Journal of Cheminformatics

RESEARCH

Open Access

MetaRF: attention-based random forest for reaction yield prediction with a few trails

Kexin Chen[1], Guangyong Chen[2*], Junyou Li[2], Yuansheng Huang[3], Ercheng Wang[2,3], Tingjun Hou[3] and Pheng-Ann Heng[1,2]

Abstract

Artificial intelligence has deeply revolutionized the field of medicinal chemistry with many impressive applications, but the success of these applications requires a massive amount of training samples with high-quality annotations, which seriously limits the wide usage of data-driven methods. In this paper, we focus on the reaction yield prediction problem, which assists chemists in selecting high-yield reactions in a new chemical space only with a few experimental trials. To attack this challenge, we first put forth MetaRF, an attention-based random forest model specially designed for the few-shot yield prediction, where the attention weight of a random forest is automatically optimized by the meta-learning framework and can be quickly adapted to predict the performance of new reagents while given a few additional samples. To improve the few-shot learning performance, we further introduce a dimension-reduction based sampling method to determine valuable samples to be experimentally tested and then learned. Our methodology is evaluated on three different datasets and acquires satisfactory performance on few-shot prediction. In high-throughput experimentation (HTE) datasets, the average yield of our methodology's top 10 high-yield reactions is relatively close to the results of ideal yield selection.

Keywords Few-shot, Yield prediction, Random forest, Meta-learning

Introduction

Computer-aided synthesis planning (CASP) [1], which aims to assist chemists in synthesizing new molecule compounds, has been rapidly transformed by artificial intelligence methods. Given the availability of large-scale reaction datasets, such as the United States Patent and Trademark Office (USPTO) [2], Reaxys [3], and SciFinder [4], CASP has become an increasingly popular topic in pharmaceutical discovery and organic chemistry with many impressive breakthroughs achieved [5]. The current CASP systems can be divided into two critical aspects, retrosynthetic planning and forward-reaction prediction [6]. Retrosynthetic planning, including template-based and template-free methods, can help generate possible synthetic routes of target molecules [7]. Forward-reaction prediction is mainly used to evaluate the strategies proposed by retrosynthetic planning and increase the likelihood of experimental success [8]. However, without considering reaction yield or reaction conditions, the synthetic strategies proposed in the CASP systems would be difficult to be implemented. It still remains a big challenge to predict the reaction yield. Due to the complexity of chemical experiments, few solid theories can help predict the reaction yield of a new chemical reaction given a specific condition, let alone optimize a reaction condition, which heavily depends on expertise, knowledge, intuition, numerous practices, extensive literature reading and even the luck of chemists [5, 9].

*Correspondence:
Guangyong Chen
gychen@zhejianglab.com
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, New Territories, Hong Kong SAR
[2] Zhejiang Lab, Zhejiang, China
[3] College of Pharmaceutical Sciences, Zhejiang University, Zhejiang, China

Some pioneer efforts have been contributed to predict the reaction yield, and then find the optimal reaction condition. Note that the optimal reaction selection problem can be naturally treated as a classical out-of-distribution (OOD) problem, since the optimal reaction is often not included in the training set. Ahneman et al. [10] reported that the random forest model achieved the best performance on OOD yield prediction due to its good generalization ability. Zuranski et al. [11] reviewed and examined the OOD performance of different machine learning algorithms and reaction embedding techniques. Dong et al. [12] used the XGBoost model and achieved satisfactory OOD performance. Zhu et al. [13] demonstrated that regression-based machine learning had great application potential in OOD yield prediction.

[R2–4] Studying the experimental results in previous work, we found that the predicting performance will deteriorate dramatically when there exists relatively large difference between training and testing data. For instance, in the experiments of [10, 14, 15], when the testing data do not contain any new reagents that are different from the training set (testing data is randomly selected from the whole dataset, and the rest of the data is used as training set), the R^2 of random forest model is 0.92. When the testing data includes new additives that are not contained in the training data (testing data includes reactions with some additives, while training data includes reactions with other additives), the R^2 of random forest model will drop to 0.19 in the worst case (the size of training set is almost same). This performance deterioration problem will be very common when using yield prediction model to explore new reaction chemical space, as the size of unknown chemical space to be predicted can be huge. Enlarging training set with huge amount of data may solve this performance deterioration problem, but it is not practical due to the high cost of experimental data and huge size of unknown chemical space.

In this paper, we follow a more relaxed but practical setting, where we are allowed to add a few data of new reagents or conditions into the training set. Considering the limited amount of reaction condition data, few-shot yield prediction has great potential in solving this problem. Few-shot yield prediction adds very few reaction samples(e.g. around five samples) from new reagents or conditions into training data. It is reasonable to hypothesize that using data of a new reagent can improve prediction results. Questions yet to be explored are how to use these new samples, which sample to select, and how much data from the new reagent leads to a satisfactory result.

To bridge this gap, we proposed MetaRF, an attention-based random forest model with a meta-learning technique applied to determine attention weights adaptively. The random forest has been proved as an ensemble method with outstanding performance on datasets with small sample size [16, 17]. Since the size of reaction condition datasets are relatively small (e.g. 781 reactions in Buchwald-Hartwig electronic laboratory notebooks dataset [18]), random forest models have shown excellent performance on reaction yield prediction task and outperformed other machine learning approaches [10, 11, 19]. Few-shot learning techniques, such as meta-learning, have great potential in helping chemists explore the new reaction chemical space. However, the structure of random forest is non-differential, which is hard to combine with the gradient-based techniques in few-shot learning. Thus random forest cannot be directly optimized by few-shot learning techniques. To solve this problem and achieve robust performance on new reagents, we design a attention-based random forest, adding attention weights to the random forest through a meta-learning framework, Model Agnostic Meta-Learning (MAML) algorithm [20]. The key idea of MAML is to train the model's initial parameters so that the model can quickly adapt to a new task after the parameters have been updated through a few gradient steps computed with few-shot data from that new task [20]. MAML is applied to determine the attention weights of decision trees in the random forest so that the model can quickly adapt to predict the performance of new reagents using few-shot training samples. In our method, Density Functional Theory [10] (DFT) descriptor is used to represent molecules due to its enhanced interpretability and feature generalization ability in yield prediction task.

Besides, the choice of few-shot training samples also has a significant influence on model performance. Few-shot learning can have better-predicting performance if it is allowed to choose the training samples [21]. To tackle this challenge, we use Kennard-Stone (KS) algorithm [22] to select the most representative samples which cover the experimental space homogeneously. Since the KS algorithm is based on Euclidean distance, which suffers from the curse of dimensionality [23], T-distributed stochastic neighbor embedding (TSNE) [24] is applied for unsupervised nonlinear dimension reduction.

Our methodology is comprehensively evaluated on Buchwald Hartwig high-throughput experimentation (HTE) dataset [10], Buchwald-Hartwig electronic laboratory notebooks (ELN) dataset [18], as well as Suzuki Miyaura HTE dataset [25]. In Buchwald-Hartwig HTE dataset, our method achieves R^2=0.648 using 2.5% of the dataset as the training set. To reach a comparable result, the baseline method (random forest) needs to use at least 20% of the dataset as the training set. With the help of 5 additional samples, our method can effectively explore

unseen chemical space and select high-yield reactions. The 10 reactions, which are predicted to have the highest yield, reach an average yield of 93.7%, relatively close to the result of ideal yield selection (95.5%). In contrast, the top 10 high-yield reactions selected by the baseline method have an average yield of 86.3%, and the average yield of random selection is 52.1%.

The overview framework of this research is presented in Fig. 1. More details of methodology are in Section-Methods. The methodology in this paper can predict the effect of a new reagent structure with few reaction data, and our sampling method can help chemists choose the order of experiments.

Methods

Reaction encoded with DFT

Density Functional Theory (DFT) descriptor is widely used in molecular embedding owing to its strong and effective feature generalization ability [26]. Previous research [11] shows that the DFT descriptor provides transferable chemical insight and sheds light on the underlying mechanism. Compared with molecular fingerprints and various learned representations, DFT descriptor is more closely associated with physical and chemical attributes of molecules, thus providing enhanced interpretability and mechanistic understandings [27]. Using DFT descriptor, chemists can draw insights about the feature importance of each atom and each functional group. Previous work [19] on experimental comparison also shows that DFT descriptor outperforms RDKit's (a cheminformatics tool) chemical reaction fingerprints [28] and deep-learning RXNFP method [29] on yield prediction task. Thus we use DFT descriptors to represent molecules in our experiments.

We followed the DFT descriptor calculation in [10], which includes molecular, atomic, and vibrational property descriptors. As in [10], we generate the numerical encoding of each reaction by concatenating the DFT descriptor of each chemical component. For example, the encoding of experiment i in Buchwald-Hartwig reaction is

$$x_i = x_{\text{Aryl halide}} \oplus x_{\text{Pd catalyst}} \oplus x_{\text{Additive}} \oplus x_{\text{Base}} \quad (1)$$

where $\oplus$ denotes concatenation and $x_{\text{Aryl halide}}$, $x_{\text{Pd catalyst}}$, x_{Additive}, x_{Base} denotes DFT descriptor vector of the corresponding Aryl halide, Pd catalyst, Additive and Base.

MetaRF: attention-based random forest

Although the random forest is a robust algorithm in yield prediction, it remains a challenge to combine random forest with few-shot learning techniques in yield prediction of new reagents or conditions. Meta-learning introduces a model that can quickly adapt to new tasks with few additional samples. Model Agnostic Meta-Learning (MAML) framework [20] is a well-known meta-learning approach with both simplicity and effectiveness. However, the non-differential characteristic of the random forest makes it difficult to integrate with the gradient-based meta-learning framework. To tackle this problem, we solve different attention weights to decision trees in the random forest using MAML framework, which consists of a meta-training phase and a few-shot fine-tuning phase.

To explore the OOD predicting ability, the testing set and validation set must include at least one unseen

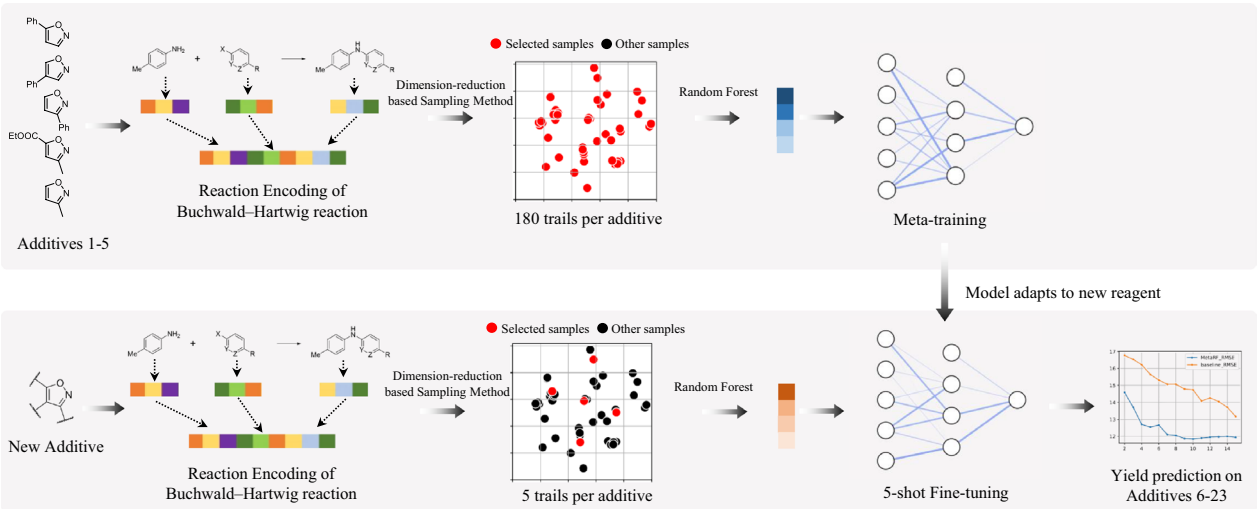

Fig. 1 Workflow of this research that includes reaction encoding, dimension-reduction based sampling method, and attention-based random forest model. Buchwald-Hartwig HTE dataset is taken as an example

reagent in the training set. For example, among the 22 different additives in Buchwald-Hartwig HTE dataset, 4 additives are used for training, 1 additive is used for validation, and 17 for testing. In this way, the training set and validation set take 22.7% of the dataset. To further reduce the size of the training set, we use the sampling method in Section-Dimension-reduction based Sampling Method. The random forest model is trained on the reduced training set.

In the random forest model, forest $\mathcal{F}$ is a collection of decision trees:

$$\mathcal{F}(\Theta) = \{h_m(\mathbf{x}; \Theta_m)\}, m = 1, 2, \ldots M \tag{2}$$

where M is the total number of decision trees, $\Theta = \{\Theta_1, \Theta_2, \ldots \Theta_M\}$ represents parameters in $\mathcal{F}$, which includes splitting variables and their splitting values. $\mathcal{F}$ is fitted by the training data $\mathcal{L} = \{(x_1, y_1), \cdots (x_N, y_N)\}$, where x_i is the embedding of reaction i (defined in the former section) and y_i represents the yield of the reaction.

The decision tree is a simple predictive model. It has the form

$$h_m(x) = \sum_{j=1}^{J} b_{jm} I(x \in R_{jm}) \tag{3}$$

where J is the number of its leaves. The tree partitions the input space into J disjoint regions $R_{1m}, \ldots, R_{Jm}$ and predicts a constant value in each region. b_{jm} is the value predicted in R_{jm}.

At each tree node, part of the variables are randomly selected as a subset. The splitting variable is chosen from this subset. This random selection of features at each node decreases the correlation between the trees in the forest and thus reduces the error rate of the random forest.

Concating the results of each decision tree $h_m(x)$, we have

$$x_i' = \begin{bmatrix} h_1(x_i) \\ h_2(x_i) \\ \vdots \\ h_M(x_i) \end{bmatrix} \tag{4}$$

Then we assign attention weights to the results of each decision tree x_i'. In this step the parameters inside these decision trees will not be changed. The attention weight of each decision tree will be updated through a meta-training phase and a few-shot fine-tuning phase.

[R2–2]In the meta-training phase (illustrated in Fig. 2A), MAML provides a good initialization of parameters in deep networks. Assume θ is the parameters that need to be optimized and f_θ is the parametrized function. In each training iteration, the updated θ is computed using one gradient descent update on task T_i, and the loss function is computed using the updated θ. Sampling task T_i includes two steps. An additive S_i is randomly sampled from the training additive set. Then K reactions with additive S_i are randomly sampled to form task T_i. More concretely, the loss function is defined as follows:

$$\min_{\theta} \sum_{T_i \sim T} L_{T_i}(f_{\theta - \alpha \nabla_\theta L_{T_i}(f_\theta)}) \tag{5}$$

where L is the mean square error between the prediction $f_\phi(x_j')$ and true value y_j in task T_i.

$$L_{T_i}(f_\phi) = \sum_{(x_j', y_j) \sim T_i} \left\| f_\phi(x_j') - y_j \right\|_2^2 \tag{6}$$

x_j' represents the list of decision tree values in former section. As in Finn et al. [20], the regressor f_θ is a neural network with 2 hidden layers of size 40 with ReLU nonlinearities. During training, Equation (5) is minimized using gradient descent algorithm Adam [30] to acquire the parameter $\theta_{\text{meta-train}}$.

[R2–2]In the few-shot fine-tuning phase (illustrated in Fig. 2B), the model is fine-tuned with a few samples from each testing additive. One iteration of gradient descent is performed to achieve $\theta_{\text{few-shot}}$ suitable for the new task T_{test}:

$$\theta_{\text{few-shot}} = \theta_{\text{meta-train}} - \alpha \nabla_{\theta_{\text{meta-train}}} L_{T_{test}}(f_{\theta_{\text{meta-train}}}) \tag{7}$$

For each additive in the testing set, the fine-tune sample in T_{test} is selected using dimension-reduction based

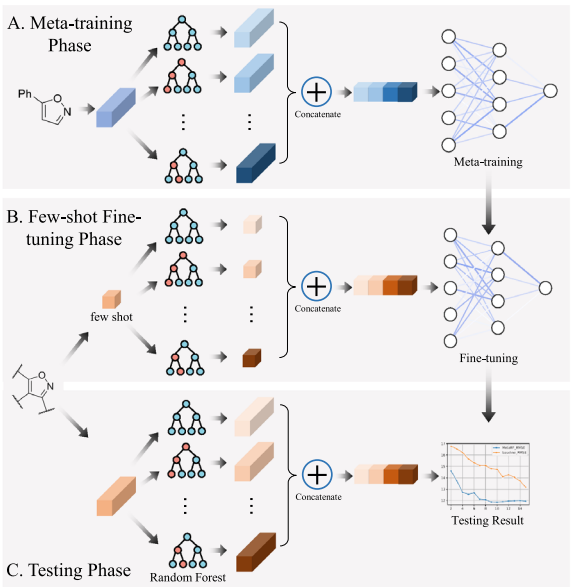

Fig. 2 **A** Meta-training Phase. **B** Few-shot fine-tuning Phase. **C** Testing Phase

sampling method in Section-Dimension-reduction based Sampling Method. The number of fine-tune samples is altered in our experiments.

Dimension-reduction based sampling method

For the few-shot learning problem, the few-shot training samples have a significant influence on the training performance. If we preferentially select the most representative samples as training samples, the performance of few-shot learning can be dramatically improved [31]. We use Kennard-Stone (KS) algorithm [22] to select the most representative samples by selecting a new sample that has relatively large distances from previously selected samples. However, the KS algorithm uses Euclidean distance to represent the distances between samples, which is less effective in the high-dimensional reaction data [32]. Thus we propose to add T-distributed stochastic neighbor embedding (TSNE) [24] before the KS algorithm to reduce the dimension of reaction data. TSNE is a widely used unsupervised nonlinear dimension reduction technique owing to its advantage in capturing local data characteristics and revealing subtle data structures [24, 33, 34].

Figure 3 use the "Swiss roll" dataset as an illustrating example for the effect of nonlinear dimension reduction method [35]. Figure 3A shows that Euclidean distance in the high-dimensional input space may not reflect the true low-dimensional geometry of the manifold. Figure 3B show the sampling result of KS algorithm without nonlinear dimension reduction method. KS algorithm is based on Euclidean distance and does not sample the central area. Figure 3C show the sampling result of KS algorithm when the dimension of data is reduced to two. KS algorithm will sample the central area after nonlinear dimension reduction. This example shows that nonlinear dimension reduction method can help our sampling method explore the intrinsic geometry of the data.

Given a set of high-dimensional reaction embedding data $x_1, x_2, \ldots, x_N$, TSNE will map the data to low dimension, while retaining the significant structure of the original data [24, 36]. It is based on probabilistic modeling of data points in the original space and the projection space [37].

The TSNE algorithm is based on the SNE framework [38], which converts high-dimensional Euclidean distances into conditional probabilities, representing similarities for every data pair. Typically the gradient descent technique is used for optimization.

After the high-dimensional reaction embedding data $x_1, x_2, \ldots, x_N$ is mapped to the low-dimensional data $z_1, z_2, \ldots, z_N$, Kennard-Stone (KS) algorithm is used to select the few-shot training samples in low-dimensional space. KS algorithm is a well-known method to select the most representative samples from the whole dataset [22, 39, 40]. The algorithm aims at choosing a subset of samples that cover the experimental space homogeneously [41]. First, the Euclidean distance between each pair of samples is calculated, and a pair of samples with the largest distance is chosen. Then the following samples are selected sequentially based on the distances to the already selected samples. The remaining sample with the largest distances is chosen and added to the subset. This procedure is repeated until a certain number of samples are selected.

From a chemical perspective, our dimension-reduction based sampling method can explore the intrinsic geometry of chemical structure and properties contained in the DFT descriptors. KS algorithm can distinguish the discrepancies and select representative samples with very different chemical structures and properties, which may shed light on the design of chemical experiments.

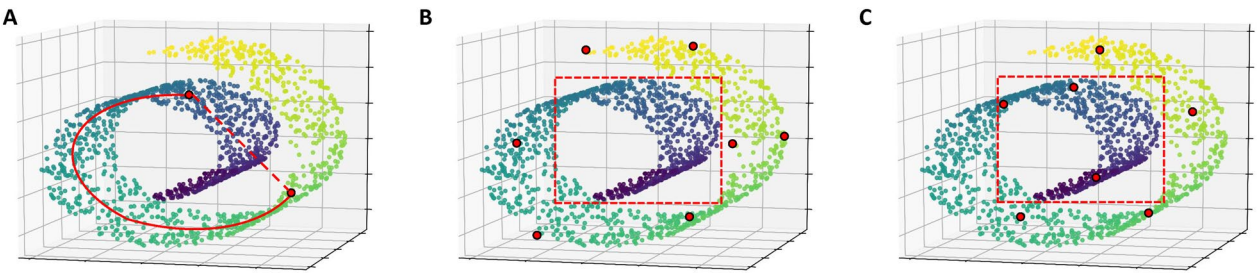

Fig. 3 Example of nonlinear dimension reduction on the "Swiss roll" dataset. **A** Euclidean distance (dashed line) in the high-dimensional input space may not reflect the true low-dimensional geometry of the manifold (Solid line). **B** Using KS algorithm to select the most representative samples on the high-dimensional input space, samples in the central area (dashed square) will not be selected. **C** Using KS algorithm after the nonlinear dimension reduction, samples in the the central area (dashed square) will be selected

Results

Performance benchmarking

We evaluate our method with Buchwald-Hartwig electronic laboratory notebooks (ELN) dataset [18], Buchwald-Hartwig HTE dataset [10] and Suzuki-Miyaura HTE dataset [25]. Buchwald-Hartwig HTE dataset is the HTE results of the Pd-catalysed Buchwald-Hartwig cross-coupling reaction. This dataset consists of 3955 reactions as shown in Fig. 4A, and the reaction space is the combination of 15 aryl halides, 4 Buchwald ligands, 3 bases, and 22 isoxazole additives. Buchwald-Hartwig ELN dataset disclosed a real-world dataset from electronic laboratory notebooks (ELN) at AstraZeneca. The dataset covers a large reaction space. 340 aryl halides, 260 amines, 24 ligands, 15 bases, and 15 solvents should have covered $4.7 * 10^8$ possible combinations. While in fact, the dataset only includes 781 reactions in Fig. 4B, resulting in a rather sparse coverage. HTE dataset greatly differs from the ELN dataset in the coverage of chemical space and characteristics. HTE dataset covers the entire search space of reaction condition while ELN dataset has a sparse coverage of wider chemical space. We also evaluate our methodology on Suzuki-Miyaura HTE dataset [25] to show that our methodology can be easily adapted

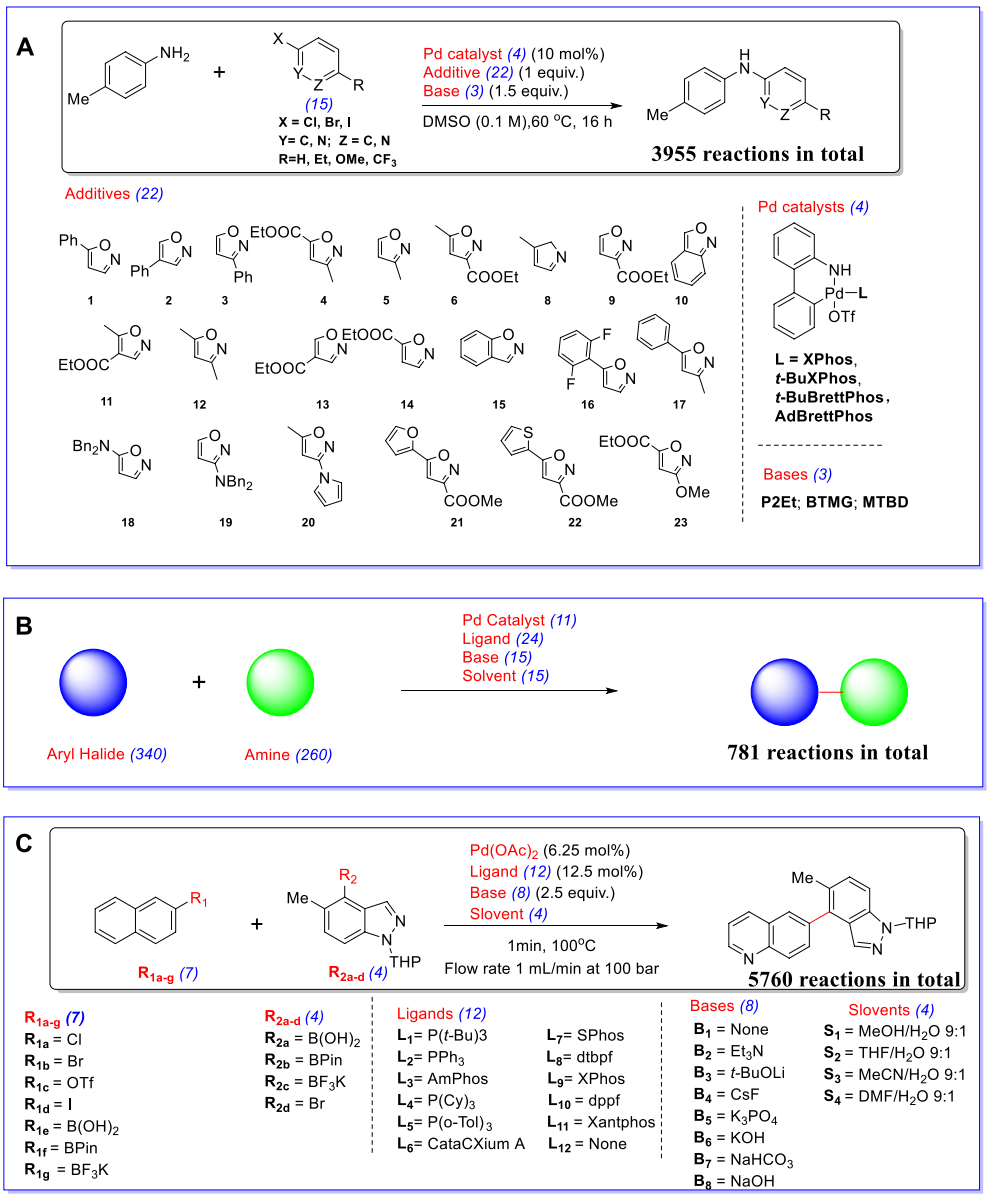

Fig. 4 **A** Buchwald-Hartwig HTE dataset. **B** Buchwald-Hartwig ELN dataset. **C** Suzuki-Miyaura HTE dataset

to other reactions. Suzuki-Miyaura reaction means that aryl halide reacts with an organoboron compound to form a new C-C bond in the presence of Pd catalyst, ligand, and base. The mechanism of the Suzuki-Miyaura reaction is close to the Buchwald-Hartwig reaction, they all include an oxidative addition step and reductive elimination step in the catalytic cycle mechanism. The dataset includes 15 pairs of electrophiles and nucleophiles(R_{1a-d} with R_{2a-c} and R_{1e-g} with R_{2d}), 12 ligands, 8 bases, and 4 solvents, resulting in 5760 reactions in Fig. 4C. Experiments show that our methodology possesses outstanding performance on few-shot yield prediction.

Tables 1,2,3 shows the performance comparision results in Buchwald-Hartwig HTE dataset, Suzuki-Miyaura HTE dataset and Buchwald-Hartwig ELN dataset, respectively. For a fair comparison, we enlarge the training set of other method with the additional fine-tune samples to guarantee that our method shares the same quantity of training data as other method. Our method has outstanding performance for few-shot yield prediction task in all three datasets. For example, in Buchwald-Hartwig HTE dataset, our method reaches R^2=0.7738 while the R^2 of the random forest [10, 16], DRFP [42], RXNFP [43, 44], neural network [45], support vector machine [46], linear model [47] and GemNet [48] are 0.6538, 0.6470, 0.0032, 0.6179, 0.5322, 0.5928, 0.5245, respectively. Among other methods, random forest has relatively good performance, which is consistent with results in previous research [10, 11, 19]. Thus random forest is chosen as the baseline method in the following analysis.

Experiments show that our method has satisfying performance when the size of training data is relatively small. As shown in Fig. 5, our model outperforms the baseline method in all three datasets when the size of training data increases gradually. Experiments show that our method possesses enhanced predictive power with markedly fewer training samples, which means that our method is an effective tool in few-shot yield prediction problem. For example, when trained on only 2.5% of Buchwald-Hartwig HTE data, MetaRF acquires comparable results with the baseline method using 20% of the same reaction data. 2.5% of the Buchwald-Hartwig HTE data includes only 90 reactions. Using 2.5% of the data as the training set, our method reaches R^2=0.648 while the R^2 of the baseline method is 0.571. When the training set increases to 20% of the data, the R^2 of the baseline method is only 0.654. This comparison is similar to the results on Buchwald-Hartwig ELN and Suzuki-Miyaura HTE datasets. These results indicate that our method has great application potential in few-shot yield prediction. In our experiments, 80 training iterations are performed, and we use one gradient update with $K = 40$ examples and learning rate $\alpha = 0.0001$. More details about the splitting of the training set, validation set, and testing set are in Section -MetaRF: Attention-based Random Forest..

Then we test our method on the ability to search for reactions with the highest yield. This ability is valuable because it helps chemists explore unseen chemical space and select high-yield reactions [49, 50]. We train our models with a relatively small training set (2.5% of the Buchwald-Hartwig HTE data, 5% of the Suzuki-Miyaura HTE data, 50% of the Buchwald-Hartwig ELN data) and use them to predict the yields of the remaining reactions. The top 10 high-yield reactions are selected according to the prediction results. Then we calculate the average and standard deviation of 10 high-yield reactions. Figure 6 presents the average and standard deviation of the yields for the top 10 reactions predicted to have the highest yields in the three datasets. Besides our method and baseline method, the result of ideal reaction selection and random reaction selection are presented. In all three datasets, our method has a higher average yield and lower standard deviation than baseline selection and random selection. For example, in the Buchwald-Hartwig HTE dataset, using MetaRF trained on 2.5% of the dataset, the predicted top 10 high-yield reactions from the remaining

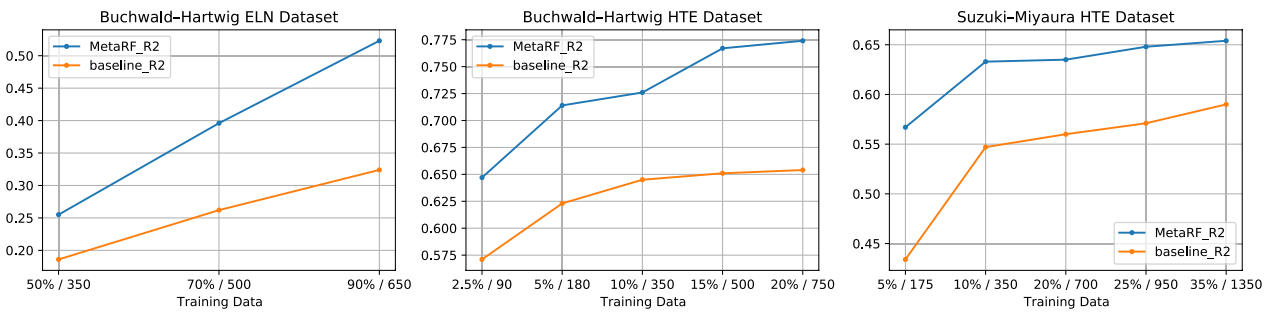

Fig. 5 Comparison of test set performance of MetaRF and baseline on three datasets. R^2 performance increases gradually as the size of training data increases. MetaRF outperforms the baseline with markedly fewer training samples

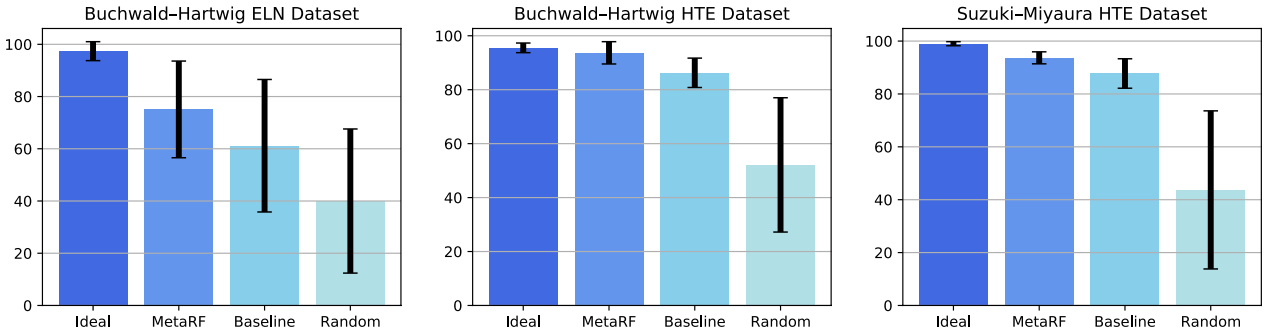

Fig. 6 Average and standard deviation of the yield for the top 10 reactions predicted to have the highest yields

dataset have an average yield of 93.7±4.1%, compared to the ideal selection of 95.5±1.8%. In contrast, baseline selection has an average yield of 86.3±5.4% and random selection has an average yield of 52.1±24.9%. The selection works similarly for the Buchwald-Hartwig ELN and Suzuki-Miyaura HTE dataset.

Ablation study

To validate the effects of each component in MetaRF, we conduct an ablation study on the Buchwald-Hartwig HTE dataset, with 20% of the data as the training set. The number of fine-tune samples is five in the ablation study. For the baseline method (random forest), five fine-tune samples are randomly selected and then added to the training set.

The first ablation replaces the dimension-reduction based sampling with random sampling. The random sampling experiment is repeated 10 times, and average performance is recorded. The second ablation removes the random forest structure, using MAML to replace the MetaRF framework. The third ablation keeps the random forest structure and uses a standard pretraining and fine-tuning framework in transfer learning [20] to replace MAML.

Table 4 presents the comparison results of predicting performance in terms of R^2 and RMSE. When dimension-reduction based sampling is replaced with random sampling, the R^2 decreases from 0.7738 to 0.7003, demonstrating the effectiveness of the dimension-reduction based sampling method. The results of the ablation study also clearly demonstrate the importance of random forest structure in MetaRF. Removing random forest causes R^2 performance to decrease from 0.7738 to 0.3730, which shows that random forest can tackle the overfitting problem in few-shot prediction. Regarding the results of the third ablation test, R^2 decreases by 10% when MAML is replaced with transfer learning, and transfer learning has minor improvement compared to the baseline.

Analysis on fine-tune sample number

The time-consuming chemical experiments raise the cost of new reaction yield data. Thus the few-shot setting and the specific number of fine-tune samples is very important in reducing the cost of empirical screening. We analyze the effect of adjusting the number of fine-tune samples on the Buchwald-Hartwig HTE dataset [10], using 20% of the data as the training set. The few-shot yield predicting ability is tested by root mean square error (RMSE) and R^2 performance.

When the number of fine-tune samples is 5, we obtain an 18.82% relative improvement in the R^2 performance and an 19.25% relative improvement in the RMSE performance. Our method reaches R^2=0.7738 and RMSE=12.6401, while the R^2 and RMSE of baseline method (random forest) is 0.6538 and 15.6535, respectively. More evaluation results of relative improvement are listed in Table 5. When the number of fine-tune samples varies, the RMSE relative improvement is still around 20%, which demonstrates the stable and satisfactory performance of MetaRF on few-shot yield prediction.

Predicting performance on each additive

For Buchwald-Hartwig HTE dataset, when using 20% of the data as the training set, the predicting performance of each additive in the testing set is shown in Fig. 7. In this experiment, the number of fine-tune samples is 5. For each additive, the predicted yield and observed yield are presented in a subplot. From Fig. 7, we can see that our model has satisfactory performance on new additives in the testing set, which shows that our model can quickly adapt with only 5 data points.

Interpretability analysis

For interpretability analysis, we visualize the most important DFT (Density Functional Theory) descriptors in the model trained on different sizes of Buchwald-Hartwig HTE data in Fig. 8. One measure of

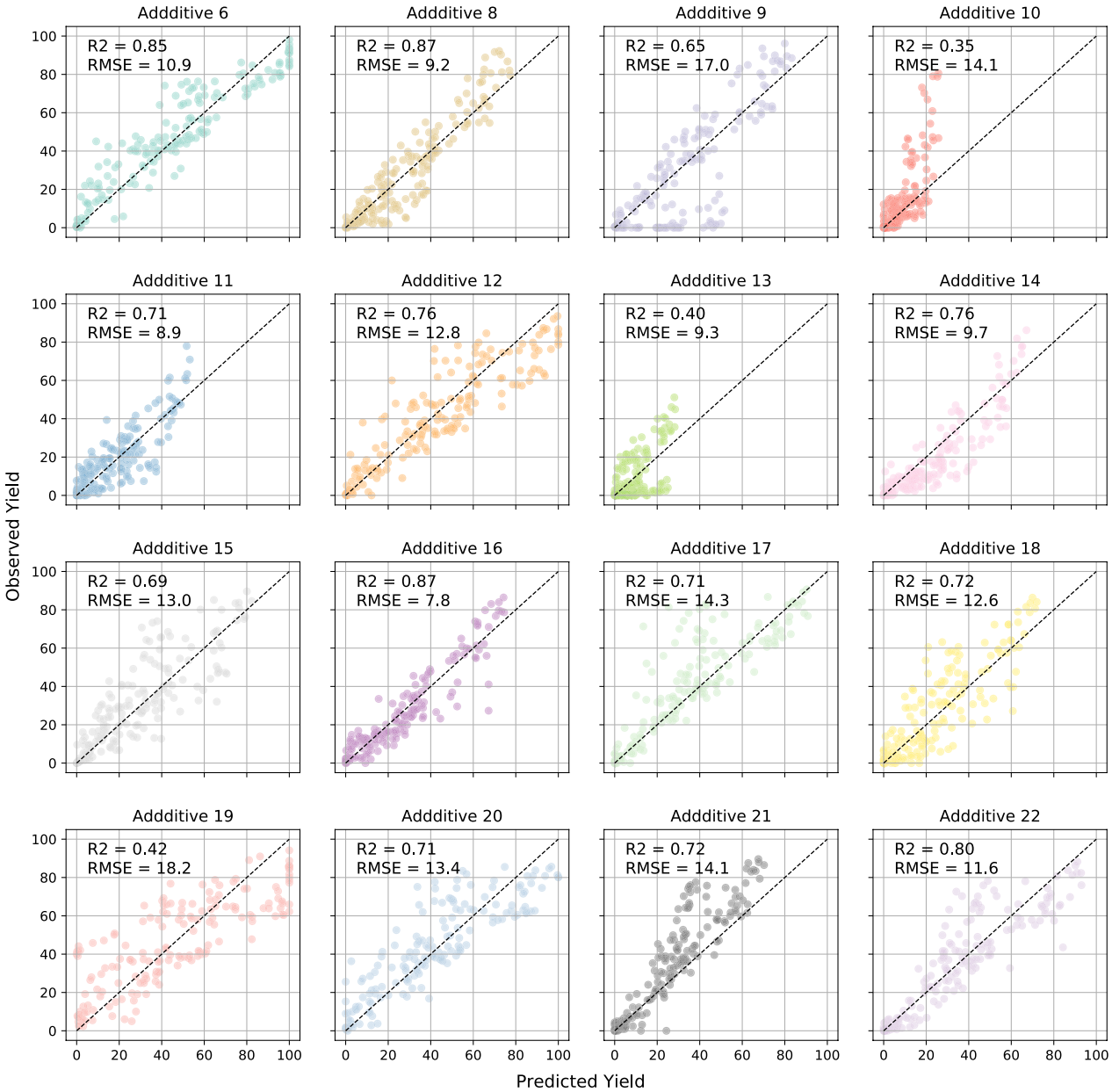

Fig. 7 The prediction results on different additives. For each additive in the testing set, the predicted yield and observed yield is presented in a subplot. The title of each subplot is the index number of the additive

feature importance is the decrease in the model's R^2 performance when the values of that feature are randomly shuffled, and the model is retrained [10]. The feature importance results of models trained on different sizes of data have a slight difference. Generally, the most important descriptors are aryl halide's *C3 nuclear magnetic resonance (NMR) shift (the asterisk indicates a shared atom), aryl halide's vibration frequency, additive's *C3 NMR shift and additive's *C3, *O1, *C4 electrostatic charges.

Discussion

The advantage of our method is that it can quickly adapt to predict the yield performance of new reagents while few additional samples are given. The underlying mechanism of this advantage is the adaption of the DFT feature importance. When MetaRF is fine-tuned with few additional samples, the importance of each decision tree model will change accordingly. Different tree models represent different distribution of DFT feature importance. Thus our model can change the DFT

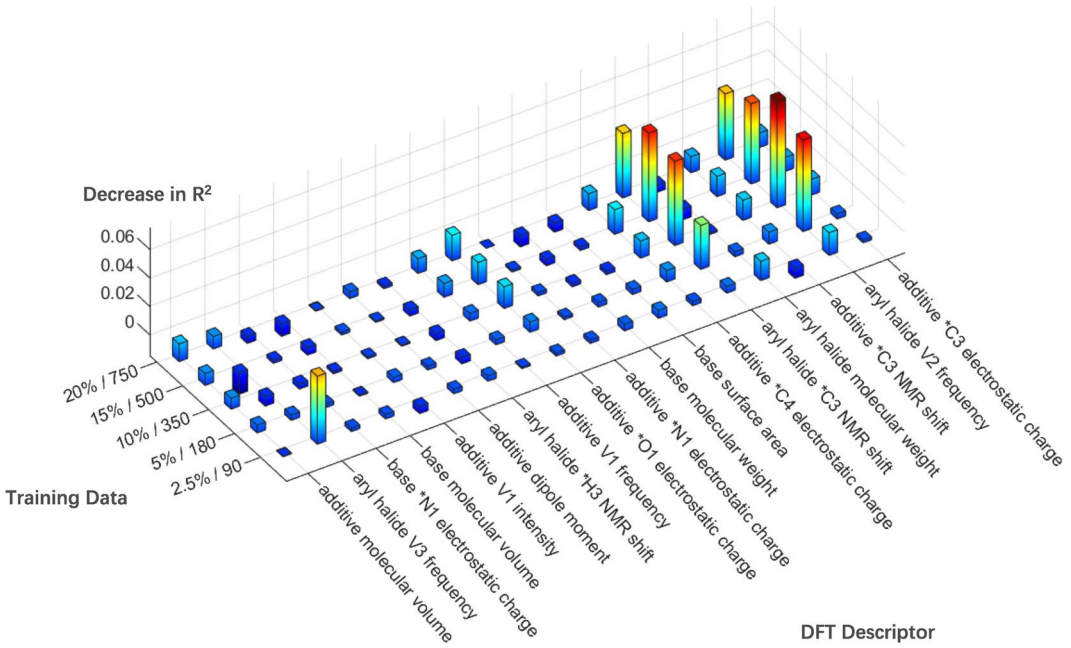

Fig. 8 Most important DFT descriptors of the model trained on different sizes of data. Feature importance is determined by the decrease in R^2 upon reshuffling the values of the feature. * indicates a shared atom. E indicates energy; HOMO indicates the highest occupied molecular orbital; V indicates vibration

Table 1 Method comparison results on Buchwald-Hartwig HTE dataset

	RMSE ↓	R^2 ↑	MAE ↓
MetaRF	**12.6401**	**0.7738**	**8.6850**
Random forest [10, 16]	15.6535	0.6538	11.1477
DRFP [42]	15.7849	0.6470	10.9302
RXNFP [43, 44]	27.4131	0.0032	23.8348
Neural network [45]	16.4294	0.6179	12.4626
Support vector machine [46]	18.1927	0.5322	14.2696
Linear model [47]	16.9791	0.5928	13.3449
GemNet [48]	18.6455	0.5245	14.0498

The bold values represent the performance of the proposed method

Table 2 Method comparison results on Suzuki-Miyaura HTE dataset

	RMSE ↓	R^2 ↑	MAE ↓
MetaRF	**18.3102**	**0.6350**	**13.8344**
Random forest [10, 16]	20.9239	0.5494	15.7090
DRFP [42]	21.0017	0.5158	15.1841
RXNFP [43, 44]	30.8297	0.0083	27.2992
Neural network [45]	23.3947	0.3961	19.8382
Support vector machine [46]	23.6997	0.3794	19.6562
Linear model [47]	24.3001	0.3480	19.9089
GemNet [48]	26.4485	0.2642	22.2841

The bold values represent the performance of the proposed method

Table 3 Method comparision results on Buchwald-Hartwig ELN dataset

	RMSE ↓	R^2 ↑	MAE ↓
MetaRF	**22.2252**	**0.3782**	**17.4771**
Random forest [10, 16]	26.0472	0.2511	20.8230
Neural network [45]	30.4522	-0.1197	23.2965
Support vector machine [46]	30.2837	-0.1283	23.0017
GemNet [48]	26.7396	0.1732	21.1857
YieldGNN [18]	/	0.1180	22.5000

The bold values represent the performance of the proposed method

Table 4 The results of ablation study(in Buchwald-Hartwig HTE dataset)

	RMSE ↓	R^2 ↑
Baseline	15.6535	0.6538
MetaRF + Dimension-reduction	**12.6401**	**0.7738**
MetaRF + Random[a]	14.5454	0.7003
MAML + Dimension-reduction[b]	21.0006	0.3730
MAML + Random	21.0204	0.3753
Transfer learning + Dimension-reduction[c]	14.4575	0.7038
Transfer learning + Random	15.2232	0.6705

The bold values represent the performance of the proposed method

[a] The first ablation test, replacing the dimension-reduction based sampling with random sampling

[b] The second ablation test, removing the random forest structure

[c] The third ablation test, replacing MAML with transfer learning framework

Table 5 The relative improvement of MetaRF compared to the baseline method(in Buchwald HTE dataset)

	RMSE ↓			R^2 ↑		
Sample[a]	MetaRF	Baseline	Margin[b]	MetaRF	Baseline	Margin[b] (%)
2	14.5912	16.7692	12.99%	0.6986	0.6026	15.94
4	12.7039	16.2212	21.68%	0.7717	0.6277	22.93
6	12.6674	15.3222	17.33%	0.7722	0.6681	15.58
8	12.0527	15.0728	20.04%	0.7929	0.6790	16.77
10	11.8515	14.7419	19.61%	0.7996	0.6922	15.52

[a] The number of fine-tune samples

[b] Relative improvement compared to the baseline method, random forest

feature importance according to few reaction samples from the new reagent.

A limitation of our method is that, it relies on historical data of the same reaction to train the model. According to the experimental results in Section-Performance Benchmarking, the training set should contain about 90 samples from the same reaction. In the actual usage, the reaction yield data from chemical literature may acts as training set. Considering the huge amount of reaction types, a possible future direction is cross-reaction prediction, which use some reaction data to predict the yield of another type of chemical reaction.

Conclusions

This paper proposes an attention-based random forest model to solve the few-shot yield prediction problem. The workflow includes using the DFT feature to encode chemical reactions and using the meta-learning framework to decide the attention weights of random forest. In the fine-tuning phase, we only need several samples to acquire satisfactory performance on new reagents. Our method obtains about 20% lower RMSE when the fine-tune sample varies from 4 to 10. The effective few-shot prediction demonstrates that our method can predict the effect of a new reactant structure with few additional data. The methodology in this paper brings benefits to future work on few-shot yield prediction.

Acknowledgements
The authors wish to acknowledge the support from Department of Computer Science and Engineering, The Chinese University of Hong Kong, New Territories, Hong Kong SAR; the support from Zhejiang Lab Zhejiang, China; and the support from College of Pharmaceutical Sciences, Zhejiang University, Zhejiang, China.

Authors' Information
K. Chen is a PHD student at the Department of Computer Science and Engineering at The Chinese University of Hong Kong. Her research interests include artificial intelligence and drug discovery.
G. Chen is a Zhejiang University - Zhejiang Lab Hundred-Program Researcher mainly based in Zhejiang Lab(https://en.zhejianglab.com/). His research interests mainly concentrate around fundamental artificial intelligence and drug discovery.
J. Li is a researcher at Zhejiang Lab. His research interests include traditional drug design/discovery, synthesis route planning and chemical synthesis.
Y. Huang is a postgraduate in College of Pharmaceutical Sciences at Zhejiang University. His research interests focus on computer-aided synthesis planning.
E. Wang is a research expert in Zhejiang Lab. His research interests include computational studies of protein-protein/ligand interactions and drug discovery.
T. Hou is a professor in College of Pharmaceutical Sciences at Zhejiang University. His research interests can be found at the website of his group:http://cadd.zju.edu.cn.
P. A. Heng is a professor at the Department of Computer Science and Engineering at The Chinese University of Hong Kong. His research interests can be found at the website of his group: http://www.cse.cuhk.edu.hk/~pheng/.

Author contributions
GC, TH and PAH conceptualized the study; KC and GC developed the methodology; KC, GC and JL performed data analysis and interpretation; KC, GC, JL, YH, EW, TH and PAH worked together to finish the paper. All authors read and approved the final manuscript.

Funding
This work is partially supported by National Key R&D Program of China (Grant No. 2022YFE0200700), National Natural Science Foundation of China (Grant No. 62006219), Hong Kong Innovation and Technology Fund (Grant No. ITS/170/20, ITS-241-21).

Availability of data and materials
All of the methods are implemented in Python. Source code is available at GitHub page: https://github.com/Nikki0526/MetaRF.

Declarations

Ethics approval and consent to participate
Not applicable.

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

Received: 3 January 2023 Accepted: 21 March 2023
Published online: 10 April 2023

References

1. Corey EJ, Wipke WT (1969) Computer-assisted design of complex organic syntheses: pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. Science 166(3902):178–192
2. Lowe DM (2012) Extraction of chemical structures and reactions from the literature. PhD thesis, University of Cambridge

3. Goodman J (2009) Computer software review: reaxys. J Chem Inf Mod 49(12):2897–2898
4. Gabrielson SW (2018) Scifinder. J Med Libr Assoc JMLA 106(4):588
5. Struble TJ, Alvarez JC, Brown SP, Chytil M, Cisar J, DesJarlais RL, Engkvist O, Frank SA, Greve DR, Griffin DJ et al (2020) Current and future roles of artificial intelligence in medicinal chemistry synthesis. J Med Chem 63(16):8667–8682
6. Coley CW, Green WH, Jensen KF (2018) Machine learning in computer-aided synthesis planning. Acc Chem Res 51(5):1281–1289
7. Fortunato ME, Coley CW, Barnes BC, Jensen KF (2020) Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. J Chem Inf Mod 60(7):3398–3407
8. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF (2017) Prediction of organic reaction outcomes using machine learning. ACS Cent Sci 3(5):434–443
9. Fortunato ME, Coley CW, Barnes BC, Jensen KF (2020) Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. J Chem Inf Mod 60(7):3398–3407
10. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG (2018) Predicting reaction performance in C–N cross-coupling using machine learning. Science 360(6385):186–190
11. Zuranski AM, Martinez Alvarado JI, Shields BJ, Doyle AG (2021) Predicting reaction yields via supervised learning. Acc Chem Res 54(8):1856–1865
12. Dong J, Peng L, Yang X, Zhang Z, Zhang P (2022) Xgboost-based intelligence yield prediction and reaction factors analysis of amination reaction. J Comput Chem 43(4):289–302
13. Zhu X, Ran C, Wen M, Guo G, Liu Y, Liao L, Li Y, Li M, Yu D (2021) Prediction of multicomponent reaction yields using machine learning. Chin J Chem 39(12):3231–3237
14. Chuang KV, Keiser MJ (2018) Comment on "predicting reaction performance in C–N cross-coupling using machine learning." Science 362(6416):8603
15. Estrada JG, Ahneman DT, Sheridan RP, Dreher SD, Doyle AG (2018) Response to comment on "predicting reaction performance in C–N cross-coupling using machine learning." Science 362(6416):8763
16. Breiman L (2001) Random forests. Mach learn 45(1):5–32
17. Luan J, Zhang C, Xu B, Xue Y, Ren Y (2020) The predictive performances of random forest models with limited sample size and different species traits. Fish Res 227:105534
18. Saebi M, Nan B, Herr J, Wahlers J, Guo Z, Zurański A, Kogej T, Norrby P-O, Doyle A, Wiest O et al (2021) On the use of real-world datasets for reaction yield prediction. ChemRxiv. https://doi.org/10.1039/D2SC06041H
19. Schleinitz J, Langevin M, Smail Y, Wehnert B, Grimaud L, Vuilleumier R (2022) Machine learning yield prediction from nicolit, a small-size literature data set of nickel catalyzed C–O couplings. J Am Chem Soc 144(32):14722–14730
20. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135
21. Zhao Z, Wallace E, Feng S, Klein D, Singh S (2021) Calibrate before use: Improving few-shot performance of language models. In: International Conference on Machine Learning, pp. 12697–12706
22. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11(1):137–148
23. Verleysen M, Francois D, Simon G, Wertz V (2003) On the effects of dimensionality on data analysis with neural networks. In: International Work-Conference on Artificial Neural Networks, pp. 105–112
24. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(11):2579–2605
25. Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, Richardson P, Sach NW (2018) A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science 359(6374):429–434
26. Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, Janey JM, Adams RP, Doyle AG (2021) Bayesian reaction optimization as a tool for chemical synthesis. Nature 590(7844):89–96
27. Żurański AM, Wang JY, Shields BJ, Doyle AG (2022) Auto-qchem: an automated workflow for the generation and storage of DFT calculations for organic molecules. React Chem Eng. https://doi.org/10.1039/D2RE00030J
28. Schneider N, Lowe DM, Sayle RA, Landrum GA (2015) Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. J Chem Inf Mod 55(1):39–53
29. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond J-L (2021) Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3(2):144–152
30. Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR)
31. Yang Y, Zhang Z, Mao W, Li Y, Lv C (2021) Radar target recognition based on few-shot learning. Multimed Syst. https://doi.org/10.1007/s00530-021-00832-3
32. Xia S, Xiong Z, Luo Y, Zhang G et al (2015) Effectiveness of the euclidean distance in high dimensional spaces. Optik 126(24):5614–5619
33. Li W, Cerise JE, Yang Y, Han H (2017) Application of t-SNE to human genetic data. J Bioinform Comput Biol 15(04):1750017
34. Kobak D, Berens P (2019) The art of using t-SNE for single-cell transcriptomics. Nat Commun 10(1):1–14
35. Tenenbaum JB, Silva Vd, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
36. Anowar F, Sadaoui S, Selim B (2021) Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Comput Sci Rev 40:100378
37. Gisbrecht A, Mokbel B, Hammer B (2012) Linear basis-function t-sne for fast nonlinear dimensionality reduction. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–8
38. Hinton GE, Roweis S (2002) Stochastic neighbor embedding. Adv Neural Inf Process Syst 15:857–864
39. Xu Y, Goodacre R (2018) On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test 2(3):249–262
40. Morais CL, Santos MC, Lima KM, Martin FL (2019) Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation kennard-stone algorithm approach. Bioinformatics 35(24):5257–5263
41. Perez-Guaita D, Ventura-Gayete J, Pérez-Rambla C, Sancho-Andreu M, Garrigues S, De La Guardia M (2012) Protein determination in serum and whole blood by attenuated total reflectance infrared spectroscopy. Anal Bioanal Chem 404(3):649–656
42. Probst D, Schwaller P, Reymond J-L (2022) Reaction classification and yield prediction using the differential reaction fingerprint DRFP. Digit Discov 1(2):91–97
43. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond J-L (2021) Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3(2):144–152
44. Schwaller P, Vaucher AC, Laino T, Reymond J-L (2021) Prediction of chemical reaction yields using deep learning. Mach Learn Sci Technol 2(1):015016
45. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851–869
46. Hasegawa K, Funatsu K (2010) Non-linear modeling and chemical interpretation with aid of support vector machine and regression. Curr Comput-Aided Drug Des 6(1):24–36
47. Yada A, Nagata K, Ando Y, Matsumura T, Ichinoseki S, Sato K (2018) Machine learning approach for prediction of reaction yield with simulated catalyst parameters. Chem Lett 47(3):284–287
48. Gasteiger J, Becker F, Günnemann S (2021) Gemnet: universal directional graph neural networks for molecules. Adv Neural Inform Process Syst 34:6790–6802
49. Schwaller P, Vaucher AC, Laino T, Reymond J-L (2021) Prediction of chemical reaction yields using deep learning. Mach Learn Sci Technol 2(1):015016
50. Granda JM, Donina L, Dragone V, Long D-L, Cronin L (2018) Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature 559(7714):377–381

Publisher's Note