ChemTransformer: A Sequence-to-Sequence Model for Organic Chemical Reaction Prediction

Mingyang Yao **Zhaoxiang Feng** **Peiyuan Sun** **Yushu Wang** **Zhiting Hu**
UC San Diego
{m5yao, zh004, p3sun, yuw170, zhh019}@ucsd.edu

Abstract

Accurate chemical reaction prediction is a fundamental task in computational chemistry with profound implications for drug discovery, material synthesis, and industrial processes. Traditional rule-based approaches often lack generalizability across diverse chemical domains, while emerging machine learning techniques show promise in capturing the intricate patterns of chemical reactions. In this study, we introduce **ChemTransformer**, a sequence-to-sequence model leveraging transformer-based architectures to predict reaction products from reactants, catalysts, and experimental conditions. By incorporating advanced tokenization strategies such as SMILES and SELFIES, as well as novel compound-based relative positional encodings, ChemTransformer effectively models the complex dependencies between reaction components. Using the Open Reaction Dataset (ORD), we evaluate multiple configurations of the model, including varying representations and embedding strategies. Our results demonstrate the superior performance of traditional positional encoding over compound-level positional encoding, highlighting the importance of preserving fine-grained spatial information within molecular representations. Despite challenges in training time and model convergence, ChemTransformer achieves competitive sequence-level accuracy, offering a scalable solution for large-scale chemical reaction prediction. These findings underline the potential of transformer architectures to advance the field of computational chemistry, paving the way for more efficient and accurate reaction modeling.

1 Introduction

Chemical reaction prediction has emerged as a significant and challenging task in computational chemistry, offering the potential to accelerate drug discovery, material synthesis, and chemical engineering. Predicting reaction outcomes prior to conducting the experiment in high precision can greatly reduce time and resource consumption for experiments in chemistry research. Traditional methods for reaction prediction often rely on empirical rules or heuristic-based approaches, which can lack generalization across diverse chemical spaces and conditions. Machine learning, particularly deep learning, has recently demonstrated promise in this domain, providing data-driven alternatives that can capture complex patterns in chemical reactions.

Among the deep learning models, transformer-based architectures have gained attention due to their powerful sequence modeling capabilities and potential to capture long-term dependency. Originally developed for natural language processing tasks, transformers excel at handling sequential data by leveraging self-attention mechanisms, which allow them to learn complicated dependencies between components in a sequence. This feature makes transformers well-suited for chemical reaction prediction, where understanding the relationships among reactants, elements composed of them, structure of compounds, and experimental conditions is crucial for accurately predicting the product.

In this research, we propose **ChemTransfromer**, a transformer encoder-decoder model that predicts the products of chemical reactions based on tokenized reaction components and experimental condi-

Preprint. Under review.

tions. Unlike rule-based systems or algorithms, our approach models each reaction as a sequence, enabling it to capture the nuanced interplay between the reaction inputs—such as reactants, catalysts, and solvents—and conditions like temperature. Also, compared with previous attempts in chemical reaction modeling using the sequence model, we also incorporate the amount of each compound involved in the reaction to enable the model to understand more accurate chemistry patterns and robust to minor amount variation in real-world scenarios. This approach allows the model to generalize across diverse reaction types, making it adaptable to various chemical domains.

This work contributes to the ongoing efforts in reaction prediction by combining modern deep learning techniques with domain-specific chemical information, offering insights that could push the boundaries of automated reaction prediction and open new avenues in the computational chemistry landscape.

2 Background & Related Work

2.1 Background

Accurate reaction-outcome prediction reduces trial-and-error in synthesis and enables faster route design in pharma and materials. Early expert systems relied on hand-coded rules and templates, which fail to generalize across reaction families and conditions. Data-driven methods replaced brittle heuristics with statistical models that learn patterns directly from reaction corpora. Among these, transformer sequence-to-sequence models have become the standard due to strong long-range dependency modeling and flexible conditioning on reagents and context. The Open Reaction Dataset (ORD) further shifts the problem from single-string encodings toward structured records that include roles and quantities, allowing models to exploit stoichiometry and conditions.

2.2 Chemical Compound Representation

Accurate representation of chemical compounds is fundamental to chemical reaction prediction, as it determines how effectively models can learn the relationships between reaction components. Traditional methods often encode compounds using simplified molecular-input line-entry systems (SMILES), which represent molecules as linear strings of atomic symbols and bonds (Schwaller et al., 2019). Another approach, the graph-based model, encodes molecules as graphs, where atoms are represented as nodes and bonds as edges. Graph-based models capture spatial and structural relationships that linear representations like SMILES may overlook (Gilmer et al., 2017; Yuan et al., 2023).

2.3 Sequence Model for Chemical Reaction

Sequence-to-sequence models, originally designed for natural language processing (NLP), have been successfully applied to chemical reaction prediction. These models treat chemical reactions as sequential data, encoding reactants, and experimental conditions as inputs and products as outputs. Sequence models excel at capturing dependencies between input components, such as reactants and catalysts, and the resulting products, even in highly complex chemical systems (Schwaller et al., 2019).

Recent advancements have further refined sequence models by addressing limitations such as invalid molecule generation and coherence in sequential prediction tasks. For instance, incorporating SELFIES representations reduces the risk of malformed outputs while maintaining robust molecular integrity (Krenn et al., 2020). Additionally, work on coherence-enhanced sequence modeling has explored hybrid strategies that combine attention mechanisms with domain-specific constraints to improve predictive performance (Karpov et al., 2019).

The Molecular Transformer, a seminal sequence-to-sequence model, utilizes SMILES representations to encode chemical reactions into tokenized sequences. By leveraging self-attention mechanisms, the transformer architecture captures long-range dependencies between reaction components, enabling accurate predictions across diverse chemical spaces (Schwaller et al., 2019).

Innovative methods also explore hybrid graph-sequence models to encode structural information while retaining sequence flexibility. For instance, Yuan et al. (2023) propose multi-agent systems where sequence models interact with graph-based reasoning agents to achieve better performance in tasks like reaction prediction. These approaches highlight the potential for combining sequence modeling with other representational techniques to improve chemical prediction tasks.

2.4 Graph-Based Model for Chemical Reaction

Graph-based models provide an alternative approach to chemical reaction prediction by explicitly representing molecules as graphs. In these models, atoms serve as nodes and bonds as edges, allowing for the direct encoding of spatial and structural relationships. Such representations are particularly effective for capturing local and global patterns in molecular structures, which are often difficult to extract using sequence-based methods alone (Gilmer et al., 2017).

One of the most prominent methods in this domain is the Message Passing Neural Network (MPNN), which propagates information between nodes (atoms) in a molecule's graph to learn embeddings for each atom and bond. These embeddings capture molecular features that are crucial for understanding chemical reactivity, making graph-based models highly effective for reaction prediction (Gilmer et al., 2017). However, these models often face challenges in encoding non-structural information, such as reaction conditions and quantities.

Recent advancements have also introduced positional encodings tailored to chemical graphs, ensuring that spatial relationships within molecules are preserved. For example, the use of relative positional encodings, as explored in Schwaller et al. (2019), improves the model's ability to focus on relevant structural patterns while mitigating biases introduced by arbitrary token orders in sequences.

3 Methodology

3.1 Dataset

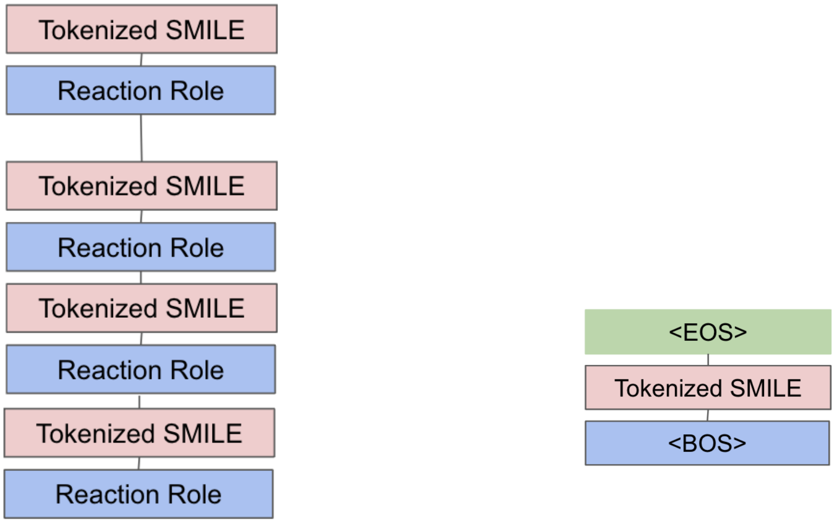
We utilize the **Open Reaction Dataset (ORD)** as the primary resource for training our sequence-to-sequence model. This extensive dataset contains millions of chemical reactions, providing detailed records of each compound involved in a reaction, their quantities (e.g., in grams or moles), and various experimental conditions, such as temperature. This comprehensive dataset enables the model to capture intricate patterns and dependencies in chemical reactions, ultimately improving prediction accuracy.

Unlike conventional datasets, where reactions are typically represented as single SMILES strings, the ORD dataset adopts a more structured format, recording each compound and its associated properties (e.g., quantity and role) separately. This granularity mirrors real-world scenarios where reaction components and conditions are independently documented, offering a more precise and flexible data representation. This structure aligns with our objective of developing a robust prediction model that can handle complex reaction scenarios.

3.2 Representation

We propose a **Structured Reaction Representation (SRR)** to encode chemical reactions in the ORD dataset. SRR divides each reaction into two distinct sequences: the input sequence and the output sequence.

The **input sequence** comprises all compounds involved in the reaction, their respective quantities, and the temperature condition. Special tokens denote reaction roles (e.g., reactant, solvent, catalyst), units of measurement, and temperature units. This structured sequence ensures precise encoding of all reaction components. The format is as follows:



(a) SSR Representation of Input Sequence

(b) SSR Representation of Target Sequence

A trivial example of an input sequence is:

```
[Reactant] [C] [(O)] [C]
[Catalyst] [C] [C][(=O)] [O]
```

The **output sequence** consists of the tokenized SMILES representation of the reaction product. A `<BOS>` token (beginning of sequence) and `<EOS>` token (end of sequence) delimit the sequence. An example output sequence is:

```
<BOS> [C] [C] [N+] [(] [C] [)] [(] [C] [)] <EOS>
```

To enhance robustness, we also experiment with **SELFIES (Self-Referencing Embedded Strings)**. SELFIES reduces the likelihood of invalid molecular structures and preserves chemical integrity, addressing SMILES limitations. Both representations employ pair encoding for atom-level tokenization, ensuring accurate context for each chemical unit.

Reactions exceeding a predefined sequence length are excluded to maintain meaningful predictions.

3.3 Model Architecture

Our model adopts a transformer architecture with 6 layers each for the encoder and decoder. The model uses an embedding dimension of 512, 8 attention heads, and feed-forward layers with 2048 neurons. Optimization is performed using the Adam optimizer with a Cosine Annealing learning rate scheduler. The initial learning rate is set to 1×10^{-4}, decreasing to 1×10^{-5}. Default PyTorch settings are used for Adam optimizer parameters, excluding the learning rate. The loss function is Cross-Entropy Loss, with `<PAD>` tokens ignored during computation. Finally, we clip the gradient and set the maximum norm of the gradient vector to be 1.0.

To align the model with chemical properties, we propose a **compound-based relative positional encoding** for input sequences. Unlike conventional token-based encodings, which assign unique positions to each token, this approach assigns the same positional bias to all tokens within a compound or condition. This mitigates the distortion caused by token-level positional encodings and emphasizes meaningful relationships among compounds.

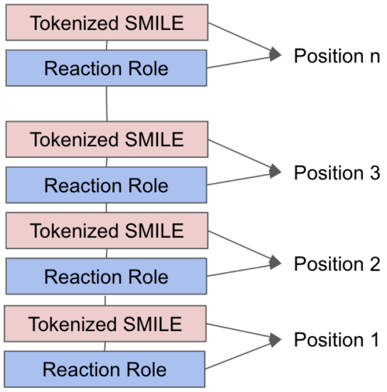


Figure 2: Compound-Based Relative Positional Encoding Visualization

An example of compound-based positional encoding is:

```
Position 1: [Reactant] [C] [(O)] [C]
Position 2: [Catalyst] [C] [C][(=O)] [O]
```

Given an embedding dimension d and a dimension index k, the relative positional encoding $P(i, j)$ for compounds c_i and c_j is defined as:

$$P_{(i,j,2k)} = \sin\left(\frac{|c_i - c_j|}{10000^{\frac{2k}{d}}}\right), \quad P_{(i,j,2k+1)} = \cos\left(\frac{|c_i - c_j|}{10000^{\frac{2k}{d}}}\right)$$

By grouping tokens by compound, this approach preserves the logical structure of chemical reactions, focusing on relationships between entire compounds. We hypothesize that this encoding scheme enhances model performance by closely aligning positional information with the chemistry of reaction components.

4 Experimentations

4.1 Experiment Outline

Due to time constraints, we explore a model with 6 layers of transformer encoders and decoders with 512 embedding dimensions and 8 heads of attention as the base model. We explore different chemical compound representations, SMILE, and SELFIES as well as different relative positional embedding, compound-level, and traditional token-level PE. To balance the hardware memory and amount of data, we set the max input length of reactions to 200 and the target length to 100, which preserves around 95% of total data in the SMILE representation and 97.4% of total data in the SELFIES representation.

In all experiments, 92.5% of data are used to train the model and the remaining are used as the test set. All models are trained on one A100 GPU with 40GB memory with a batch size of 350 and each epoch takes around 1 hour to complete. We report the loss, token-level accuracy, and sequence-level accuracy for both the training and validation sets in all experiments. A sequence prediction is considered correct when its prediction starts from BOS to EOS and is all correct and padding tokens are excluded in both loss and accuracy calculation.

4.2 Experiment 1

The SMILE representation with compound-level relative positional embedding takes 32 epochs to converge. The following are the statistics of the model.

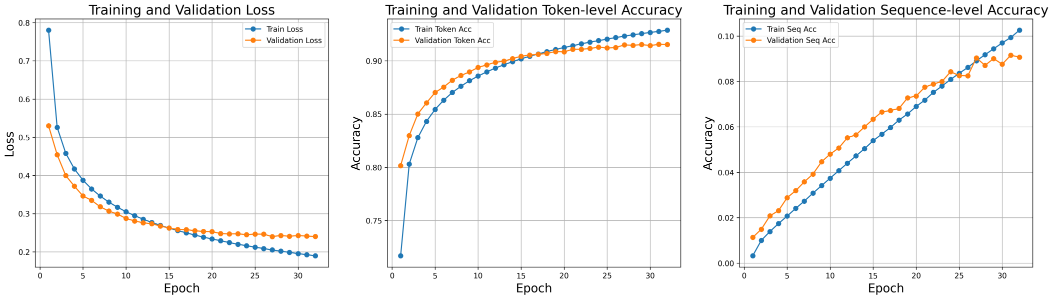


Figure 3: Compound-Level PE and SMILE Representation

4.3 Experiment 2

The SELFIES representation with compound-level relative positional embedding takes 40 epochs to converge with slightly worse performance than the SMILE representation.

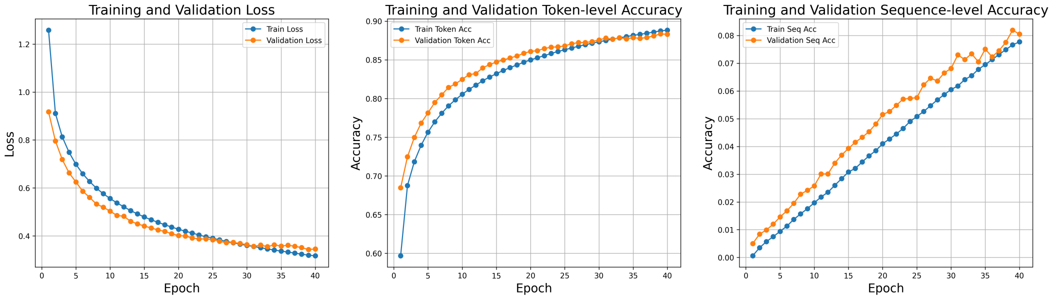


Figure 4: Compound-Level PE and SELFIES Representation

4.4 Experiment 3

We then attempt SELFIES with traditional relative positional embedding. This approach takes 46 epochs to converge.

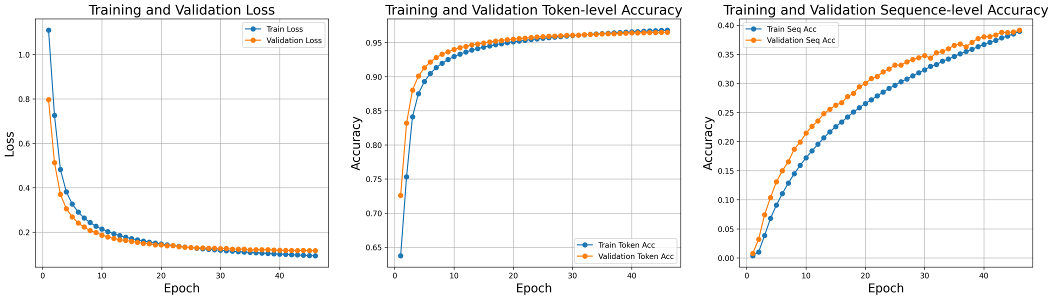


Figure 5: Token-Level PE and SELFIES Representation

5 Conclusion

5.1 Discussion of Results

Contrary to our hypothesis that one compound should have the same positional information, at least on SELFIES representation, within one compound, attention weight difference still exists. This conclusion is supported by poor results on two compound-level positional embedding models (less than 10% accuracy in sequence for both training and validation).

The nature of SELFIES representation likely causes it for its ability to preserve compound spatial features instead of solely a compound annotation. Due to the long time taken for model training, the baseline model using SMILE representation and traditional positional embedding has not run yet. However, based on current results, it's extremely likely that traditional positional embedding would outperform compound-level positional embedding. The best model, using SELFIES representation and traditional positional embedding, reaches 96.5% accuracy by tokens and 39.2% accuracy by sequence. A100 GPU with 40GB memory takes less than 1 minute to make inferences on around 80000 chemical reactions and the model has around 45 million parameters, which is not hardware-consuming in the inference stage.

5.2 Future Work

To investigate the robustness and power of transformer architecture, a large-scale model with 8 layers of encoders and decoders might be worth trying, and using a smaller model fits its behavior to distill the large network. Also, to make the model more robust, augmentation such as messing compound's order in the sequence may be necessary.

Graph-based representation is also a promising direction for chemical reaction prediction. However, considering the fact that chemical compounds are spatial and graph-like data but a reaction resembles a sequence, combining the graph-based and sequence-based representation and model architecture might be necessary to reach better performance.

References

[1] Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. S., & Hernández-Lobato, J. M. (2019) A model to search for synthesizable molecules. In *Advances in Neural Information Processing Systems*, 32, pp. 7937–7949.

[2] Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., & Barzilay, R. (2019) A graph-convolutional neural network model for the prediction of chemical reactivity. In *Advances in Neural Information Processing Systems*, 32, pp. 9373–9383.

[3] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017) Message passing neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 70, pp. 1263–1272.

[4] Jin, W., Coley, C. W., Barzilay, R., & Jaakkola, T. (2017) Predicting organic reaction outcomes with Weisfeiler-Lehman network. In *Advances in Neural Information Processing Systems*, 30, pp. 2607–2616.

[5] Karpov, P., Godin, G., & Tetko, I. V. (2019) A reinforcement learning approach to improve chemical reaction prediction. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5138–5146.

[6] Kayala, M. A., & Baldi, P. (2011) A machine learning approach to predict chemical reactions. In *Advances in Neural Information Processing Systems*, 24, pp. 747–755.

[7] Krenn, M., Häse, F., Nigam, A., Friederich, P., & Aspuru-Guzik, A. (2020) Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. In *Nature Machine Intelligence*, 2(10), pp. 487–496.

[8] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019) Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. In *ACS Central Science*, 5(9), pp. 1572–1583.

[9] Segler, M. H. S., Preuss, M., & Waller, M. P. (2018) Planning chemical syntheses with deep neural networks and symbolic AI. In *Nature*, 555(7698), pp. 604–610.

[10] Yuan, Y., Jiang, Z., Zhu, X., Wang, X., Zhang, P., & Wang, S. (2023) Multi-agent framework for chemical reaction prediction. In *arXiv preprint arXiv:2310.13590*.