

RESEARCH ARTICLE

Enhancing Generic Reaction Yield Prediction through Reaction Condition-Based Contrastive Learning

Xiaodan Yin^{1,4†}, Chang-Yu Hsieh^{2*†}, Xiaorui Wang^{1,4}, Zhenxing Wu^{2,4}, Qing Ye^{2,4}, Honglei Bao¹, Yafeng Deng⁴, Hongming Chen⁵, Pei Luo¹, Huanxiang Liu³, Tingjun Hou^{2*}, and Xiaojun Yao^{3*}

¹Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery, State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macao 999078, China. ²Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China. ³Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China. ⁴CarbonSilicon AI Technology Co. Ltd, Hangzhou, Zhejiang 310018, China. ⁵Center of Chemistry and Chemical Biology, Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou 510530, China.

*Address correspondence to: xjyao@mpu.edu.mo (X.Y.); tingjunhou@zju.edu.cn (T.H.); kimhsieh@zju.edu.cn (C.-Y.H.)

†These authors contributed equally to this work.

Deep learning (DL)-driven efficient synthesis planning may profoundly transform the paradigm for designing novel pharmaceuticals and materials. However, the progress of many DL-assisted synthesis planning (DASP) algorithms has suffered from the lack of reliable automated pathway evaluation tools. As a critical metric for evaluating chemical reactions, accurate prediction of reaction yields helps improve the practicality of DASP algorithms in the real-world scenarios. Currently, accurately predicting yields of interesting reactions still faces numerous challenges, mainly including the absence of high-quality generic reaction yield datasets and robust generic yield predictors. To compensate for the limitations of high-throughput yield datasets, we curated a generic reaction yield dataset containing 12 reaction categories and rich reaction condition information. Subsequently, by utilizing 2 pretraining tasks based on chemical reaction masked language modeling and contrastive learning, we proposed a powerful bidirectional encoder representations from transformers (BERT)-based reaction yield predictor named Egret. It achieved comparable or even superior performance to the best previous models on 4 benchmark datasets and established state-of-the-art performance on the newly curated dataset. We found that reaction-condition-based contrastive learning enhances the model's sensitivity to reaction conditions, and Egret is capable of capturing subtle differences between reactions involving identical reactants and products but different reaction conditions. Furthermore, we proposed a new scoring function that incorporated Egret into the evaluation of multistep synthesis routes. Test results showed that yield-incorporated scoring facilitated the prioritization of literature-supported high-yield reaction pathways for target molecules. In addition, through meta-learning strategy, we further improved the reliability of the model's prediction for reaction types with limited data and lower data quality. Our results suggest that Egret holds the potential to become an essential component of the next-generation DASP tools.

Introduction

Efficient chemical synthesis is crucial to satisfying the future demands for pharmaceuticals, materials, and energy [1]. Corey and Wipke [2] first proposed the concept of computer-aided synthesis planning (CASP) in the 1960s. CASP programs take the target molecule as input and return a series of single-step reactions that decompose the target molecule into a set of commercially available starting compounds or simple precursors that

can be easily synthesized [3]. A feasible synthesis plan may dramatically accelerate the synthesis of desired molecules [4]. In recent years, with the development of data science, deep learning (DL) algorithms, and computing power, DL-assisted synthesis planning (DASP) has gained considerable interest [5–17]. Modern DASP programs can quickly plan multiple potential retrosynthetic pathways for a given target molecule according to the constraints set by the user for the retrosynthetic search (such as the overall search time and number of single-step

Citation: Yin X, Hsieh CY, Wang X, Wu Z, Ye Q, Bao H, Deng Y, Chen H, Luo P, Liu H, et al. Enhancing Generic Reaction Yield Prediction through Reaction Condition-Based Contrastive Learning. *Research* 2024;7:Article 0292. <https://doi.org/10.34133/research.0292>

Submitted 14 September 2023

Accepted 6 December 2023

Published 12 January 2024

Copyright © 2024 Xiaodan Yin et al. Exclusive licensee Science and Technology Review Publishing House. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY 4.0).

expansion steps) [18]. However, these theoretically feasible reaction pathways often become impractical because of such factors as incomplete conversion of reactants, side reactions, or inadequate purification [19]. Therefore, retrosynthetic route planning is only a major component of a successful DASP system [20]. To provide feasible suggestions that can be implemented by chemists in the laboratory, it is necessary to identify the optimal reaction conditions for the retrosynthetic route [21] and evaluate the quality of the overall synthesis route, and reaction yield is one of the most scientific and intuitive metrics for screening reaction conditions and evaluating synthesis pathway [22,23].

Reaction yield refers to the percentage of reactants that are successfully converted to the desired product [24]. Models that can reliably predict actual yields not only serve as scoring functions of DASP but also help chemists evaluate the overall yield of complex reaction pathways, giving priority to high-yield reactions to save time and cost in wet experiments [25]. However, because of the complexity of molecular structures, the multidimensionality of chemical reactions, and the limited availability of data, it is still a great challenge to predict the yields of chemical reactions under specific conditions [26]. The current yield prediction models are mainly built on high-throughput experimental (HTE) datasets, and Buchwald–Hartwig reactions [26–28] and Suzuki–Miyaura reactions [29,30] are the 2 most well-studied HTE yield datasets (Fig. 1A and B). Early studies utilized computed

physicochemical descriptors [26], one-hot encoding of reactions [27], or structure-based molecular fingerprints [28] to predict the yields for these 2 datasets. Recently, the DL fingerprint rxnfp developed by Schwaller et al. [31] and the differential reaction fingerprint drfp developed by Probst et al. [32] have substantially outperformed previous methods. Rxnfp and drfp respectively achieved the best performance on test sets of Buchwald–Hartwig dataset and Suzuki–Miyaura dataset (70:30 random split), with coefficient of determination (R^2) scores of 0.95 and 0.85. However, by studying previous experimental results, we found that most of the aforementioned methods did not achieve the ideal predictive performance on the out-of-sample test sets of Buchwald–Hartwig reactions containing additional additives. This indicates the limitations of using HTE datasets for yield prediction. HTE datasets usually involve specific classes of reactions and focus on a narrow chemical space. When using yield prediction models to explore unknown chemical spaces, this performance degradation problem will be prevalent because the unknown chemical space to be predicted can be very large [33]. Yield prediction models trained on HTE datasets cannot be applied in the real-world scenarios aimed at predicting the yields for a broad variety of reactions. Therefore, curating generic reaction yield datasets that are not limited to specific reaction classes is the first step to promote the practical application of yield prediction models.

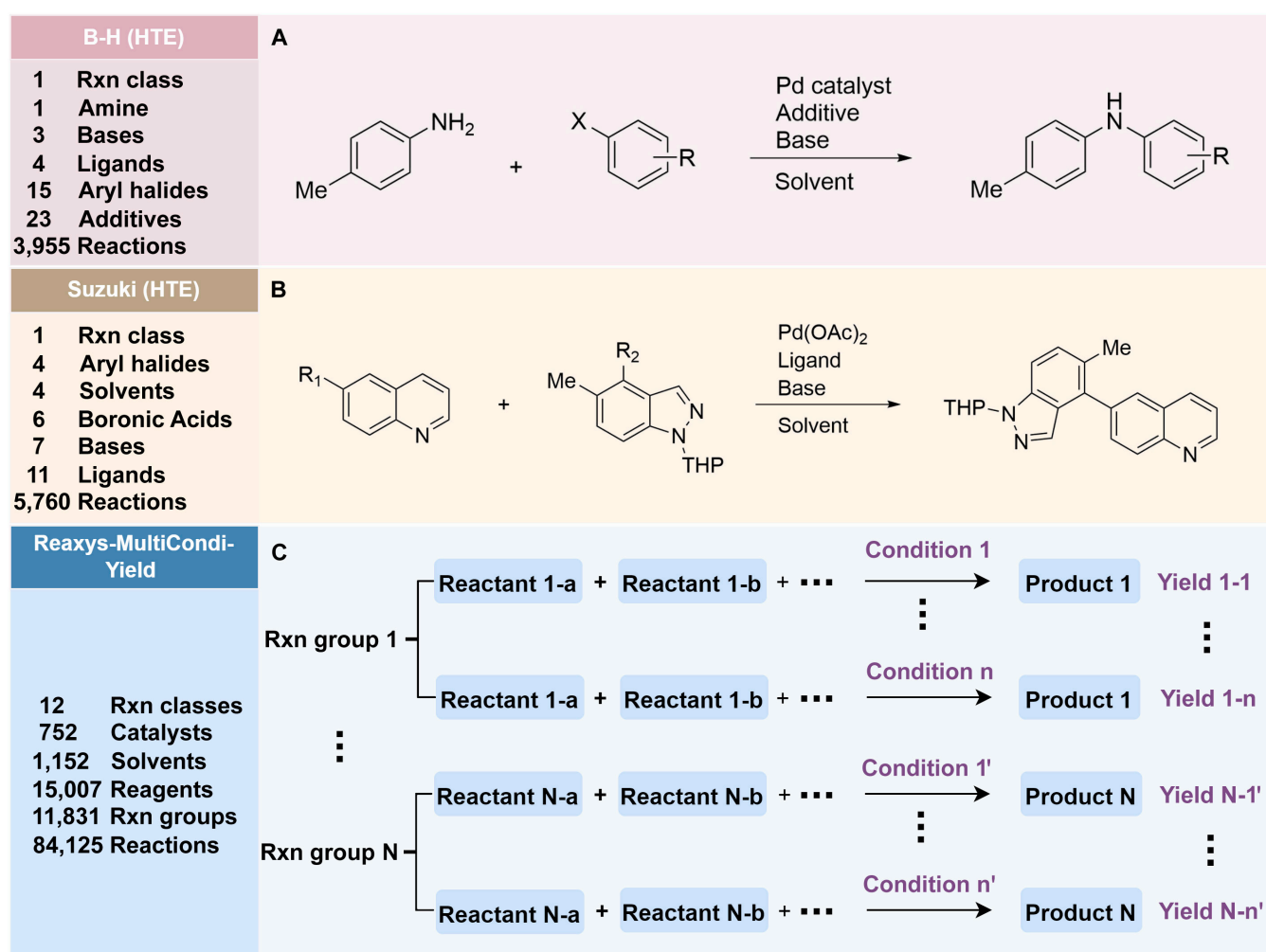


Fig. 1. Overall reaction and variables for the Buchwald–Hartwig (B-H) (A), Suzuki–Miyaura (B), and Reaxys-MultiCondi-Yield (C) datasets.

Another key point to apply yield prediction models to chemical synthesis practice is to adopt effective modeling methods for generic reaction yield datasets. Reaction Simplified Molecular Input Line Entry System (SMILES) is a simplified chemical language for representing chemical reactions [34]. Therefore, SMILES-based yield prediction can be viewed as a natural language processing (NLP) problem, extracting molecular features directly from reaction SMILES without relying on any manually generated feature. In 2017, Vaswani et al. [35] proposed the transformer architecture for handling various NLP tasks, which achieved excellent feature extraction capability through the self-attention mechanism. In recent years, many pretraining language models such as bidirectional encoder representations from transformers (BERT) [36] and generative pretrained transformer (GPT) [37] based on transformer have been proposed. For example, Schwaller et al. [38] fine-tuned a BERT model pretrained on a large unlabeled reaction corpus on downstream yield prediction tasks and achieved satisfactory performance, demonstrating the advantage of the BERT-based pretraining for chemical reaction yield prediction. In chemical experiments, even for the same set of reactants and target products, the reaction yields may vary widely because of different reaction conditions [39]. Appropriate reaction conditions are key to achieving high-yield chemical reactions [40]. However, because of the limitations of current available reaction yield datasets and modeling methods, existing yield prediction methods have not fully considered the impact of varying reaction conditions on the yields when reactants and products are identical. This leads to the inability of yield prediction models to accurately predict the changes of reaction yields based on different reaction conditions. To address this issue, we can explore the use of contrastive learning strategy [41,42] to enhance the model's sensitivity to reaction conditions on the specifically curated generic reaction yield dataset with rich reaction condition information. Furthermore, some reaction classes in generic reaction yield datasets may not have sufficient samples. In this regard, meta-learning strategies can be adopted to improve the yield prediction performance under the setting of few-shot learning.

The contributions of this work can be briefly summarized as follows:

1. We extracted and curated a high-quality generic reaction yield dataset named Reaxys-MultiCondi-Yield from the Reaxys database (Fig. 1C). Compared to HTE datasets, Reaxys-MultiCondi-Yield encompasses a broader chemical space, including 12 reaction types, 752 catalysts, 1,152 solvents, 15,007 reagents, and 84,125 reactions. Specifically, this dataset consists of 11,831 reaction groups, which are sets of reactions with the same reactants and products but varying yields due to different reaction conditions.
2. To implement a general yield prediction model, we designed a pretraining framework named Egret (Fig. 2), which is based on BERT and includes 2 pretraining tasks: masked language modeling (MLM) and reaction-condition-based contrastive learning. Egret performed comparably or even better than the previous best models on 4 benchmark datasets and achieved optimal performance on the Reaxys-MultiCondi-Yield dataset.
3. We proposed a yield-incorporated scoring for multistep retrosynthesis planning, and the results indicated that the yield-incorporated scoring can indeed prioritize literature-supported high-yield synthesis routes for target molecules.
4. Finally, we used a meta-learning strategy to model the low-sample-size or low-quality data of 5 reaction classes in the Reaxys-MultiCondi-Yield dataset, resulting in a significant improvement in prediction accuracy. Specifically, the accuracy of the reaction class 10 has increased by 33.33%.

Results

Dataset

Pretraining dataset and benchmark datasets

Reaction data in Pistachio [43] dataset were used for the pretraining of Egret. Here, 4 benchmark datasets were prepared. Among them, Buchwald–Hartwig dataset [26] consists of 3,955 reactions whose reaction space contains 1 amine, 3 bases, 4 ligands, 15 aryl halides, and 23 isoxazole additives (Fig. 1A). Suzuki–Miyaura dataset [44] consists of 5,760 reactions, and its chemical space contains 4 aryl halides, 4 solvents, 6 boronic acids, 7 bases, and 11 ligands (Fig. 1B). US Patent and Trademark Office (USPTO) yield dataset [31] was split into subgram dataset and gram dataset by Schwaller et al. according to product mass. However, these reaction data were originally obtained through text mining from open-access texts from the US Patents database, and it may lack complete experimental details, such as reaction conditions. The data split methods of these 4 benchmark datasets are consistent with previous studies, and we also resplit the 2 USPTO yield datasets according to the ratio of training:validation:test = 6:2:2.

Reaxys-MultiCondi-Yield

According to the specific data processing procedure that is easy to reproduce (Section S1.2), we exported and curated a generic reaction yield dataset called Reaxys-MultiCondi-Yield from the Reaxys database. Within this dataset, a reaction group refers to a collection of reactions characterized by identical reactants and products but different yields, which result from different reaction conditions (Fig. 1C). The dataset comprises 11,831 reaction groups and 84,125 reactions, and it encompasses a chemical space comprising 15,007 reagents, 1,152 solvents, and 752 catalysts. Moreover, we utilized a neural network classifier based on rxnfp, trained on the Pistachio dataset, to assign reaction categories to the cleaned Reaxys-MultiCondi-Yield dataset, and the accuracy of the reaction classifier on the Pistachio test set achieved 97.8%. As a result, the reaction data in the Reaxys-MultiCondi-Yield dataset were classified into 12 categories. As illustrated in Fig. 3A, aside from the reactions for which the reaction classifier could not predict the type (unrecognized), the most common type of reaction in the Reaxys-MultiCondi-Yield dataset is oxidation, followed by reduction reactions. The remaining types of reactions, in descending order of their proportions, are C–C bond formation, functional group interconversion, heteroatom alkylation and arylation, acylation and related processes, heterocycle formation, deprotections, functional group addition, protections, and resolutions. This demonstrates that the Reaxys-MultiCondi-Yield dataset includes a rich variety of chemical transformations, which is beneficial for developing generalized models for predicting reaction yields. In addition, we further visualized and discussed the distribution and characteristics

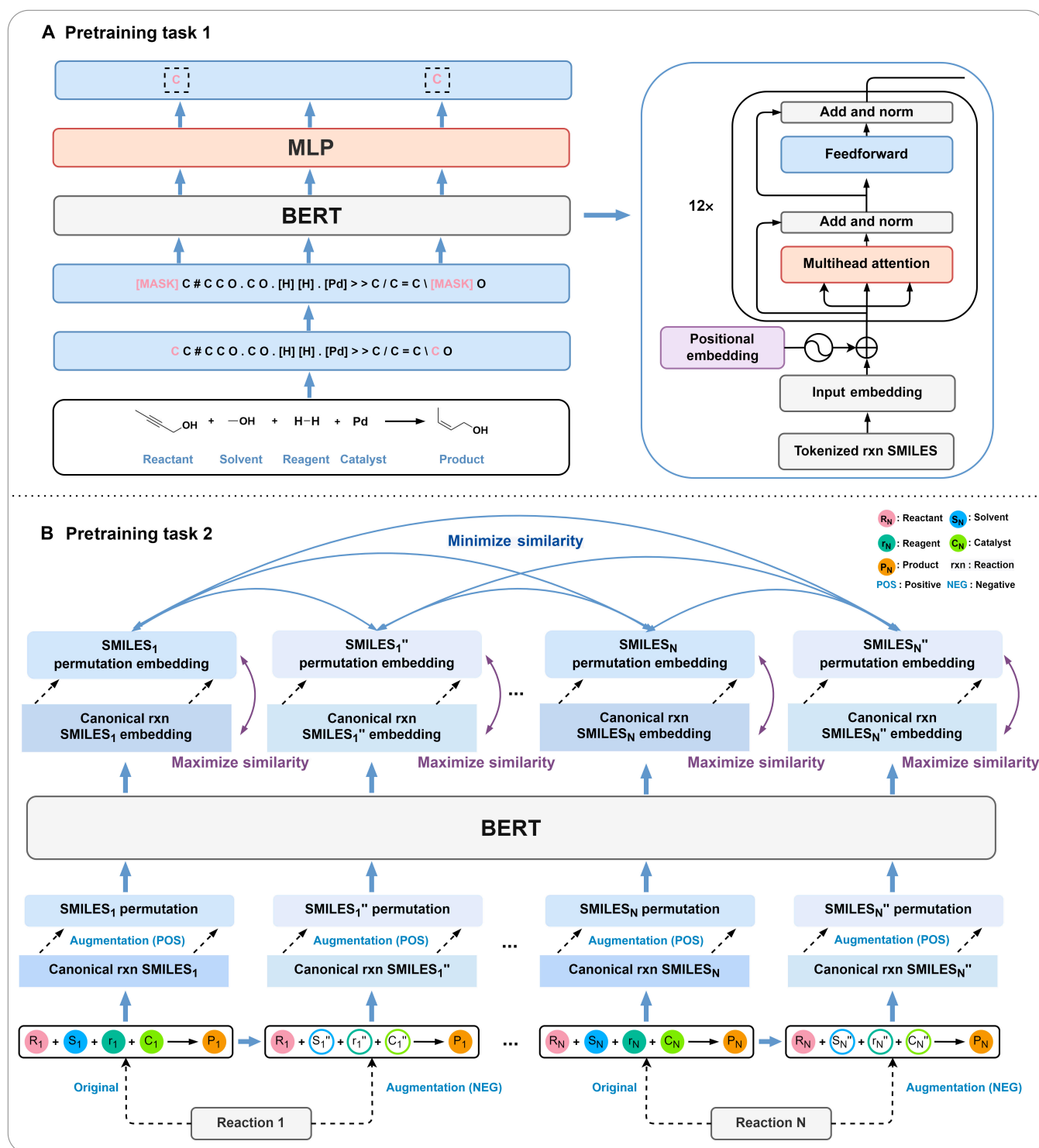


Fig. 2. Overview of Egret. (A) Pretraining task 1: MLM task; (B) pretraining task 2: reaction-condition-based contrastive learning task.

of the reaction data in the Reaxys-MultiCondi-Yield dataset in Section S1.3. The Reaxys-MultiCondi-Yield dataset was split into the training, validation, and test sets at the 8:1:1 and 6:2:2 ratios using stratified sampling based on reaction groups.

Model architecture

BERT has achieved impressive successes in various NLP tasks by pretraining the transformer encoder [45–48], and the BERT

models trained on chemical reaction data may embed any reaction data (in the form of SMILES) in an abstract vector space, without first preprocessing the reaction data, such as atom mapping. Encouraged by the convenience of such a deep learning paradigm that may autonomously extract and learn useful relations from a large-scale reaction data, we proposed a generic pretrained yield predictor, BERT (Egret; see Fig. 2), which can be broadly applied to any reaction records for yield estimation without being restricted

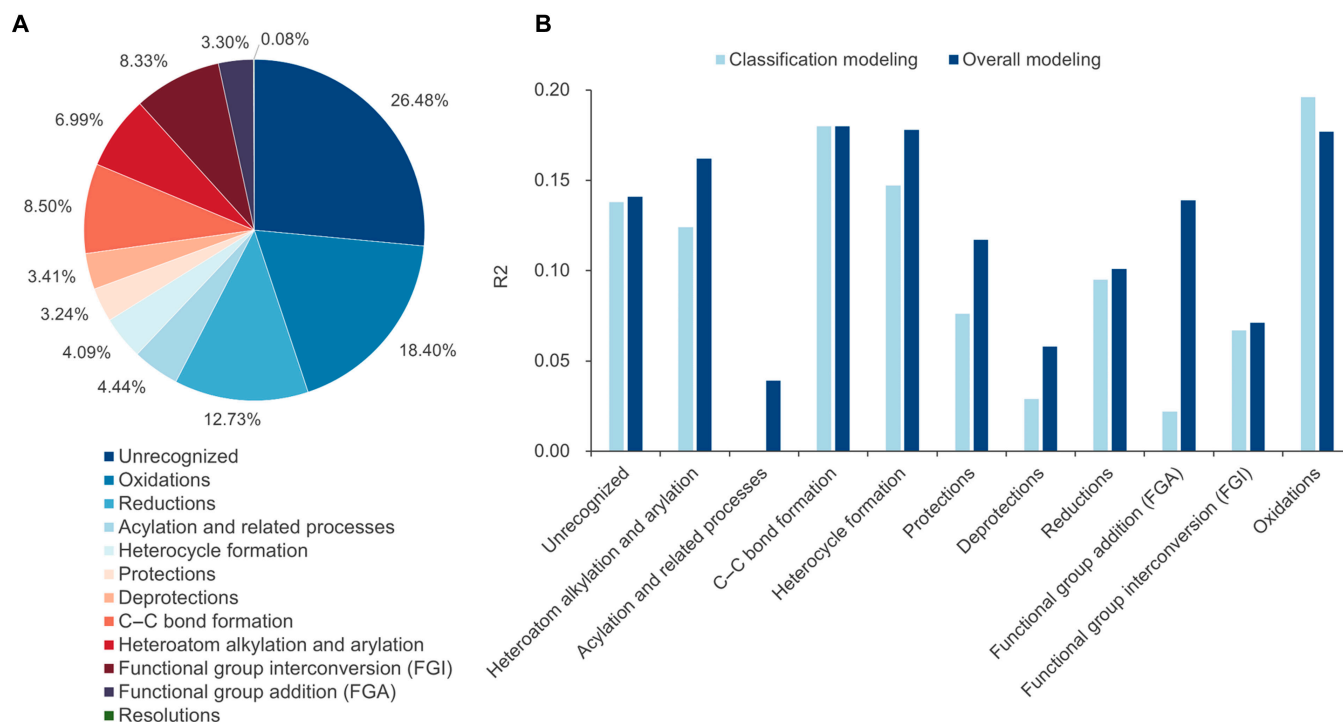


Fig. 3. Visualization of the Egret model performance based on modeling strategies. (A) Reaction category composition of the Reaxys-MultiCondi-Yield dataset; (B) prediction performance on the Reaxys-MultiCondi-Yield dataset.

to only some specific reaction classes. Egret consists of 2 pretraining tasks: MLM (task 1) and contrastive learning (task 2).

MLM task

The canonical reaction SMILES were tokenized using the procedure proposed by Schwaller et al. [49] for the MLM task. The goal is to train the model to predict individual tokens that have been randomly masked in a given reaction SMILES as shown in Fig. 2A. MLM helps the model to better learn the syntax and semantics of chemical reactions in a self-supervised manner.

Contrastive learning task

For a canonical reaction SMILES, a negative reaction SMILES was generated by randomly replacing some of its reaction conditions (reagents, solvents, and catalysts), and then the positive reaction SMILES of the reaction SMILES was generated by SMILES permutation (Fig. 2B), and we have described the process of obtaining the positive and negative reaction samples in detail in Section S1.4. The objective of this task is primarily twofold: (a) to enable the model to distinguish negative reaction samples by minimizing the embedding similarity between reactions involving the same reactants and products but under different conditions, thereby obtaining distinctive feature representations for negative samples; 2) to achieve a robust representation of chemical reactions in the model by maximizing the embedding similarity of different SMILES forms of the same reaction (positive reaction samples), thereby reducing the uncertainty caused by the diversity of reaction SMILES. Therefore, the contrastive learning task uses the following loss function:

$$\text{CL loss} = \alpha \sum_{n=1}^N \sum_{i \in I_n} \cos(V_{n,c}, V_{n,i}) + \sum_{n=1}^N \sum_{p \in P_n} \frac{1}{2} (1 - \cos(V_{n,c}, V_{n,p})), \alpha \in \{0.02, 0.05, 1\}$$

$$\cos(V_a, V_b) = \frac{V_a \cdot V_b}{\|V_a\| \|V_b\|}$$

where n denotes a reaction record and N is the number of reactions in a batch. I_n represents the negative reaction SMILES of reaction n in a batch, and P_n represents the positive reaction SMILES of reaction n . $V_{n,c}$ is the embedding of the canonical reaction SMILES of reaction n generated by Egret, and $V_{n,i}$ and $V_{n,p}$ are the embeddings generated by Egret for the canonical reaction SMILES of reactions i and p , respectively. α is a balance coefficient, and we tested 3 numbers (0.02, 0.05, and 1). When α is 0.02, the model converged most stably. $\cos()$ denotes the cosine similarity function, which assesses the similarity of 2 embeddings. $\|V_a\|$ is the Euclidean norm of the embedding

$$V_a = (V_{a,1}, V_{a,2}, \dots, V_{a,p}), \text{ defined as } \sqrt{V_{a,1}^2 + V_{a,2}^2 + \dots + V_{a,p}^2}.$$

In addition, in the second stage of Egret pretraining, the loss includes both the MLM pretraining and the contrastive learning, and the overall loss function reads,

$$\text{Loss} = \sum_{n=1}^N \text{CrossEntropyLoss}(t_j, \hat{t}_j) + \text{CL loss}$$

where t_j is the predicted value of the masked token and $\hat{t}_j$ is the ground truth of the masked token.

Model performance

Benchmark datasets

For the Buchwald–Hartwig dataset, we tested the performance of Egret on 11 different data partitioning schemes. These schemes included 7 partitions with varying relative sizes of the training set, ranging from 70% down to 2.5%, as well as 4 out-of-sample

partitions based on the addition of isoxazole as additives. Each of the first 7 data partitions has 10 random splits, and we optimized the hyperparameters on the training set of the first random split and applied them to the remaining splits. The obtained performance (R^2) was compared with 3 competitive methods: Yield-BERT based on the DL fingerprints [31], XGBoost based on drfp (DRFP) [32], and a classical-descriptor-based method: random forest based on the density functional theory (DFT)-derived fingerprints [26]. The results are shown in Table 1, Egret achieved the best R^2 in 6 of 11 data splits, highlighting the effectiveness of our pretraining strategy. Across the 7 different data partitions with varying proportions of the training set, Egret outperformed DFT and Yield-BERT, while being slightly inferior to DRFP. It is gratifying to note that Egret achieved the best performance in 3 of 4 out-of-sample tasks. To further evaluate the performance of Egret and DRFP, we compared the mean absolute error (MAE) and root mean squared error (RMSE) of the 2 models on the above 11 splits. As shown in Table S4, Egret's MAE was better than DRFP in 8 of the 11 splits, and Egret achieved better MAE and RMSE than DRFP in 3 of the 4 out-of-sample tasks. In the out-of-sample tasks, the splits were defined by the presence of isoxazole additives that strongly influence reactivity. Therefore, the model had to appropriately extrapolate to unseen additives to perform well, indicating that Egret possesses better robustness and generalization capabilities. Figure 4 is the regression graph of Egret on 11 data split methods. It can be observed from Fig. 4A to H that the model has learned to reasonably predict yields even when the training set was markedly smaller (20%). Moreover, the model's performance gradually improves as the proportion of the training set increases, and the performance difference on the test set decreases with further augmentation of training data. Figure 4I to L demonstrates that in the out-of-sample tests 1 to 4, Egret demonstrates a satisfactory predictive capability for test 1 and test 2, but the prediction performance on test 3 and test 4 needs to be further improved. The performance of Egret was further tested on the 70:30 (train:test) split of the Suzuki–Miyaura dataset. We conducted 10 different random folds and utilized the same hyperparameter optimization method as the Buchwald–Hartwig

dataset. The test results, as shown in Table 2, show that Egret achieved the best performance at the dataset. Regarding the 2 USPTO yield datasets, in addition to maintaining the data split consistency with prior works [31,32], we redivided the training, validation, and test sets at a 6:2:2 ratio to validate the model's stability. Table 2 indicates that Egret achieved the best performance on the 2 datasets under 2 data segmentation approaches, except for the USPTO gram dataset with reference split. The above results confirm that Egret has competitive predictive power in yield prediction on the HTE datasets. Moreover, the impressive performance on the USPTO yield datasets showcases the promising predictive potential of Egret for generic reaction yield datasets.

Reaxys-MultiCondi-Yield dataset

Compared to the HTE yield datasets and USPTO yield datasets, Reaxys-MultiCondi-Yield offers a broader chemical space and more comprehensive information about reaction conditions, which strives to resemble the real-world scenarios for chemical synthesis. Here, we conducted 2 subtasks on this dataset—regression for continuous yields and multiclassification for yield categories—and compared the performance of Egret with Yield-BERT and DRFP. Unfortunately, as shown in Table 3, although we tried different dataset segmentation methods, none of the models were able to make accurate predictions on the Reaxys-MultiCondi-Yield dataset on the regression task. Even if Egret performed best, its performance was still not satisfactory. Therefore, we explored the influence of modeling strategies on the predictive performance. As described above, the reactions in the Reaxys-MultiCondi-Yield dataset were classified into 12 categories (Fig. 3A). Except for the category resolutions (due to insufficient data), we separately fine-tuned Egret on the training set of each reaction category and conducted testing on their respective test sets (classification modeling). In addition, we also tested Egret, which was fine-tuned on the entire training set of Reaxys-MultiCondi-Yield, on the test set of each reaction category (overall modeling), and compared the performance of these 2 modeling strategies. From Fig. 3B, it can be observed that Egret exhibited varying predictive

Table 1. R^2 of Egret on the Buchwald–Hartwig dataset and comparison with other methods (boldface indicates the best performance, and the same applies to the following tables)

R^2	Methods			
	DFT	Yield-BERT	DRFP	Egret
Rand 70/30	0.92	0.95 ± 0.01	0.95 ± 0.01	0.94 ± 0.01
Rand 50/50	0.9	0.92 ± 0.01	0.93 ± 0.01	0.93 ± 0.01
Rand 30/70	0.85	0.88 ± 0.01	0.89 ± 0.01	0.89 ± 0.01
Rand 20/80	0.81	0.86 ± 0.01	0.87 ± 0.01	0.86 ± 0.02
Rand 10/90	0.77	0.79 ± 0.02	0.81 ± 0.01	0.80 ± 0.01
Rand 5/95	0.68	0.61 ± 0.04	0.73 ± 0.02	0.73 ± 0.02
Rand 2.5/97.5	0.59	0.45 ± 0.05	0.62 ± 0.04	0.50 ± 0.12
Test 1	0.8	0.84 ± 0.01	0.81 ± 0.01	0.84 ± 0.01
Test 2	0.77	0.84 ± 0.03	0.83 ± 0.003	0.88 ± 0.03
Test 3	0.64	0.75 ± 0.04	0.71 ± 0.001	0.65 ± 0.06
Test 4	0.54	0.49 ± 0.05	0.49 ± 0.004	0.54 ± 0.06

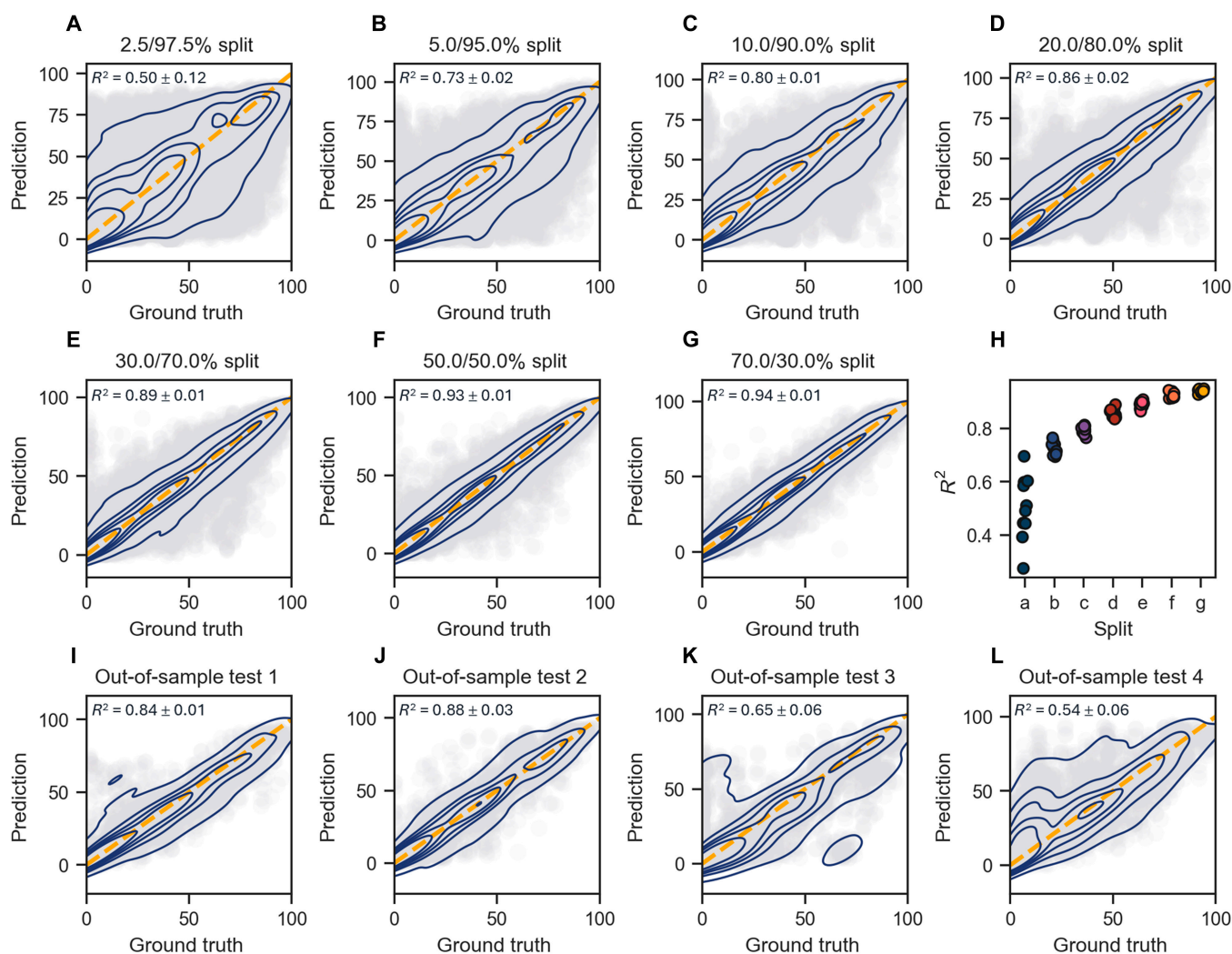


Fig. 4. Regression plots of the Buchwald-Hartwig dataset using 11 different data partitioning methods. (A to G) Seven splits with training set proportions ranging from 2.5% to 70%; (I to L) 4 out-of-sample test splits.

performance on different reaction categories. However, the overall modeling outperformed the classification modeling on 9 of 11 reaction categories. This indicates that integrating all types of reaction data for training is meaningful, and the model appears to effectively learn deeper and transferable knowledge across different types of chemical reactions.

Previous studies have shown that in generic reaction yield datasets, although the reactions of the same category exhibit consistent yield trends, different subreactions within the same category often have different yield values [31]. This results in excessive local noise, which affects the performance of the yield prediction models. To fully consider the influence of reaction conditions on reaction yield, the Reaxys-MultiCondi-Yield dataset was constructed with 11,831 reaction groups. Each reaction group consists of a series of reactions with identical reactants and products but varying yields due to different reaction conditions. Inevitably, in addition to the inherent noise in the reaction data, this also increases the local noise and affects the performance of continuous yield regression predictions. In contrast to modeling HTE reaction yield data (stringent control of variables), the task of modeling general reaction yield data is substantially more complex because of not only the characteristics of the recorded reaction data but also unrecorded factors such

as laboratory conditions, instrument precision, and differences in personnel operations. These variables substantially increase the difficulty of the modeling process, leading to a pronounced disparity in the model's predictive performance on HTE reaction data versus general reaction data.

In fact, the yield of a reaction is not an exact constant in the chemistry laboratory. It is influenced not only by the reaction conditions but also by various factors such as the laboratory environment, the operator's technical skills, and the purity of the reactants. As a result, the yield often fluctuates within a certain range. Therefore, predicting a reasonable yield interval for chemical reactions is also of great significance for guiding chemical synthesis. Here, on the basis of the common yield ranges in chemistry: 80% to 100%, 50% to 80%, 30% to 50%, and 0 to 30%, we categorized the yield data from Reaxys-MultiCondi-Yield into 4 categories: high, medium, low, and extremely low. We evaluated the performance of the multiclass-Egret (MC-Egret) on this dataset and compared it with Yield-BERT and DRFP. As shown in Table 3, MC-Egret achieved a prediction accuracy of 60.16%, higher than both BERT and DRFP. This indicates that Egret exhibited the best predictive performance on the Reaxys-MultiCondi-Yield dataset.

Table 2. Performance of Egret on the Suzuki–Miyaura dataset and the USPTO datasets and comparison with other methods

Methods	Metric	Dataset				
		Suzuki–Miyaura	USPTO (grams)		USPTO (subgrams)	
		70/30 ^a	Reference split ^b	6:2:2	Reference split ^c	6:2:2
Yield-BERT	R^2	0.81 ± 0.01	0.117	0.112	0.195	0.171
	MAE	0.08 ± 0.003	0.157	0.156	0.196	0.198
	RMSE	0.12 ± 0.004	0.197	0.196	0.239	0.242
DRFP	R^2	0.85 ± 0.01	0.130	0.118	0.197	0.183
	MAE	0.07 ± 0.002	0.156	0.156	0.197	0.198
	RMSE	0.11 ± 0.004	0.195	0.196	0.239	0.240
Egret	R^2	0.85 ± 0.01	0.128	0.118	0.206	0.202
	MAE	0.07 ± 0.002	0.154	0.156	0.194	0.196
	RMSE	0.11 ± 0.004	0.195	0.196	0.237	0.238

^aTrain/test = 70/30, using 10 different random folds, validation set split from training set.^bTrain/test = 158,095/39,524, validation set split from training set.^cTrain/test = 241,632/60,408, validation set split from training set.**Table 3.** Performance of Egret on the Reaxys-MultiCondi-Yield dataset and comparison with other methods

Methods	Regression						Multiclassification
	8:1:1			6:2:2			8:1:1
	R^2	MAE	RMSE	R^2	MAE	RMSE	Accuracy
Yield-BERT	0.1439	0.1317	0.1769	0.1249	0.1363	0.1782	0.5836
DRFP	0.1458	0.1306	0.1767	0.1262	0.1320	0.1780	0.5849
Egret	0.1529	0.1300	0.1760	0.1429	0.1306	0.1764	0.6016

Contrastive learning pretraining task can enhance the model's sensitivity to reaction conditions

As mentioned above, the effectiveness of our pretraining framework has been proven. Next, we explored the role of the contrastive learning pretraining task, for which we constructed an Egret without the contrastive learning pretraining task (Egret-WCL) for fair comparison. We collected the reaction IDs of the reaction groups in the Reaxys-MultiCondi-Yield training set that have more than 5 reactions. On the basis of these reaction IDs, we selected their corresponding reactions from the Reaxys-MultiCondi-Yield test set to create a test set for the reaction groups with more than 5 reactions. Following the same approach, we collected the test sets for the reaction groups with more than 10, 15, 20, 25, 30, 35, 40, 50, and 60 reactions, respectively. In addition, we evaluated the yield prediction performance of Egret and Egret-WCL, which were fine-tuned on the Reaxys-MultiCondi-Yield training set, on the test sets of the aforementioned reaction groups. At the same time, we also measured the performance of DRFP on the test sets of these reaction groups. The results are shown in Fig. 5, and we can see that Egret outperformed Egret-WCL and DRFP in the test sets of different reaction groups. These test sets contain a substantial number of reactions with varying yields due to different reaction conditions. This demonstrates that the contrastive learning pretraining

based on reaction conditions effectively enhances the model's sensitivity to reaction conditions. Egret excels in distinguishing reactions with identical reactants and products but differing reaction conditions, and it also exhibits a better understanding of different forms of the same reaction SMILES. Furthermore, by observing the magnitude of performance improvement, we found that as the number of reactions in the reaction group increases, Egret exhibits a more significant improvement in performance compared to Egret-WCL and DRFP. This indicates that when the dataset has a more comprehensive sampling of reaction-condition-yield combinations for the same reaction, Egret captures a finer characterization of subtle differences among these cases featuring the same reactions under different conditions. This not only further enhances the model's performance in predicting yields but also indicates the exceptional potential of Egret in handling large-scale, high-quality reaction yield datasets.

Generic yield prediction model enables retrosynthetic planning tool to prioritize high-yield synthetic routes

The generic yield prediction model, established on the basis of a generic reaction yield dataset, thoroughly learns the relationship

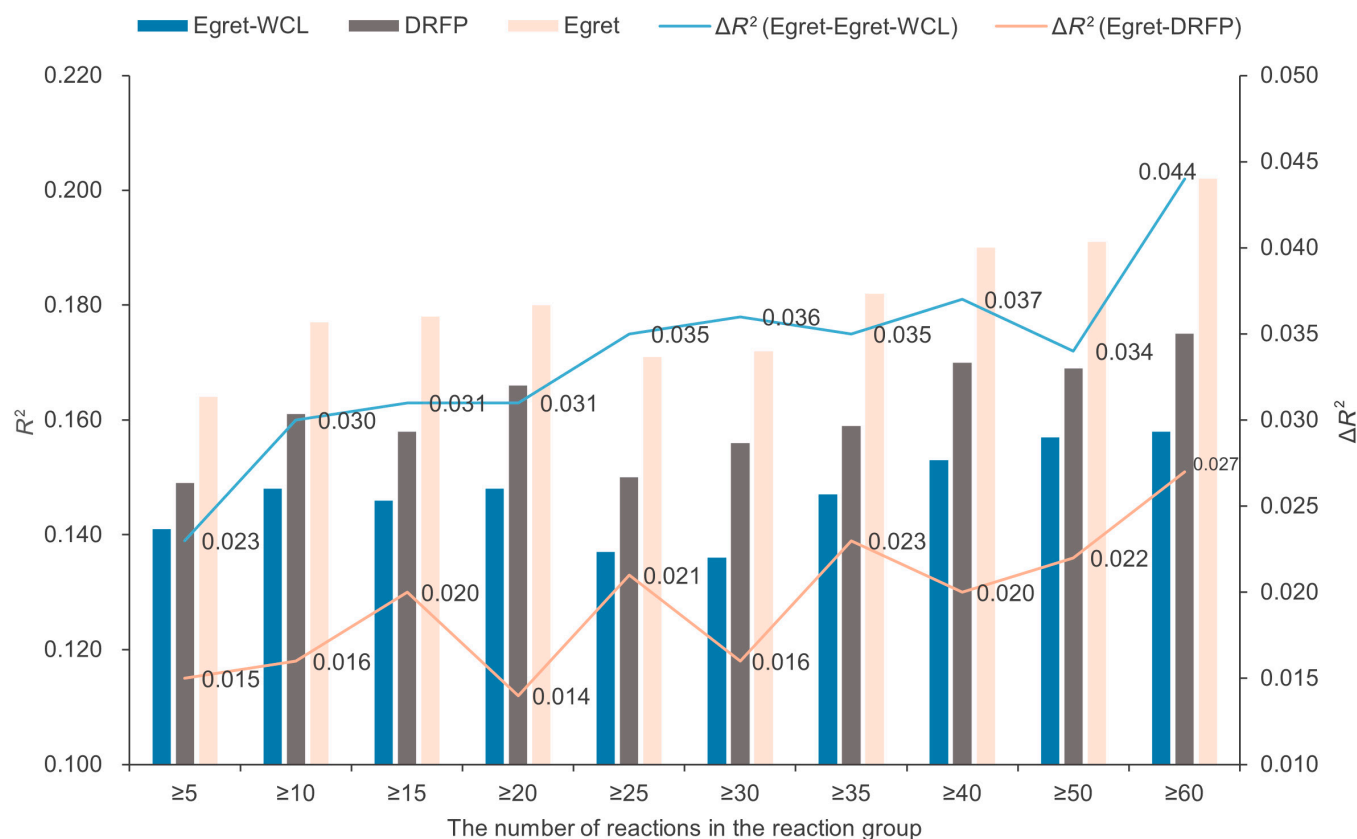


Fig. 5. Visualization of the effects of the reaction-condition-based comparative learning task.

between reaction conditions and yields. It possesses the capability to predict yields or yield intervals for any given reaction, without being constrained by specific reaction types. To investigate the application potential of the generic yield prediction model, we proposed a new scoring that incorporates the yield prediction model into a synthetic planning tool and investigated the effects of yield-augmented scoring on synthesis planning. As shown in Fig. 6, we incorporated the MC-Egret trained on Reaxys-MultiCondi-Yield and the reaction condition recommendation model developed by Gao et al. [40] into AiZynthFinder (AZ), a multistep retrosynthesis tool developed by Genheden et al. [11]. MC-Egret was used as part of the prior scores of reaction templates in AZ's single-step predictions. For an input target molecule, the process begins by utilizing the template-based model within AZ to predict the single-step routes for it (Fig. 6A). Subsequently, the reaction condition recommendation model is used to recommend the reaction conditions for each predicted single-step route (Fig. 6B). The single-step reaction SMILES, which includes the predicted reaction conditions, is then fed into MC-Egret (Fig. 6C). In addition, in the subsequent route prediction, priority is given to expanding molecular nodes within single-step routes that have higher-yield-incorporated prior score.

We randomly selected 50 bioactive small molecules from the PubChem database and conducted the tests both on the original version of AZ and the AZ embedded with MC-Egret (AZ-Egret), and we set a max depth of 6, a search time of 120 s, and an iteration of 300 rounds. As shown in Table 4, we selected the compounds that were successfully solved by both methods, and then we found the optimal (higher-yield) synthetic routes for these compounds by diligently consulting

the literature. In Table 4, "step" indicates the number of the synthesis steps for the target molecule, while "recommended number" indicates the number of the optimal synthetic route planned by synthesis planning methods. It can be seen that when both AZ-Egret and AZ successfully found the optimal synthesis routes for target molecules, in most cases, the first route planned by AZ-Egret is the optimal synthetic route, while AZ is not. For instance, the optimal synthetic route of compound 7 [50] reported in the literature is shown in Fig. 7A, and the recommended numbers of AZ-Egret and AZ are 1 and 4, respectively. Furthermore, for compounds 13, 14, 15, 17, 21, and 22, only AZ-Egret successfully found the literature-supported synthesis pathways for them. Figure 7B and C is the optimal synthetic routes of compounds 15 [51,52] and 22 [53–55], respectively. In fact, the integration of the yield prediction model extends the search time for single-step prediction. Therefore, within the same search time, the average number of the search iterations of AZ is more than that of AZ-Egret. However, AZ-Egret can still preferentially plan superior synthesis routes for the target molecules with relatively fewer search iterations. The above results indicate that yield-incorporated scoring can facilitate synthesis planning tools to recommend higher-yield and more reasonable synthesis routes for target compounds. Moreover, we have further discussed the diversity and synthetic complexity of the compounds in Table 4 in Section S4, to gain a more comprehensive understanding of the capabilities and characteristics of AZ-Egret. The details about the AZ-Egret and AZ's planned synthetic routes for the 50 test molecules can be found at <https://github.com/xiaodanyin/yield-score-analysis/tree/main/aizynthfinder/aizynthfinder/interfaces>.

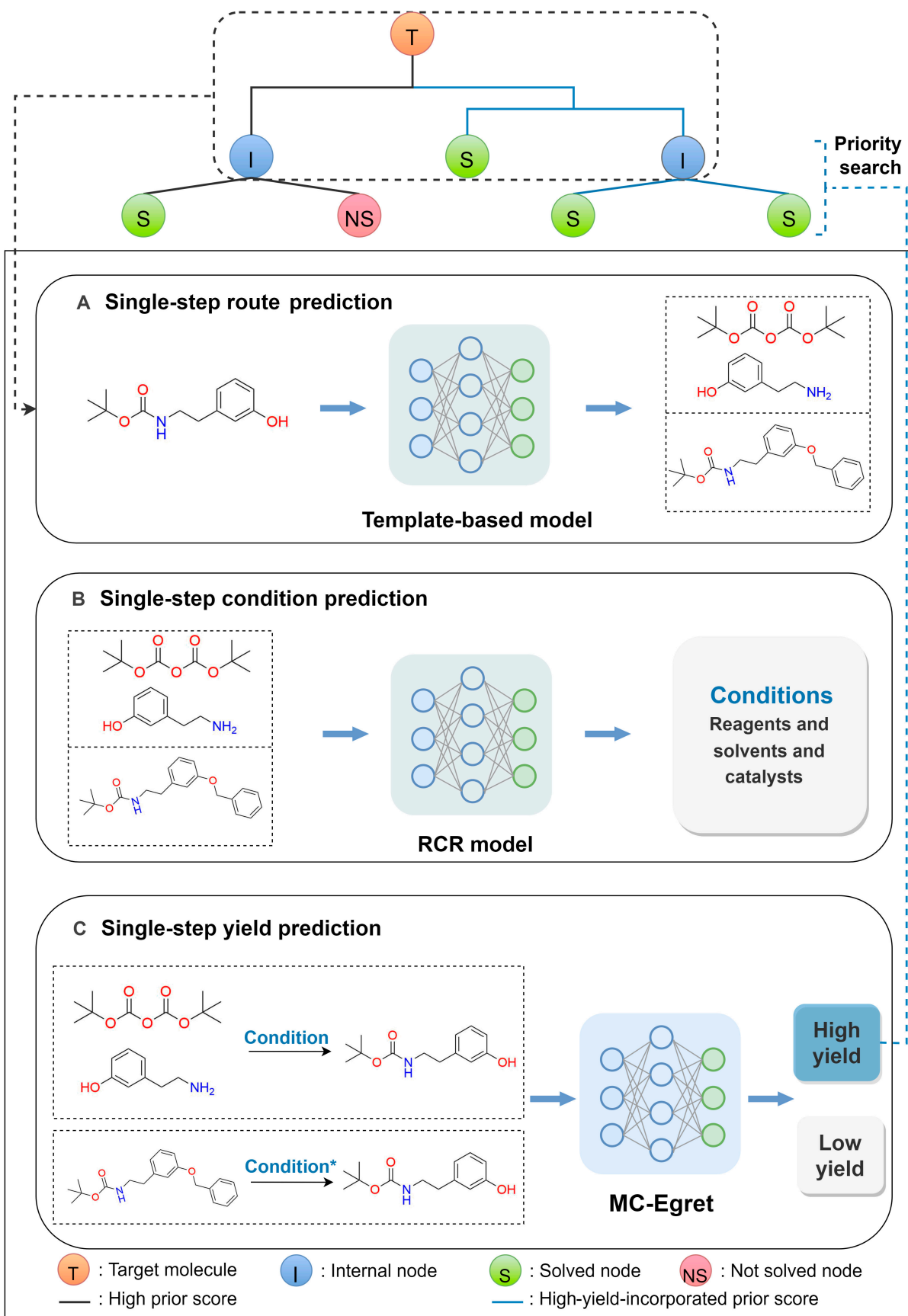
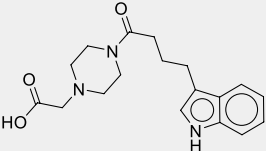
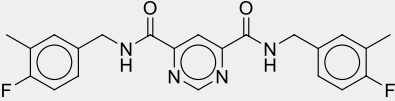
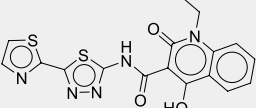
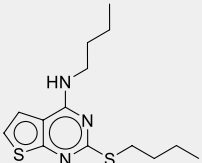
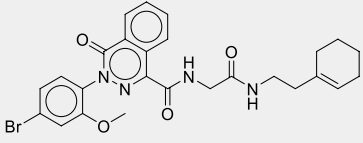
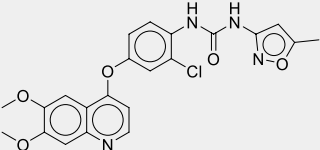
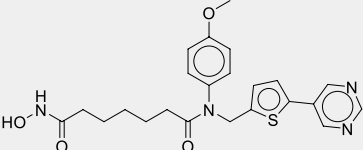
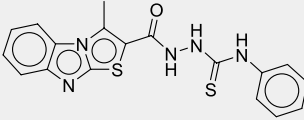


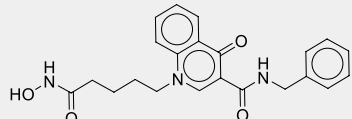
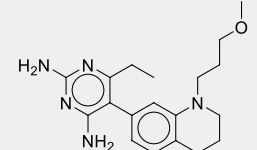
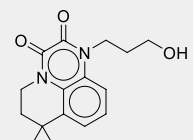
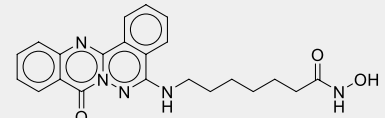
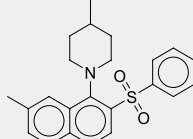
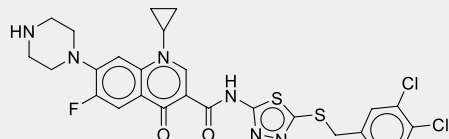
Fig. 6. Workflow of the multistep retrosynthetic pipeline embedded with MC-Egret. (A) Single-step route prediction; (B) single-step reaction condition prediction; (C) single-step reaction yield prediction.

Table. (Continued)

Compound	Recommended number		
	Step	AZ	AZ-Egret
 <chem>O=C(O)CN1CCN(C(=O)CCCC2c[nH]c3ccccc23)CC1</chem> (10)	2	2	1
 <chem>Cc1cc(CNC(=O)c2cc(C(=O)NCc3ccc(F)c(C)c3)ncn2)ccc1F</chem> (11)	2	4	1
 <chem>CCn1c(=O)c(C(=O)Nc2nnc(-c3nccs3)s2)c(O)c2ccccc21</chem> (12)	2	5	1
 <chem>CCCCNc1nc(SCCCC)nc2sccc12</chem> (13)	3	-	1
 <chem>COc1cc(Br)ccc1-n1nc(C(=O)NCCC(=O)NCCC2=CCCCC2)c2ccccc2c1=O</chem> (14)	3	-	1
 <chem>COc1cc2nccc(OC3ccc(NC(=O)Nc4cc(C)on4)c(Cl)c3)c2cc1OC</chem> (15)	3	-	1
 <chem>COc1ccc(N(Cc2ccc(-c3cnnc3)s2)C(=O)CCCCC(=O)NO)cc1</chem> (16)	3	2	1
 <chem>Cc1c(C(=O)NNC(=S)Nc2ccccc2)sc2nc3ccccc3n12</chem> (17)	3	-	1

(Continued)

Table. (Continued)

Compound	Recommended number		
	Step	AZ	AZ-Egret
 <chem>O=C(CCCCN1cc(C(=O)NCc2ccccc2)c(=O)c2ccccc21)NO</chem> (18)	4	3	1
 <chem>CCc1nc(N)nc(N)c1-c1ccc2c(c1)N(CCCOC)CCC2</chem> (19)	4	7	1
 <chem>CC1(C)CCn2c(=O)c(=O)n(CCCO)c3ccccc1c32</chem> (20)	4	4	1
 <chem>O=C(CCCCCCNc1nn2c(=O)c3ccccc3nc2c2ccccc12)NO</chem> (21)	4	-	1
 <chem>Cc1ccc2ncc(S(=O)(=O)c3ccccc3)c(N3CCC(C)CC3)c2c1</chem> (22)	5	-	1
 <chem>O=C(Nc1nnc(SCc2ccc(Cl)c(Cl)c2)s1)c1cn(C2CC2)c2cc(N3CCNCC3)c(F)cc2c1=O</chem> (23)	5	5	2

Meta-learning strategy improves the predictive performance of Egret for reaction types with limited data quantity and low quality

As mentioned earlier, we have experimentally confirmed the positive impact of the generic yield prediction model on synthetic planning. However, the performance of the generic yield prediction model is severely limited because of the lack of a large number of high-quality annotated training data. As described in the “Dataset” section, we utilized a high-accuracy reaction classifier trained on the Pistachio dataset to categorize the reaction data in the Reaxys-MultiCondi-Yield dataset into 12 classes. On this basis, we conducted an analysis by reaction category and evaluated the yield prediction accuracy of MC-Egret on each category of the reaction data in the Reaxys-MultiCondi-Yield

validation set and found that the model suffers performance drop for the following reaction types: unrecognized (0), deprotections (6), functional group interconversion (9), functional group addition (10), and resolutions (11). From Fig. 3A, it is obvious that the reaction types 6, 9, 10, and 11 possessed relatively few reaction data. Category 0, on the other hand, contains data for which the reaction classifier is unable to predict the reaction type. Despite having a larger data volume, this category contains more complex and lower-quality reaction data. This suggests that the current DL-based reaction yield prediction suffers a low-data problem.

The problem of modeling a small number of training data has been actively studied by the few-shot learning community, and the paradigm of meta-learning has emerged as a noteworthy solution [56–59]. Meta-learning is an advanced machine learning



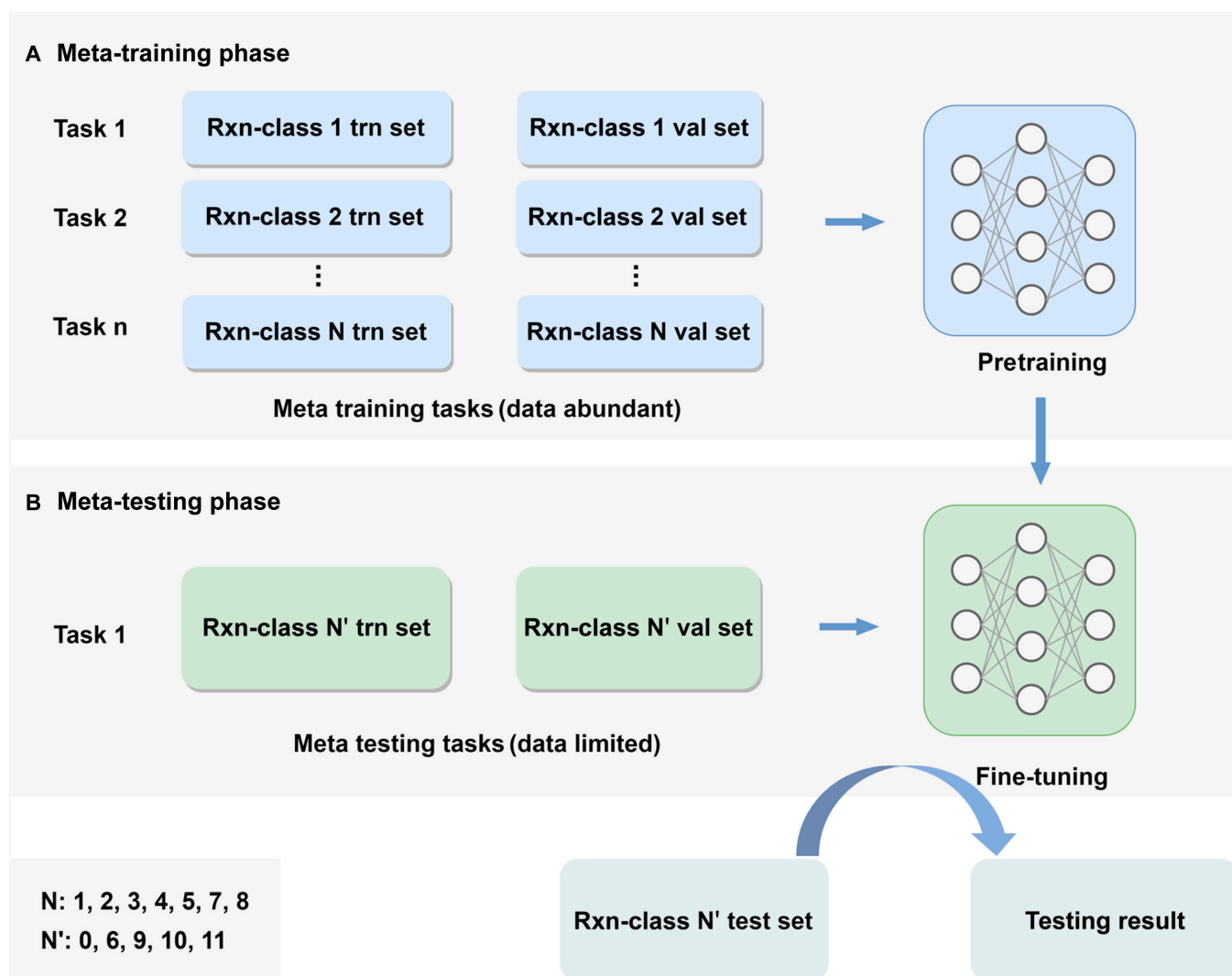


Fig. 8. Schematic diagram of Meta-Egret. (A) Meta-training phase; (B) meta-testing phase.

MC-Egret and Meta-Egret on the respective test sets of the aforementioned 5 reaction categories and found that the meta-learning strategy improves the model's prediction accuracy for the long-tail yield labels in these datasets. Figure S3 visualizes the yield distribution of the training set for the reaction category **10**, where it can be observed that the yield data follows a long-tail distribution, with the long-tail data including the reaction data with the “low” and “extremely low” yield labels. On the test set of the reaction category **10**, MC-Egret lacked the ability to predict the “low” yield label (accuracy is 0), but Meta-Egret achieved an accuracy of 15.63% in predicting the “low” yield label. This indicates that the meta-learning strategy also helps alleviate the problem of poor model performance caused by imbalanced data distribution.

Discussion

On the basis of the well-curated generic reaction yield dataset Reaxys-MultiCondi-Yield, we have successfully developed a generic yield prediction model that can improve the efficiency of synthesis planning. However, it is worth noting that there are still some limitations to our work at present. First, because

Table 5. Accuracy of MC-Egret and Meta-Egret on the test sets of the reaction categories **0, 6, 9, 10, and 11**

Reaction category	MC-Egret	Meta-Egret
0	0.5958	0.6010
6	0.5596	0.5927
9	0.5921	0.6184
10	0.5323	0.5703
11	0.5556	0.8889

of the absence of information regarding the time, temperature, and pressure of the chemical reactions in the pretraining dataset Pistachio, the reaction condition information in this work only considers the reagents, solvents, and catalysts. However, the influence of reaction time, temperature, and pressure on reaction yield should not be ignored. Therefore, in future work, we will further incorporate some factors that affect reaction

yield, such as time, temperature, and pressure, into the design of yield prediction model. Second, despite the fact that the data in the Reaxys database is sourced from high-quality literature and undergoes rigorous quality control and verification, there are still some unavoidable human errors and a scarcity of reported data on low-yield reactions [60]. Therefore, we also look forward to more unbiased reporting of reaction yield data and the development of more scientific data management tools to promote the advancement of chemical reaction yield prediction tasks.

Conclusion

In this study, considering the lack of high-quality generic reaction yield datasets, we meticulously curated the Reaxys-MultiCondi-Yield dataset, which contains abundant information on reaction condition and is not limited to a single type of reaction. On the basis of our well-designed pretraining strategies, we proposed a novel yield prediction model called Egret, which achieved the best performance on the Reaxys-MultiCondi-Yield dataset. In addition, Egret exhibited comparable or even superior performance on the 2 HTE yield datasets and 2 USPTO yield datasets when compared to other methods. This indicates the effectiveness of the pretraining strategies of Egret, where contrastive learning enabled the model not only to understand different forms of the same reaction SMILES but also to differentiate the same reaction with varying reaction conditions, thereby enhancing sensitivity to reaction conditions. Consequently, the model could extract chemical reaction features more efficiently and accurately, making it applicable to yield datasets with distinct characteristics. To explore the practical value of the generic yield prediction model, we embedded the generic yield prediction model trained on the Reaxys-MultiCondi-Yield dataset into a multistep retrosynthetic tool. The results demonstrated that yield-incorporated scoring can facilitate synthetic planning tools to prioritize the recommendation of literature-supported reaction pathways with higher yields for target molecules. Furthermore, we further improved the yield prediction accuracy for reaction types with limited data and lower data quality in the Reaxys-MultiCondi-Yield dataset using meta-learning strategies. Specifically, the accuracy for reaction class 10 has been increased by 33.33%. In the future, as more high-quality reaction yield data become available, we believe that Egret and algorithms inspired by Egret will play crucial roles in the development of a comprehensive solution for the DL-based synthesis planning.

Methods

Dataset curation

The details of the processing methods for the pretraining dataset and Reaxys-MultiCondi-Yield dataset can be found in Sections S1.1 and S1.2, respectively.

Model construction and evaluation

Pretraining of Egret

This process consists of 2 stages. In the first stage, pretraining task 1 was carried out. After the model's embedding capability for chemical reactions was sufficiently optimized, the second stage of pretraining began. At this stage, both pretraining tasks were performed simultaneously. The 2 phases were pretrained

for 50 and 10 epochs, respectively. The transformer encoder used in Egret has 12 layers, each layer contains 4 attention heads, and the hidden size is 256. All codes were implemented in the Python 3.7 environment, the RDKit [61] cheminformatics toolkit was used for data processing, and the model was developed based on the PyTorch library.

Fine-tuning of Egret

The DL model's workflow can be broadly divided into 2 key parts: autonomous feature extraction and downstream task predictions. Egret has learned how to extract the reaction features from the reaction SMILES through pretraining. Then, we kept the same framework to load all the encoding layers in the pretraining model and retrained feedforward neural networks for regression or classification on specific yield prediction datasets.

Model evaluation

To evaluate the performance of Egret on 4 benchmark datasets, we compared its performance with the benchmark methods reported previously on these datasets. For the Buchwald–Hartwig dataset, DFT [26], Yield-BERT [31], and DRFP [32] were used as the baselines, and Yield-BERT and DRFP were used as the baselines for the Suzuki–Miyaura dataset and the 2 USPTO yield datasets. Since Yield-BERT and DRFP are the only 2 methods that have been tried on the generic reaction yield datasets (USPTO yield datasets), we used them as the baselines on Reaxys-MultiCondi-Yield too and adopted the regression and multiclassification tasks to evaluate the performance of these methods. The regression tasks were evaluated by the coefficient of determination (R^2), MAE, and RMSE, and the classification tasks were evaluated by the accuracy. Section S2 summarizes the detailed hyperparameters of the Egret, Yield-BERT, and DRFP models and the calculation method of R^2 , MAE, RMSE, and accuracy.

Yield-incorporated synthesis planning analysis

To investigate the application potential of the generic yield prediction model, as shown in Fig. 6, we incorporated MC-Egret into the prior scores of reaction templates in the retrosynthesis planning tool AZ's single-step predictions to explore the influence of the generic yield prediction model on retrosynthesis planning. The details of how to incorporate MC-Egret into AZ can be found in Section S3.

Code availability

All the codes of Egret are available at <https://github.com/xiaodanyin/Egret>, and the code of Egret-incorporated synthesis planning analysis can be found at <https://github.com/xiaodanyin/yield-score-analysis>.

Acknowledgments

Funding: This work was supported by grants from the Science and Technology Development Fund, Macau SAR (file nos. 0056/2020/AMJ, 0114/2020/A3, and 0015/2019/AMJ) and Dr. Neher's Biophysics Laboratory for Innovative Drug Discovery (file no. 002/2023/ALC).

Author contributions: X. Yin contributed to the conception and implementation of the manuscript and the primary work. C.-Y.H. provided the key ideas for the work and participated in manuscript revisions. X.W. collected some of the data. Z.W., Q.Y., H.B., and Y.D. participated in discussions regarding the work's approach. H.C. provided some of the data. P.L. supervised the progress of the work. H.L. provided computing resources.

T.H. participated in manuscript revisions and supervised the progress of the work. X. Yao provided funding support and participated in manuscript revisions while also supervising the progress of the work.

Competing interests: The authors declare that they have no competing interests.

Data Availability

Benchmark datasets are available at <https://github.com/xiaodanyin/Egret/tree/main/dataset>, and we provide the “Reaction ID” and “Links to Reaxys” for all the reactions in Reaxys-MultiCondi-Yield. On the basis of this information, these reactions can be exported from Reaxys to reproduce the dataset. Pistachio [43] dataset can be obtained from NextMove software.

Supplementary Materials

Sections S1 to S5

Figs. S1 to S3

Tables S1 to S5

References

- Lin Y, Zhang R, Wang D, Cernak T. Computer-aided key step generation in alkaloid total synthesis. *Science*. 2023;379:453–457.
- Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*. 1969;166:178–192.
- Schreck JS, Coley CW, Bishop KJ. Learning retrosynthetic planning through simulated experience. *ACS Cent Sci*. 2019;5:970–981.
- Wang X, Li Y, Qiu J, Chen G, Liu H, Liao B, Hsieh C-Y, Yao X. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chem Eng J*. 2021;420:Article 129845.
- Cook A, Johnson AP, Law J, Mirzazadeh M, Ravitz O, Simon A. Computer-aided synthesis design: 40 years on. *Wiley Interdiscip Rev Comput Mol Sci*. 2012;2(1):79–107.
- Coley CW, Rogers L, Green WH, Jensen KF. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent Sci*. 2017;3(12):1237–1245.
- Coley CW, Thomas DA III, Lummiss JA, Jaworski JN, Breen CP, Schultz V, Hart T, Fishman JS, Rogers L, Gao H, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*. 2019;365(6453):eaax1566.
- Dai H, Li C, Coley C, Dai B, Song L. Retrosynthesis prediction with conditional graph logic network. *Adv Neural Inf Process Syst*. 2019;32.
- Johansson S, Thakkar A, Kogej T, Bjerrum E, Genheden S, Bastys T, Kannas C, Schliep A, Chen H, Engkvist O. AI-assisted synthesis prediction. *Drug Discov Today Technol*. 2019;32–33:65–72.
- Chen B, Li C, Dai H, Song L. Retro*: Learning retrosynthetic planning with neural guided A* search. In: *Proceedings of the 37th international conference on machine learning*. PMLR; 2020. pp. 1608–1616.
- Genheden S, Thakkar A, Chadimová V, Reymond J-L, Engkvist O, Bjerrum E. Aizynthfinder: A fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform*. 2020;12:70.
- Mikulak-Klucznik B, Gołębiewska P, Bayly AA, Popik O, Klucznik T, Szymkuć S, Gajewska EP, Dittwald P, Staszewska-Krajewska O, Beker W, et al. Computational planning of the synthesis of complex natural products. *Nature*. 2020;588:83–88.
- Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, Bekas C, Iuliano A, Laino T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem Sci*. 2020;11:3316–3325.
- Somnath VR, Bunne C, Coley C, Krause A, Barzilay R. Learning graph models for retrosynthesis prediction. *Adv Neural Inf Process Syst*. 2021;34:9405–9415.
- Skoraczynski G, Kitlas M, Miasojedow B, Gambin A. Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning. *J Cheminform*. 2023;15:6.
- 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;11:5575.
- Karpov P, Godin G, Tetko IV. A transformer model for retrosynthesis. In: *International conference on artificial neural networks*. Cham: Springer; 2019. vol. 11731. pp. 817–830.
- Behl T, Chadha S, Sachdeva M, Sehgal A, Kumar A, Venkatachalam T, Hafeez A, Aleya L, Arora S, Batiha GE-S. Understanding the possible role of endocannabinoid system in obesity. *Prostaglandins Other Lipid Mediat*. 2021;152:Article 106520.
- Yarish D, Garkot S, Grygorenko OO, Radchenko DS, Moroz YS, Gurbych O. Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. *J Comput Chem*. 2023;44(2):76–92.
- Struble TJ, Alvarez JC, Brown SP, Chytil M, Cisar J, Desjarlais RL, Engkvist O, Frank SA, Greve DR, Griffin DJ, et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J Med Chem*. 2020;63:8667–8682.
- Wang X, Hsieh C-Y, Yin X, Wang J, Li Y, Deng Y, Jiang D, Wu Z, Du H, Chen H, et al. Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research*. 2023;6:0231.
- Rouh H, Tang Y, Xu T, Yuan Q, Zhang S, Wang J-Y, Jin S, Wang Y, Pan J, Wood HL, et al. Aggregation-induced synthesis (AIS): Asymmetric synthesis via chiral aggregates. *Research*. 2022;2022:9865108.
- Wu G, Liu Y, Yang Z, Katakam N, Rouh H, Ahmed S, Unruh D, Surowiec K, Li G. Multilayer 3D chirality and its synthetic assembly. *Research*. 2019;2019:6717104.
- Schwaller P, Vaucher AC, Laino T, Reymond J-L. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *ChemRxiv*. 2020. <https://doi.org/10.26434/chemrxiv.13286741.v1>.
- Zhu X-Y, Ran C-K, Wen M, Guo G-L, Liu Y, Liao L-L, Li Y-Z, Li M-L, Yu D-G. Prediction of multicomponent reaction yields using machine learning. *Chin J Chem*. 2021;39(12):3231–3237.
- Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*. 2018;360(6385):186–190.
- Chuang KV, Keiser MJ. Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science*. 2018;362(6416):eaat8603.
- Sandfort F, Strieth-Kalthoff F, Kühnemund M, Beecks C, Glorius F. A structure-based platform for predicting chemical reactivity. *Chem*. 2020;6(6):1379–1390.

29. Granda JM, Donina L, Dragone V, Long D-L, Cronin L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*. 2018;559:377–381.
30. Fu Z, Li X, Wang Z, Li Z, Liu X, Wu X, Zhao J, Ding X, Wan X, Zhong F, et al. Optimizing chemical reaction conditions using deep learning: A case study for the Suzuki–Miyaura cross-coupling reaction. *Org Chem Front*. 2020;7(16):2269–2277.
31. Schwaller P, Vaucher AC, Laino T, Reymond J-L. Prediction of chemical reaction yields using deep learning. *Mach Learn Sci Technol*. 2021;2:Article 015016.
32. Probst D, Schwaller P, Reymond J-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit Discov*. 2022;1:91–97.
33. Chen K, Chen G, Li J, Huang Y, Wang E, Hou T, Heng P-A. MetaRF: Attention-based random forest for reaction yield prediction with a few trails. *J Cheminform*. 2023;15(1):43.
34. Weininger D. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28:31–36.
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.
36. Devlin J, Chang M-W, Lee K, Toutanova K, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. <https://doi.org/10.48550/arXiv.1810.04805>.
37. Radford A, Narasimhan K, Salimans T, Sutskever I. *Improving language understanding by generative pre-training*. San Francisco (CA): OpenAI; 2018.
38. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond J-L. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell*. 2021;3:144–152.
39. Andronov M, Voinarovska V, Andronova N, Wand M, Clevert D-A, Schmidhuber J. Reagent prediction with a molecular transformer improves reaction data quality. *Chem Sci*. 2023;14:3235–3246.
40. Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent Sci*. 2018;4:1465–1476.
41. Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. *Nat Mach Intell*. 2022;4:279–287.
42. Wu Z, Jiang D, Wang J, Zhang X, Du H, Pan L, Hsieh C-Y, Cao D, Hou T. Knowledge-based BERT: A method to extract molecular features like computational chemists. *Brief Bioinform*. 2022;23(3):bbac131.
43. Ottaviano M, Giunta EF, Tortora M, Curvietto M, Attademo L, Bosso D, Cardalesi C, Rosanova M, De Placido P, Pietroluongo E, et al. Braf gene and melanoma: Back to the future. *Int J Mol Sci*. 2021;22(7):3474.
44. Perera D, Tucker JW, Brahmabhatt S, Helal CJ, Chong A, Farrell W, Richardson P, Sach NW. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*. 2018;359:429–434.
45. Honda S, Shi S, Ueda HR. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery. arXiv. 2019. <https://doi.org/10.48550/arXiv.1911.04738>.
46. Ying C, Cai T, Luo S, Zheng S, Ke G, He D, Shen Y, Liu T-Y. Do transformers really perform badly for graph representation? *Adv Neural Inf Process Syst*. 2021;34:28877–28888.
47. Irwin R, Dimitriadis S, He J, Bjerrum EJ. Chemformer: A pre-trained transformer for computational chemistry. *Mach Learn Sci Technol*. 2022;3:Article 015022.
48. Zhang X-C, Wu C-K, Yi J-C, Zeng X-X, Yang C-Q, Lu A-P, Hou T-J, Cao D-S. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. *Research*. 2022;2022:0004.
49. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent Sci*. 2019;5(9):1572–1583.
50. Oshiro Y, Sato S, Kurahashi N, Tanaka T, Kikuchi T, Tottori K, Uwahodo Y, Nishi T. Novel antipsychotic agents with dopamine autoreceptor agonist properties: Synthesis and pharmacology of 7-[4-(4-phenyl-1-piperazinyl) butoxy]-3,4-dihydro-2 (1H)-quinolinone derivatives. *J Med Chem*. 1998;41(5):658–667.
51. Chen C, Huang X, Fan Z. Synthesis process of VEGFR inhibitor tivozanib. 2022; CN114213404A:2022-03-22.
52. Chen H, Shang Y, Li J, Zhu F, Shi P, Wu Q, Shi B, Li Y. Method for preparing isocyanate by salifying-atomizing phosgenation method. 2023; CN111825572A:2020-10-27.
53. Inaoka KD, Sakura T, Takaya K, Nakamura KR. Preparation of quinolone derivative having antimalarial activity. 2022; WO2022250104:WO2022-JP21483.
54. Dabholkar VV, Sanghvi AS. Synthesis of oxazoles, thiazoles and benzothiazines by microwave technique. *Ind J Het Chem*. 2006;16:105–108.
55. Coantic S, Mouysset D, Mignani S, Tabart M, Stella L. The use of N-sulfonylimines in the β -lactam synthon method: Staudinger reaction, oxidation of the cycloadducts and ring opening of β -lactams. *Tetrahedron*. 2007;63(15):3205–3216.
56. Wang J, Zheng S, Chen J, Yang Y. Meta learning for low-resource molecular optimization. *J Chem Inf Model*. 2021;61:1627–1636.
57. Gao Y, Gao Y, Fan Y, Zhu C, Wei Z, Zhou C, Chuai G, Chen Q, Zhang H, Liu Q. Pan-peptide meta learning for T-cell receptor–antigen binding recognition. *Nat Mach Intell*. 2023;5:236–249.
58. Lv Q, Chen G, Yang Z, Zhong W, Chen CY-C. Meta learning with graph attention networks for low-data drug discovery. *IEEE Trans Neural Netw Learn Syst*. 2023;1–13.
59. Vella D, Ebejer J-P. Few-shot learning for low-data drug discovery. *J Chem Inf Model*. 2022;63(1):27–42.
60. Saebi M, Nan B, Herr JE, Wahlers J, Guo Z, Zurański AM, Kogej T, Norrby P-O, Doyle AG, Chawla NV, et al. On the use of real-world datasets for reaction yield prediction. *Chem Sci*. 2023;14(19):4997–5005.
61. Landrum G. Rdkit: Open-source cheminformatics. 2006. Google Scholar 2006.