ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

M2Echem: A multilevel dual encoder-based model for predicting organic chemistry reactions

Linxing Zhu[1†], Jing Wang[1†], Jiashuang Huang[1], Yifan Jiang[2*], and Shu Jiang[1*]

[1]Department of Computer Science, School of Artificial Intelligence and Computer Science, Nantong University, Nantong, Jiangsu, China

[2]Department of Electrical and Computer Engineering, State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, Macau, China

Abstract

Chemical reaction prediction is a vital application of artificial intelligence. While Transformer models are widely used for this task, they often overlook deeper-level semantic information. In addition, the traditional Transformer model suffers from a decline in prediction performance and shows poor generalization when faced with different representations of the same molecule. To address these challenges, we propose a dual encoder-based reaction prediction method tailored for multilevel organic chemistry. Our approach began with the introduction of synergistic dual-encoder architecture: The atomic encoder focused on inter-atomic attention weights. In contrast, the molecular encoder employed a molecular maximum dimension reduction algorithm to identify key chemical features. We then performed multilevel feature fusion by combining the outputs from both the atomic and molecular encoders. Finally, we applied an optimized contrast loss to enhance the model's robustness. The results indicated that this method outperformed existing models across all four datasets, significantly improving generalization performance and contributing to advancements in artificial intelligence-driven drug development and research.

Keywords: Forward reaction prediction; Multilevel feature fusion; Machine learning; Simplified molecular input line entry system code; Transformer

†These authors contributed equally to this work.

***Corresponding authors:**
Yifan Jiang
(yc27495@umac.mo)
Shu Jiang
(jshmjs45@ntu.edu.cn)

Citation: Zhu L, Wang J, Huang J, Jiang Y, Jiang S. M2Echem: A multilevel dual encoder-based model for predicting organic chemistry reactions. *Artif Intell Health*. 2026;3(1):88-103. doi: 10.36922/AIH025260058

Received: June 26, 2025

1st revised: July 15, 2025

2nd revised: July 21, 2025

Accepted: July 23, 2025

Published online: August 5, 2025

1. Introduction

As early as the 1960s, organic chemists attempted to use modern computers for planning synthetic pathways[1]. However, these early efforts relied heavily on reaction templates[1-3]—manual rules developed by experts—or other forms of chemical knowledge.[4-6] Such approaches struggled to address the full complexity of organic chemistry prediction problems.

In the 1980s, the Simplified Molecular Input Line Entry System (SMILES)[7] was developed to represent the structure of chemical compounds, allowing computers to manage chemical reactions in a linear format. Consequently, chemical reactions could be

regarded as a specific language and modeled symbolically for computer-assisted reaction prediction. Many text processing methods have been utilized in chemical reaction prediction, including the sequence-to-sequence model,[8] with the Transformer model[9] gaining widespread use in neural machine translation.

Traditional chemical reaction prediction methods use SMILES representations to segment chemical formulas at the atomic level and apply the Transformer model.[10] However, these approaches can only capture attention weights between individual atoms, thereby overlooking important intermolecular interactions. Moreover, the Transformer model struggles with handling long sequences,[11,12] which frequently occur in chemical reaction text.

We analyzed the vocabulary size and sequence length of natural language and chemical reaction texts, highlighting their significant differences (Table 1). The units in the table are based on words. Chemical reaction texts exhibit longer sequences than natural language. While attention is calculated between any two tokens, long sequences can lead to insufficient encoding. In natural language processing, missing parts of text often have minimal effect on comprehension. In contrast, missing or altered structures in a chemical molecular formula may severely compromise the overall representation.

In this study, we propose a dual-encoder method for multilevel organic chemical reaction prediction—M2Echem—to address the problems of limited attention to molecular interactions, insufficient encoding, and poor generalization ability of the Transformer model. First, the Transformer model is enhanced using two encoders to simultaneously construct representations of atoms and molecules in the input chemical formula. The atomic encoder utilizes the self-attention mechanism of the Transformer model to capture the interrelationships between atoms. In contrast, the molecular encoder employs the molecular maximum dimension reduction algorithm to generate molecular embeddings. The outputs of the atomic and molecular encoders undergo multilevel feature fusion, and the fused feature representations are input into the decoder. Finally, the optimized loss function enhances the model's understanding of different SMILES representations.

Table 1. Comparison between natural language and chemical reaction text

Dataset	Corpus size	Vocabulary size	Length
WMT17-en[a]	4,004,240	40,394	28.44
WMT17-en[b]	1,104,577	35,483	28.43
USPTO (original)[c]	479,035	448	375.00
Single reactant	12,581	175	87.51
Multiple reactants	1,091,996	574	114.17

Notes: [a]WMIT17-en refers to the English corpus from the Second Conference on Machine Translation, [b]WMT17-en (random) refers to randomly selected sentences from WMT17-en for a fair comparison, and [c]USPTO (canonical) refers to the canonical version of the tokenized chemical dataset USPTO-Jin.
Abbreviation: USPTO: United States Patent and Trademark Office.

The main contributions of this study are as follows:

(i) The proposed model utilizes molecular and atomic encoders to extract information at both molecular and atomic levels. Key features in molecules are identified using a molecular maximum dimension reduction algorithm. Compared to traditional methods, this approach effectively captures hierarchical correlations between atoms and molecules, thereby improving the encoding capabilities of the Transformer model for processing long-sequence data.

(ii) A fusion layer with automatic weight updating is proposed for multilevel feature fusion. This layer uses linear concatenation to fuse the outputs of the atomic and molecular encoders. It also incorporates batch normalization and a softmax activation function to obtain weight parameters. Unlike simple concatenation, this design enables deeper extraction of complementary information and promotes the integration of multiple features.

(iii) The proposed model generates augmented SMILES representations with implicit positive and negative labels for contrastive learning. Dimension compression is then applied to retain all the essential features required for this learning process. The contrastive learning loss is combined with the cross-entropy loss function, enabling automatic updates of the weight parameters.

2. Related work

This section includes three parts, focusing on essential methods relevant to this study, such as chemical product prediction methods, multilevel feature fusion approaches, and comparative learning.

2.1. Chemical product prediction methods

Large language models effectively capture long-distance dependencies through a self-attention mechanism, enabling stronger feature representation and parallel computing capabilities.[13-15] These models have shown excellent performance in predicting organic chemistry outcomes.[16] However, their performance is influenced by both the quality and quantity of training data. Therefore, extracting more accurate and detailed chemical information is essential for improving model predictions,

particularly when computational resources and chemical datasets are limited.

Researchers have developed various methods[17,18] to enhance or extend the SMILES language. Lo *et al.*[19] proposed SELFIES disambiguation to obtain valid SMILES molecules. Ucak *et al.*[20] proposed atom-in-SMILES disambiguation, where the model learns chemical information within the radius of covalent bonding between atoms to reduce molecular labeling duplication. While these methods provide in-depth chemical insights and enhance the accuracy of chemical reaction predictions, the large word lists used during model training often overlook the overall information contained within the molecule.

In contrast, models can learn more chemical information through pre-training and fine-tuning phases or multi-task training. For example, Wu *et al.*[21] proposed the knowledge-enabled language representation model, which acquires chemical information through atomic feature prediction, molecular feature prediction, and comparison learning. Chen and Jung[22] introduced the LocalRetro model, which utilizes mechanisms of local reactivity and global attention. Liu *et al.*[23] developed the MolXPT model, which combines scientific text with SMILES representations of chemical molecules for pre-training to improve the performance. Lu and Zhang[24] created the T5Chem model, which takes advantage of mutual learning among related tasks. Although the models extract deeper information, they all employ a single encoder-decoder architecture, which makes it difficult to optimize the extraction of information on different features simultaneously.

2.2. Multilevel feature fusion approaches

Multilevel feature fusion methods create a unified representation by mapping unimodal representations into a shared semantic subspace, allowing for the integration of multimodal features.[25] Joint representations can be categorized into two types: Feature-level fusion and model-level fusion.

Feature-level fusion integrates features from different modalities into a unified representation. For instance, Ma *et al.*[26] proposed an early fusion unified encoder model, Flat-Transformer, which concatenates context and source sentence features in the same space. Zhang *et al.*[27] proposed the multi-grained Bidirectional Encoder Representations from Transformers (BERT) model, which utilizes a dual encoder to extract information from chemical SMILES sequences. The two feature matrices generated by the encoders are combined and fed into a decoder, enhancing the performance of natural language understanding tasks. However, directly fusing features from different modalities does not effectively capture complex relationships.

Model-level fusion refers to the simultaneous processing of inputs from several models. Zhu *et al.*[28] enhanced machine translation performance by integrating BERT features into the encoder and decoder layers of the neural machine translation model using the attention mechanism. This approach often requires vital computational resources and longer training times.

2.3. Comparative learning

Several studies have shown that machine learning-based natural language processing models are vulnerable to minor disturbances.[29,30] To prevent the model from being affected by synonyms and different SMILES representations, contrastive learning can help models identify semantic similarities among different SMILES sequences representing the same chemical formula.

Gao *et al.*[31] proposed the Simple Contrastive Learning of Sentence Embeddings framework based on BERT, which generates positive samples in an unsupervised manner using dropout and optimizes contrastive loss to enhance sentence embeddings. Wu *et al.*[21] proposed that the knowledge-enabled language representation model enhances the dataset by dividing positive and negative samples, which are then input into the model for contrastive learning. While these methods generate pairs of positive and negative samples that retain semantic information, the samples are typically augmented independently, and feature dimensionality reduction using token classification is commonly employed in BERT models. In contrast, Chen *et al.*[32] proposed a contrastive learning loss mechanism that obtains local and global losses through average pooling-based feature dimensionality reduction for scientific literature-related work generation tasks. Improving the relevance of generated text requires sampling negative examples from numerous non-references, as average pooling can lead to information loss.

3. Methods

This section introduces a multilevel dual encoder model—M2Echem—designed for predicting organic chemistry reactions. The M2Echem model is built upon the T5chem framework, as shown in Figure 1. Compared to T5chem, the M2Echem model incorporated three significant modifications: A feature extraction module, a multilevel feature fusion module, and a fused loss function module.

3.1. Feature extraction module

The M2Echem model employed character-level tokenization to segment reaction SMILES into individual alphabet letters, digits, or special symbols. The processed

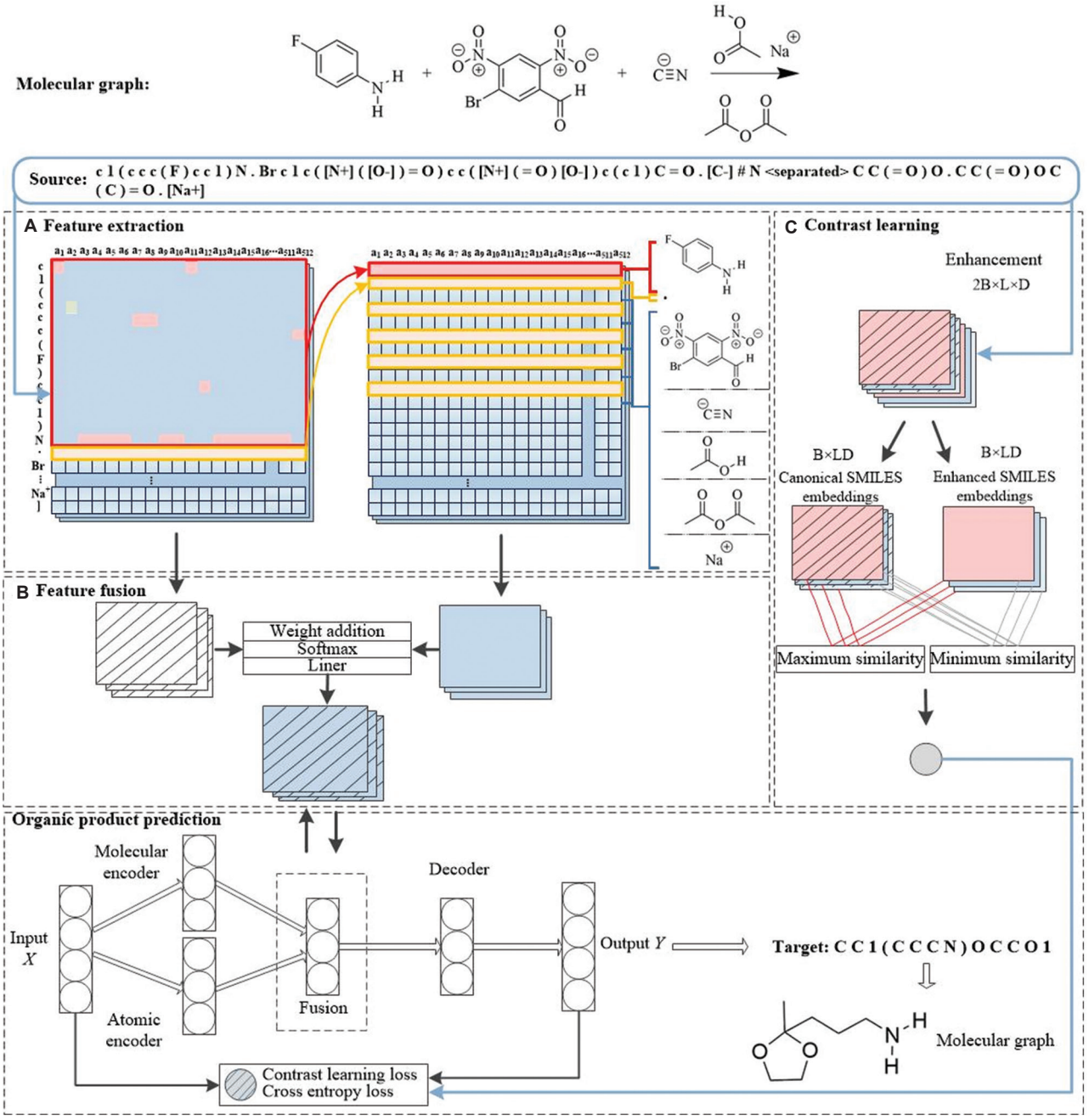

Figure 1. Model diagram of M2Echem. (A) Feature extraction: extraction of atomic-level and molecular-level features. (B) Feature fusion: Multi-level fusion of atomic and molecular characteristic information. (C) Contrastive learning: fusion of contrastive learning loss and cross-entropy loss from different representations of the same SMILES to train the model.
Abbreviation: SMILES: Simplified molecular input line entry system.

data were then input into the atomic and molecular encoders to extract features, with the two encoders not sharing weights.

Atomic encoding used the multi-head attention mechanism to project the query vector Q, the key vector K, and the value vector V h times to dimensions d_q, d_k, and d_v, respectively. The attention function was then executed in parallel to produce output values of dimension d_v. The final results were generated by concatenating these output values and projecting them again. The scaled dot-product attention was computed using the formula in Equation I:

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

Multi-head attention combined multiple scaled dot-product attention operations, as defined by Equations II and III:

$$Atom\ (H) = Concant\ (head_1,\ldots,\ head_h)W^p \quad \text{(II)}$$

$$head_i = Attention\ (HW_1^O, HW_1^K, HW_1^V) \quad \text{(III)}$$

Where H is an input representation or the hidden state in the encoder.

The method for extracting features from the molecular encoder is illustrated in Figure 1A, while the molecular maximum dimension reduction algorithm is outlined in Algorithm 1. Following Algorithm 1, feature fragments were embedded within the molecules of each chemical reaction system. Dimensional reduction was then conducted using a maximum aggregation operation alongside conditional structural parsing, enabling directional extraction and enhanced characterization of key features. The resulting point numbers and special symbols were retained to construct a molecular embedding representation. Finally, the molecular embeddings and the new filling matrix were fed into the encoder within the multi-head attention mechanism.

Specifically, the input to the model consisted of a SMILES sequence, in which different reactants were separated by a period (.) and reactants were separated from reagents using a greater-than symbol (>). For example, if the SMILES input is "CC.CCO>NCO</s>," Algorithm 1 first maps this SMILES sequence to an embedding matrix of size (1 × 11 × 256). It then identifies the positions of specific markers within the sequence: The fill position is at 1, the "." is at 2, the ">" position is at 6, and "</s>" is at 10. Using these position numbers, the embedding matrices for each reactant and reagent were obtained, resulting in matrices of sizes (2 × 256) for the first reactant, (3 × 256) for the second reactant, and (3 × 256) for the reagent. The maximum values of these features were reduced to a matrix of size (1 × 256), which encapsulated the important molecular information. Finally, the reactants, reagents, and characteristic symbols were combined into a new size embedding matrix of (6 × 256). This approach differs from traditional pooling-based methods through its integration of conditional structural parsing and a residual maximization mechanism.

3.2. Multilevel feature fusion module

The outputs from the molecular and atomic encoders were further processed through context-gating mechanisms.[33]

Algorithm 1: Molecular maximum dimension reduction algorithm

Input: Dataset *Z*, dictionary corresponding to the source dataset *n*.

Output: The embedding matrix of the final molecular-atomic combination *U'*

```
src_dict=n;
for i=0 to Z - 1 do
  Extract dot_p and sep_p according to src_dict; // dot_p represents dots, and sep_p stands for a greater-than symbol.
  Extract last_src, pad_p; // last_src is the final value in the list, and pad_p indicates whether the sentence has padding.
  if sep_p.numel() > 0 then
    if dot_p.numel() > 0 then
      v=dot_p [-1];
      for j, dot in enumerate (dot_p) do
        if dot < sep_p and dot==v then
          The embedding between dot and sep_p and sep_p and last_src is maximum;
        else if start < sep_p < dot then
          sep_p is maximum before and after embedding;
        else
          Handle embedding between start and dot;
        end
        start=dot + 1;
        Residual embedding is maximum;
      end
    else
      Handles embedding between > and last_src;
    end
  else
    if dot_p.numel() > 0 then
      for each dot in dot_p do
        Handles embedding between start and dot;
      end
      Handles embedding between start and last_src;
    End
    Finalize and store U' for the sample;
End
```

This enabled the dynamic modulation of atomic-level feature representations y_a and molecular-level context embeddings f_a, facilitating fine-grained interactions between hierarchical feature spaces. The design of the multilevel feature fusion module is illustrated in Figure 1B, with the accompanying formulas provided in Equations IV-VI:

$$L = W \times Concant\ (y_a, f_a) \quad \text{(IV)}$$

$$\lambda_a = softmax\ (L, din = -1) \quad (V)$$

$$\bar{y}_{a=}\lambda_a y_a + (1-\lambda_a) f_a \quad (IV)$$

Where W represents the parameter matrix and y_a denotes the final hidden representation of the fusion of the two encoders.

3.3. Fused loss functions module

The fused loss functions module introduces a novel combination of cross-entropy loss[34] and contrastive loss.

For the cross-entropy component, let the target labels be represented as $y = [y_1, y_2,\ldots, y_N]$, and the probability distribution predicted by the model as $Y = [Y_1, Y_2,\ldots, Y_N]$. Given that the predicted labels in the SMILES sequence span multiple categories, the multi-category cross-entropy loss was adopted. The loss is calculated using the following formula in Equations VII and VIII:

$$\bar{Y} = \ln(\sigma(Y,a)) \quad (VII)$$

$$L_{con} = -\sum_{i=1}^{N} y_i^{bi} \log(\bar{Y}^{bi}) \quad (VIII)$$

Where b is the target class of sample, $\bar{Y}$ denotes the predicted probability of the processed tensor from the i-th sample on the target label y_i, and N represents the number of samples in a batch.

Contrastive loss learning focuses on enhancing consistency among positive samples while reducing similarity among negative samples in the representation space. The contrastive learning approach used in this study is illustrated in Figure 1C. When the model's data loader retains the original samples, the sample size doubles using the molecular order exchange enhancement method. In this process, a positive sample pair consists of one standard and one enhancement sequence, both of which are labeled accordingly.

In addition, the sequence tensor comprises three dimensions: Batch size, sample length, and feature dimension. Before conducting contrastive learning, batch flattening was applied to merge the sample length and feature dimension into a single dimension while preserving the integrity of the batch dimension. The relevant formulas are given in Equations IX and X:

$$\cos(\tau_a, \tau_b) = \frac{\tau_a \cdot \tau_b}{\tau_a \tau_b} \quad (IX)$$

$$L_{contrast} = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{exp(\cos(\tau_{n,c}, \tau_{n,d}))/\zeta}{exp(cos(\tau_{n,c}, \tau_{n,d}))/\zeta + \sum_{m \epsilon D_n} exp(\cos(\tau_{n,c}, \tau_{n,m}))/\zeta} \quad (X)$$

Where:

(i) cosrepresents the cosine similarity function.
(ii) $\tau_{n,c}$ represents the embedding of the canonical SMILES for molecule n.
(iii) $\tau_{n,d}$ refers to the SMILES sequence enhanced from the canonical SMILES.
(iv) $\tau_{n,m}$ represents the embedding of the negative samples.
(v) D_n represents the set of molecules in the batch, excluding the canonical SMILES.

Figure 2 shows that a temperature parameter (ζ) of 0.05 yields the best product prediction results. Therefore, ζ was fixed at 0.05 for all experiments.

For the overall loss function, after introducing the two individual loss functions, the fused loss function L_z is defined as in Equations XI and XII:

$$W_s = \frac{E_\alpha}{E_{total}} \quad (XI)$$

$$L_z = W_s L_{con} + (1-W_s) L_{contrast} \quad (XII)$$

The value of E_a increases with the number of training epochs, while E_{total} was fixed at 500. During both the early and late stages of training, the model adaptively adjusted the weights assigned to two loss functions. In the early phase, greater emphasis was placed on learning the fundamental features captured by the L_{con} loss function. In the later phase, the focus shifted to fine-tuning the details of the $L_{contrast}$ loss function to improve the flexibility and effectiveness of the training process.

4. Experiments

4.1. Datasets

The datasets, as shown in Table 2, were derived from two different sources: The chemical journals with high impact factors (CJHIF) dataset, a large-scale compilation of reactions extracted from reports in high-impact chemical

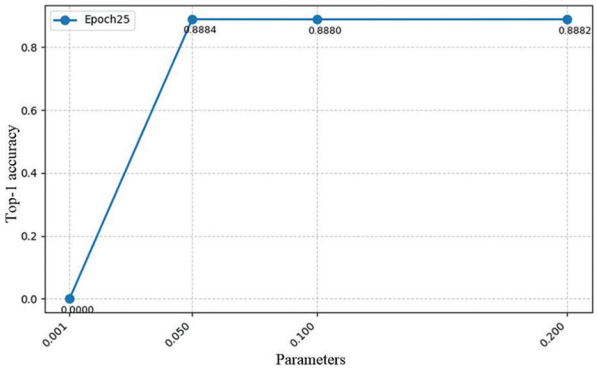

Figure 2. M2Echem prediction results under varying temperature parameter (ζ) at Epoch 25

Table 2. Number of reactions across four datasets

Dataset	Training set	Validation set	Testing set	Total
USPTO-50k	40,029	5,004	5,004	50,037
USPTO-Jin	409,035	30,000	40,000	479,035
USPTO-Schwaller	902,581	50,131	50,258	1,002,970
CJHIF	2,894,430	3,000	3,000	2,900,430

Abbreviations: CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

journals,[35] and the United States Patent and Trademark Office (USPTO) dataset, a public collection of chemical reactions mined from United States patent grants spanning from 1976 to September 2016.[36]

(i) USPTO-50k: This dataset contains 50,037 reactions categorized into 10 different types.[37] Training, validation, and testing sets comprised 40,029, 5,004, and 5,004 reactions, respectively.

(ii) USPTO-Jin: Jin *et al.*[38] extracted data from the USPTO dataset and retained 479,035 reactions without stereochemical information or atom-mapping. They combined the reagents and reactants into a source string, delimiting them with a period (.).

(iii) USPTO-Schwaller: Schwaller *et al.*[8] removed duplicates and non-canonicalizable items, ultimately retaining 1,002,970 reactions that contained only a single product.

(iv) CJHIF: Since the reagent information in the raw CJHIF data was provided by name rather than in the SMILES representation, we utilized the PUG-View application programming interface to search for reagents in PubChem and obtained their SMILES expressions. For reagents that were not found, we excluded them. Subsequently, we converted the reactions in the CJHIF dataset to follow the reactants > reagents form. The target dataset contained the corresponding products. Finally, we retained 638,597 reactions that were normalizable and non-duplicated, and selected 3,000 reactions for each validation and test set.

4.2. Related parameter settings

M2Echem was developed using Python 3.7, RDKit version 2022.9.5 (Novartis, Basel), and Hugging Face Transformers version 4.10.2 (Thomas, America). The model architecture includes encoders and a decoder consisting of four identical layers, with a multi-head attention layer employing eight attention heads. The hidden dimension was set to 256, and the intermediate feed-forward layer size was set to 2,048. The Adam optimizer was used, and a beam search size of five was implemented. The batch size, total number of training epochs, and training equipment were configured to 8, 30, and A6000, respectively.

4.3. Evaluation metrics

Three commonly used evaluation metrics were considered: Bilingual evaluation understudy (BLEU),[39] accuracy (Top-ς), and *t*-test ($\left|\hat{T}\right|$). These are standard external methods for evaluating models.

The BLEU metric measures similarity between sentences by calculating the maximum matches of *n*-grams from the target sequence within the prediction sequence. The formula for *n*-grams is presented in Equation XIII:

$$T_n = \frac{\sum_i^E \sum_k^K min(M_k(c_i), max_{j\epsilon\omega} M_k(s_{i,j}))}{\sum_i^E \sum_k^K min(M_k(c_i))} \quad \text{(XIII)}$$

Where ω refers to the number of target sequences associated with the same chemical reaction text. *E* denotes the collection of target sequences in the dataset, *k* is the *k*-th phrase, $M_k(c_i)$ depicts the number of times the *k*-th phrase occurs in the prediction sequence, while $M_k(S_{i,j})$ represents its occurrences in the target sequence. We used a 4-gram BLEU calculation, and its formula is presented in Equation XIV:

$$BLEU = \frac{1}{N} \sum_{i=1}^N T_n \quad \text{(XIV)}$$

Top-ς accuracy gauges the percentage of predictions with the correct label among the Top-ς results. Higher accuracy generally corresponds to improved prediction performance. The calculation formula is presented in Equation XV:

$$Top-\varsigma = \frac{1}{N} \sum_{i=1}^N \pi(s_i, c_{i,s}) \quad \text{(XV)}$$

Where π (·) indicates a value of 1 if the target sequence matches the prediction sequence; otherwise, it indicates a value of 0. $c_{i,s}$ represents the topresults in the Top-ς results in the prediction results, with $\varsigma\epsilon[1, 2, 3, 4, 5]$.

$\left|\hat{T}\right|$ assesses whether the difference between the two group means is significantly greater than the random error. The greater the value, the more significant the difference. The formula is given in Equation XVI:

$$\left|\hat{T}\right| = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{(XVI)}$$

Where $\bar{X}_1$ and $\bar{X}_2$ represent the sample means, while S^2 denotes the variance.

4.4. Comparative analysis

Table 3 presents the Top-1, Top-3, and Top-5 accuracy rates, BLEU scores, $|\hat{T}|$ values, and *p*-values for the baseline and M2Echem models across four datasets, with the best results indicated by superscripted lowercase "a" ([a]). If the average values of the two groups are the same, the probability of obtaining the observed *p*-value was calculated. If this probability is <0.05, the difference is considered statistically significant and is marked with an asterisk (*). If this probability is <0.001, it is considered highly statistically significant and is marked with a double asterisk (**). "Baseline" refers to the T5chem model, which utilizes a single encoder-decoder architecture and optimizes using cross-entropy loss.

Based on the experimental results, M2Echem outperformed the baseline model by the following margins on four different datasets: It surpassed the baseline by 1.3, 1.3, 1.5, and 0.36 points across four metrics on the USPTO-50k dataset. These results indicate that the fusion loss mitigates small-sample overfitting. For the USPTO-Jin dataset, M2Echem outperformed the baseline model by 0.3, 0.1, and 0.07 points across three metrics. These findings suggest that the model enhances information complementarity and improves prediction performance by capturing atomic details, key molecular features, and implementing a fusion strategy. In addition, for the USPTO-Schwaller dataset, the proposed model surpassed the baseline model by 0.3, 0.3, 0.4, and 0.21 points across four metrics. Finally, for the CJHIF dataset, the proposed model exceeded the baseline by 4.1, 5.0, 5.0, and 4.1 points across four metrics. Both USPTO-Schwaller and CJHIF datasets provide stereochemical information. The improved experimental results demonstrate the effective use of stereochemical information in model design. Moreover, the independent sample $|\hat{T}|$ demonstrated that the prediction accuracy of the M2Echem model on the four datasets was significantly better than that of the baseline model, with $p<0.05$. This indicates a statistically significant difference, suggesting that our method achieved improved prediction accuracy.

To assess the effectiveness of the proposed model for product prediction, we compared it with several existing models, including S2S,[8] WLDN5,[22] Fairseq,[40] Molecular Transformer,[10] and T5chem. As shown in Table 4, the M2Echem model significantly outperformed these traditional models in both accuracy and BLEU scores on the USPTO-Jin and USPTO-Schwaller datasets. In addition, we conducted a *t*-test to compare the validation accuracy during the training process of the M2Echem model and several traditional models, including T5chem, Molecular Transformer, Fairseq, WLDN5, and S2S, resulting in *p*-values of 0.035010, 0.033011, 0.000091, 0.000042, and 0.000039, respectively. These *p*-values were well below the conventional significance threshold of 0.05, confirming that the accuracy improvement achieved by the M2Echem model is statistically significant.

4.5. Ablation experiments

To analyze the role of the molecular maximum dimension reduction algorithm and the fused loss method, ablation experiments were conducted on three datasets, and the results are presented in Table 5. The label "MAX" refers to the molecular maximum dimension reduction algorithm, while "NTL" represents the fused loss method.

Based on Table 5, the molecular maximum dimension reduction algorithm plays a crucial role in enabling the model

Table 3. Baseline and M2Echem models' performances across four datasets

Dataset	Model	Accuracy (%)			BLEU	$\lvert\hat{T}\rvert$	*p*-value	Time (h)
		Top-1	Top-3	Top-5				
USPTO-50k	Baseline	40.60	61.50	68.40	88.18	2.601	0.01093*	3.40
	M2Echem	41.90[a]	62.80[a]	69.90[a]	88.54[a]			3.92
USPTO-Jin	Baseline	89.40	95.10	96.10	98.29	2.108	0.03501*	5.31
	M2Echem	89.70[a]	95.10[a]	96.20[a]	98.36[a]			7.08
USPTO-Schwaller	Baseline	77.00	85.80	87.60	94.45	2.191	0.03452*	35.27
	M2Echem	77.30[a]	86.10[a]	88.00[a]	94.66[a]			36.45
CJHIF	Baseline	56.20	69.10	73.40	87.81	4.503	0.00001**	119.01
	M2Echem	60.30[a]	74.10[a]	78.40[a]	90.22[a]			120.00

Notes: Superscripted lowercase "a" ([a]) indicates the best results. An asterisk (*) represents a statistically significant difference at $p<0.05$, whereas a double asterisk (**) denotes a highly statistically significant difference at $p<0.001$.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

Table 4. Comparison of the M2Echem model with several existing models

Model	USPTO-Jin					USPTO-Schwaller				
	Accuracy (%)			BLEU	p (10^{-5})	Accuracy (%)			BLEU	p (10^{-5})
	Top-1	Top-3	Top-5			Top-1	Top-3	Top-5		
S2S	80.30	86.20	87.50	-	3.9**	65.40	74.10	-	-	27.1**
WLDN5	80.60	-	93.40	-	4.2**	-	-	-	-	-
Fairseq	82.42	89.85	90.74	90.74	9.1**	69.69	77.33	78.92	92.50	52.3**
Molecular Transformer	88.80	92.60	94.40	96.27	3301.1*	76.17	82.86	83.69	85.17	3107.5*
T5chem	89.40	95.10	96.10	98.29	3501.0*	77.00	85.80	87.60	94.45	3452.0*
M2Echem	89.70[a]	95.10	96.20[a]	98.36[a]	-	77.30[a]	86.10[a]	88.00[a]	94.66[a]	-

Notes: Superscripted lowercase "a" ([a]) indicates the best results. An asterisk (*) represents a statistically significant difference at $p<0.05$, whereas a double asterisk (**) denotes a highly statistically significant difference at $p<0.001$.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

Table 5. Results of the ablation experiments

Dataset	Molecular maximum dimension reduction algorithm method	Fused loss method	Accuracy (%)			BLEU
			Top-1	Top-3	Top-5	
USPTO-Jin	/	/	89.40	95.10	96.10	98.29
	MAX	/	89.60	95.10	96.10	98.30
	MAX	NTL	89.70	95.10	96.20	98.36
USPTO-Schwaller	/	/	77.00	85.80	87.60	94.45
	MAX	/	77.20	85.80	87.60	94.53
	MAX	NTL	77.30	86.10	88.00	94.66
CJHIF	/	/	56.20	69.10	73.40	87.81
	MAX	/	58.80	71.80	76.10	89.23
	MAX	NTL	60.30	74.10	74.10	90.22

Notes: "MAX" refers to the molecular maximum dimension reduction algorithm, while "NTL" represents the fused loss method.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

to learn key molecular features, thereby improving the accuracy of chemical product predictions. When this algorithm is removed, there is a noticeable decrease in evaluation metrics across all three datasets. In addition, the fused loss method contributed positively to training the M2Echem model. Removing this component negatively affects the evaluation metrics across the three datasets. The results of the ablation experiments demonstrate that the molecular maximum dimension reduction algorithm and the fused loss method together enhance the proposed model's performance.

4.6. Model understanding of different SMILES representations

A molecule represented by "C\C(COC1CCCCO1)=C/I," which was not included in the pre-training model, was selected to generate four different SMILES using non-standard SMILES data augmentation.[41,42] Figure 3 illustrates the generation of atom embeddings using both the baseline and M2Echem models, visualized through t-distributed stochastic neighbor embedding. The baseline model exhibited a broad and scattered distribution of atoms at Epoch 1 as the model had not yet been sufficiently trained. By Epoch 30, the baseline model gradually learnt to interpret the different SMILES representations of the same molecule. In contrast, the M2Echem model at Epoch 30 produced a more aggregated distribution of the different SMILES of the same molecule and was able to distinguish between iodine atoms in different chemical environments while maintaining their similarity. Therefore, the M2Echem model demonstrated superior capability in recognizing the same atomic labeling in different SMILES at Epochs 1 and 30, while the baseline showed weaker performance.

4.7. Model understanding of long SMILES embeddings

Figure 4 illustrates the embedding of Tanimoto coefficients[43] derived from three different SMILES

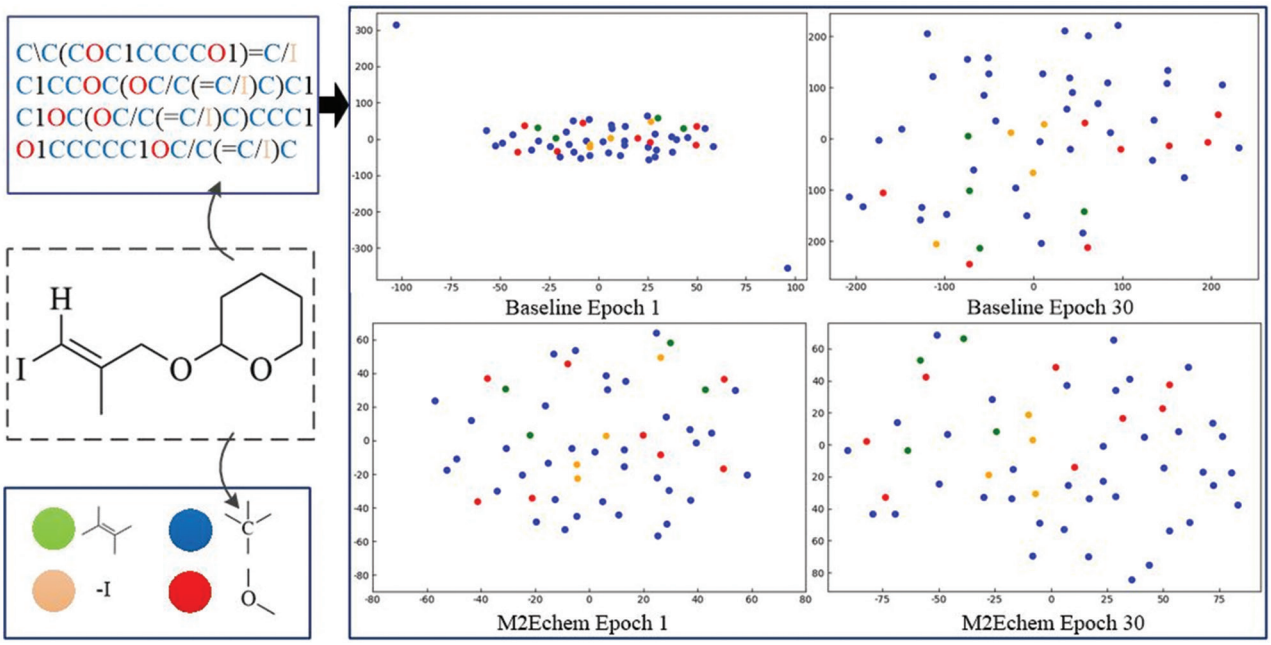

Figure 3. Visualization of atom embeddings for different SMILES formats using *t*-distributed stochastic neighbor embedding

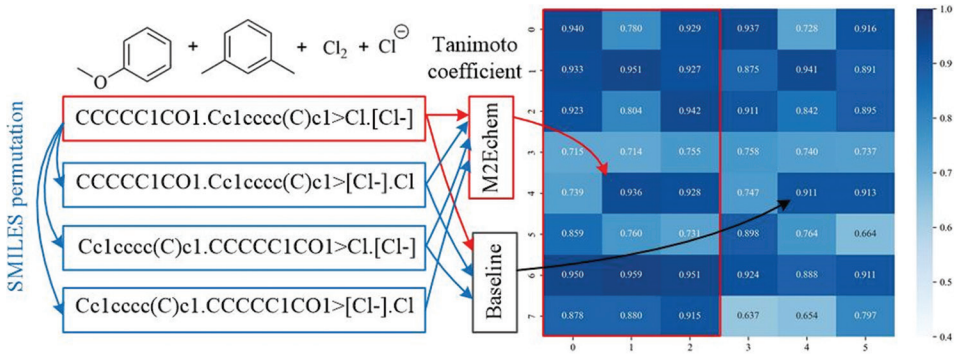

Figure 4. Tanimoto coefficient embeddings for reactants. The left panel (red box) displays the Tanimoto coefficients for M2Echem, while the right panel displays those for the baseline model.

embeddings of eight molecules that were not included in the pre-trained dataset, along with the corresponding canonical SMILES embeddings. The left panel (red box) displays the Tanimoto coefficients for M2Echem, while the right panel displays those for the baseline model. The Tanimoto coefficient approaches 1 as the SMILES embeddings increase in similarity. In Table 6, "Tanimoto coefficient 1" indicates the average value for different SMILES representations of the same molecule in the M2Echem model, while "Tanimoto coefficient 2" refers to the corresponding average in the baseline model. The graph and table illustrate the performance of both models at Epoch 30, and the vertical coordinates numbered 0–7 in Figure 4 correspond to the SMILES serial numbers 0–7 in Table 6. These findings suggest that, for SMILES sequences ranging in length from 32 to 90, the proposed model demonstrated greater embedding similarity compared to the baseline model.

To assess the model's ability to understand long sequences, we categorized three datasets according to sequence length: 0–55 as short sequences, 55–110 as medium sequences, and those longer than 110 as long sequences. Table 7 presents the accuracy rates and BLEU scores for these three sequence-length categories across the three datasets on different models. The molecular maximum dimension reduction algorithm demonstrated significant benefits for long sequences in the USPTO-Jin dataset, showing improvements of 3.15 points in BLEU

Table 6. Sequence lengths and Tanimoto coefficients for eight simplified molecular input line entry system embeddings

SMILES	Sequence length	Tanimoto coefficient 1	Tanimoto coefficient 2
0	32	0.883065[a]	0.860441
1	86	0.937071[a]	0.902324
2	47	0.889812[a]	0.882767
3	50	0.728019	0.744914[a]
4	59	0.867746[a]	0.744914
5	90	0.783501[a]	0.775470
6	85	0.953346[a]	0.907687
7	57	0.890992[a]	0.696146

Note: Superscripted lowercase "a" ([a]) indicates that the average embedding values for different SMILES of the same molecule exhibit high similarity.
Abbreviation: SMILE: Simplified molecular input line entry system embeddings.

score and 2.13 points in Top-1 accuracy compared to the baseline. The M2Echem model provided consistent performance across all three sequence lengths in the datasets. In particular, for the USPTO-Jin dataset, the Top-3 accuracy for long sequences increased by 2.13 points, while the BLEU score and accuracy for long sequences in the CJHIF dataset improved by 1.35, 2.94, 5.95, and 7.28 points, respectively, compared to the baseline.

4.8. Model performance in product prediction

Reaction types were randomly selected from the test set to predict organic chemical products using both the M2Echem and the baseline model. The results are illustrated in Figure 5, where red markings indicate changes in core atoms or atomic groups. The organic reactions in Figures 5A and 5B are both substitution reactions, and although the reactions are simple, the baseline model predicted incorrect products, thereby violating the principle of elemental conservation. The baseline model shown in Figure 5A predicted products with fewer oxygen atoms, while the model in Figure 5B predicted products with fewer carbon and fluorine atoms.

In addition, Figure 5C depicts a mono-disubstituted reduction reaction in which an NH group first undergoes acid-base neutralization with formic acid to form NOOCH, and then NOOCH is reduced to NCH_3. This transformation is correctly predicted by the M2Echem model, whereas the baseline model generates unreasonable products. In addition, Figure 5D illustrates a nucleophilic substitution reaction in which negatively charged carbon atoms attack positively charged carbon atoms. This interaction disrupts the C-Cu coordination bond and the C-Br bond, leading to the formation of the final product.

However, the predictions made by the baseline model were incorrect. Figure 5E shows a displacement reaction. The product predicted by the M2Echem model was close to the correct answer, whereas the baseline model fails to account for the metal sodium atom. These findings suggest that the model presented in this study effectively predicts the outcomes of fundamental reaction types, including displacement and substitution.

4.9. Attention weight

Two SMILES sequences were randomly selected from the USPTO-Jin test set. Figure 6 illustrates the relationship between attention weights for reactant-product mappings under both the baseline model and the M2Echem model for the two selected SMILES sequences.

In the attention matrix, the horizontal axis represents the reactants and reagents, while the vertical axis represents the products. Figure 6A presents the attention maps for both the baseline and M2Echem models during a reaction in which the aldehyde group "-COH" is reduced to an alcohol. In comparison to the baseline model, the M2Echem model demonstrated a more continuous and concentrated focus within the blue box. This suggests that the M2Echem model can accurately and consistently capture the mapping relationship between reactants and products, and that there are fewer noise points in the non-core areas. Figure 6B illustrates an esterification reaction. The M2Echem model better identified the esterification process involving acid dehydroxylation and alcohol dehydrogenation within the first blue box.

4.10. Computational efficiency of the model

The time and space complexity analyses of the baseline and M2Echem models is as follows, where the longest sequence length is L, the number of heads in multi-head attention is H, the dimension of the hidden layer is d_{hid}, the dimension of the feedforward layer is d_{ff}, the batch size is B, the number of layers is L_{layer}, the total number of samples is N, the average number of points in the sequence is P, and the model parameters are denoted as ϑ.

The baseline model consists of a single encoder-decoder structure. The time complexity of the encoder primarily arises from the multi-head attention mechanism and the feedforward network, which have time complexities of $O(d_{hid}HL^2)$ and $O(d_{hid}d_{ff}L)$, respectively. Therefore, the overall time complexity of the encoder is $O(BL_{layer}d_{hid}HL^2)$, and that of the decoder is $O(L_{layer}L^3d_{hid}H)$. The total time and space complexities of the baseline model are $O(L_{layer}d_{hid}H(B/L+1)L^3)$ and $O(2\vartheta + BLd_{hid})$, respectively.

Table 7. Accuracy rates and bilingual evaluation understudy scores across three sequence-length categories for different models on three datasets

Dataset	Sample size	Molecular maximum dimension reduction algorithm method	Fused loss method	Accuracy (%)			BLEU
				Top-1	Top-3	Top-5	
USPTO-Jin-pre	31,225	/	/	88.43	94.59	95.67	98.02
		MAX	/	88.62 94.56 95.73 98.02	94.56	95.73 [a]nd 98.02	98.02
		MAX	NTL	88.62	94.60[a]	95.83[a]	98.12[a]
USPTO-Jin-mid	8,728	/	/	92.99	96.77	97.49[a]	99.26
		MAX	/	92.98	96.84	97.39	99.23
		MAX	NTL	93.27[a]	96.91[a]	97.48	99.27[a]
USPTO-Jin-last	47	/	/	82.98	91.49	91.49	95.32
		MAX	/	85.11[a]	91.49	93.62	98.47
		MAX	NTL	76.60	93.62[a]	93.62[a]	97.42[a]
USPTO-Schwaller-pre	37,052	/	/	76.40	85.02	87.13	93.61
		MAX	/	76.34	85.29	87.18	93.61
		MAX	NTL	76.96[a]	85.71[a]	87.75[a]	93.93[a]
USPTO-Schwaller-mid	12,845	/	/	78.74	87.24	89.16	96.89
		MAX	/	79.17[a]	87.44	89.05	96.86
		MAX	NTL	79.05	87.68[a]	89.27[a]	96.94[a]
USPTO-Schwaller-last	350	/	/	60.94	72.85	75.35	93.70
		MAX	/	61.11	73.71[a]	75.56	93.86
		MAX	NTL	61.43[a]	72.22	76.29[a]	94.19[a]
CJHIF-pre	2,025	/	/	57.63	70.67	75.11	87.31
		MAX	/	59.51	72.94	77.58	88.96
		MAX	NTL	61.33[a]	75.37[a]	79.62[a]	89.84[a]
CJHIF-mid	732	/	/	49.32	62.84	67.21	88.19
		MAX	/	53.96	66.94	70.77	89.84
		MAX	NTL	55.12[a]	68.70[a]	73.27[a]	91.16[a]
CJHIF-last	127	/	/	50.19	62.20	64.57	89.39
		MAX	/	51.18	64.57	68.50	90.09
		MAX	NTL	53.33[a]	68.15[a]	71.85[a]	90.74[a]

Notes: "MAX" refers to the molecular maximum dimension reduction algorithm, while "NTL" represents the fused loss method. Superscripted lowercase "a" ([a]) indicates that the three models yield the best results on the same dataset within the same sequence length category.
Abbreviations: BLEU: Bilingual evaluation understudy; CJHIF: Chemical journals with high impact factors; USPTO: United States Patent and Trademark Office.

In contrast, the M2Echem model introduced additional components, including a molecular encoder, an algorithm, multi-level feature fusion, and an optimization loss, which have time complexities of $O(2BL_{layer}d_{hid}HL^2)$, $O(2NP^2)$, $O(2Nd_{hid}B)$, and $O(2d_{hid}B)$, respectively. The space complexities for these added components are $O(2BLd_{hid})$, $O(NL)$, and $O(2Bd_{hid})$. As a result, the total time and space complexities for the M2Echem model are $O(2BL_{layer}d_{hid}HL^2 + 2NP^2 + 2NBd_{hid} + 2L_{layer}d_{hid}HL^3)$, and $O(4\vartheta + 2BLd_{hid} + NL + 2Bd_{hid})$, respectively.

In this study, both models utilized eight multi-head attention heads, a hidden dimension of 256, four layers, and a feedforward layer dimension of 2,048. Consequently, as shown in Table 8, the time and space complexities for the baseline model are $O(8192BL^2+8192L^3)$ and $O(2\vartheta + 256B)$, while the time and space complexities for the M2Echem model are $O(2NP^2 + 512NBP^2 + 16384BL^2 + 16384L^3)$ and $O(4\vartheta + 512BL + NL + 512B)$, respectively.

The time and space complexities of the M2Echem model were approximately twice those of the baseline model. Furthermore, training time records (Table 3) indicated that the training time of the M2Echem model is 1–3 h longer than that of the baseline model. Nevertheless, as shown in Figure 7, M2Echem achieved

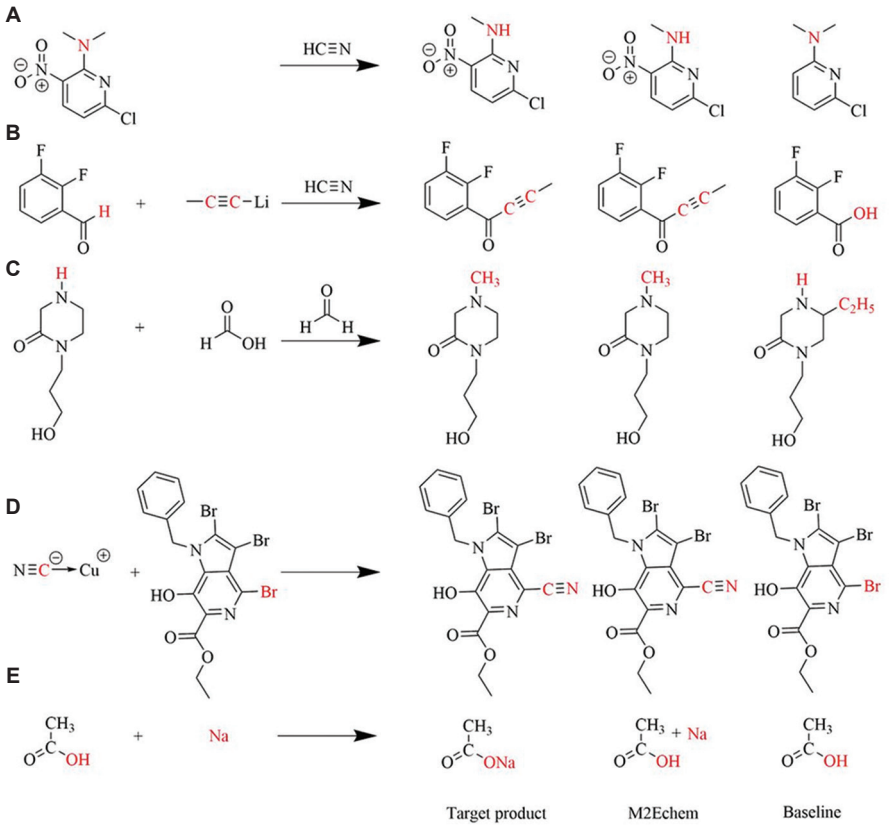

Figure 5. (A-E) Prediction results of the M2Echem model versus the baseline model across various reaction types

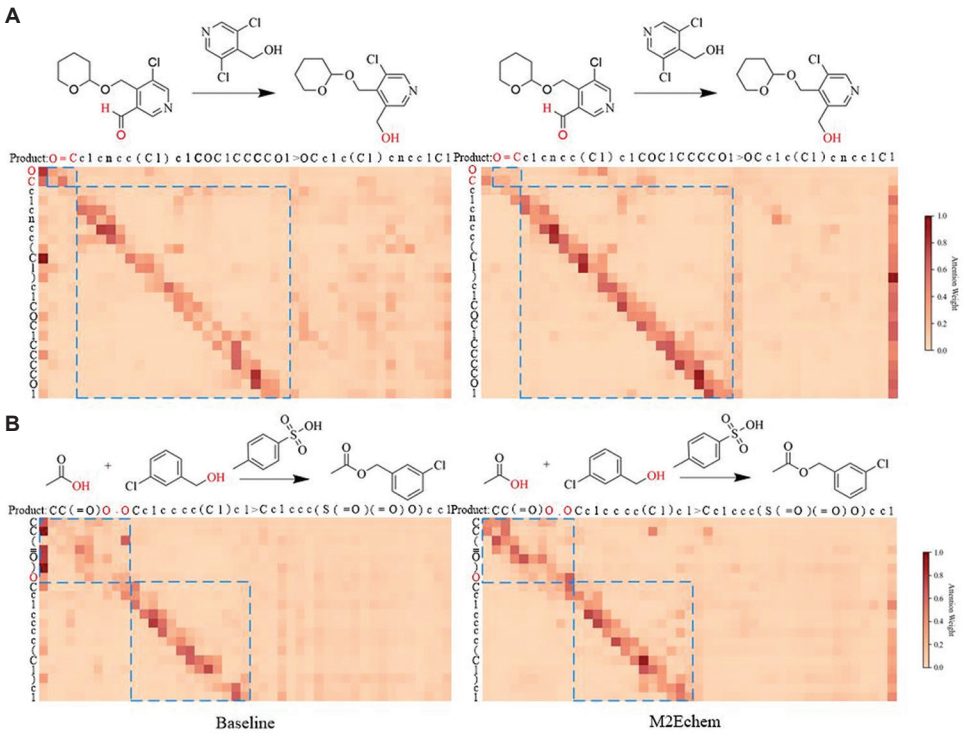

Figure 6. (A and B) Comparative analysis of attention weights for reactant-product interactions in chemical reactions between the baseline and M2Echem models

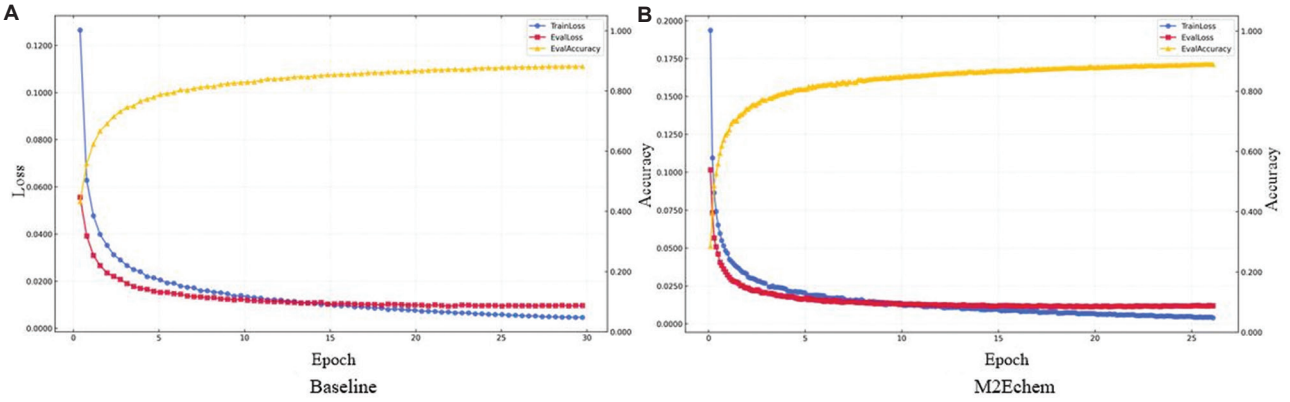

Figure 7. (A and B) Convergence curves for the baseline and M2Echem models

Table 8. Time and space complexities of the baseline and M2Echem models

Model	Time complexity	Space complexity	Model parameters
Baseline	O ($8192BL^2$+$8192L^3$) O ($8192BL^2$+$8192L^3$)	O (2ϑ + 256BL)	24.2 million
M2Echem	O ($2NP^2$+$512NBP^2$+ $16384BL^2$+$16384L^3$)	O (4ϑ + 512BL+ NL+512B)	37.8 million

faster convergence in product prediction. Although the time and space complexities of our model increased, the model extracted richer features and converged more rapidly during training, thereby improving prediction accuracy for product prediction.

5. Conclusion

The M2Echem model was developed to address complex organic chemistry forward prediction tasks, enhancing the performance of the T5chem model. This model utilizes molecular and atomic encoders to extract crucial feature information from atoms and molecules, enabling multi-level feature fusion that is subsequently fed into the decoder. The fusion loss component helps the model learn SMILES similarity more effectively. The findings demonstrate that the M2Echem model surpasses the baseline model across all four datasets. It extracts more comprehensive feature information, improves generalization across different representations of the same molecular structure, enhances the encoding of long sequence features, and contributes to artificial intelligence-driven drug development and research. However, we also noted that the model's effectiveness declines when dealing with complex or rare reaction types. Future research should focus on exploring a broader range of feature extraction methods and on improving the model's generalization ability for complex or rare reaction types.

Acknowledgments

None.

Funding

This research was funded by the National Natural Science Foundation of China (62406153, 62471259, and 62371261), the General Program of the Natural Science Research of Higher Education of Jiangsu Province (23KJB520031), and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (SJCX25_2007).

Conflict of interest

Jiashuang Huang is the Youth Editorial Board Member of this journal, but was not in any way involved in the editorial and peer-review process conducted for this paper, directly or indirectly. Separately, other authors declared that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author contributions

Conceptualization: Shu Jiang, Yifan Jiang
Formal analysis: Linxing Zhu
Investigation: Yifan Jiang, Jiashuang Huang
Methodology: Shu Jiang, Jing Wang, Linxing Zhu
Writing–original draft: Jing Wang
Writing–review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data are publicly accessible in an open repository. The USPTO-50k dataset can be found at https://yzhang.hpc.nyu.edu/T5Chem/, while the USPTO-Schwaller and USPTO–Jin datasets are available at https://ibm.ent.box.com/v/ReactionSeq2SeqDataset. However, the CJHIF dataset is proprietary and cannot be accessed due to commercial restrictions.

References

1. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. *Science*. 1969;166(3902):178-192.

   doi: 10.1126/science.166.3902.178

2. Satoh H, Funatsu K. Further development of a reaction generator in the SOPHIA system for organic reaction prediction. Knowledge-guided addition of suitable atoms and/or atomic groups to product skeleton. *J Chem Inform Comput Sci*. 1996;36(2):173-184.

   doi: 10.1021/ci950058a

3. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Cent Sci*. 2017;3(5):434-443.

   doi: 10.1021/acscentsci.7b00064

4. Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems*. Volume 2. California: Curran Associates, Inc.; 2015:2224-2232.

5. Raccuglia P, Elbert KC, Adler PD, *et al*. Machine-learning-assisted materials discovery using failed experiments. *Nature*. 2016;533(7601):73-76.

   doi: 10.1038/nature17439

6. Segler MH, Waller MP. Modelling chemical reasoning to predict and invent reactions. *Chemistry*. 2017;23(25):6118-6128.

   doi: 10.1002/chem.201604556

7. Weininger DJ. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Infrom Comput Sci*. 1988;28(1):31-36.

   doi: 10.1021/ci00057a005

8. Schwaller P, Gaudin T, Lanyi D, Bekas C, Laino TJ. "Found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem Sci*. 2018;9(28):6091-6098.

   doi: 10.1039/c8sc02339e

9. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is All you Need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017:6000-6010.

10. Schwaller P, Laino T, Gaudin T, *et al*. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent Sci*. 2019;5(9):1572-1583.

    doi: 10.1021/acscentsci.9b00576

11. Tang G, Müller M, Rios A, Sennrich R. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2018:4263-4272.

    doi: 10.18653/v1/d18-1458

12. Wu F, Fan A, Baevski A, Dauphin YN, Auli M. Pay Less Attention with Lightweight and Dynamic Convolutions. arXiv. Preprint posted online 2019.

    doi: 10.48550/arXiv.1901.10430

13. Schwaller P, Probst D, Vaucher AC, *et al*. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell*. 2021;3(2):144-152.

    doi: 10.1038/s42256-020-00284-w

14. Mellah Y, Kocaman V, Haq HU, Talby D. Efficient schema-less text-to-SQL conversion using large language models. *Artif Intell Health*. 2024;1(2):96-106.

    doi: 10.36922/aih.2661

15. Mumtaz U, Ahmed A, Mumtaz S. LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. *Artif Intell Health.* 2024;1(2):16-28.

    doi: 10.36922/aih.2558

16. Bran AM, Schwaller P. Transformers and large language models for chemistry and drug discovery. In: *Drug Development Supported by Informatics*. Berlin: Springer; 2024. p. 143-163.

17. Leon M, Perezhohin Y, Peres F, Popovič A, Castelli M. Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. *Sci Rep.* 2024;14(1):25016.

    doi: 10.1038/s41598-024-76440-8

18. Xiong J, Zhang W, Wang Y, *et al*. Bridging chemistry and artificial intelligence by a reaction description language. *Nat Mach Intell*. 2025;7(5):782-793.

    doi: 10.1038/s42256-025-01032-8

19. Lo A, Pollice R, Nigam A, White AD, Krenn M, Aspuru-Guzik AJ. Recent advances in the self-referencing embedded strings (SELFIES) library. *Dig Discov*. 2023;2(4):897-908.

    doi: 10.1039/D3DD00044C

20. Ucak UV, Ashyrmamatov I, Lee J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. *J Cheminform*. 2023;15(1):55.

    doi: 10.1186/s13321-023-00725-9

21. Wu Z, Jiang D, Wang J, *et al*. Knowledge-based BERT: A method to extract molecular features like computational chemists. *Brief Bioinform*. 2022;23(3):bbac131.

doi: 10.1093/bib/bbac131

22. Chen S, Jung Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*. 2021;1(10):1612-1620.

doi: 10.1021/jacsau.1c00246

23. Liu Z, Zhang W, Xia Y, *et al*. MolXPT: Wrapping Molecules with Text for Generative Pre-training. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics; 2023:1606-1616.

doi: 10.18653/v1/2023.acl-short.138

24. Lu J, Zhang Y. Unified deep learning model for multitask reaction predictions with explanation. *J Chem Inform Model*. 2022;62(6):1376-1387.

doi: 10.1021/acs.jcim.1c01467

25. Guo W, Wang J, Wang S. Deep multimodal representation learning: A survey. *IEEE Access*. 2019;7:63373-63394.

doi: 10.1109/ACCESS.2019.2916887

26. Ma S, Zhang D, Zhou M. A Simple and Effective Unified Encoder for Document-Level Machine Translation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:3505-3511.

doi: 10.18653/v1/2020.acl-main.321

27. Zhang X, Li P, Li H. AMBERT: A Pre-trained Language Model with Multi-Grained Tokenization. *arXiv*. Preprint posted online 2020.

doi: 10.48550/arXiv.2008.11869

28. Zhu J, Xia Y, Wu L, *et al*. Incorporating BERT into Neural Machine Translation. *arXiv*. Preprint posted online 2020.

doi: 10.48550/arXiv.2002.06823

29. Jin D, Jin Z, Zhou JT, Szolovits P. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020;34(05):8018-8025.

doi: 10.1609/aaai.v34i05.6311

30. Singh S, Shingatgeri V, Srivastava P. Revolutionizing new drug discovery: Harnessing AI and machine learning to overcome traditional challenges and accelerate targeted therapies. *Artif Intell Health*. 2024;2(2):29-40.

doi: 10.36922/aih.4423

31. Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv*. Preprint posted online 2021.

doi: 10.48550/arXiv.2104.08821

32. Chen X, Alamro H, Li M, *et al*. Target-aware Abstractive Related Work Generation with Contrastive Learning. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery*; 2022:373-383.

doi: 10.1145/3477495.3532065

33. Miculicich L, Ram D, Pappas N, Henderson J. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Belgium: Association for Computational Linguistics; 2018:2947–2954.

34. Mao A, Mohri M, Zhong Y. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In: *Proceedings of the 40th International Conference on Machine Learning*. 2023:23803-23828.

35. Jiang S, Zhang Z, Zhao H, *et al*. When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access*. 2021;9:85071-85083.

doi: 10.1109/ACCESS.2021.3083838

36. Lowe D. Chemical reactions from US patents (1976-Sep2016). Published online 2017.

doi: 10.6084/M9.FIGSHARE.5104873

37. Liu B, Ramsundar B, Kawthekar P, *et al*. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent Sci*. 2017;3(10):1103-1113.

doi: 10.1021/acscentsci.7b00303

38. Jin W, Coley C, Barzilay R, Jaakkola TJ. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017:2604-2613.

39. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Association for Computational Linguistics*; 2001:311-318.

doi: 10.3115/1073083.1073135

40. Wang T. Research on chemical reaction prediction model based on Fairseq. In: *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*. IEEE; 2021:167-171.

doi: 10.1109/icbase53849.2021.00039

41. Bjerrum EJ. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. arXiv. Preprint posted online 2017.

doi: 10.48550/arXiv.1703.07076

42. Tetko IV, Karpov P, Van Deursen R, Godin G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat Commun*. 2020;11(1):5575.

doi: 10.1038/s41467-020-19266-y

43. Khalifa AA, Haranczyk M, Holliday J. Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J Chem Inf Model*. 2009;49(5):1193-1201.

doi: 10.1021/ci8004644