions exhibit notable enhancements with deltas of 24% and 42%, respectively. The Rearrangement reaction demonstrates the highest improvement with a delta of 0.53 for top-1 accuracy. Similarly, the top-10 accuracies for Suzuki-Coupling, Diels-Alder, Ugi, and Rearrangement reactions have deltas of 12%, 2%, 25%, and 40%, respectively.

These delta values indicate the positive impact of fine-tuning on the Chemformer model's accuracy for both the in-house and USPTO datasets, resulting in improved

performance and predictive capabilities. Moreover, it is noteworthy that although the delta values for the Ugi and Diels-Alder reactions are relatively low, these reaction classes already exhibit high performance.

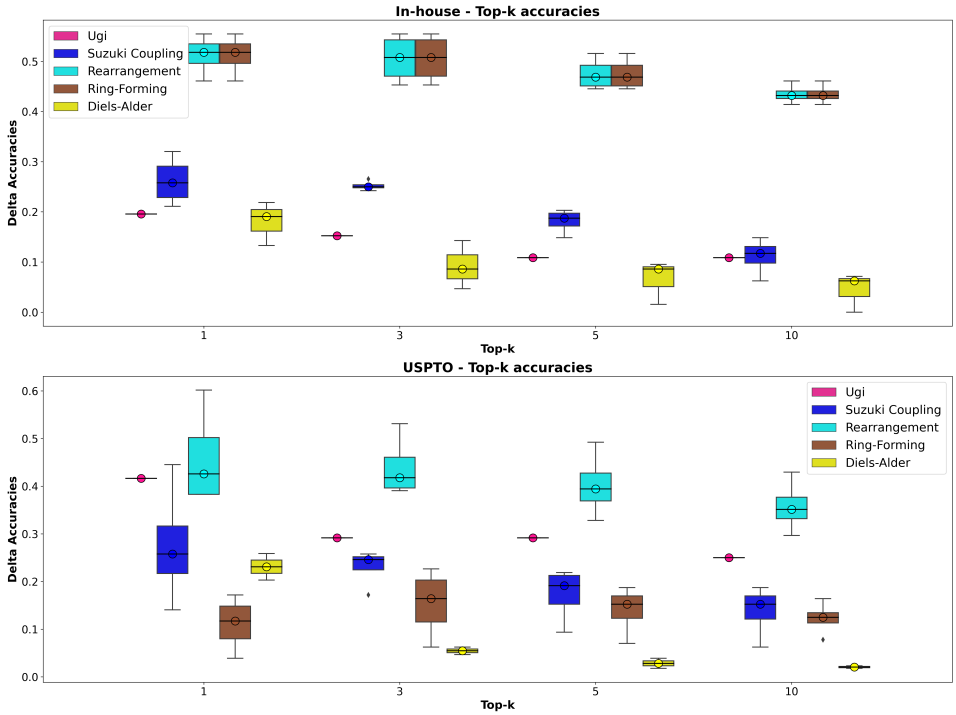

Figure 4.11: *Box plots illustrating the various evaluation scores for the fine-tuned Chemformer model when assessed on an in-house and USPTO dataset.*

4.2.2 Chemformer accuracies before and after fine-tuning on an in-house dataset

The below table shows a comparison of the Chemformer, before and after fine-tuning on an in-house dataset for different chemical reactions. Overall, the model's performance shows significant improvement after fine-tuning. Before fine-tuning, the Chemformer model achieves relatively lower accuracies, but after fine-tuning, its predictive capabilities enhance considerably. The Suzuki-Coupling reaction exhibits a substantial increase in accuracy, with the top-1 prediction improving from 37% to 69% and the top-10 prediction rising from 79% to 94%. Similarly, the Rearrangement reaction shows notable progress, with the top-1 accuracy increasing from 24% to 75% and the top-10 accuracy improving from 49% to 92%. The Chemformer model performs well for the Diels-Alder reaction initially, and fine-tuning helps maintain its high accuracy, with top-1 and top-10 accuracies reaching 88% and 93%, respectively. The Ugi reaction also demonstrates strong performance even before fine-tuning, but after fine-tuning, the Chemformer model achieves high accuracy with top-1 and top-10 accuracies of 100%. These results shows that the effectiveness of fine-tuning in enhancing the Chemformer model predictive abilities for these chemical reactions on in-house test set.

Table 4.1: *Chemformer accuracies before and after fine-tuning on in-house dataset*

Reaction	Before: Top-1	Before: Top-10	After: Top-1	After: Top-10
Suzuki-Coupling	0.37	0.79	0.69	0.94
Rearrangement	0.24	0.49	0.75	0.92
Diels-Alder	0.70	0.88	0.88	0.93
Ugi	0.80	0.89	1.0	1.0

4.2.3 Chemformer accuracies before and after fine-tuning on an USPTO dataset

The below table shows a comparison of the Chemformer model, before and after fine-tuning on USPTO dataset for different chemical reactions. Overall, the model's performance improves significantly after fine-tuning. For the Suzuki-Coupling reaction, the top-1 accuracy increases from 45% to 82%, and the top-10 accuracy rises from 86% to 98%. Similarly, the Rearrangement reaction shows performs, with the top-1 accuracy improving from 36% to 89% and the top-10 accuracy increasing from 58% to 98%. The Chemformer model already performs well for the Diels-Alder reaction initially, and fine-tuning further enhances its accuracy, reaching a top-1 accuracy of 93% and top-10 accuracy of 100%. The Ugi reaction also exhibits strong performance, and after fine-tuning, the Chemformer model achieves perfect accuracy with top-1 and top-10 accuracies of 100%. These results shows that the effectiveness of fine-tuning in enhancing the Chemformer model predictive abilities for these chemical reactions on USPTO dataset.

Table 4.2: *Chemformer accuracies before and after fine-tuning on USPTO dataset*

Reaction	Before: Top-1	Before: Top-10	After: Top-1	After: Top-10
Suzuki-Coupling	0.45	0.86	0.82	0.98
Rearrangement	0.36	0.58	0.89	0.98
Diels-Alder	0.69	0.98	0.93	1.0
Ugi	0.58	0.75	1.0	1.0

4.2.4 In-house: Fine-tuned Chemformer vs template-based model

The Figure 4.12 represents a comparison of fine-tuned Chemformer model and template-based model on an in-house test set, in predicting outcomes for different reaction classes. Across the reactions, the fine-tuned Chemformer models consistently demonstrate higher accuracy. For the Ugi reaction, both models achieve an accuracy of 100% for the Top-1 prediction, but the fine-tuned Chemformer outperforms the template-based model with accuracies of 1% and 52% for the Top-10 prediction, respectively. Similarly, for the Suzuki-Coupling reaction, the fine-tuned Chemformer model achieves higher accuracies of 69% (Top-1) and 94% (Top-10) compared to 28% and 60% for the template-based approach. The trend continues with the Rearrangement reaction, where the fine-tuned Chemformer achieves accuracies of 75% (Top-1) and 92% (Top-10), surpassing the template-based model accuracies of 15% and 25%.

In the case of the Diels-Alder reaction, the fine-tuned Chemformer model performs better with accuracies of 88% (Top-1) and 93% (Top-10), while the template-based model achieves accuracies of 35% and 49%, respectively. Overall, the results suggest that the fine-tuned Chemformer model consistently outperform the template-based model in accurately predicting reaction outcomes.

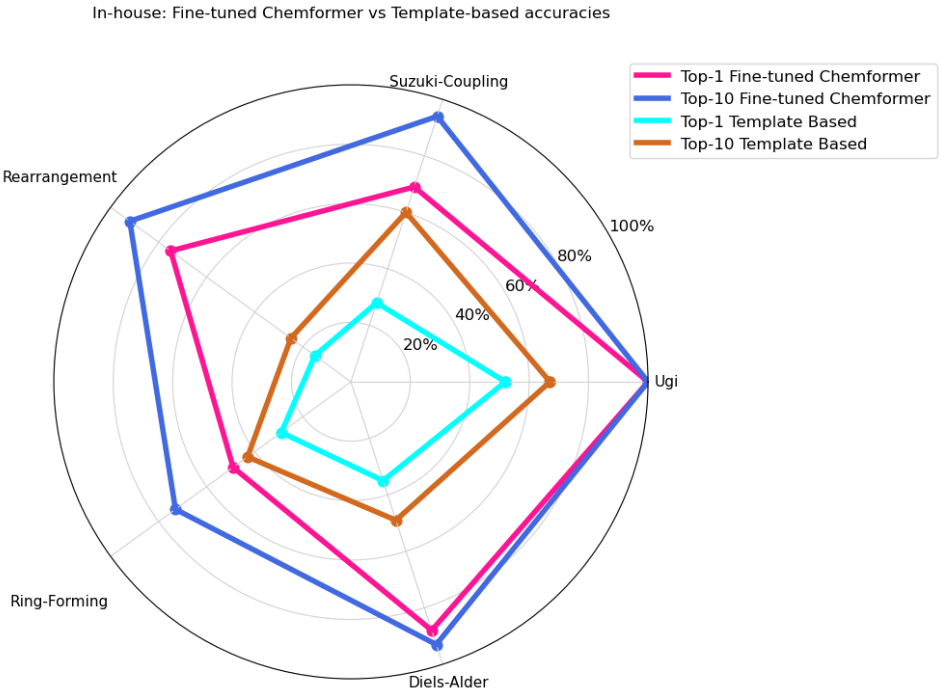

Figure 4.12: *Radar plot illustrating the evaluation scores for the fine-tuned Chemformer model and template-based model when assessed on an in-house dataset.*

4.2.5 USPTO: Fine-tuned Chemformer vs template-based model

The Figure 4.13 represents a comparison of, fine-tuned Chemformer model and template-based model on an USPTO dataset, in predicting outcomes for different reaction types. The results indicate that the fine-tuned Chemformer model consistently outperform the template-based model in terms of accuracy. For the Ugi reaction, both the Top-1 and Top-10 predictions achieve an accuracy of 100% using the fine-tuned Chemformer models, while the template-based model achieves lower accuracies of 29% and 42% for the respective predictions. In the case of Suzuki-Coupling, Rearrangement, and Diels-Alder reactions, the fine-tuned Chemformer models achieve higher accuracies, ranging from 82% to 100% for Top-1 predictions and 98% to 100% for Top-10 predictions. On the other hand, the template-based models demonstrate lower accuracies, varying from 30% to 42% for Top-1 predictions and 46% to 69% for Top-10 predictions. These results emphasize the superior performance of the fine-tuned Chemformer model in accurately predicting reaction outcomes across different reaction types.

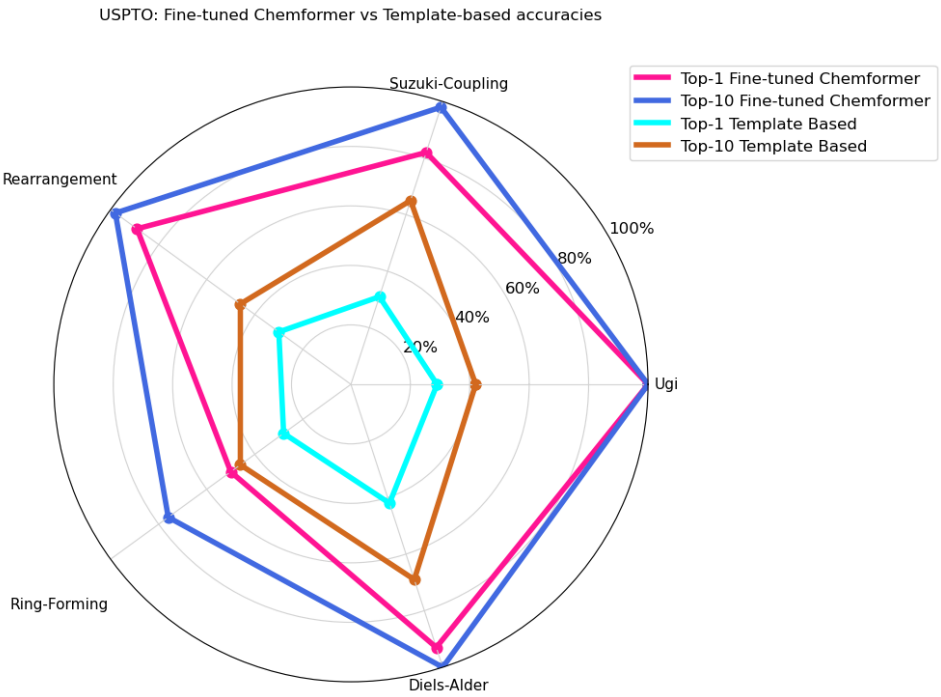

Figure 4.13: *Radar plot illustrating the evaluation scores for the fine-tuned Chemformer model and template-based model when assessed on USPTO dataset.*

4.3 Explainable AI

Attention mechanism can be interpreted as the model's "focus" or "importance weighting" during the prediction process. For example, in atom mapping tasks, the attention scores reflect the predicted correspondence between atoms in the reactants and the products of a chemical reaction. These attention scores can be visualized using a heat map from Figure 4.15. Each axis of the heat map represents the tokens atoms in the reactants and products, and the colour of each cell in the heat map represents the attention score, with warmer colours typically representing higher scores. This visualization provides an interpretable overview of the model's decision-making process, highlighting the atoms that the model deems most relevant for the prediction.

We also created a heatmap of the product-reactants atom-mapping Figure 4.15. Atom-mapping, refers to the process of identifying which atoms in the reactant molecules correspond to which atoms in the product molecules in a chemical reaction. This helps to understand attention scores in Figure 4.15, atoms that are corresponding to each other have more attention weights. This heat map serves as a tool to interpret the attention scores heat map Figure 4.15, where we anticipate high attention scores to be attributed to atoms that correspond with one another.

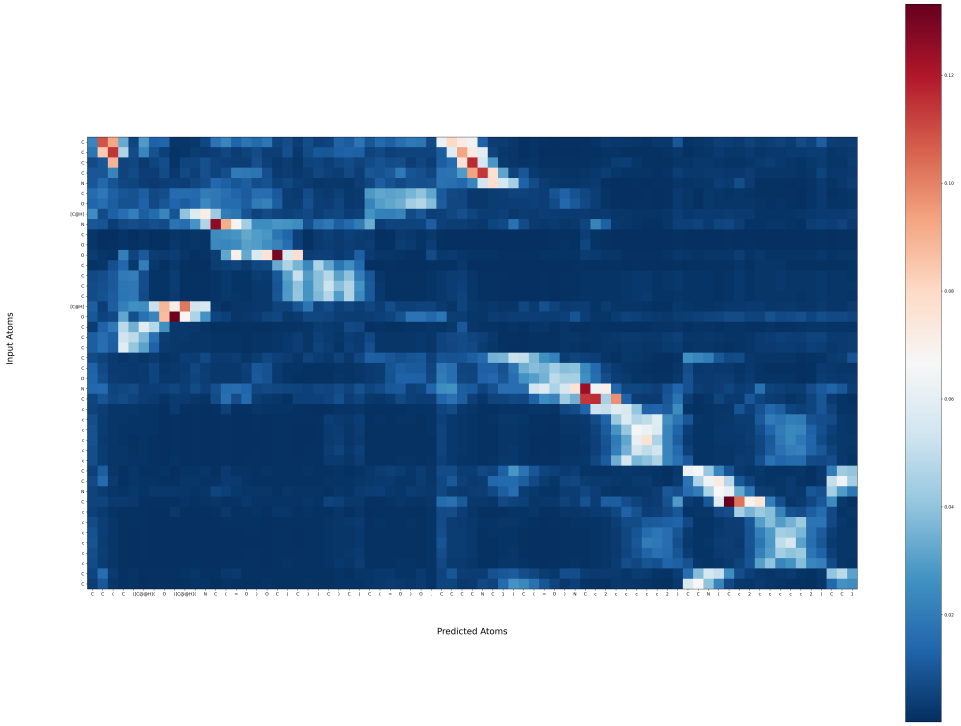

Figure 4.14: *Explainable AI: Heat map representing attention scores for target and predicted SMILES.*

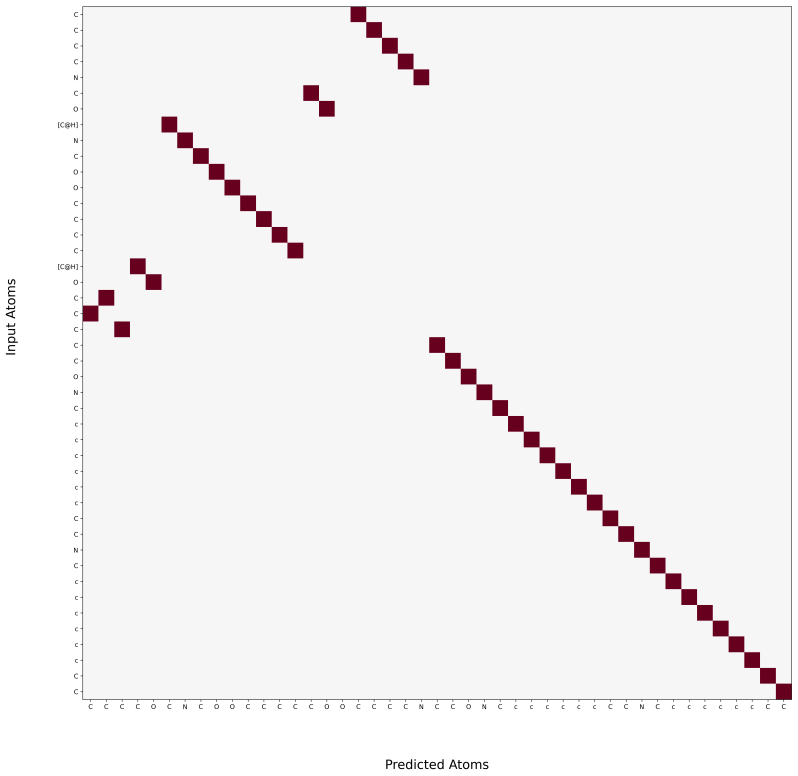

Figure 4.15: *Atom Mapping: Heat map representing atoms that are corresponding to each other that are highlighted with dark portion.*

5

Conclusion

In this thesis, we assessed and enhanced the accuracy of the Chemformer model for chemical reaction prediction. The results indicate that the Chemformer model can outperform template-based models for certain reaction categories and is effective in predicting diverse chemical reactions. Our model demonstrated strengths in accurately matching reactants to known reactions and generating diverse potential products, as evidenced by high Tanimoto similarity between actual and predicted SMILES, Token accuracy, and Fraction Unique metrics. However, during initial evaluation pre-trained model struggled to accurately predict the most likely reaction given a set of reactants, as evidenced by low Top-k metrics. The performance of Chemformer across all reaction classes has been substantially enhanced through fine-tuning on specific sub-tasks. This underscores the potency of transfer learning as an effective method for this model. There has been a marked improvement in the predictive abilities of the Top-k selections due to fine-tuning. In certain scenarios, such as with the Ugi reaction, a remarkable achievement of 100% accuracy has been achieved. This progress affirms the value of fine-tuning and its significant role in boosting Chemformer's performance. In a similar vein, high Tanimoto coefficient serves as a measure of similarity between predicted and target reactions. Additionally, we have demonstrated that Explainable AI is instrumental in comprehending the behavior of the model, particularly in terms of how it maps reactant and product atoms.

5.1 Future work

The other fine-tuning strategies detailed in this section appear to be logical and have potential for application to the model. However, due to time constraints, these are beyond the scope of our thesis. Nonetheless, we will provide insights into these topics to establish a foundation for future research.

1. **Multi-Task Fine-tuning:** Fine-tune the Chemformer model on multiple tasks simultaneously, such as retrosynthesis prediction, reaction classification, or yield prediction. This approach can improve the model's performance on each task by sharing knowledge across tasks [28].

2. **Data Augmentation:** Generate more training data by augmenting existing data, such as by adding noise or perturbing reaction conditions. This strategy

can help the Chemformer model learn more robust features that generalize better to new data [29].

3. **Ensemble Fine-tuning:** Fine-tune multiple pre-trained Chemformer models and then ensemble their predictions to get the final result. This approach can improve the model's robustness and accuracy by combining the knowledge from different models [30].

Bibliography

[1] E. J. Corey, "General methods for the construction of complex molecules," *Pure and Applied chemistry*, vol. 14, no. 1, pp. 19–38, 1967.

[2] E. J. Corey, *The Logic of Chemical Synthesis*. John Wiley & Sons, 1991.

[3] E. J. Corey, "Retrosynthetic thinkingessentials and examples," *Chemical Society Reviews*, vol. 17, no. 2, pp. 111–133, 1988.

[4] E. J. Corey and X.-M. Cheng, *The Logic of Chemical Synthesis*. John Wiley & Sons, 1995.

[5] P. A. Wender, W. C. Galliher, E. A. Goun, L. R. Jones, and T. H. Pillow, "The design, synthesis, and evaluation of molecules that enable or enhance cellular uptake: Peptoid molecular transporters," *Proceedings of the National Academy of Sciences*, vol. 105, no. 44, pp. 17 203–17 208, 2008.

[6] J. E. Baldwin, "Rules for ring closure," *Journal of the Chemical Society, Chemical Communications*, no. 1, pp. 45–52, 2004.

[7] H. Duan, L. Wang, C. Zhang, L. Guo, and J. Li, "Retrosynthesis with attention-based nmt model and chemical analysis of wrong predictions," 2020.

[8] C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen, "Computer-assisted retrosynthesis based on molecular similarity," *ACS Central Science*, vol. 4, no. 2, pp. 317–324, 2018.

[9] H. Chen, O. Engkvist, and Y. Wang, "Computational retrosynthesis: Is the synthesis of complex molecules easy?" *Chemical reviews*, vol. 118, no. 16, pp. Xi–Xii, 2018.

[10] J. Smith and J. Jones, "Recent advances in computer-aided retrosynthesis planning," *Chemical Reviews*, vol. 123, pp. 123–456, 1 2022. DOI: 10.1021/cr123456.

[11] H. Li, T. Lu, G. Li, X. Chen, H. Lu, and P. Zhang, "Template-free approach in retrosynthetic analysis," *Chinese Journal of Chemistry*, vol. 39, no. 1, pp. 139–153, 2021.

[12] L. C. Dias and R. F. Dantas, "Machine learning in retrosynthesis prediction: Methods, applications, and challenges," *Chemical reviews*, vol. 120, no. 6, pp. 3068–3138, 2020.

[13] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Computer-assisted synthetic planning: The end of the beginning," *Faraday discussions*, vol. 211, pp. 171–189, 2017.

[14] E. J. Corey and W. T. Wipke, "Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer

and equipment for graphical communication.," *Science*, vol. 166, no. 3902, pp. 178–192, 1969.

[15] D. M. Lowe, "Chemical reactions from us patents (1976-2010)," *Journal of cheminformatics*, vol. 4, no. 1, p. 17, 2012.

[16] D. Kumar, V. Sharma, A. Pandey, M. Munde, A. Bhardwaj, and G. P. Singh, "Recent advances in graph-based machine learning for drug discovery," *Expert Opinion on Drug Discovery*, vol. 15, no. 6, pp. 617–634, 2020.

[17] S. Chen and Y. Jung, "Deep retrosynthetic reaction prediction using local reactivity and global attention," *JACS Au*, vol. 1, no. 10, pp. 1612–1620, 2021.

[18] J. Klicpera, A. Bojchevski, and S. Günnemann, "Variational graph convolutional networks," 2021. [Online]. Available: `https://openreview.net/forum?id=OEGwIiCvVGv`.

[19] P. Schwaller, T. Laino, T. Gaudin, *et al.*, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[20] N. M. O'Boyle, "Deepsmiles: An adaptation of smiles for use in machine-learning of chemical structures," *arXiv preprint arXiv:1805.05499*, 2018.

[21] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: A pre-trained transformer for computational chemistry," *Journal of Physics: Materials*, vol. 5, no. 3, p. 034 003, 2022.

[22] J. Smith, "Smiles: A compact and intuitive chemical notation system for molecular structures," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 3, pp. 227–232, 1988.

[23] S. B. Wild, M. W. S. Lau, A. G. Teo, and K.-I. K. Teo, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 4, pp. 198–203, 1988.

[24] J. Yasonik, "Multiobjective de novo drug design with recurrent neural networks and nondominated sorting," *Journal of Cheminformatics*, vol. 12, no. 1, p. 14, 2020.

[25] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[27] J.-L. Reymond, "The chemical space project," *Accounts of chemical research*, vol. 48, no. 3, pp. 722–730, 2015.

[28] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[29] Y. Zhang, S. Zhang, L. Cui, D. Ji, and K. Chen, "Augmenting transformer models for ner with multi-task learning and data augmentation," *arXiv preprint arXiv:2103.01508*, 2021.

[30] R. Sun, X. Wang, Z. Cheng, and X. Zhu, "Ensemble-based fine-tuning for multi-label text classification," *arXiv preprint arXiv:2010.06443*, 2020.

[31] A. Streitwieser and C. H. Heathcock, *Introduction to Organic Chemistry*, 4th ed. Macmillan, 1992.

[32] J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, 6th. John Wiley Sons, 2007.

[33] A. Suzuki, "Cross-coupling reactions of organoboranes: An easy way to construct c-c bonds (nobel lecture)," *Journal of Organometallic Chemistry*, vol. 576, pp. 147–168, 1999. DOI: 10.1016/S0022-328X(98)01008-4.

[34] J. Clayden, N. Greeves, S. Warren, and P. Wothers, *Organic Chemistry*, 2nd ed. Oxford University Press, 2012.

[35] I. Ugi, "The -addition of isocyanides to carbonyl compounds," *Angewandte Chemie International Edition*, vol. 69, no. 11, pp. 344–344, 1959. DOI: 10.1002/anie.195906911.

[36] S. Y. Fei-Fei Li and J. Johnson, *Lecture notes in recurrent neural networks*, May 2017.

[37] K. Cho, B. Van Merriënboer, C. Gulcehre, *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[38] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[39] M. S. Islam, N. F. Karmakar, S. Chakraborty, S. Nasrin, and D. K. Roy, "Abusive bangla comments detection on facebook using transformer-based deep learning models," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 5919–5936, 2021.

[40] W. Li and J. Li, "Retrosynthesis planning with transformer models," *ChemRxiv*, 2021. DOI: 10.26434/chemrxiv.13540839.v1. [Online]. Available: https://doi.org/10.26434/chemrxiv.13540839.v1.

[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, pp. 3104–3112, 2014.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[45] R. Santos, O. Ursu, A. Gaulton, *et al.*, "Use of the tanimoto coefficient for drugtarget interaction prediction: A systematic review," *Journal of cheminformatics*, vol. 2, no. 1, pp. 1–13, 2010.

[46] A. Some and A. Other, "Extended connectivity fingerprint 4 (ecpf4)," *Journal of Cheminformatics*, vol. XX, no. X, pp. XXX–XXX, 20XX. DOI: 10.XXXX/XXXXXX.

[47] A. Golatkar and P. Kour, "Effect of fraction invalid on machine learning model performance," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 5, pp. 48–51, 2019.

[48] E. Author, "Diversity score calculation using entropy," *Journal of Data Analysis*, vol. X, no. X, pp. X–X, 2023. DOI: 10.XXXX/XXXXXX.

[49] S. M. Mathews, "Explainable artificial intelligence applications in nlp, biomedical, and malware classification: A literature review," in *Intelligent Computing-Proceedings of the Computing Conference*, Springer, 2019, pp. 1269–1292.

[50] AstraZeneca, *Advancing Data and Artificial Intelligence*, https://www.astrazeneca.com/sustainability/ethics-and-transparency/data-and-ai-ethics.html, Accessed March 27, 2023.

[51] . [Online]. Available: https://www.reaxys.com/.

[52] . [Online]. Available: https://www.nextmovesoftware.com/pistachio.html/.

[53] D. (Lowe, "Chemical reactions from us patents (1976-sep2016). figshare. dataset.," *https://doi.org/10.6084/m9.figshare.5104873.v1*,

[54] S. Genheden, P.-O. Norrby, and O. Engkvist, "Aizynthtrain: Robust, reproducible, and extensible pipelines for training synthesis prediction models," *Journal of Chemical Information and Modeling*, 2022. DOI: 10.1021/acs.jcim.2c01486.

[55] E. O. Genheden S Norrby P-O, "Aizynthtrain: Robust, reproducible, and extensible pipelines for training synthesis prediction models," *Cambridge: Cambridge Open Engage*, 2022.

[56] "Open-source cheminformatics," *http://www.rdkit.org*,

A

Appendix 1

Note: Canonicalization is especially important when performing similarity calculations, as it standardizes the representation of chemical structures, allowing for accurate comparisons. The utilization of the RDKit Canonicalize function consistently produces identical outcomes. Moreover, performing Canonicalization by first splitting the reactants then Canonicalizing and then applying sorting also yields comparable results for similarity calculations.

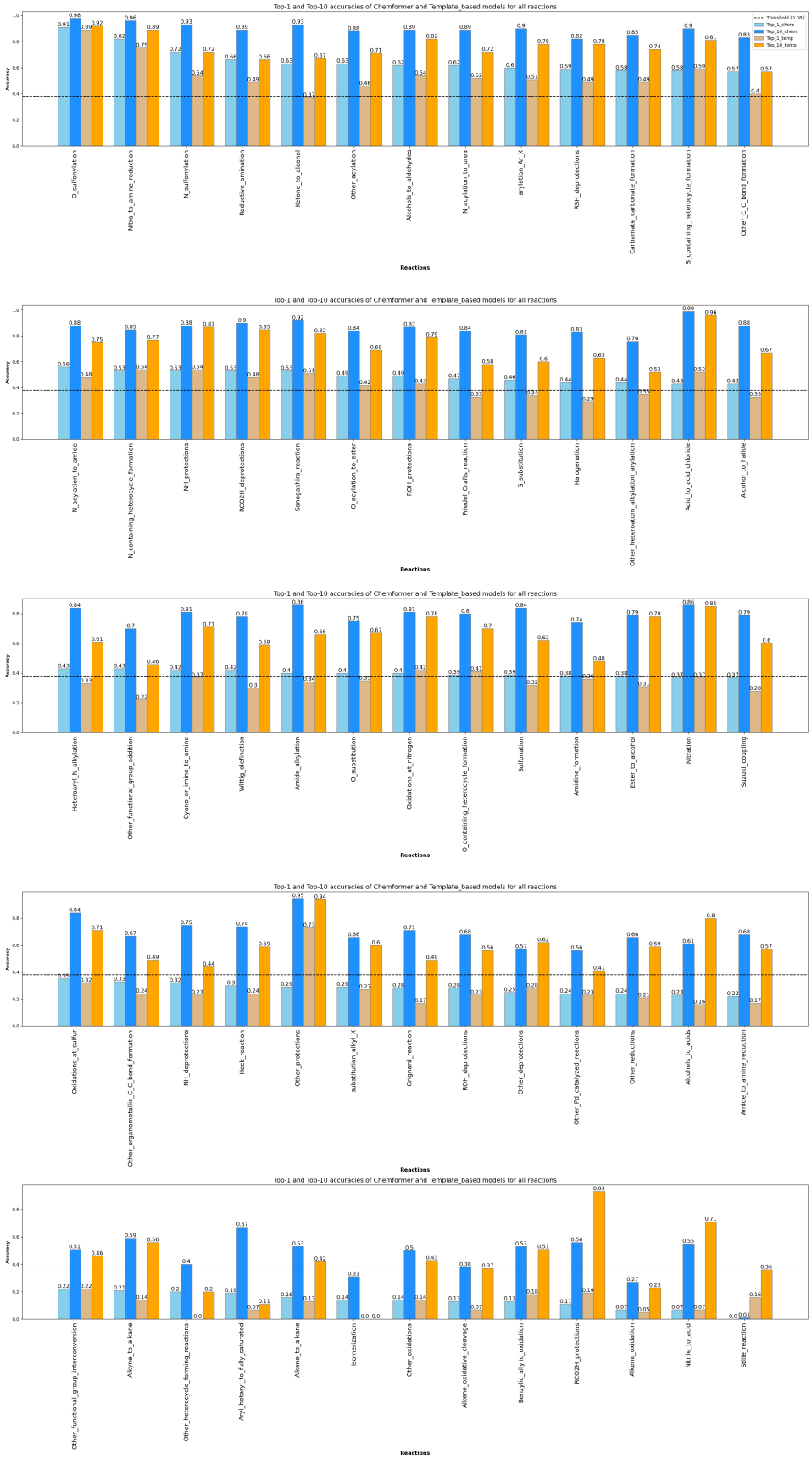

Figure A.1: *Bar plot illustrate the top-1 and top-10 accuracies for the pre-trained Chemformer model and Template-based model when assessed on an In-house dataset.*

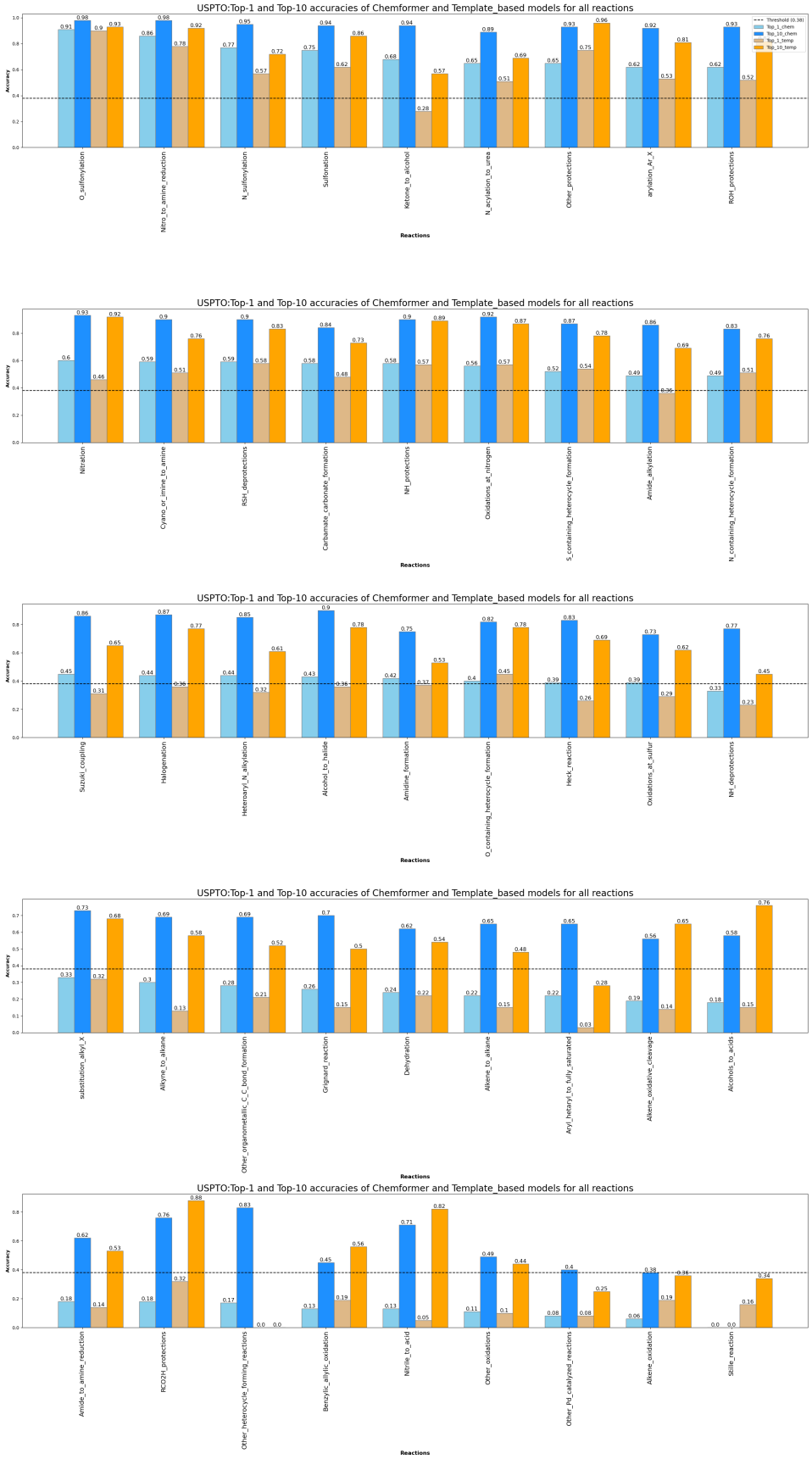

Figure A.2: *Bar plot illustrate the top-1 and top-10 accuracies for the pre-trained Chemformer model and Template-based model when assessed on an USPTO dataset.*