Received March 25, 2021, accepted May 16, 2021, date of publication May 26, 2021, date of current version June 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3083838

When SMILES Smiles, Practicality Judgment and Yield Prediction of Chemical Reaction via Deep Chemical Language Processing

SHU JIANG[1,2], ZHUOSHENG ZHANG[1,2], HAI ZHAO[1,2], JIANGTONG LI[1,2], YANG YANG[1,2], (Member, IEEE), BAO-LIANG LU[1,2], (Fellow, IEEE), AND NING XIA[3]

[1]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[3]Chemical.AI, Shanghai 200240, China

Corresponding author: Hai Zhao (zhaohai@cs.sjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0304100; and in part by the Key Projects of National Natural Science Foundation of China under Grant U1836222, Grant 61733011, and Grant 61972251.

ABSTRACT Simplified Molecular Input Line Entry System (SMILES) provides a text-based encoding method to describe the structure of chemical species and formulize general chemical reactions. Considering that chemical reactions have been represented in a language form, we present a symbol only model to generally predict the yield of organic synthesis reaction without considering complex quantum physical modeling or chemistry knowledge. Our model is the first deep neural network application that treats chemical reaction text segments as embedding representation to the most recent deep natural language processing. Experimental results show our model can effectively predict chemical reactions, which achieves a high accuracy of 99.76% on practicality judgment and the Root Mean Square Error (RMSE) is around 0.2 for yield prediction. Our work shows the great potential for automatic yield prediction for organic reactions under general conditions and further applications in synthesis path prediction with the least modeling cost.

INDEX TERMS Chemical reaction, practicality judgment, yield prediction, machine learning.

I. INTRODUCTION

Organic reaction including addition reactions [1], elimination reaction [2], substitution reactions [3]–[5], pericyclic reactions [6], rearrangement reactions [7], [8], redox reaction [9] have been studied for hundreds of years. Owing to the development of organic methodology [10], hundreds of millions of reactions have been practiced, and more and more compounds have been produced. Nevertheless, the mechanism of organic reactions has not been completely understood. A new organic reaction's practicality still mainly relies on humans to judge it is positive or negative from the expertise and the eventual exploratory synthesis verification, let alone the more difficult yield prediction.

Modeling the organic reactions through traditional physical-level method may lead to a too complicated huge model but with poor informative representation [11]. Even for simple reaction containing only several atoms, it is practically difficult due to the need to consider the combinatorial component arrangement using quantum chemistry method. Predicting a chemical reaction from a given complicated reaction, especially in a particular reaction condition, is even much more challenging [12] because it requires considering every transition-state and the combination between molecules and their given environment.

Instead, in the latest symbol model for the chemical reaction, the chemical elements and molecules are regarded as symbols, and the chemical reactions are identified as text with chemical information. Consequently, most text processing methods, including machine learning, especially deep learning, can be applied to the chemical language text. Support Vector Machine (SVM) has been proved to be useful to predict the result of crystallization of templated vanadium selenites [13]. However, it requires a complicated manual feature selection based on necessary chemical knowledge, making it much more like human classification.

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

Information retrieval is also an effective way to predict the products of organic reactions [14], [15], which presented a limited candidate set for ranking without yield prediction. Continuous representation of molecules [16] provides a convenient method to automatically generating chemical structures. More recently, some researchers [17] cast the reaction prediction task as a translation problem by introducing a template-free sequence-to-sequence model, trained end-to-end and entirely data-driven and achieved a top-1 accuracy above 90% on a common benchmark data set without relying on auxiliary knowledge such as reaction rules. Recently, Abigail Doyle *et al.* [18] proposed the random forest algorithm, which can accurately predict the yield of Buchwald Hartwig cross-coupling reactions with many detailed features of materials in reactions. However, this model can only process a kind of reaction and needs too much information about the reactions.

Existing work using machine learning for chemical information processing falls either relying on the strong chemical knowledge source or focusing on specific types of reactions. Distinct from previous works, we provide a cutting-edge symbol alone model on the chemical text of organic reactions from a general background. A complete data-driven method is proposed for open-type organic chemical reactions, releasing the inconvenient prerequisite with prior chemical knowledge. Our approach can automatically discover the salient features and reaction patterns for reaction practicality and yield prediction without the complex parameter setting or manual chemical knowledge-based feature selection.

In recent years, natural language processing (NLP) has popularly adopted embedding representation for text units, a sort of low-dimensional continuous representation learned from neural networks. Following NLP's latest advance, embedding is also used to represent chemical text segments and later learning. Using a data-driven mode, our model will directly learn from a large scale of available reaction data. We use reaction formulae collected from the publications for about 1.7 million reactions. To form a negative set for discriminative learning, we collect about 12K invalid reactions [19] from experiments in Chemical.AI's laboratory, which keeps the balance between the positive and the negative cases. Given the reactants and products, along with the reaction conditions, our model can accurately judge the reaction practicality and predict the yield. Our model pipeline is demonstrated as follows. First, we preprocess the SMILES sequence of each reactant and product in atom-wise form and tokenize the resulting text into segments in a natural language processing way. Then, we extract and tag a series of text segments by *edit distance* detection operation on the reactants and the products as the *Reaction Symbol Edit Operation* to represent the text difference between both sides. Finally, we convert all the resulting chemical text segments into embedding form for more effective representation and feed the reactants, products, reactions, and reaction conditions in vectors to a Siamese neural network for *practicality judgment* and *yield prediction*. The sources of our model are available on https://github.com/jshmjs45/data_for_chem.

II. DATA POSSESSING AND MODEL CONSTRUCTION

For chemical reaction prediction, the critical point is to effectively capture the internal relationships between a reactant and the corresponding product representation, along with Reaction Symbol Edit Operation (RSEO) and conditions. This task is formulized into language processing over the corresponding chemical text. Figure 1 shows that the text segments of reactants and products are represented as vectors of low-dimensional embedding representation. We train a deep neural network (i.e., the Siamese) to learn the chemical principles of reactions by transforming the feature representation of reactants and products. After training, given a specific reaction, the model will judge the practicality and predict the yield.

A. UNSUPERVISED TOKENIZATION

SMILES is a line notation for entering and representing molecules and reactions using short ASCII strings initiated by David Weininger at the USEPA Mid-Continent Ecology Division Laboratory in Duluth in the 1980s [20]. SMILES is more useful than an extended connection table because it is a linguistic construct rather than a computer data structure. SMILES is a true language, albeit with a simple vocabulary (atom and bond symbols) and only a few grammar rules. Some examples of SMILES are shown in Table 1.

At the very beginning, we remove the hydrogen atoms and the atom mappings from the reaction string and canonicalize the molecules. We treat the chemical reaction described by SMILES as a kind of text in natural language. Considering that chemical elements (atoms) and various SMILES bond symbols are "*characters*" in the chemical language, we regard a sequence of SMILES which stands for a chemical compound as the corresponding sentence. Therefore we need to mine the sequence to find a basic meaningful linguistic unit, "*word*". However, the SMILES encoding text does not provide a word segmentation with definite chemical meaning to facilitate the chemical text processing. Thus, we turn to an unsupervised tokenization solution in the existing natural language processing [21]. We adopt a goodness measure based method to tokenize each reactant, product, or condition text in SMILES into a sequence of "*words*". Let $W = \{\{w_i, g(w_i)\}_{i=1,\ldots,n}\}$ be a list of character n-grams (namely, word candidates) each associated with a goodness score for how likely it is to be a true word from a linguistic/chemical perspective, where w_i is a word candidate and $g(w_i)$ is its goodness function.

The adopted segmentation algorithm is a greedy maximal-matching one with respect to a goodness score.

$$\{w^*, t^*\} = \underset{w_1 \ldots w_i \ldots w_n = T}{\arg\max} \sum_{i=1}^{n} g(w_i) \tag{1}$$

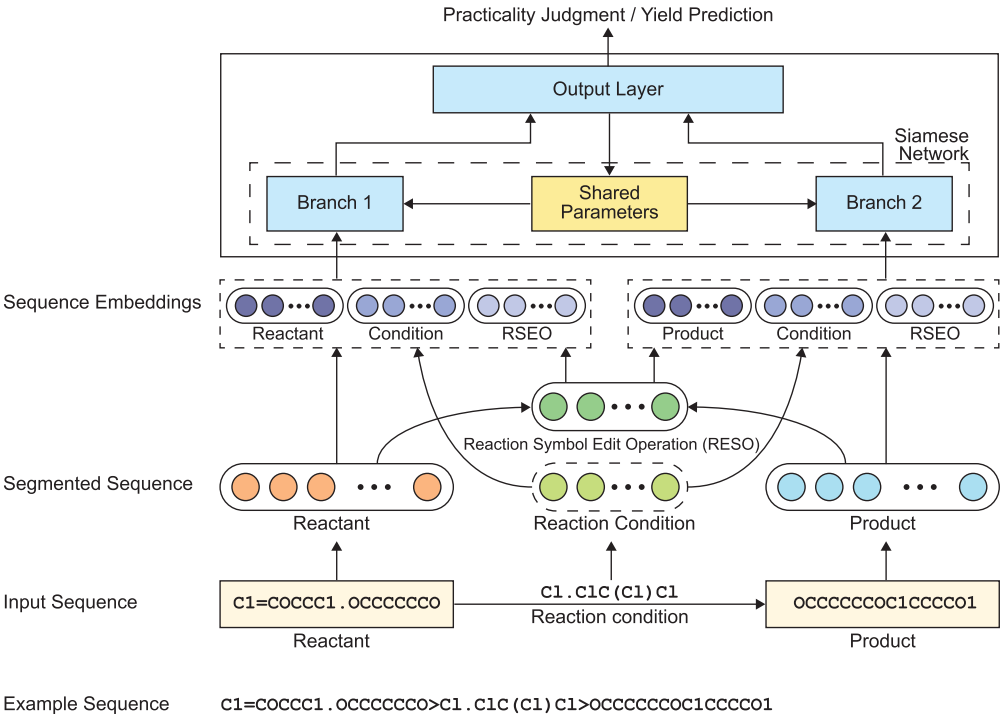


FIGURE 1. The Model for *practicality judgment* and *yield prediction*.

TABLE 1. Examples of SMILES.

SMILES	Name	SMILES	Name
CC	ethane	[OH3+]	hydronium ion
O=C=O	carbon dioxide	[2H]O[2H]	deuterium oxide
C#N	hydrogen cyanide	[235U]	uranium-235
CCN(CC)CC	triethylamine	F/C=C/F	E-difluoroethene
CC(=O)O	acetic acid	F/C=C\F	Z-difluoroethene
C1CCCCC1	cyclohexane	N[C@@H](C)C(=O)O	L-alanine
c1ccccc1	benzene	N[C@H](C)C(=O)O	D-alanine

It works on T to output the best current word w^* repeatedly with $T = t^*$ for the next round as follows, with each $\{w, g(w)\} \in W$.

In our work, we use *Description Length Gain (DLG)* as the goodness measurement for a candidate character n-gram from the chemical text. In principle, the higher goodness score for a candidate, the more likely it is to be a true word. *DLG* was proposed by Kit and Wilks [22] for compression-based unsupervised segmentation. The DLG extracts all occurrences of $x_{i..j}$ from a corpus $X = x_1 x_2 \dots x_n$, and selects the top 100,000 items as the vocabulary in our model sorted by its DLG goodness score, which is defined as:

$$g_{DLG}(x_{i..j}) = L(X) - L(X[r \to x_{i..j}] \oplus x_{i..j}), \tag{2}$$

where $X[r \to x_{i..j}]$ represents the resultant corpus by replacing all items of $x_{i..j}$ with a new symbol r throughout X, and $\oplus$ denotes the concatenation operator. $L(\cdot)$ is the empirical description length of a corpus in bits that can be estimated by the Shannon-Fano code or Huffman code as below, following classic information theory [23],

$$L(X) \doteq -|X| \sum_{x \in V} \hat{p}(x) \log_2 \hat{p}(x), \tag{3}$$

where $|\cdot|$ denotes the string length, V is the character vocabulary of X and $\hat{p}(x)$ is x's frequency in X.

The preprocessing steps, together with examples, are summarized in Table 2. The same preprocessing steps were applied to all datasets.

B. REACTION SYMBOL EDIT OPERATION (RSEO) GENERATION

To formally represent the text difference from reactants to products in a reaction formula, we introduce the formal concept of *Reaction Symbol Edit Operation (RSEO)*. It indicates how the source chemical text can be transformed into the target one through a series of symbol inserting and deleting operations. The text operation series can be decoded from the *edit distance* [24], which quantifies how different two strings are from one another by counting the minimum number of operations required to transform one string into the other.

For a source sequence $S = s_1\ s_2 \dots s_n$ and the target sequence $T = t_1\ t_2 \dots t_m$, the RSEO sequence $R = r_1\ r_2 \dots r_n$ is encoded by the following tags:

- **AD** indicates a string should be added right before the corresponding location.

TABLE 2. Data processing steps. The tokens are separated by a space and individual molecules by a point token.

Step	Example (reactants > reagents > products)
1) Original string	[C:1]([C:3]1[CH:8]=[CH:7][CH:6]=[CH:5][C:4]=1[OH:9])#[N:2].[CH2:10]([CH:12]1[O:14][CH2:13]1)Cl>N1CCCCC1>[O:14]1[CH2:13][CH:12]1[CH2:10][O:9][C:4]1[CH:5]=[CH:6][CH:7]=[CH:8][C:3]=1[C:1]#[N:2]
2) Atom-mapping removal and canonicalization	ClCC1CO1.N#Cc1ccccc1O>N1CCCCC1>N#Cc1ccccc1OCC1CO1
3) Tokenization atom-wise	Cl C C 1 C O 1 . N # C c 1 c c c c c 1 O > N 1 C C C C C 1 > N # C c 1 c c cc c 1 O C C 1 C O 1
4) DLG segmentation	ClC C 1 CO1. N#Cc 1cccc c1 O > N1CC CCC 1> N#Cc 1cccc c1 O C C 1 C O 1

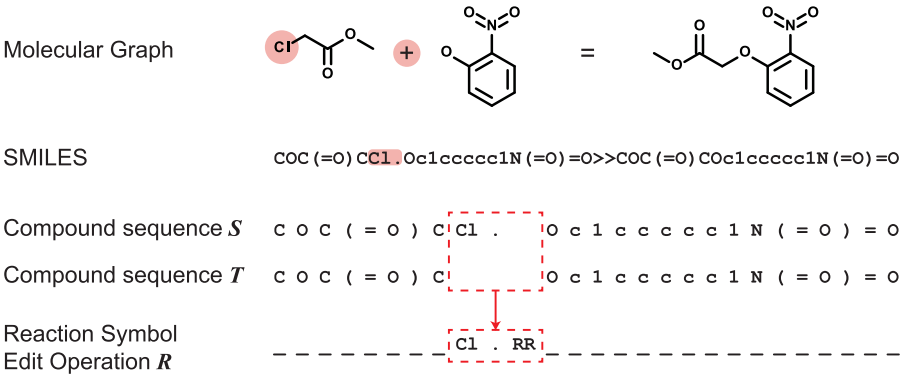


FIGURE 2. Generation of Reaction Symbol Edit Operation (RSEO).

- **AR** indicates the corresponding symbol should be replaced by the given string with the tag.
- **RR** the corresponding symbol should be deleted.
- _ means that there is no operation at the location.

All compound sequences S and T are split into elements, and the resulting RSEO from S to T are illustrated in Figure 2. The data processing steps together with examples are summarized in Table 2. The same preprocessing steps were applied to all datasets.

C. EMBEDDING

In our adopted neural model, an embedding layer is used to map each element or segmented *"word"* from a sequence into a vector with dimension d. Our model takes three types of inputs, reactant, RSEO, and product. After embedding, the reactant sequence with n words is represented as $\mathbb{R}^{d\times n}$. Similarly, we obtain the embeddings of the reactant sequence $\mathbf{R}$, the RSEO sequence $\mathbf{S}$ and the product sequence $\mathbf{P}$. Then, the input sequences are subsequently aggregated into two compact representations through projection and concatenating:

$$\mathbf{M_1} = \begin{bmatrix} \mathbf{R}_1^1 \oplus \mathbf{S}_1^1 \\ \vdots \\ \mathbf{R}_h^1 \oplus \mathbf{S}_h^1 \end{bmatrix}, \quad \mathbf{M_2} = \begin{bmatrix} \mathbf{P}_1^1 \oplus \mathbf{S}_1^1 \\ \vdots \\ \mathbf{P}_h^1 \oplus \mathbf{S}_h^1 \end{bmatrix} \tag{4}$$

D. SIAMESE NETWORK

To learn the optimal representations of chemical reactants $\mathbf{M_1}$ and products $\mathbf{M_2}$ with RSEO, we propose to use a pair-based network structure called Siamese network. It has been proven as a useful framework for image matching [25] and sequence similarity comparison tasks [26], [27]. Since the negative reaction instances are extremely insufficient, and most reported yields concentrate in a narrow range, common neural networks suffer from the imbalance of learning difficulty. The structure of the Siamese network consists of two identical branches that share weights and parameters. Each branch poses a deep neural network for feature learning. In this work, we adopt Long-Short Term Memory (LSTM) Network [28] as the branch component due to its advance for sequence modeling. Figure 3 shows an LSTM based branch architecture.

The LSTM unit is defined as follows.

$$\mathbf{i_t} = \sigma(\mathbf{W}_w^i \mathbf{x}_t + \mathbf{W}_h^i \mathbf{h}_{t-1} + \mathbf{b}_i), \tag{5}$$

$$\mathbf{f_t} = \sigma(\mathbf{W}_w^f \mathbf{x}_t + \mathbf{W}_h^f \mathbf{h}_{t-1} + \mathbf{b}_f), \tag{6}$$

$$\mathbf{u_t} = \sigma(\mathbf{W}_w^u \mathbf{x}_t + \mathbf{W}_h^u \mathbf{h}_{t-1} + \mathbf{b}_u), \tag{7}$$

$$\mathbf{c_t} = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_w^c \mathbf{x}_t + \mathbf{W}_h^c \mathbf{h}_{t-1} + \mathbf{b}_c), \tag{8}$$

$$\mathbf{h_t} = \tanh(\mathbf{c}_t) \odot \mathbf{u}_t, \tag{9}$$

where σ stands for the sigmoid function and the $\odot$ represents element-wiselayer to form a final representation. multiplication. $\oplus$ denotes vector concatenation. $\mathbf{i_t}$, $\mathbf{f_t}$, $\mathbf{u_t}$, $\mathbf{c_t}$, $\mathbf{h_t}$ are the input gates, forget gates, memory cells, output gates and the current states, respectively. For a sequence input, the network computes the hidden sequence $\mathbf{h}_t$ by applying the formulation for each time step.

After embedding, the vectorized inputs M_1 and M_2 are separately fed to forward LSTM and backward LSTM (BiLSTM) to obtain two directions' internal features. The output for

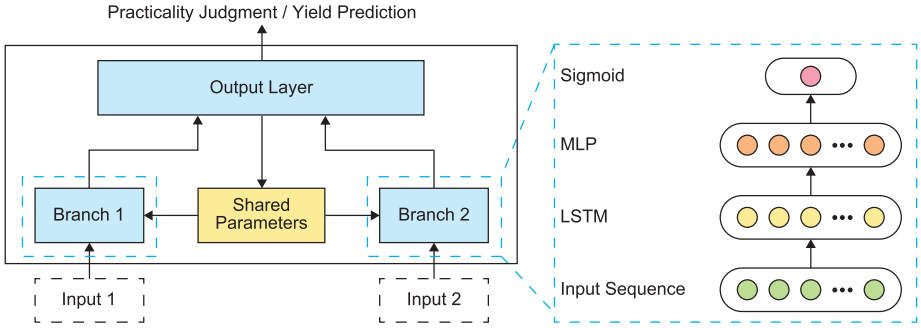


FIGURE 3. Siamese Network with LSTM based branch architecture.

each input is the concatenation of the two vectors from both directions: $\mathbf{h}_t = \overrightarrow{\mathbf{h}_t} \parallel \overleftarrow{\mathbf{h}_t}$. Hence, we have the processed representations of the reactant and product with RSEO, $\hat{M}_1 = BiLSTM(M_1)$ and $\hat{M}_2 = BiLSTM(M_2)$. Then, our model concatenates the representation of $\hat{M}_1$ and $\hat{M}_2$ to a Multi-Layer Perception (MLP) layer to form a final representation. The output of the model is activated by a *sigmoid* function to ensure the prediction is in [0,1].

$$\mathbf{y} = \frac{1}{1 + e^{\mathbf{x}}} \tag{10}$$

where $\mathbf{x}$ is the output of MLP and $\mathbf{y}$ is the prediction.

E. TRAINING OBJECTIVES

For *practicality judgment*, we use binary cross entropy as the loss function.

$$\mathcal{L} = -\frac{1}{N}\sum_{t=1}^{n}\left[\mathbf{y}_t \mathbf{log}\hat{\mathbf{y}}_t + (1 - \mathbf{y}_t)\mathbf{log}(1 - \hat{\mathbf{y}}_t)\right] \tag{11}$$

where $\hat{\mathbf{y}}_t$ denotes the prediction, $\mathbf{y}_t$ is the target and t denotes the data index.

For *yield prediction*, the loss function is the mean square error (MSE).

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2 \tag{12}$$

III. REACTION DATA FOR MODEL EVALUATION

The reaction data for our model evaluation has three sources,[1] (1) a public chemical reaction dataset USPTO, (2) a large scale of reaction dataset extracted from reports of *Chemical Journals with High Impact factors* (CJHIF) (3) real chemical reactions verified by Chemical.AI laboratory.[2]

A. STATISTICS

- **Reactions from USPTO (USPTO)**
  This public chemical reaction dataset was extracted from the US patents grants and applications dating from 1976 to September 2016 [29] by Daniel M. Lowe [30].[3]

[1]https://figshare.com/articles/dataset/original_data_zip/14647095

[2]http://www.chemical.ai

[3]https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873

TABLE 3. Statistics of data from 3 sources.

Data	Original size	Filtered size
USPTO	1,808,938	269,132
CJHIF	3,219,165	1,763,735
CA	12,225	8,797
CA-test	25,001	24,528

The portion of *granted patents* contains 1,808,938 reactions described using SMILES with `TextMindYield` and `CalculatedYield`.[4] Such reaction strings are composed of three groups of molecules: the *reactants*, the *reagents*, and the *products*, separated by a '>' sign. After filtering out the items with inadaptable yield [5] and cleaning with RDKit [31] [6] it remained 269,132 items at last.

- **Reactions from the chemical journals with high impact factor (CJHIF)**
  3,219,165 reactions mined from high impact factor journals[7] with *reagent*, *solvent* and *catalyst* information, in addition with *yield*, besides the *reactants* and the *products*. After data cleaning and selection, we used the remaining 1,763,735 items at last.
- **Reactions from Chemical.AI (CA)**
  12,225 real negative reactions from failed chemistry experiments, and 8,797 reactions remained after data deduplication and canonicalization.
- **Reactions from Chemical.AI for test (CA-test)**
  24,528 real reactions from another filtered chemical experiment record of Chemical.AI partner laboratories, in which there are 16,151 positive reactions and 8,377 negative reactions%.

In our experiments, we set the ratio of the training set and the test as 9:1, and 10% of the training set is held out as development (dev) set.[8] For *practicality judgment*, we use

[4]The `TextMindYield` is where the experimental paragraph explicitly stated the yield. In some cases, only the mass of product is given where the mass or amount (mols) are known for the reactants and products. So `CalculatedYield` can be derived from these quantities.

[5]For example, the yield contains the symbol '>', '~', '<=', only with the value range, or is out of range [0, 100].

[6]An open-source cheminformatics and machine learning tool: http://www.rdkit.org/

[7]The journal list is in the Appendix.

[8]The dev set is used to supervise the training process in case of over-fitting or under-fitting in the deep learning scenario.

TABLE 4. Distributions of positive and negative reaction from the train/dev/test sets in two combinations.

Data	Case	train	dev	test
CJHIF + CA	Positive	1,428,614	158,735	176,386
	Negative	7,137	793	867
USPTO + CA	Positive	218,001	24,222	26,909
	Negative	7,123	791	883
CA-test	Positive ‘	-	-	16,151
	Negative	-	-	8,377

the positive reactions from the USPTO and the CJHIF to pair the negative samples from the CA since there are no negative samples from them. The *yield prediction* task is based on the positive reactions of the USPTO and the CJHIF.

For *practicality judgment*, we let the two positive sets collocate with the negative datasets to form two combinations. The distributions of positive and negative reaction from the train/dev/test sets are in Table 4.

B. SETUP

Considering the calculation efficiency, we specify a max length of 100 words for each SMILES sequence and apply truncating or zero-padding when needed. The embedding weights are randomly initialized with the uniformed distribution in the interval [-0.05, 0.05].

C. EVALUATION METRICS

Because of the unbalance on the positive data and the negative data, we adopt the following metrics: Accuracy, Precision, Recall, and F1-score to evaluate the *practicality judgment*. F1-score has been widely used in the natural language processing literature, which considers both precision and recall and can better reflect the model performance on the positive and negative set individually.

Four types of predictions are as shown in Table 5. Accordingly, we can calculate the performance of Accuracy, Precision, Recall, and F1-score as follows.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

For *yield prediction*, we adopt Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for evaluation.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\mathbf{y} - \hat{\mathbf{y}})^2} \tag{17}$$

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|\mathbf{y} - \hat{\mathbf{y}}\right| \tag{18}$$

TABLE 5. Possible prediction results.

	Predicted Positive	Predicted Negative
True Positive	TP	FN
True Negative	FP	TN

where $\mathbf{y}$ is the actual yield from chemical experiments and $\hat{\mathbf{y}}$ is the predicted one.

IV. EXPERIMENT

A. PRACTICALITY JUDGMENT

Given input sequences, the model will output the reaction success probabilities ranging from [0,1]. To evaluate the result, we require a threshold to distinguish from positive or negative predictions. According to our preliminary experiments, we set the thresholds to 0.5.[9] The experimental result[10] is shown in Table 6. We observe the positive reactions could be recognized essentially (nearly 99%). Though the proportion of positive and negative cases is over 30:1, our model also ensures a high negative F1-score of more than 72%.

1) EFFECT OF RECURRENT NEURAL NETWORK TYPES

We also compare the Siamese network with the different standard recurrent neural networks - LSTM, BiLSTM, GRU, and BiGRU. The comparison of the results is demonstrated in Table 7.

Obviously, the Siamese network outperforms all the others, especially in the negative cases, which shows the Siamese network could effectively handle the data imbalance issue.

2) GENERALIZATION ABILITY

To demonstrate the generation ability of our learning model, we report the judgment results on the CA-test with different training models in Table 8.

As the CA-test comes from the true laboratory record, our model prediction is evaluated in these real chemical experiments. Note that the negative reactions in this dataset were expected to work by experienced chemists, which implies they are correct in chemistry rules literally. But the chemists encounter difficulties in predicting the practicality of these reactions indeed. Therefore, when our model gives correct practicality prediction over these failed reactions, it means that our model performs better than human experts in these cases.

3) INCREMENTAL EXPERIMENT

Different datasets may have different statistical distribution characteristics on reaction types. To thoroughly examine the capacity of our model, we conduct a series of incremental experiments by mixing a small part of the different datasets to the original one and using the rest as the test set.

[9]This is also the common setting for binary classification tasks.

[10]Due to the lack of reaction condition records for negative samples, we exclude the reaction condition embeddings for both positive and negative cases in the *practicality judgment* experiments.

TABLE 6. Results for *practicality judgment* in different datasets (%).

Data (Positive + Negative)	Case	Precision	Recall	F1-score	Accuracy
CJHIF + CA	P	99.82	99.95	99.88	99.76
	N	86.09	63.73	72.24	
USPTO + CA	P	98.83	99.21	99.02	97.92
	N	72.45	63.83	67.88	

TABLE 7. Comparison of F1-score for *practicality judgment* on USPTO + CA (%).

Model	Case	Precision	Recall	F1-score	Accuracy
Siamese	Positive	98.83	99.21	**99.02**	97.92
	Negative	72.45	63.83	**67.88**	
LSTM	Positive	98.78	98.92	98.85	97.77
	Negative	65.24	60.49	63.83	
BiLSTM	Positive	98.87	99.04	98.96	97.97
	Negative	68.87	65.34	67.06	
GRU	Positive	99.13	98.53	98.83	97.74
	Negative	61.92	73.43	67.19	
BiGRU	Positive	98.83	99.18	99.01	97.93
	Negative	71.78	64.08	67.71	

TABLE 8. Results for *practicality judgment* in CA-test dataset (%) using different models.

Training Data	Case	Precision	Recall	F1-score	Accuracy
CJHIF + CA	P	66.00	85.81	74.61	61.49
	N	34.42	14.42	20.32	
USPTO + CA	P	66.19	26.31	37.65	42.98
	N	37.65	34.15	73.99	

TABLE 9. Results for *yield prediction* in different datasets (%).

Dataset	Train Num	Test Num	RMSE	MAE
CA	1,587,359	176,375	19.05	14.85
USPTO 1 (`TextMindYield`)	224,823	24,991	23.00	19.32
USPTO 2 (`CalculatedYield`)	224,823	24,991	23.76	19.95

We divided the CA-test dataset into two parts, the incremental set, and the test set, in the ratio of 1:1. Then, we add different sized parts of the incremental set with ratios [0.1, 0.2, ..., 0.9, 1.0] to the training set (*USPTO + CA*) and conduct the experiments, respectively.

The results in Figure 4 show that even if there is only a small amount of data added into the training set as the same source as the test set, the judgment results will be improved greatly.

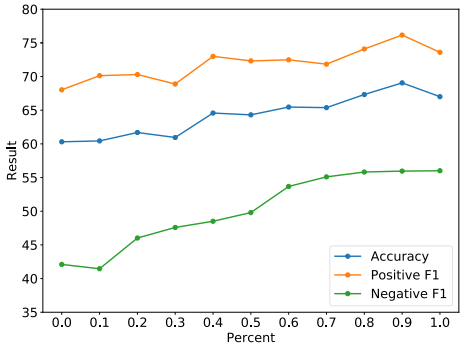


FIGURE 4. The result in our incremental experiment.

B. YIELD PREDICTION

The data statistics and results[11] on the test set are shown in Table 9.[12]

1) EFFECT OF SEGMENTATION

In our model, we take the DLG segmentation to represent the resulting text fragment (i.e., "word") as embedding. We also compare the performance of using DLG or atom-wise for tokenization in Table 10. The comparison indicates a word-like segmentation over chemical text indeed improves the performance on *yield prediction*.

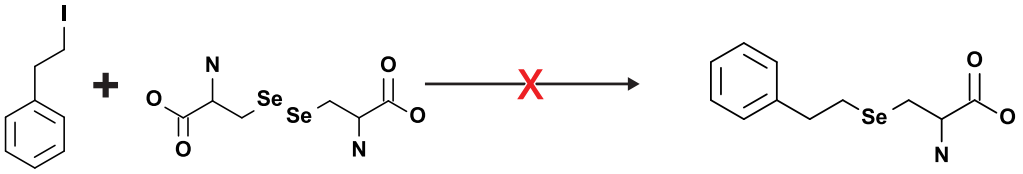
The adopted unsupervised tokenization over the SMILES text is based on "words" with significant DLG scores in terms of the goodness measure methods. Despite its usefulness in our computational process, we also observe their chemical meaning. Table 11 lists a small part of the "words" with high DLG scores. For example, it is not strange to any chemists that the metal complex structure is the key to many organic reactions. In the word list, we find the number and the metallic element are always put in the same fragment, which means that the ligands' position information is attached to the metallic element for better, more sufficient, and more

TABLE 10. Comparison of the results with and without DLG segmentation (%).

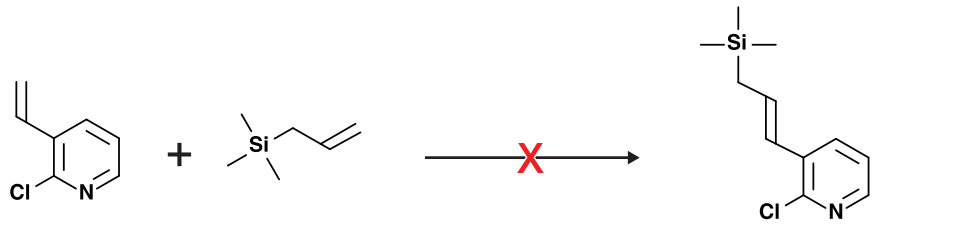
	CA		USPTO 1		USPTO 2	
Tokenization	RMSE	MAE	RMSE	MAE	RMSE	MAE
DLG	19.05	14.85	23.11	19.15	23.54	19.43
Atom-wise	19.71	15.38	23.29	19.24	23.75	19.45

[11] The yield is scaled to [0,1] for better readability.

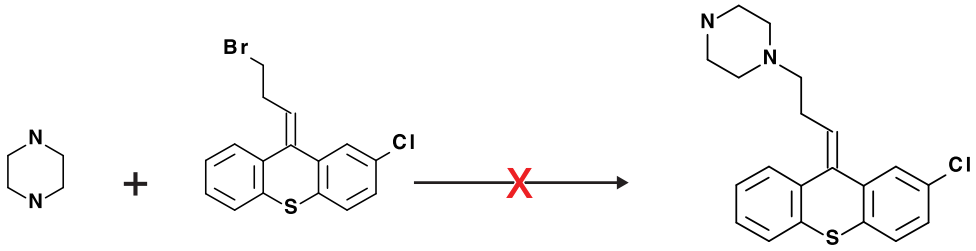
[12] In the following tables, "USPTO 1" means USPTO data with `TextMindYield` and "USPTO 2" means USPTO data with `CalculatedYield`.



ICCc1ccccc1.NC(C[Se][Se]CC(N)C(=O)O)C(=O)O
>>NC(C[Se]CCc1ccccc1)C(=O)O



C=CC[Si](C)(C)C.C=Cc1cccnc1Cl>>C[Si](C)(C)C/C=C/c1cccnc1Cl



C1CNCCN1.Clc1ccc2c(c1)/C(=C\CCBr)c1ccccc1S2
>>Clc1ccc2c(c1)/C(=C\CCN1CCNCC1)c1ccccc1S2

FIGURE 5. Positive-like cases in the test set and "X" means it cannot react actually.

TABLE 11. Examples of SMILES words from DLG segmentation.

Word	DLG Score	Word	DLG Score
[Rh]789%10	78.18	[Ru++]5678	74.61
Mo+6]89%10	68.77	3[Zn++]579	66.74
Mg]Br)cc1.	52.68	ccc3)[Ru++	49.01
C#C[Mg]Br	47.60	C=OBr[Mg]	40.02
\C=C/I	20.09	(C#N)	7.68

accurate embedding representation in our model. In the meanwhile, most ordinary functional groups are also in the same fragment, like C=C, C#C, C=O and C#N, which means the model regards them as a group to process the reaction like what organic chemists do in their research. We also find that the ring structure in a molecule is always divided into different fragments. Though seemingly irrational, the model actually recognizes different functional parts in a ring for more targeted processing, which is indeed helpful to extract the reaction pattern in the later process.

2) ABLATION STUDY

We investigate the effect of different features by removing them one by one. As shown in Table 12, all the features contribute to the performance of our final system. If we remove either *RSEO* or *Reaction Conditions*, the performance drops. This result indicates that both features play an important and complementary role in the feature representation.

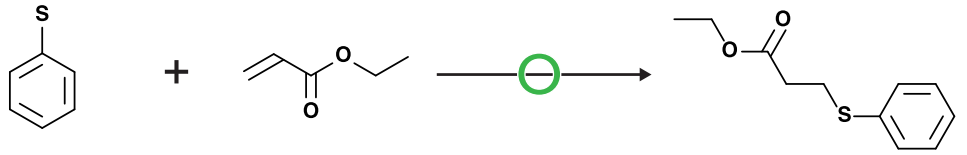
3) PERFORMANCE ON SPECIFIC REACTIONS

To explore the model's ability on specific reactions, we split the CJHIF dataset according to three types of conditions, including *catalyst*, *reagent*, and *solvent*. For each condition, we select the largest five subsets for evaluation. It is observed that our model performs even better in terms of different types of conditions, which shows our model is reasonably capable of learning multiple types of reactions, despite it also learns well from general conditioned reactions.

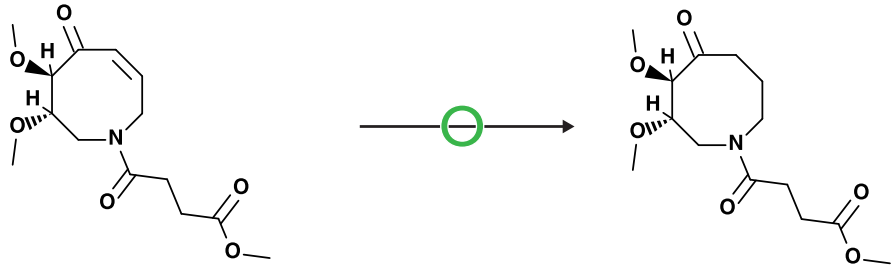
C. EXAMPLES

During the test of *practicality judgment*, we find that the model can recognize some reactions which seem to be impractical but can react actually and some reactions which seem to be reactive but cannot react actually. Figure 5 and Figure 6 show such highly confused examples.

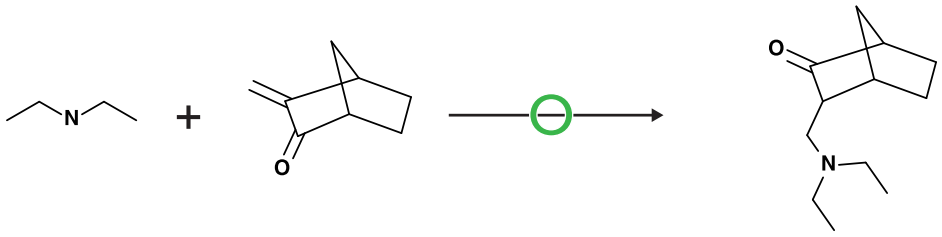
To demonstrate the model performance on *yield prediction* more specifically, we illustrate 10 strong examples and 10 weak examples from CJHIF are in Figure 7 and Figure 8.



CCOC(=O)C=C.SC1=CC=CC=C1>>CCOC(=O)CCSC1=CC=CC=C1



CO[C@H]1CN(C\C=C/C(=O)[C@@H]1OC)C(=O)CCC(=O)OC
>>CO[C@H]1CN(CCCC(=O)[C@@H]1OC)C(=O)CCC(=O)OC



C=C1C2CCC(C2)C1=O.CCNCC>>CCN(CC)CC1C2CCC(C2)C1=O

FIGURE 6. Negative-like cases in the test set and "O" means it can react actually.

TABLE 12. Ablation study for *yield prediction* (%).

	CJHIF		USPTO 1		USPTO 2	
Features	RMSE	MAE	RMSE	MAE	RMSE	MAE
Full	19.05	14.85	23.11	19.15	23.54	19.43
w/o *RSEO*	19.10	14.67	23.22	19.08	23.79	19.07
w/o *Conditions*	19.07	15.22	23.25	19.11	23.76	19.70
w/o *RSEO* & *Conditions*	19.41	15.57	23.24	19.22	23.64	19.42

TABLE 13. Results on different types of reactions (%).

	Name	Train size	Test size	RMSE	MAE
catalyst	palladium on activated charcoal	15,707	1,731	17.75	13.51
	palladium diacetate	46,969	5,213	18.11	14.15
	tetrakis(triphenylphosphine) palladium	28,590	3,158	19.03	15.05
	bis-triphenylphosphine-palladium(ii) chloride	16,559	1,828	18.11	14.45
	copper(I) iodide	37,559	4,162	18.23	14.45
reagent	triethylamine	102,970	11,425	18.97	14.67
	potassium carbonate	68,207	7,572	19.11	14.78
	hydrogenchloride	43,035	4,766	19.08	14.98
	pyridine	47,028	5,213	19.21	14.73
	palladium diacetate	46,973	5,207	17.98	13.99
solvent	tetrahydrofuran	398,696	44,275	18.05	14.03
	dichloromethane	306,391	34,019	18.05	13.87
	water	178,799	19,849	19.00	14.72
	toluene	143,386	15,911	18.46	14.14
	N,N-dimethyl-formamide	125,178	13,897	18.92	14.82

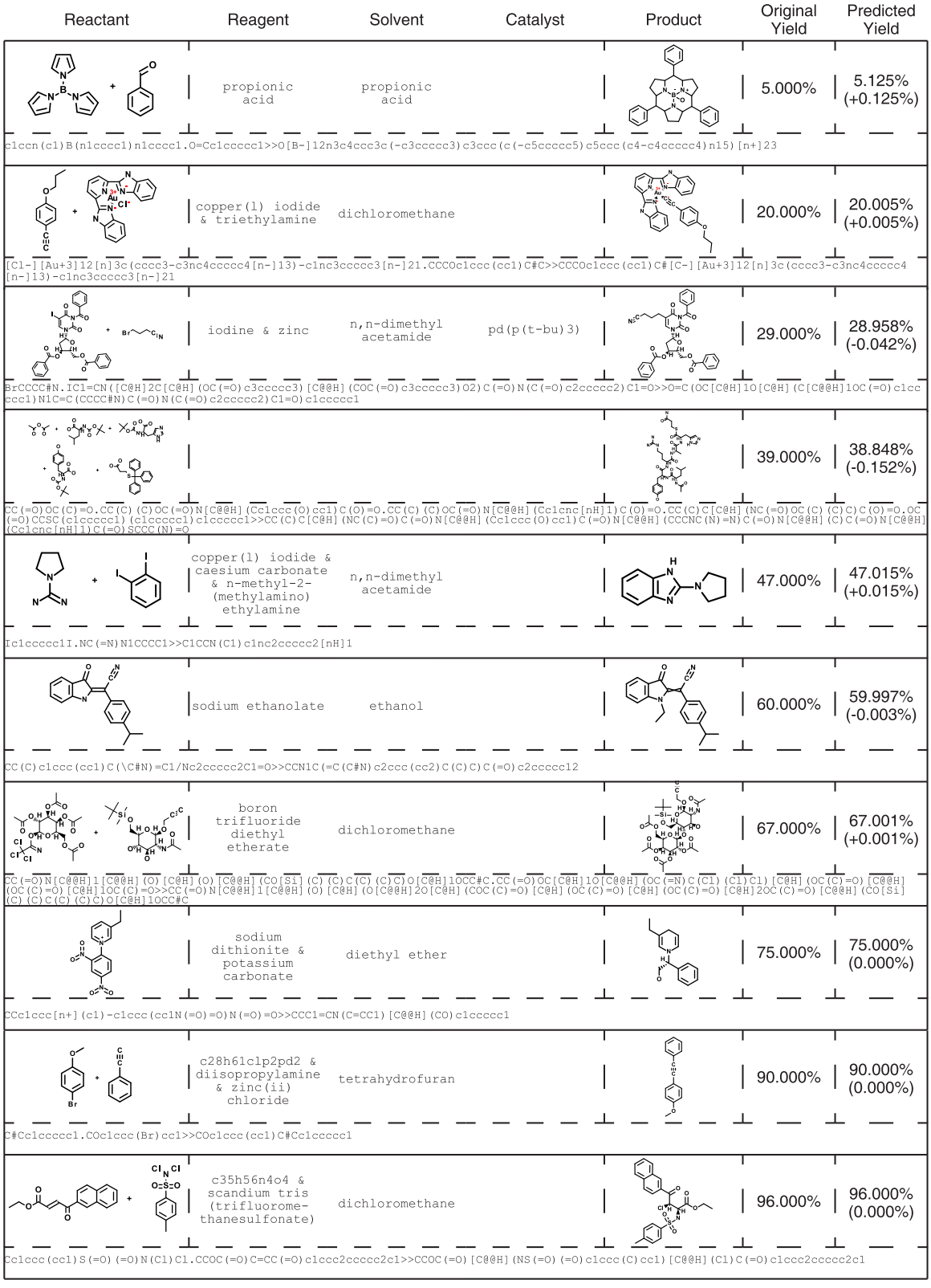


FIGURE 7. Strong examples on *yield prediction*.

We select the most strong and most weak examples every 10% yield range by sorting the absolute error between the original yield and the predicted yield and guarantee the diversity of the reaction types simultaneously. It shows that our model would not perform well on specific reaction categories.

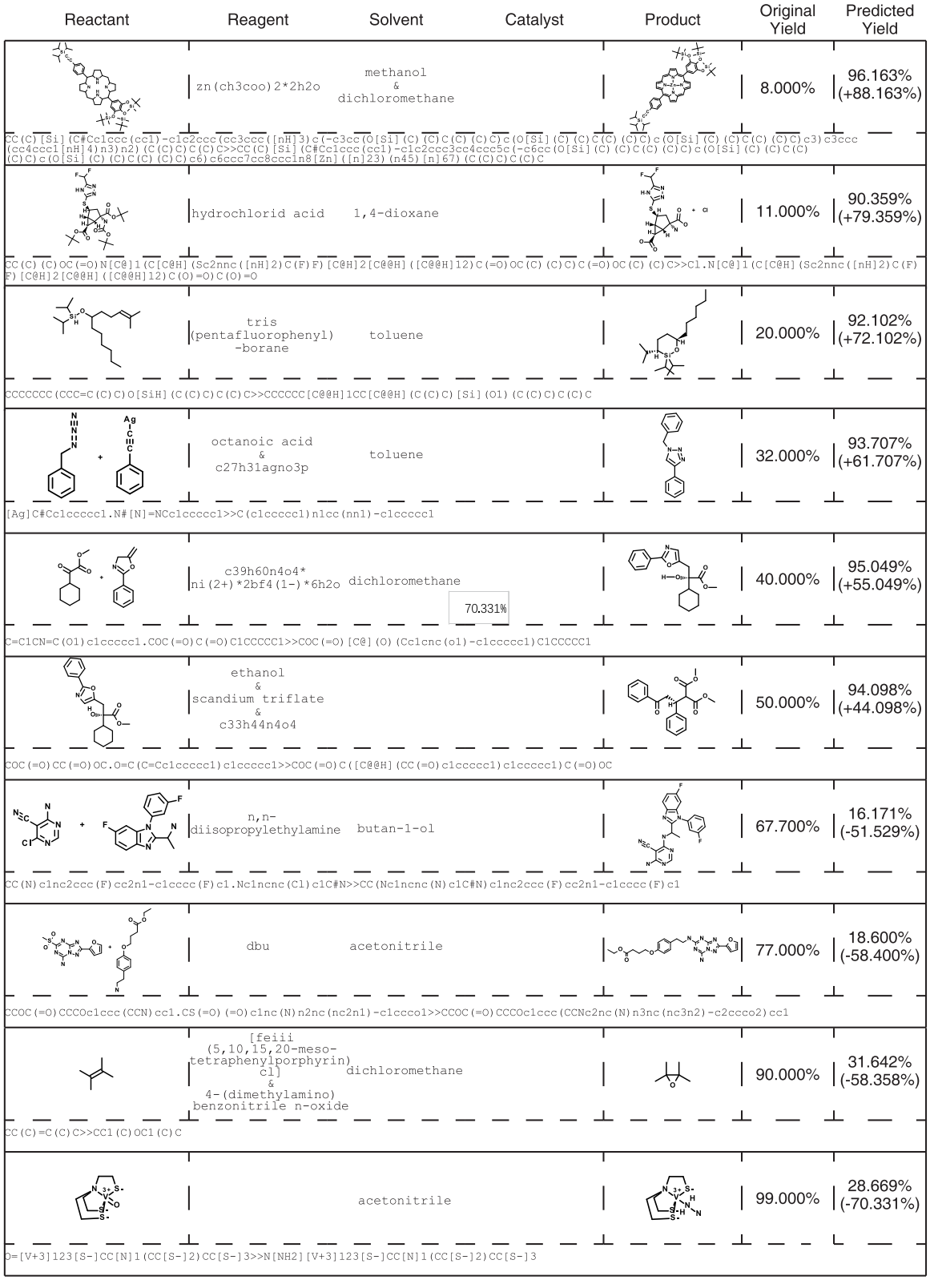


FIGURE 8. **Weak examples on *yield prediction*.**

V. DISCUSSION

Of course, the practicality and yield of chemical reactions are affected by various environmental conditions, such as temperature, pressure, catalysts, and solvent. But our model provides a heuristic method that treats the chemical reactions as the plain text and offers a quick judgment by deep learning

mechanism. This model is sensitive to the order and form of the reactants due to adopting the NLP method. Moreover, the scores gained from the models trained by the different training sets have relatively large differences. In our future work, we will improve performance and fix the problems.

VI. CONCLUSION

We present a deep learning model to model real-world chemical reactions and unearth the factors governing reaction outcomes only from symbol representation of chemical information. Compared to conventional methods that require massive manual features or are only evaluated on small datasets for specific reaction types, our approach is much more straightforward, end-to-end, and effective. In a distinctive perspective, this work reveals the great potential to employ the deep learning method to help chemists judge the practicality of chemical reactions and develop more efficient experimental strategies to reduce the cost of invalid experiments. The resultant model can be used more than yield prediction but has the potential to help effective synthesis route design by simply searching the highest yield reactions among a large scale of automatically generated synthesis schemes, which has been an ongoing task in our current study and chemical practice.[13]

ACKNOWLEDGMENT

(Shu Jiang and Zhuosheng Zhang contributed equally to this work.)

APPENDIX

JOURNAL LIST FOR CJHIF DATASET

- Journal of Medicinal Chemistry
- Journal of Organic Chemistry
- Chemistry - A European Journal
- Journal of the American Chemical Society
- New Journal of Chemistry
- Organic Letters
- European Journal of Organic Chemistry
- Chemical Communications
- Angewandte Chemie - International Edition

REFERENCES

[1] G. Casiraghi, L. Battistini, C. Curti, G. Rassu, and F. Zanardi, "The vinylogous aldol and related addition reactions: Ten years of progress," *Chem. Rev.*, vol. 42, no. 36, pp. 3076–3154, 2011.

[2] Z. A. El-Rub, E. A. Bramer, and G. Brem, "Review of catalysts for tar elimination in biomass gasification processes," *Ind. Eng. Chem. Res.*, vol. 43, no. 22, pp. 6911–6919, Oct. 2004.

[3] S. Caron, N. M. Do, J. E. Sieser, D. C. Whritenour, and P. D. Hill, "Preparation of a corticotropin-releasing factor antagonist by nucleophilic aromatic substitution and copper-mediated ether formation," *Organic Process Res. Develop.*, vol. 13, no. 2, pp. 324–330, Mar. 2009.

[4] H. Yao, H. Zhang, L. Ye, W. Zhao, S. Zhang, and J. Hou, "Dialkylthio substitution: An effective method to modulate the molecular energy levels of 2D-BDT photovoltaic polymers," *ACS Appl. Mater. Interfaces*, vol. 8, no. 6, pp. 3575–3583, Feb. 2016.

[13]Our prediction models have been online, the link is http://bcmi.sjtu.edu.cn/ dl4chem

[5] G. C. Fu, "Transition-metal catalysis of nucleophilic substitution reactions: A radical alternative to S_N1 and S_N2 processes," *ACS Central Sci.*, vol. 3, no. 7, pp. 692–700, Jul. 2017.

[6] O. Wiest, D. C. Montiel, and K. N. Houk, "Quantum mechanical methods and the interpretation and prediction of pericyclic reaction mechanisms," *J. Phys. Chem. A*, vol. 101, no. 45, pp. 8378–8388, Nov. 1997.

[7] Z.-L. Song, C.-A. Fan, and Y.-Q. Tu, "Semipinacol rearrangement in natural product synthesis," *Chem. Rev.*, vol. 111, no. 11, pp. 7523–7556, Nov. 2011.

[8] J. A. Rincón, M. Barberis, M. Gonzálezesguevillas, M. D. Johnson, J. K. Niemeier, and W. M. Sun, "Safe, convenient *ortho*-claisen thermal rearrangement using a flow reactor," *Organic Process Res. Develop.*, vol. 15, no. 6, pp. 1428–1432, 2011.

[9] Z. Flisak and W.-H. Sun, "Progression of diiminopyridines: From single application to catalytic versatility," *ACS Catalysis*, vol. 5, no. 8, pp. 4713–4724, Aug. 2015.

[10] S. E. Denmark, "The interplay of invention, discovery, development, and application in organic synthetic methodology: A case study," *J. Organic Chem.*, vol. 74, no. 8, pp. 2915–2927, Apr. 2009.

[11] A. Streitwieser, "Perspectives on computational organic chemistry," *J. Organic Chem.*, vol. 74, no. 12, pp. 4433–4446, Jun. 2009.

[12] R. E. Plata and D. A. Singleton, "A case study of the mechanism of alcohol-mediated Morita Baylis–Hillman Reactions. The importance of experimental observations," *J. Amer. Chem. Soc.*, vol. 137, no. 11, pp. 3811–3826, Mar. 2015.

[13] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, "Machine-learning-assisted materials discovery using failed experiments," *Nature*, vol. 533, no. 7601, pp. 73–76, May 2016.

[14] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of organic reaction outcomes using machine learning," *ACS Central Sci.*, vol. 3, no. 5, pp. 434–443, May 2017.

[15] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, "Neural networks for the prediction of organic chemistry reactions," *ACS Central Sci.*, vol. 2, no. 10, pp. 725–732, Oct. 2016.

[16] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Sci.*, vol. 4, no. 2, pp. 268–276, 2018.

[17] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, and T. Laino, "'Found in Translation': Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models," 2017, *arXiv:1711.04810*. [Online]. Available: http://arxiv.org/abs/1711.04810

[18] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, "Predicting reaction performance in C–N cross-coupling using machine learning," *Science*, vol. 360, pp. 186–190, Apr. 2018. [Online]. Available: http://science.sciencemag.org/content/early/2018/02/14/science.aar5169

[19] M. H. S. Segler and M. P. Waller, "Neural-symbolic machine learning for retrosynthesis and reaction prediction," *Chem. A, Eur. J.*, vol. 23, no. 25, pp. 5966–5971, May 2017, doi: 10.1002/chem.201605499.

[20] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chem. Inf. Model.*, vol. 28, no. 1, pp. 31–36, Feb. 1988.

[21] H. Zhao and C. Kit, "An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework," in *Proc. 3rd Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2008, pp. 9–16. [Online]. Available: http://www.aclweb.org/anthology/I08-1002

[22] C. Kit and Y. Wilks, "The virtual corpus approach to deriving ngram statistics from large scale corpora," in *Proc. Int. Conf. Chin. Inf. Process.*, 1998, pp. 223–229.

[23] C. E. Shannon, *A Mathematical Theory of Communication*, vol. 27. Oxford, U.K.: Blackwell, 1948. [Online]. Available: http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x

[24] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, Mar. 1970.

[25] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 378–383.

[26] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2786–2792.

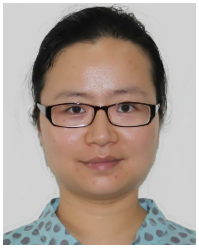
[27] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.
[28] S. Hochreiter and J. Schmidhuber, *Long Short-Term memory*. Berlin, Germany: Springer, 1997.
[29] D. Lowe. (Jun. 2017). *Chemical Reactions From U.S Patents (1976-sep2016)*. Chemical_reactions_from_US_patents_1976-Sep2016_
[30] D. M. Lowe, "Extraction of chemical structures and reactions from the literature," Ph.D. dissertation, Dept. Chem., Univ. Cambridge, Cambridge, U.K., 2012.
[31] G. Landrum *et al.*, "rdkit/rdkit: 2017_09_1 (Q3 2017) release," Zenodo, Tech. Rep., Oct. 2017, doi: 10.5281/zenodo.1004356.

SHU JIANG received the B.S. degree in instrument science and technology from Southeast University, China, in 2015. She is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University. Her research interests include machine translation and natural language processing.

ZHUOSHENG ZHANG received the bachelor's degree in Internet of Things from Wuhan University, in 2016, and the M.S. degree in computer science from Shanghai Jiao Tong University, in 2020. He is currently pursuing the Ph.D. degree in computer science with the Center for Brain-Like Computing and Machine Intelligence, Shanghai Jiao Tong University. He was an Internship Research Fellow with NICT, from 2019 to 2020. His research interests include natural language processing, machine reading comprehension, dialogue systems, and machine translation.

HAI ZHAO received the B.Eng. degree in sensor and instrument engineering and the M.Phil. degree in control theory and engineering from Yanshan University, Qinhuangdao, China, in 1999 and 2000, respectively, and the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2005. He was a Research Fellow with the City University of Hong Kong, from 2006 to 2009, a Visiting Scholar with Microsoft Research Asia, in 2011, and a Visiting Expert with NICT, Japan, in 2012. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, after he joined the university, in 2009. His research interests include natural language processing and related machine learning, data mining, and artificial intelligence. He is also an ACM Professional Member. He has served as an Area Co-Chair for ACL 2017 on Tagging, Chunking, Syntax and Parsing, and a (Senior) Area Chair for ACL 2018, 2019 on Phonology, Morphology and Word Segmentation.

JIANGTONG LI received the bachelor's degree in chemistry from Shanghai Jiao Tong University, in 2019. He is currently pursuing the Ph.D. degree in computer science with the Center for Brain-Like Computing and Machine Intelligence, Shanghai Jiao Tong University. He was a Research Intern with Tencent AI Lab, from January 2019 to September 2019. His research interests include natural language processing, computer version, multi-modal retrieval, multi-modal reasoning, and semantic segmentation.



YANG YANG (Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, China, in 2009. She was a Visiting Scholar with the University of California, Riverside, from 2012 to 2013. She is currently an Associate Professor with the Department of Computer Science, Shanghai Jiao Tong University. Her research interests include machine learning and bioinformatics.

BAO-LIANG LU (Fellow, IEEE) received the B.S. degree in instrument and control engineering from the Qingdao University of Science and Technology, Qingdao, China, in 1982, the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 1989, and the D.Eng. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1994. He was with the Qingdao University of Science and Technology, Qingdao, from 1982 to 1986. From 1994 to 1999, he was a Frontier Researcher with the Bio-Mimetic Control Research Center, Institute of Physical and Chemical Research (RIKEN), Nagoya, Japan, and a Research Scientist with the RIKEN Brain Science Institute, Wako, Japan, from 1999 to 2002. Since 2002, he has been a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He has also been an Adjunct Professor with the Laboratory for Computational Biology, Shanghai Center for Systems Biomedicine, Shanghai, since 2005. His current research interests include brain-like computing, neural networks, machine learning, computer vision, bioinformatics, brain–computer interface, and affective computing. He is also a Board Member of APNNA. He was the President of the Asia–Pacific Neural Network Assembly (APNNA) and the General Chair of the 18th International Conference on Neural Information Processing, in 2011. He is also an Associate Editor of the IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS AND NEURAL NETWORKS.

NING XIA received the Ph.D. degree in organic chemistry from Montpellier University, France. He worked in chemoinformatics in different companies for more than ten years. In 2018, he founded Chemical.AI and the company mainly focuses on computer-aided synthesis planning systems.