REACTCLASS: Cross-Modal Supervision for Subword-Guided Reactant Entity Classification

Xuan Wang[1], Vivian Hu[1], Minhao Jiang[1], Yu Zhang[1], Jinfeng Xiao[1], Danielle Cherrice Loving[2], Heng Ji[1], Martin Burke[2], Jiawei Han[1]
[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[2]Department of Chemistry, University of Illinois at Urbana-Champaign
[1]{xwang174,vivianh2,minhaoj,yuz9,jxiao13,hengji,hanj}@illinois.edu, [2]{dloving,mdburke}@illinois.edu

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995489

***Abstract*—We propose REACTCLASS that automatically maps the low-level concrete chemical entities into the high-level reactant groups without human effort for training data annotation. REACTCLASS is designed to take two special characteristics of the chemical molecules into consideration. The first characteristic is that each chemical molecule can be represented in two modalities: a chemical name in the text and a molecular structure in the graph. We propose to use cross-modal supervision to automatically create the training data for chemical name classification in the text via molecular structure matching in the graph. The second characteristic is that there is a knowledge-aware subword correlation between the surface names of the chemical entities to be classified and that of the reactant groups as class labels. We propose to train a classification model based on the subword cross-attention map between each chemical name and the corresponding reaction group. Experiments demonstrate that REACTCLASS is highly effective, achieving state-of-the-art performance in classifying the chemical names into human-defined reactant groups without requiring human effort for training data annotation.**

***Index Terms*—Chemistry Text Mining; Cross-Modal Supervised Learning; Attention Map Representation**

I. INTRODUCTION

Scientific knowledge can be described on various levels of abstractions: from high-level categorical concepts to low-level concrete entities. For example, in Figure 1, the Csp^3-Csp^3 Suzuki cross-coupling reaction is defined by chemists as a process involving a pair of high-level reactant groups (i.e., the M-side reactant group "primary alkyl boronate" and the X-side reactant group "primary alkyl halide"). While in the chemistry literature, this chemical reaction can also be described as a process involving two low-level concrete chemical entities (e.g., "1-bromododecane" and "B-n-octyl-9-BBN"). This gap between high-level and low-level abstractions of scientific knowledge is a common phenomenon in various domains such as biology, chemistry, and physics.

We propose REACTCLASS that bridges this gap by automatically mapping the knowledge descriptions from low-level concrete entities to high-level categorical concepts without requiring human effort for training data annotation. REACTCLASS benefits various downstream applications, such as chemistry knowledge base completion [11], chemistry information retrieval [4], [9], [14], and prediction of chemical reactions, products, and properties [1], [2], [5], [12].

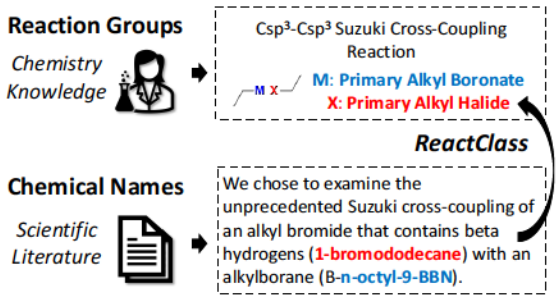

Fig. 1: REACTCLASS automatically classifies the concrete chemical entities in the text into the high-level reactant groups defined by the chemical scientists.

REACTCLASS is designed to take two special characteristics of the chemical molecules into consideration. First, we propose to use cross-modal supervision to automatically create the training data for chemical name classification in the text via molecular structure matching in the graph. Second, we propose to train a classifier based on the knowledge-aware subword cross-attention map between each chemical name and its corresponding reaction group. REACTCLASS is highly effective, achieving state-of-the-art performance in classifying the chemical names into human-defined reactant groups without requiring human effort for training data annotation.

II. REACTCLASS: METHODOLOGY

REACTCLASS consists of two parts: (1) automatic training data creation with molecular structure matching, (2) classification model with knowledge-aware subword cross-attention.

A. Cross-Modal Supervision of Molecular Structure Matching

By definition in chemistry knowledge, the training data for each reactant group can be automatically created by finding the chemical names with graph representations that match the graph representation of the reactant group.

To convert the chemical names into their graph representations, we first collect a large number of candidate chemical names [e.g., "tert-butyl (4-bromo-2-nitrophenyl) carbamate" in Figure 2] from both the chemical reaction knowledge base (Reaxys[1] [6]) and the chemi-

[1]https://www.reaxys.com/#/search/quick

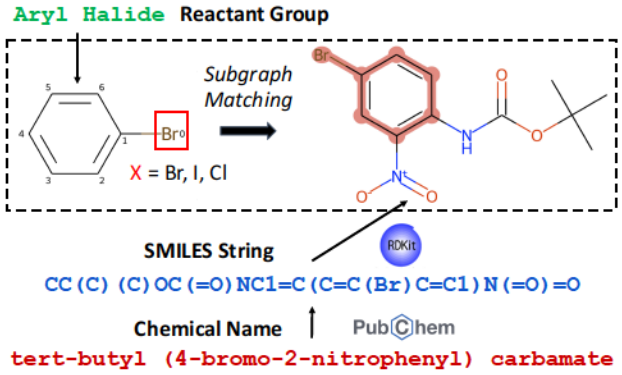

Fig. 2: Illustration of cross-modal supervision of molecular structure matching.

cal named entity recognition (ChemNER [13]) results in the chemistry literature. Then we convert the chemical names into their corresponding SMILES strings [e.g., "CC(C)(C)OC(=O)NC1=C(C=C(Br)C=C1)N(=O)=O" in Figure 2], a character-based sequence representation of the chemical molecules. This chemical name to SMILES string conversion is automatically done by linking the chemical names to a chemistry knowledge base (PubChem[2] [8]) where we can directly find their corresponding SMILES strings. Finally, the SMILE strings can be converted into molecular structures for the next step of subgraph matching, using an open-source cheminformatics software RDKit[3].

To convert the human-defined reactant groups into their graph representations, we first get the ten reactant groups (e.g., "Aryl Halide" in Figure 2) for Suzuki cross-coupling reactions from chemists. Then we convert the reactant group into a subgraph regular expression. For example, in Figure 2, the reactant group "Aryl Halide" can be converted into a subgraph regular expression of a benzene ring with a Br connected to carbon #1, where this Br can be replaced with either I or Cl. By definition in chemistry knowledge, a candidate chemical name belongs to a reactant group if any subgraphs in its molecular structure match the subgraph regular expression of that reactant group. So any chemical names (e.g., "tert-butyl (4-bromo-2-nitrophenyl) carbamate") with a molecular structure that can match the graph representation of "Aryl Halide" can be classified into the "Aryl Halide" reactant group. The subgraph regular expressions of the ten reactant groups are also defined by chemists.

After we get the graph representations of both the chemical names and the reactant groups, we use the RDKit software to conduct the subgraph matching in the molecular structures.

B. Subword Cross-Attention-Guided Chemical Classification

Based on the training data obtained from the previous step of subgraph matching, we observe a knowledge-aware subword correlation between the chemical names to be classified and the reactant groups as class labels. For example, in the bottom part of Figure 3, we construct a subword cross-attention map between the chemical name "tert-butyl (4-bromo-2-nitrophenyl) carbamate" and its corresponding reactant group "Aryl Halide". Looking at this subword cross-attention map, we observe that the subword string "phenyl" in the chemical name is highly correlated with the subword string "Aryl" in the reactant group. This is well-aligned with the chemistry knowledge: "Aryl" means any species created by removing a hydrogen atom from an aromatic hydrocarbon and "phenyl" is a specific aryl radical, which is created by removing a hydrogen atom from a benzene ring. Similarly, we can observe that the subword string "bromo" in the chemical name is highly correlated with the subword string "halide" in the reactant group. This observation of knowledge-aware subword correlation also motivates us to train a classifier based on the subword cross-attention maps between the chemical names and the reaction groups.

We first describe how we construct the cross-attention map. Taken each chemical name $e_i = \langle w_1, w_2, ..., w_{|e_i|} \rangle$ and reactant group $g_j = \langle w_1, w_2, ..., w_{|g_j|} \rangle$ as a sequence of subword tokens w_k, we first extract the representation for each subword token from last hidden states of the fine-tuned ChemBERT model. Then we calculate the cosine similarities between the representations of subword tokens in the chemical names and the reactant groups. Specifically, we obtain the cross-attention matrix as follows.

$$q = \mathbf{x}_{\text{group}}\mathbf{W}_q, k = \mathbf{x}\mathbf{W}_k, v = \mathbf{x}\mathbf{W}_v \tag{1}$$

$$\mathbf{A} = \text{softmax}(qk^T/\sqrt{C/h}) \tag{2}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C\times(C/h)}, C, h$ are the embedding dimension and the number of heads, $\mathbf{x}$ is representation of the input chemical name, and $\mathbf{x}_{\text{group}}$ is the representation of the input reactant group. The attention matrix $\mathbf{A}$ will be used as our input feature for the classification model training.

Taken the attention matrix $\mathbf{A}$ constructed above, we then describe our design of the classification model. We consider each cross-attention map as a single-channel image and encode it with a three-layer CNN, transforming the text classification task into an image classification task for the final prediction. For each chemical name, we first create positive training data by constructing a cross-attention matrix between the chemical name and its corresponding reaction group from subgraph matching. We then create negative training data by constructing a cross-attention matrix between the chemical name and other group names. Our learning task is a binary classification task. For the learning objectives, we adopt binary-class cross-entropy loss for simplicity with our created training data. Thus, the training loss function for the classification model can be formulated as

$$\mathcal{L} = -\frac{1}{|\mathbb{D}|}\sum_{x_i,y_i\in\mathbb{D}} y_i \log(f(x_i)) \tag{3}$$

$$+ (1-y_i)\log(1-f(x_i)) \tag{4}$$

where $\mathbb{D} = \{(x_i, y_i)\}$ is the training dataset, x_i is the attention matrix, and $y_i \in \{+1, -1\}$.

[2]https://pubchem.ncbi.nlm.nih.gov/
[3]https://www.rdkit.org/

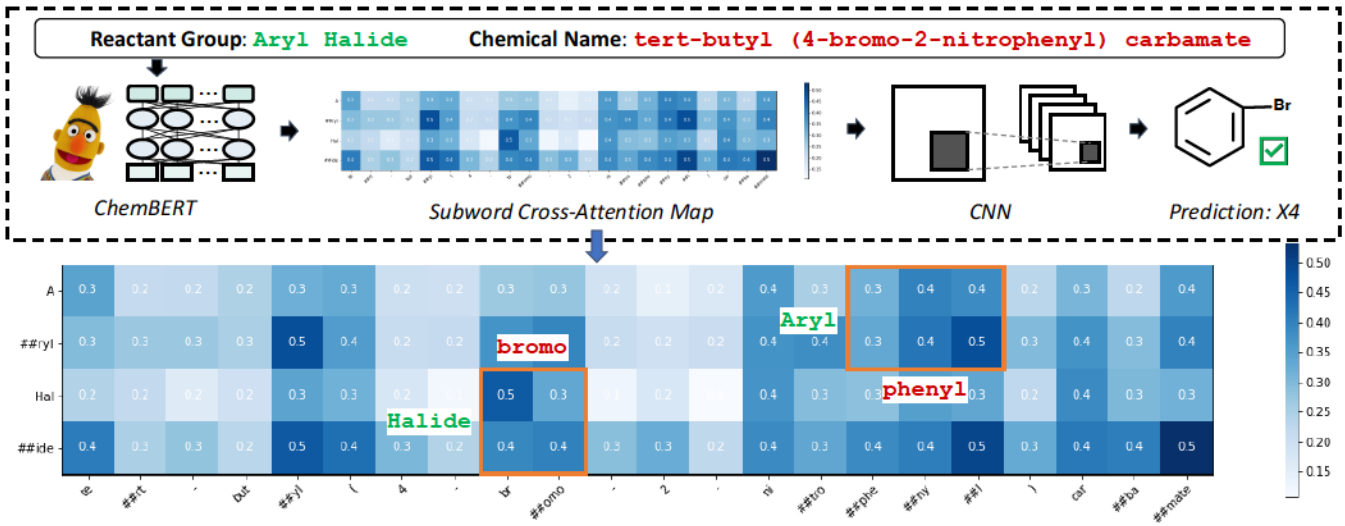

Fig. 3: Illustration of subword cross-attention-guided name classification.

TABLE I: Dataset Statistics

Dataset	Suzuki Coupling
# Training Samples	30,488
# Validation Samples	3,855
# Testing Samples	3,858
# Groups	11

During inference, we compute the scores of the attention matrices between each chemical name and all the ten reactant groups to find the reactant group with the highest probability.

III. Experiments

A. Dataset

We create a dataset for our task of chemical name classification. We first get ten reactant groups from chemistry experts to serve as our class labels. Each reactant group has a corresponding reactant group name (e.g., "M1" is "Primary Boronate") that can be used in our experiments. Then we collect the chemical names to be classified from both the reaction database (Reaxys [6]) and the named entity recognition (ChemNER [13]) results in chemistry literature. Last, following the training data creation process described in Section II-A, we automatically create around 38K training data for the ten reactant groups plus an "Other" class with cross-modal supervision of molecular structure matching. We split the 38K training data into training/validation/test sets with a ratio of 8:1:1. The dataset details can be found in Table I.

B. Baselines

We compare the performance of ReactClass with several baseline methods as follows.

- **BERT/BioBERT/ChemBERT + Softmax**: This is a simple baseline method that directly uses the output states of a pre-trained language model plus a linear layer for prediction. We explored various pre-trained language models in different domains (e.g., BERT [3] in the general domain, BioBERT [10] in the biomedical domain, and ChemBERT [7] in the chemistry domain).
- **ChemBERT + Triplet Loss**: To tackle the class imbalance problem in our training data, we tried the triplet loss that is less sensitive to the imbalanced training data compared to the softmax loss.
- **ChemBERT + Softmax (Oversampling)**: To further tackle the class imbalance problem in our training data, we leverage the weighted bootstrapping strategy to ensure that the models receive about the same number of data in each class during training.
- **Subword + CNN + Softmax**: This is our proposed method that takes the subword cross-attention map between each chemical name and the corresponding reaction group as the input feature and then encodes it with a 3-layer CNN for the final prediction.
- **Subword + CNN + Softmax (Oversampling)**: This is our final model that has the same architecture as Subword + CNN + Softmax, only adding the weighted bootstrapping strategy to further tackle the class imbalance problem in our training data.

We use the micro-F1 and macro-F1 scores[4] as the evaluation metrics for our performance comparison.

C. Main Results

Table II shows the main results on the test set of our chemical name classification dataset. Comparing different pre-trained language models (BERT/BioBERT/ChemBERT + Softmax), the domain-specific pre-trained language model achieves better performance than that is trained in the general domain. Comparing ChemBERT + Softmax, ChemBERT + Triplet Loss, and ChemBERT + Softmax (Oversampling), both the triplet loss and the oversampling strategy are effective in

[4]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

TABLE II: Main Results (F1 Scores in %)

Baseline	Micro F1	Macro F1
BERT + Softmax	97.72	82.83
BioBERT + Softmax	97.85	85.00
ChemBERT + Softmax	97.95	84.20
ChemBERT + Triplet Loss	98.16	88.25
ChemBERT + Softmax (Oversampling)	98.16	89.46
Subword + CNN + Softmax (**REACTCLASS**)	98.28	83.44
Subword + CNN + Softmax (**REACTCLASS** + Oversampling)	**98.56**	**90.76**

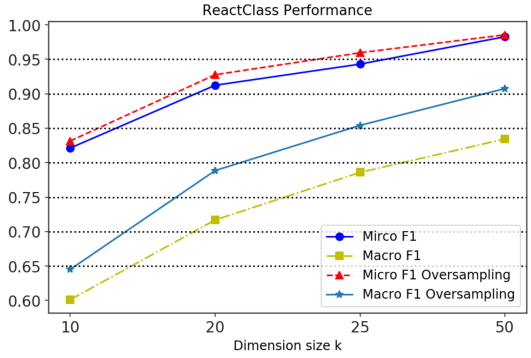

Fig. 4: Parameter Studies on Dimension K

dealing with the class imbalance problem in our automatically created training data. The oversampling strategy is the most effective one that brings the most performance improvement. Our final model (REACTCLASS + Oversampling) achieves 98.56% micro-F1 and 90.76% macro-F1 scores with significant performance improvements compared with all the baseline methods. It demonstrates the effectiveness of our proposed method that takes the subword cross-attention maps between the chemical names and the reaction groups as the input feature for classification.

D. Parameter Studies

We perform several experiments on the hyperparameters in our framework to study the efficacy of our classification model. One important hyperparameter is the dimension of our cross-attention matrix K during training and inference. Due to a large amount of computation in calculating cross-attention maps, large dimensions result in much longer computation time while small dimensions may cause information loss. We conducted experiments on a dimension size K from 10 to 50. In Figure 4, we observe that a larger dimension K will lead to better performance. We use $K = 50$ in all our experiments.

IV. CONCLUSIONS

In this work, we propose a highly effective method, REACTCLASS, for reactant entity classification without requiring human effort for training data annotation. REACTCLASS is designed to take two special characteristics, multi-modal representation and knowledge-aware subword correlation, of the chemical molecules into consideration. Our method achieves state-of-the-art performance in classifying the chemical names into ten Suzuki cross-coupling reactant groups.

ACKNOWLEDGMENTS

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

[1] 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*, 360(6385):186–190, 2018.

[2] 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 science*, 3(5):434–443, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[4] C. Edwards, C. Zhai, and H. Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, 2021.

[5] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, and P. Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 3(3):442–452, 2018.

[6] J. Goodman. Computer software review: Reaxys, 2009.

[7] J. Guo, A. S. Ibanez-Lopez, H. Gao, V. Quach, C. W. Coley, K. F. Jensen, and R. Barzilay. Automated chemical reaction extraction from scientific literature. *Journal of Chemical Information and Modeling*, 2021.

[8] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 11 2020.

[9] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, and A. Valencia. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117(12):7673–7761, 2017.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[11] E. K. Mallory, M. de Rochemonteix, A. Ratner, A. Acharya, C. Re, R. A. Bright, and R. B. Altman. Extracting chemical reactions from text using snorkel. *BMC bioinformatics*, 21:1–15, 2020.

[12] H. Wang, W. Li, X. Jin, K. Cho, H. Ji, J. Han, and M. Burke. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*, 2022.

[13] X. Wang, V. Hu, X. Song, S. Garg, J. Xiao, and J. Han. ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[14] Y. Zhou, B. Zhou, S. Jiang, and F. J. King. Chemical-text hybrid search engines. *Journal of chemical information and modeling*, 50(1):47–54, 2010.