Intermediate Knowledge Enhanced the Performance of N-Acylation Yield Prediction Model

Chonghuan Zhang,[†,¶] Qianghua Lin,[†,¶] Hao Deng,[‡] Yaxian Kong,[‡] Zhunzhun Yu,[*,†] and Kuangbiao Liao[*,†]

†*Guangzhou National Laboratory, No. 9 Xingdaohuanbei Road, Guangzhou International Bio Island, Guangzhou, Guangdong, PR China, 510005*

‡*AIChemEco Inc., Guangzhou, Guangdong, PR China, 510005*

¶*These authors contributed equally to this work.*

E-mail: yu_zhunzhun@gzlab.ac.cn; liao_kuangbiao@gzlab.ac.cn

Abstract

Acylation is an important reaction widely applied in medicinal chemistry. However, yield optimization remains a challenging issue due to the broad conditions space. Recently, accurate condition recommendations via machine learning have emerged as a novel and efficient method to achieve the desired transformations without a trial-and-error process. Nonetheless, accurately predicting yields is challenging due to the complex relationships involved. Herein, we present our strategy to address this problem. Two steps were taken to ensure the quality of the dataset. First, we skillfully selected substrates to ensure diversity and representativeness. Second, experiments were conducted using our in-house high-throughput experimentation (HTE) platform to minimize the influence of human factors. Additionally, we proposed an intermediate knowledge-embedded

strategy to enhance the model's robustness. The performance of the model was first evaluated at three different levels—random split, partial substrate novelty, and full substrate novelty. All model metrics in these cases improved dramatically, achieving an R^2 of 0.89, MAE of 6.1%, and RMSE of 8.0%. Moreover, the generalization of our strategy was assessed using external datasets from reported literature. The prediction error for nine reactions among 30 was less than 5%, and the model was able to identify which reaction in a reaction pair with a reactivity cliff had a higher yield. In summary, our research demonstrated the feasibility of achieving accurate yield predictions through the combination of HTE and embedding intermediate knowledge into the model. This approach also has the potential to facilitate other related machine learning tasks.

Introduction

N-acylation (referred to as acylation hereafter) reaction is one of the most important reactions in drug discovery.[1–3] It plays a crucial role in the synthesis of various pharmaceutical compounds and has significant implications in this field. For example, surveys in pharmaceutical chemistry have shown that the frequency of acylation reactions often dominates among all reaction types. Despite their importance, yield optimization still remains challenging due to the multitude of variables such as reagents, solvents, temperatures, and catalysts that must be considered to achieve high yields. Traditional strategies for optimization typically rely on prior knowledge and involve manual trial-and-error, which is a time-consuming and labor-intensive process.

To address these challenges, scientists have turned to machine learning (ML) models to predict reaction yields and streamline the optimization process. Recent publications underscore the importance of ML in reaction yield prediction.[4–7] For instance, Doyle *et al.* demonstrated the application of Random Forest algorithms to predict yields in C–N cross-coupling reactions, highlighting ML's potential to reduce experimental workloads and accelerate discovery.[8] Similarly, Schwaller *et al.* utilized neural networks to predict reaction

outcomes based on textual descriptions of experimental procedures, offering a novel approach to yield prediction.[9] However, both data quality and model could effect the prediction result, thus, yield prediction remains challenging up to now.

The first acylation reaction yield prediction model by Isayev *et al.* utilized literature-based reactions curated from Reaxys to build predictive models but highlighted the inherent difficulties of using such data. Literature reactions often suffer from inconsistencies in reporting, variability in experimental conditions, and a lack of comprehensive datasets, making it challenging to build robust and generalizable models.[10] Additionally, literature sources typically report only successful reactions with high yields, neglecting low-yield and negative data that are crucial for creating well-distributed and accurate predictive models.[11] It is essential to curate relevant datasets for model development and to identify and control factors that complicate yield prediction. The variability in data sources, reaction scales, and structural diversity reported in the literature further complicates the development of reliable models.[12] In summary, the quality of reaction datasets is paramount for building robust and high-performance prediction models.

High-throughput experimentation (HTE)[13] has emerged as a powerful alternative to traditional literature-based approaches for building reaction yield prediction models.[14–22] HTE techniques generate large datasets through automated, parallelized experiments, offering a more consistent, comprehensive, and controlled data source with a broader range of reaction conditions, including low-yield and negative outcomes. This systematic approach helps in developing more robust and generalizable models. With our in-house automated HTE platform, we have successfully optimized reaction conditions, explored the reaction space, and collected standardized experimental data for AI studies, resulting in a series of related publications.[17–23]

Despite these advantages, many HTE-based models achieve high accuracy but are limited to a narrow range of substrate and reaction condition spaces. Additionally, a common issue with these models is the evaluation methodology. Often, models are tested using data splits

that include test substrates seen by the model in the training set, resulting in overly optimistic performance metrics. However, when evaluated using a strict test set — where the model must predict yields for entirely new combinations of substrates—the performance typically drops. This strict testing better reflects real-world applications where chemists need to predict reaction yields for novel substrate pairs. Therefore, creating a reaction dataset with a diversified substrate and conditions and implementing rigorous testing protocols and curating relevant datasets is crucial for developing reliable and accurate predictive models.

In this context, we aim to build a high-quality dataset on acylation and develop a high-performance yield prediction model that can accurately recommend optimal conditions for novel substrate pairs in the training dataset. In this work, we first demonstrate our efforts to prepare the dataset. We selected substrate pairs according to structures reported in the USPTO reaction dataset[24] to ensure potential application and structural diversity. Second, our in-house HTE platform was utilized to collect data, ensuring repeatability. With the dataset in hand, we then focused on developing a robust prediction model. Given the challenge to develop a robust model under 95 conditions, we transformed our goal into an iterative prediction task across the 95 conditions list. Meanwhile, intermediate knowledge was embedded into the model to enhance its performance. The distinguishing feature of this strategy is that the model does not need to learn the relationships among different conditions, while still retaining condition information, thereby providing better performance with high probability. The results of a series of studies revealed that the generalization ability of the model could be significantly improved after applying this strategy (Figure 1). The yield prediction error for 9 out of 30 reaction samples, which were from literatures entirely unseen in the training dataset, was less than 5%, indicating the potential application of our work. Additionally, our strategy achieved satisfactory prediction results for reaction pairs with reactivity cliffs, delivering an accuracy of 0.73 in binary classification.

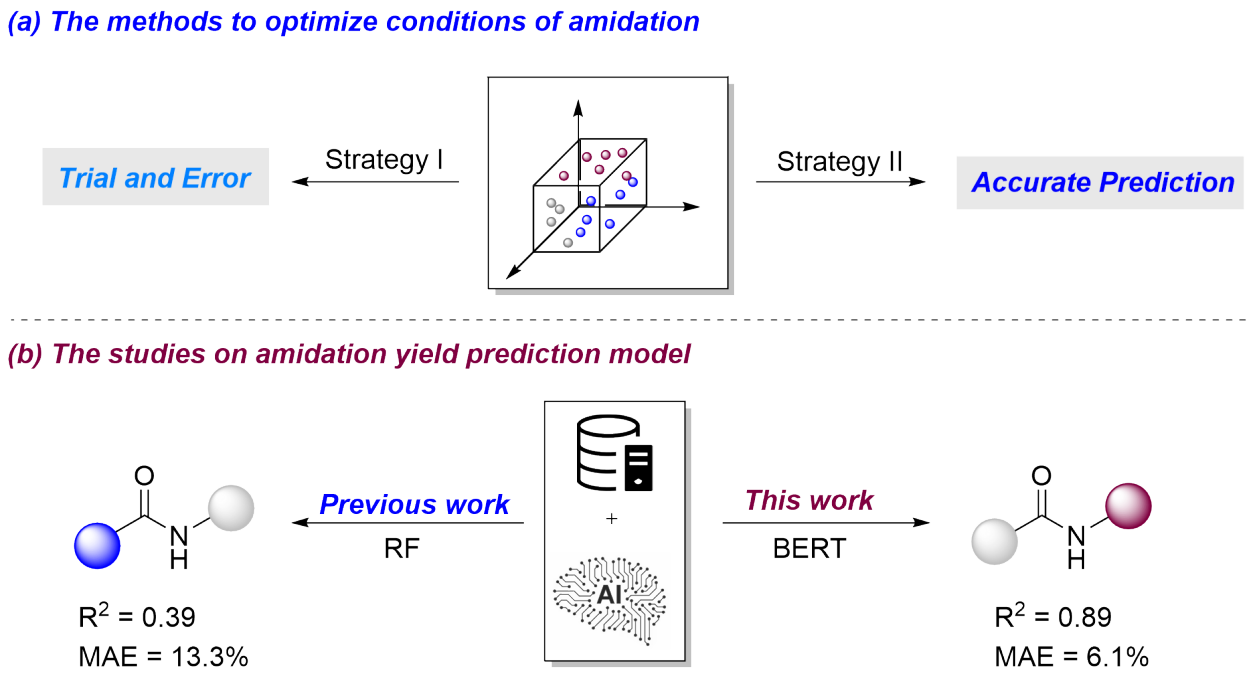

Figure 1: Studies on conditions optimization of acylation.

Results and discussion

HTE substrates selections

From the USPTO reaction dataset,[24] encompassing over 50,000 synthetic reactions derived from published US patents, we initiated our study by compiling a comprehensive dataset focusing on acylation reactions. The USPTO dataset is notable for its open-source availability, enabling reproducibility by others. We emphasize the biological activity and practical applications of the products derived from these reactions, underscoring their potential significance in various fields. To do this, as shown in Figure 2a, by following the general equation of acylation reaction, we first composed a reaction template in SMiles ARbitrary Target Specification (SMARTS) syntax.[25] We used RDKit[26] to filter acylation reactions from the USPTO dataset, which found 11663 entries of acylation reaction. Each entry in the dataset includes detailed reaction conditions, substrates, and products information. The product SMILES strings were then converted into extended connectivity fingerprints (1024-bit ECFP),[27] with 1024 bits and radius of 2, which serve as numerical representation of the

molecular structures. To manage the high-dimensional nature of the Morgan fingerprints and facilitate analysis, we employed a unsupervised learning technique, Principal Component Analysis (PCA)[28] to reduce the dimensionality of the data, preserving the variance inherent in the molecular descriptors. The data was reduced to a two-dimensional space, optimizing for the visualization and clustering of the chemical space.

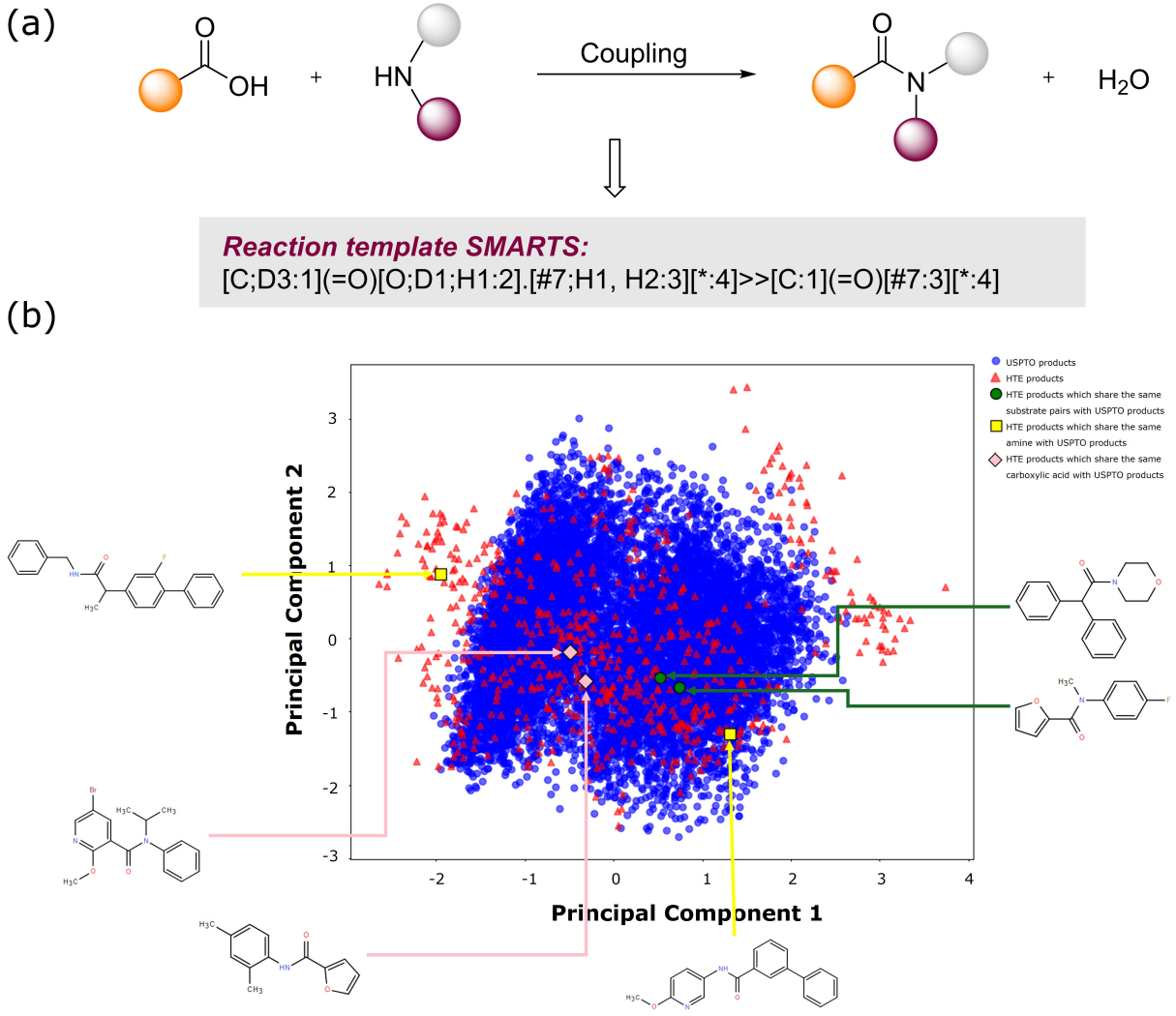

Figure 2: (a) acylation reaction general equation and SMILES template, (b) chemical space of patented acylation reaction data compared with self developed HTE data using the PCA of acylation reaction product

From the USPTO chemical space, we strategically selected substrates from readily available molecules to conduct high-throughput experimentation (HTE) due to the synthesis difficulty

of many USPTO substrates, which commonly originate from patented molecules. To ensure broad coverage across the chemical space, we identified 3,212 amines and 3,816 carboxylic acids from our in-house commercial molecule database curated from various chemical providers. We first selected a subset of commercially available substrates and enumerated the corresponding virtual amide products. These virtual products were then projected into the PCA-reduced space. Using MaxMin sampling, we started with a random substrate pair and then selected subsequent pairs that maximize the minimum distance to previously chosen pairs, ensuring diversity. This technique allowed us to select diverse product-substrate pairs, ensuring that a few substrate pairs were already present in the USPTO dataset. These selected pairs were combined with those derived from the USPTO data and divided into strata based on their PCA-reduced coordinates. Within each stratum, we conducted random sampling to ensure coverage of both dense and sparse regions of the chemical space. This approach allowed us to curate a balanced and representative set of substrate pairs for our high-throughput experimentation, effectively spanning the entire chemical space of interest.

Figure 2b illustrates the distribution of acylation products within a two-dimensional chemical space, revealing a meaningful distribution of data. Several example molecules are highlighted, demonstrating how different types of acylation products are organized into distinct distributions. These clusters also provide insights into the acylation substrates, as the substrates involved in each reaction can be inferred from the reaction templates. Only a few of the acylation products came from substrates seen in the USPTO, shown in green, yellow, and pink in Figure 2. However, the comparison of this chemical space indicates that our self-developed HTE reactions encompass a breadth of chemical diversity comparable to that of the USPTO dataset, which is recognized for its extensive coverage in reaction modeling. Although the USPTO dataset does not capture the entire chemical space historically explored by chemists,[29] it represents a robust and comprehensive starting point due to its open-access nature. This facilitates the replication of our study and underscores that our strategies to select substrates effectively mirrors the diversity and scope of chemical spaces reported in the

literature, which would be benefit to develop a robust model.

HTE conditions selection

According to the results of previous related work, we prepare 95 different conditions for HTE. The details of the 95 conditions are shown in Table S1 of the Supporting Information (SI). It should be noted that all commercial coupling reagents were involved in conditions set except for acyl chloride, because it is not compatible with DMF solvent. With the conditions and substrates pair in hand, we performed the HTE to collect reaction yield data. As a result, more than 47000 yield data were collected except for those discard data. Overlapping of chromatography peak and difficulty in NMR analysis usually result in failure to obtain the corresponding yield data. The details of our HTE platforms and experimentation procedures can be found from our previous publications.[17,18,20,23] This dataset was designed to be rich and diverse, providing a robust foundation for training machine learning models. The aim was to ensure that our HTE data would be sufficiently comprehensive to enable the models to understand and predict the yields of acylation reactions across the entire chemical space covered by our study.

Multi-condition model development and assessment

In this section, we describe the development and assessment of our multi-condition models for predicting acylation reaction yields. As shown in Figure 3, we employed several machine learning algorithms, including XGBoost,[30] Support Vector Machine (SVM),[31] Random Forest,[32] and AutoGluon,[33] utilizing 1024-bit ECFP fingerprints as descriptors.[27] Additionally, we explored the use of advanced deep learning textual methods such as yield-BERT[9] and T5-Chem,[34] which leverage reaction SMILES strings for yield prediction.

First, we generated 1024-bit ECFP descriptors from the SMILES strings of reactants and products. These ECFP descriptors, with 1024 bits and a radius of 2, capture the structural features of the molecules involved in the reactions. These fingerprints of substrates and

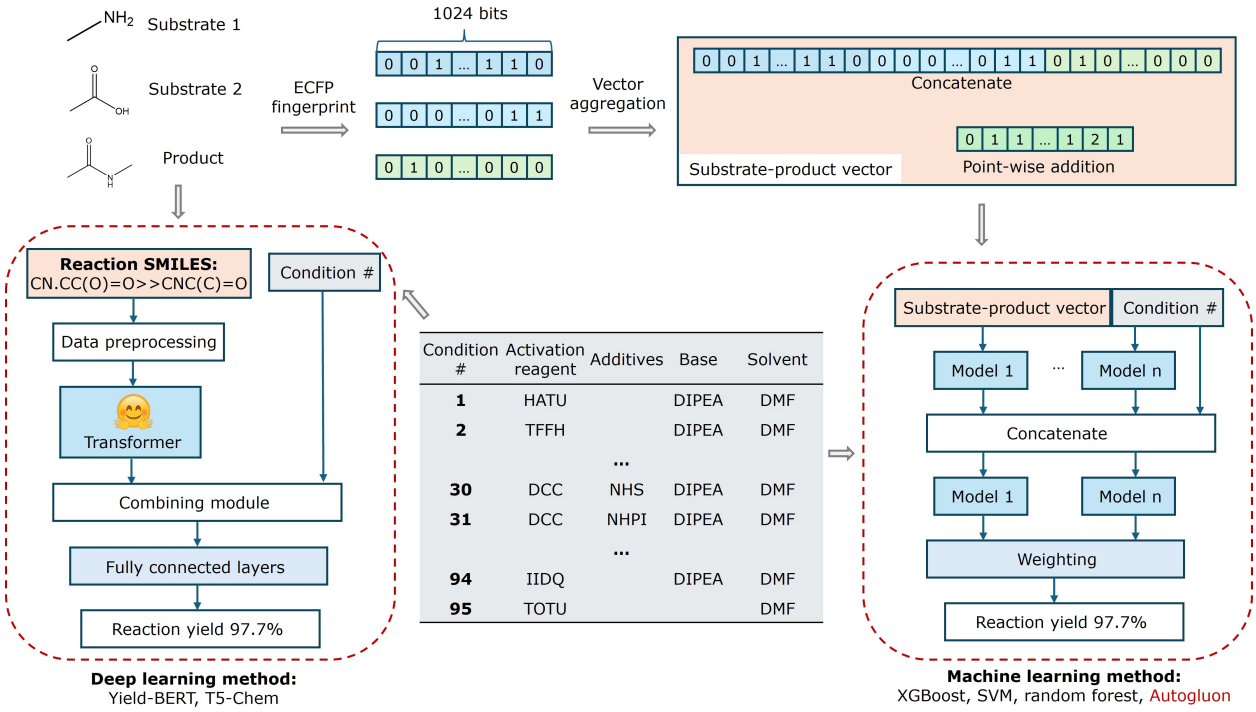

Figure 3: Schematics of multi-conditional model workflow, using methylamine reacting with acetic acid as an example.

product of a reaction were concatenated to create a vector of 3072 or point-wise added to keep size of 1024, as the reaction fingerprint. Each of the different reaction conditions was encoded in a unique integer (1-95). We then used these descriptors to train the XGBoost,[30] SVM,[31] Random Forest,[32] and AutoGluon[33] models. Each model was fine-tuned to optimize hyperparameters, ensuring the best performance for yield prediction. AutoGluon, a robust ensemble model, combines the strengths of various machine learning algorithms to improve predictive performance and model robustness.[35] Figure 3 uses the modelling workflow of AutoGluon to exemplify the machine learning approaches.

In parallel, we implemented deep learning yield-BERT[9] and T5-Chem[34] models. Yield-BERT, based on the BERT architecture[36] was trained on reaction SMILES to predict yields by understanding the sequence-to-sequence relationships within the reaction data. Similarly, T5-Chem, a variant of the T5 transformer model,[37] was also trained to capture contextual information from reaction SMILES strings, enabling it to predict yields by considering the

entire reaction context. To do this, we first tokenized the reaction SMILES strings, converting them into a sequence of tokens that represent the individual components of the molecules. These tokens were then fed through the transformer processes, which include multiple layers of attention mechanisms and feed-forward neural networks. Next, we incorporated the categorical features of reaction conditions into the model. These categorical features were combined with the text features output by the transformer through a combining module. This module integrated the encoded textual information from the SMILES strings with the reaction condition data. Finally, the combined features were passed through fully connected layers to predict the reaction yield. These layers consisted of several dense neural network layers that progressively refined the combined features into a single output value representing the predicted yield. In the above methods, we tried using fingerprints or reaction SMILES to represent the reaction conditions, but these did not represent the conditions well. Since we do not intend to predict outside these conditions, we opted for a categorical encoding approach to maintain clarity and consistency.

For the assessment of HTE-based reaction models, the conventional approach involves using a random split to build a test dataset. The HTE-based reaction data come from three dimensions: substrate 1 (amine), substrate 2 (carboxylic acid), and reaction conditions (encoded from 1 to 95). Randomly splitting the reaction data into training and test sets can result in some test substrates being seen by the model during training. For instance, the test dataset might include two familiar substrates with an unfamiliar condition or a familiar amine with an unfamiliar carboxylic acid, leading to overly optimistic performance metrics. However, when evaluated using a strict test set with entirely new substrate combinations (or at least one novel substrate), performance typically drops. This strict testing better reflects real-world applications, where chemists need to predict reaction yields for novel substrate pairs not previously encountered by the model.

To further assess the model, we developed three levels of test sets, as shown in Figure 4. The "random split" approach involves randomly dividing the dataset into training and testing

sets, ensuring that the model is exposed to a broad range of substrates and reaction conditions. The "partial substrate novelty" test is a subset of the Random Split, excluding any test cases where the reaction conditions are novel but involve substrates seen during training. In this test, one of the substrates is novel to the model, whilst the other one is known. The "full substrate novelty" test set is another subset of random split, and consists entirely of new combinations of substrates that has not encountered during training, providing a more rigorous evaluation of the model's predictive capabilities in real-world scenarios. These three levels of testing — random split, partial substrate novelty, and full novelty — offer a comprehensive assessment of the model's performance. Through this comprehensive approach, we have developed a suite of multi-condition models that leverage both traditional molecular descriptors and advanced textual methods, providing a robust framework for predicting acylation reaction yields across a wide range of chemical spaces.

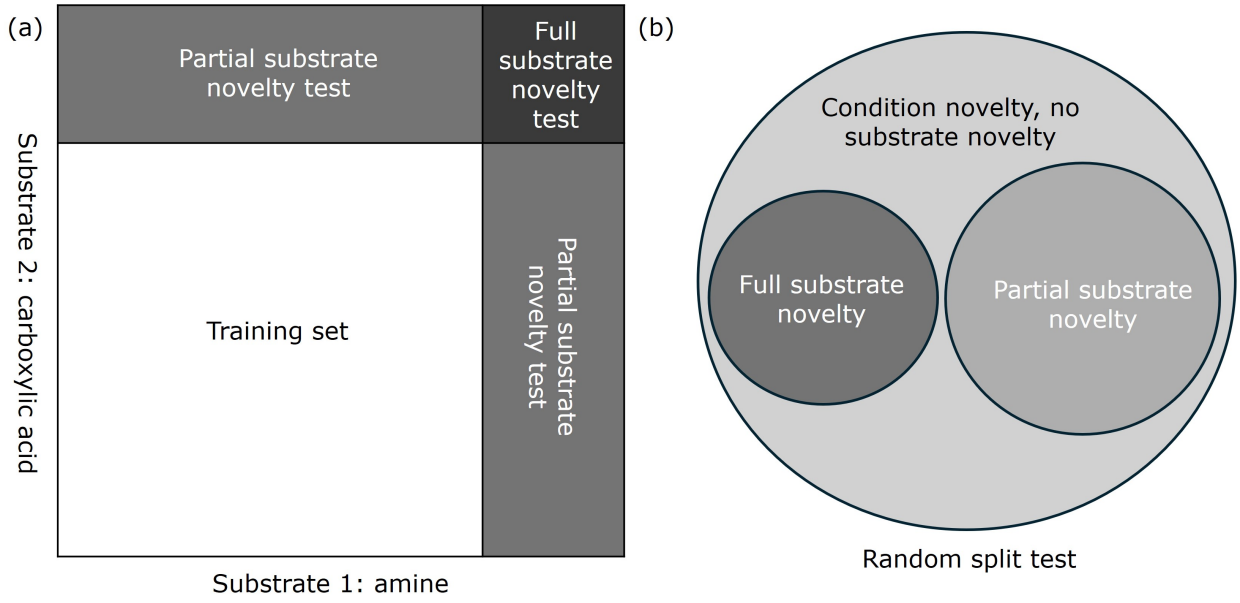

Figure 4: Schematics of three level of test sets - random split, partial substrate novelty, and full novelty test: (a) split of training and test sets in the dimension of two substrates, (b) vienn diagram of the three level of test sets

The results indicate that models performed better on the random split and partial substrate novelty test sets compared to the full substrate novelty test set. This is evident from the lower MAE and RMSE values and higher R^2 values for the first two splits (shown

in Table 1). These findings align with our expectations that models trained on datasets where they are exposed to a broad range of substrates and reaction conditions perform better on familiar substrates, but their performance drops when predicting yields for entirely new substrate pairs.

Table 1: Model evaluation under HTE data set.

Test data split	Metrics	XGBoost	SVM	RF	AutoGluon	Yield-BERT	T5-Chem
Random split	R^2	0.76	0.21	0.54	0.36	0.66	0.53
	RMSE	12%	23%	18%	24%	15%	22%
	MAE	9.00%	19.00%	17%	19%	10%	16%
Partial substrate novelty	R^2	0.71	0.19	0.52	0.3	0.68	0.58
	MAE	13%	24%	19%	26%	14%	20%
	RMSE	11%	21%	17%	21%	10%	15%
Full substrate novelty	R^2	0.65	0.22	0.51	-0.1	0.63	0.58
	MAE	14%	22%	19%	32%	15%	22%
	RMSE	12%	18%	17%	26%	11%	17%

In the random split dataset, XGBoost and BERT showed relatively better results with R^2 values of 0.76 and 0.66, respectively. These models outperformed SVM, Random Forest, T5, and AutoGluon. Even when evaluated on the full novelty substrate dataset, XGBoost and BERT still maintained higher R^2 values of 0.65 and 0.63, respectively, compared to the other models. This indicates that XGBoost and BERT are more robust and capable of generalizing better across different substrate combinations.

Single condition model development and assessment

The results from models trained under 95 different conditions demonstrated their reliability in accurately recommending conditions to facilitate optimal reactions, especially as model performance improved further. Given the complex structure-yield relationship (SYR) and the cost of data collection via HTE, we decided not to generate more data to enhance model performance. Inspired by the concepts of knowledge embedding[38] and dimensionality reduction, we proposed to achieve our goal through single condition model development presented as follows.

We could transform the yield prediction under multiple conditions into an iterative prediction within a condition list, a method we termed single condition prediction. In this approach, all reaction data within a single model were generated under the same set of conditions, thus eliminating reaction contexts such as condensation reagents, catalysts, bases, and solvents. This allowed the model to focus solely on the relationship between substrates and products, leading to improved learning and predictive accuracy. However, a significant challenge with this method is the potential loss of critical reaction condition information. Since reaction conditions play a crucial role in determining the outcome of chemical reactions, ignoring them can lead to incomplete models that do not accurately reflect real-world scenarios. To address this issue, we incorporated intermediate information based on reaction mechanisms into our model.

To evaluate our concept, we chose six different conditions with various coupling reagents that are frequently used in the literature and have well-defined intermediates. For single condition selection, we performed a statistical analysis of the literature-reported acylation reactions curated from Reaxys.[39] We identified the 25 most frequently used conditions, as shown in Table S6 of the SI, and selected six for model development and assessment, as shown in Table 2. These conditions were chosen based on their prevalence in the dataset, ensuring that our HTE conditions were both representative and relevant to a wide range of acylation reactions. In our investigation of single-condition models, we followed the meticulous

approach in preparing our dataset, ensuring it mirrored the model's rigor through three distinct datasets: random split, partial substrate novelty, and full substrate novelty.

Table 2: Details of the six reaction conditions

Condition #	Activation reagent	Additive	Base	Solvent
1	HATU		DIPEA	DMF
6	TBTU		DIPEA	DMF
13	EDC.HCl	HOBT	DIPEA	DMF
21	HBTU		DIPEA	DMF
34	PyBOP		DIPEA	DMF
79	DCC	HOBT		DMF

To simplify the complexity of the reaction system in multi-condition acylation reaction modeling, we developed multiple single-condition models by removing condition variables as mentioned above. Meanwhile, we incorporated intermediate information based on reaction mechanisms by using reaction SMARTS templates to represent the formation of activated acid intermediates. For example, in the presence of HATU as a condensation reagent, the transformation of an acid to its activated intermediate was represented by the following template shown in Figure 5(a).

This template converts the acid into the activated acid. We applied specific SMARTS templates for all the six conditions, which are detailed in the code repository. Next, we added the intermediate information into the reaction contexts, allowing the model to learn the effect of intermediates on the reaction outcome. To generate descriptors, we experimented with three approaches for generating the reaction context for the single-condition model, using the patterns of:

1. no intermediate
2. amine + acid + intermediate → amide
3. amine + intermediate → amide

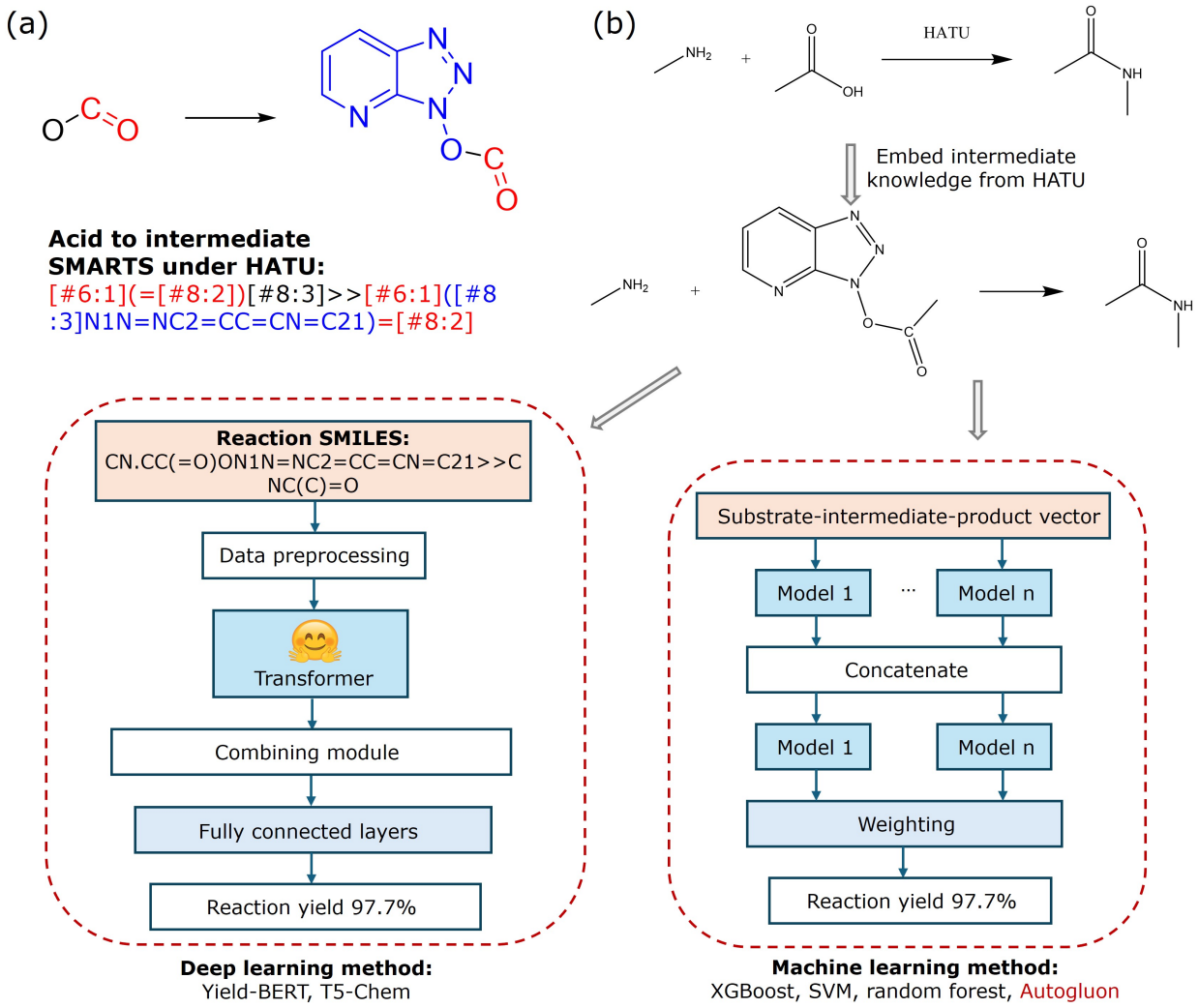

Figure 5: (a) Transformation of acid to intermediate SMARTS pattern under HATU as activation reagent, and (b) schematics of single-conditional model workflow

As shown in Figure 5(b), the reaction contexts were vectorized into ECFP fingerprints and also converted into reaction SMILES, effectively capturing the structural features of the reactants and products, along with crucial intermediate information. This approach ensured that the model considered the essential reaction conditions indirectly through the intermediate representation. The ECFP fingerprints and reaction SMILES were then used to train single-condition models using similar machine learning and deep learning algorithms respectively as those employed for multi-condition models (Figure 3). However, in this case, the reaction conditions were no longer concatenated with the reaction vector. We employed the same

rigorous testing protocols as used for the multi-condition models, evaluating performance across random split, partial substrate novelty, and full substrate novelty datasets.

Our results reveal that the BERT model trained on the random split dataset usually delivered superior performance, characterized by lower Mean Absolute Error (MAE) and Mean Squared Error (MSE), alongside higher R^2 values. This trend indicates that fewer variables enhance model accuracy. Moreover, descriptors incorporating intermediate information indeed enhanced performance. Specifically, under HATU and TBTU conditions, R^2 values surged from 0.69 and 0.71 to 0.86 and 0.84, respectively, with corresponding decreases in errors, underscoring the efficacy of our intermediate strategy. This robust performance of intermediate-inclusive descriptors persisted even in the full substrate novelty dataset, where the BERT model maintained an R^2 value of approximately 0.8 across all reaction conditions, albeit with slight reductions (Table 3). Among the intermediate-inclusive descriptors, the amine + intermediate approach usually outperformed the amine + acid + intermediate strategy across all reaction conditions when using the BERT algorithm (more metric details can be found in Table S7 of the SI). This observation aligns with the reaction mechanism, where amines and acids form intermediates before converting to products. Since our intermediate is represented as activated acid, it already encapsulates acid information, making the amine + acid + intermediate descriptors redundant. Consequently, the more precise amine + intermediate descriptors yield better results by avoiding redundant information and focusing on the critical reaction components.

Contrastingly, the XGBoost algorithm's performance lagged behind the multi-condition model, and the inclusion of intermediate descriptors did not enhance results, resulting in marginal declines (details on metrics, please see Table S4 and S7 in SI). This discrepancy between BERT and XGBoost is likely attributable to algorithmic differences. XGBoost, a machine learning algorithm, excels in learning simple reactions but struggles with the complexity added by intermediate descriptors. In contrast, the deep learning-based BERT algorithm thrives on this additional complexity, leveraging it to improve predictive accuracy.

Besides XGBoost, we also investigated other algorithms' performance after intermediate knowledge was embedded into the model. However, no better result was obtained in all cases (details on metrics please see Table S7 in SI).

By incorporating intermediate information, our single-condition models demonstrated significantly improved performance. The intermediate-powered model achieved an R^2 of 0.86, compared to an R^2 of 0.69 for the model without intermediate incorporation. This innovative strategy not only enhanced model accuracy but also provided a balanced approach that integrates condition-specific data with broader chemical knowledge, ultimately improving the robustness and generalizability of yield predictions for acylation reactions. This comprehensive approach ensures that our single-condition models are capable of accurately predicting reaction yields while considering the crucial role of reaction conditions through intermediate representations, thereby providing reliable and practical tools for chemists in optimizing acylation reactions.

Six-condition model development and assessment

Having identified the power of intermediate knowledge embedded in a model, we next aim to determine whether our strategy would also work well in the case of combining data from all six different conditions into one dataset . Indeed, the performance of the BERT model enhanced by intermediate knowledge improved, but the growth rate in performance was less than that observed in single-condition predictions, especially in cases of complete novelty splitting, as shown in Table 4. This may be because the model needs to learn the relationships among the six different conditions, but the data is insufficient for the model to learn these relationships effectively.

In summary, it is evident that intermediate-inclusive descriptors yield better results not only in single-condition predictions but also in multi-condition predictions. The absence of intermediate descriptors leads the model to erroneously assume that reactions depend solely on substrates, ignoring the significant impact of reaction conditions. This misassumption

Table 3: Performances of the BERT model in single condition prediction. The results with embedded intermediate knowledge are outside the parentheses, while the results with no intermediate knowledge are inside the parentheses.

Test data split	Metrics	TBTU	HATU	PyBOP	DCC	HBTU	EDC
Random split	R^2	0.84 (0.71)	0.86 (0.69)	0.90 (0.80)	0.86 (0.80)	0.89 (0.83)	0.89 (0.82)
	RMSE	10% (13%)	9.0% (14%)	8.0% (11%)	9.0% (11%)	9.0% (11%)	8.0% (11%)
	MAE	7.0% (10%)	6.0% (10%)	5.0% (8.0%)	7.0% (8.0%)	6.0% (8.0%)	6.1% (7.0%)
Partial substrate novelty	R^2	0.77 (0.57)	0.78 (0.53)	0.82 (0.63)	0.81 (0.74)	0.86 (0.72)	0.88 (0.79)
	MAE	12% (16%)	12% (17%)	10% (14%)	11% (13%)	9.0% (13%)	9.0% (12%)
	RMSE	8.0% (12%)	8.0% (13%)	7.0% (11%)	8.0% (9.0%)	7.0% (10%)	6.0% (8.0%)
Full substrate novelty	R^2	0.85 (0.66)	0.84 (0.39)	0.89 (0.40)	0.67 (0.1)	0.83 (0.68)	0.75 (0.46)
	MAE	9.0% (13%)	7.0% (14%)	8.0% (18%)	7.0% (12%)	10% (14%)	14% (18%)
	RMSE	7.0% (11%)	6.0% (11%)	6.0% (12%)	5.0% (10%)	7.0% (8.0%)	11% (13%)

Table 4: Performances of BERT under dataset of six different conditions

Intermediate information	Test data split	R^2	RMSE	MAE
Without intermediate knowledge	Random split	0.77	12%	9.00%
	Partial novelty	0.71	14%	10%
	Full novelty	0.62	10%	8.00%
With embedded intermeidate knowledge	Random split	0.85	10%	7.00%
	Partial novelty	0.8	11%	8.00%
	Full novelty	0.65	9.00%	8.00%

explains why models without intermediate descriptors perform well in the random split and known single-substrate datasets but experience sharp performance drops in the full novelty substrate dataset.

Prediction of external literature reactions

In order to demonstrate the generalization ability of our model enhanced by intermediate knowledge, we subsequently evaluated its performance on an unseen dataset originating from the literature. We collected data from 30 reactions related to medicinal chemistry to showcase the potential application of our model in this field. Five reactions were extracted from the literature for each of the six conditions mentioned in the proof of concept section. We then used the corresponding single-condition prediction models, which had the best performance in the full novelty dataset, to predict the yields of these reactions. The performance of the models was quite good, resulting in a mean absolute error of 12.2% (for details on prediction results, please see SI, Table S10). Excitingly, the prediction errors for 9 reactions were less than 5% (as shown in Figure 6).

For some substrates, the amidation transformations were challenging, and the relationships were complex due to the presence of heterocycles or special functional groups, such as amine group (4) and boronic acid group (7). Thus, achieving accurate yield predictions for these transformations was not easy. However, the model provided rather accurate predictions, indicating that it had indeed learned the complex structure-yield relationships. Given the structural diversity and very accurate predictions for some reactions, the model appears to have achieved a considerable balance between sensitivity and robustness.

Model performance evaluation in a benchmark dataset

After gaining a clear understanding of the generalization ability of the model trained on the HTE dataset, we aimed to evaluate the performance of the BERT model enhanced with intermediate knowledge on a benchmark dataset from Isayev's work. We prepared the dataset

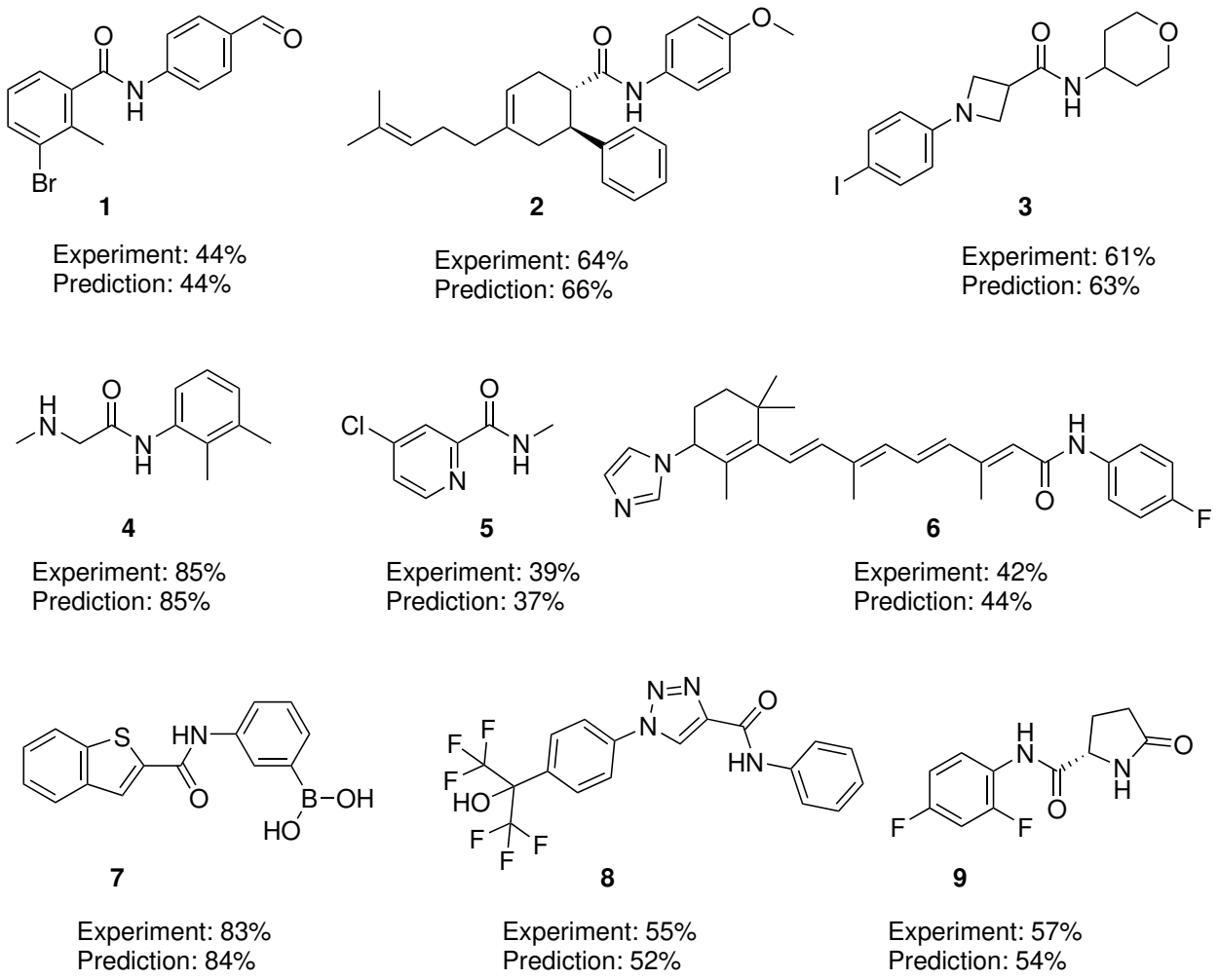

Figure 6: Prediction results of some external literature reaction examples

according to the list of reaction IDs provided in the report. The model's performance was evaluated via random split. Initially, we studied the performance of the standard BERT framework, which delivered modest metrics with an R^2 of 0.37, MAE of 13%, and RMSE of 18%. Howeever, the BERT model enhanced by intermediate knowledge improved performance to some extent, achieving an R^2 of 0.42, MAE of 12%, and RMSE of 16%. Isayev disclosed that reactivity cliffs were a reason for the poor performance of the model. Reactions were considered "cliffs" when their similarity surpassed 0.9, yet the yield difference was greater than 30. We were curious about whether our model's performance was affected by reactivity cliffs. Therefore, we predicted the yield of reactions identified as reactivity cliffs in Isayev's

work. The prediction error averaged 0.34, indicating the reactivity cliff may also weaken the performance of our model. Although the performance of BERT in regression was not optimal, the model could extrapolate which reaction from a reaction pair with a reactivity cliff would achieve a greater yield, with a classification accuracy of 0.73 (details on prediction results, please see the Table S11 in SI). To our delight, the model provided very satisfactory prediction results for reaction pairs that best fit the concept of reactivity cliffs (Figure 7).

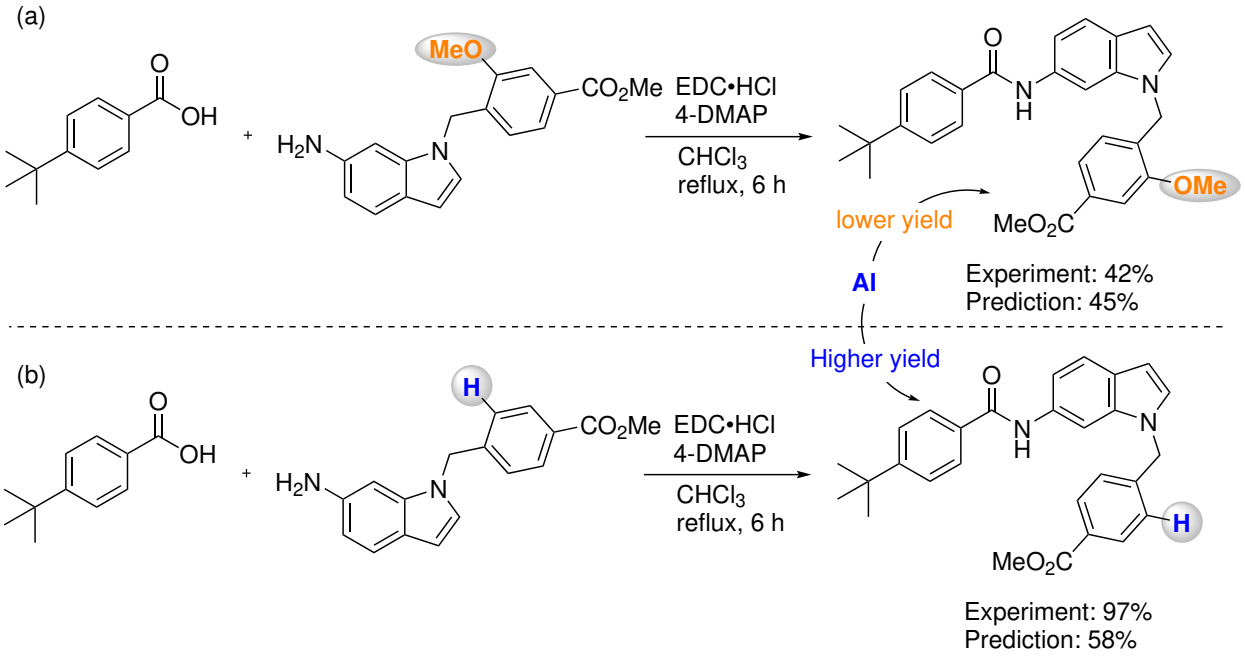

Figure 7: Prediction performance toward a reaction pair with reactivity cliff.

Conclusions

Accurate yield prediction is a crucial objective among many reaction-related prediction tasks, as several tasks can be viewed as yield prediction problems, including selectivity, conditions recommendation, catalysts design, ligands design, and more. Despite its importance, it remains a challenging issue due to the impact of both data quality and the generalization ability of the model. During the process of substrate pair selection, our goal was to match the diversity found in literature-reported reactions. This targeted method ensured that our

data collection was comprehensive and purposeful, rather than arbitrary. The data was then collected utilizing our in-house high-throughput experimentation (HTE) platform to ensure its quality. Our model's performance was validated through three levels of test sets — from random splits to strict tests — and further calibrated using recent unbiased external literature datasets. To address the challenges observed with strict test results, we proposed a strategy that enhances model performance by embedding domain-specific knowledge about reaction intermediates and dimension reduction. We evaluated our concept from different aspects, and the results revealed the importance of intermediate knowledge in elevating the model's performance. Excitingly, the model could even provide quite accurate predictions for some useful reactions reported in the literature. In summary, we developed an acylation yield prediction model with high performance by embedding intermediate knowledge into the model and employing dimension reduction, using computationally economical SMILES as input. Our strategy can also be applied to other related machine learning tasks to enhance model performance.

Corresponding authors

Kuangbiao Liao - Guangzhou National Laboratory

No. 9 Xingdaohuanbei Road, Guangzhou International Bio Island

Guangzhou 510005, Guangdong Province, China

liao_kuangbiao@gzlab.ac.cn

Zhunzhun Yu - Guangzhou National Laboratory

No. 9 Xingdaohuanbei Road, Guangzhou International Bio Island

Guangzhou 510005, Guangdong Province, China

orcid.org/0000-0002-0903-0818

yu_zhunzhun@gzlab.ac.cn

Authors

C. Zhang - Guangzhou National Laboratory

No. 9 Xingdaohuanbei Road, Guangzhou International Bio Island

Guangzhou 510005, Guangdong Province, China

Q. Lin - Guangzhou National Laboratory

No. 9 Xingdaohuanbei Road, Guangzhou International Bio Island

Guangzhou 510005, Guangdong Province, China

H. Deng - AIChemEco Inc., Guangzhou

Guangdong, PR China, 510006

Y. Kong - AIChemEco Inc., Guangzhou

Guangdong, PR China, 510005

Author contributions

Z. Yu and K. Liao conceived and supervised the project. C. Zhang, Q. Lin, H. Deng, Z. Yu built the model and performed the evaluation. The manuscript was written through contributions of C. Zhang, Q. Lin, Z. Yu, K. Liao. / All authors have given approval to the final version of the manuscript.

Acknowledgement

We acknowledge National Natural Science Foundation of China (No. 22393892 and No. 22071249) for the financial support for this work.

Data & code availability

The data and code related to model development and evaluation could be found at following link: `https://www.github.com/aichemeco/Acylation/tree/main`.

References

(1) Brown, D. G.; Bostrom, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? Miniperspective. *Journal of medicinal chemistry* **2016**, *59*, 4443–4458.

(2) Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery* **2018**, *17*, 709–727.

(3) Pattabiraman, V. R.; Bode, J. W. Rethinking amide bond synthesis. *Nature* **2011**, *480*, 471–479.

(4) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Accounts of Chemical Research* **2021**, *54*, 1856–1865, PMID: 33788552.

(5) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; others On the use of real-world datasets for reaction yield prediction. *Chemical science* **2023**, *14*, 4997–5005.

(6) Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. V. When yield prediction does not yield prediction: an overview of the current challenges. *Journal of Chemical Information and Modeling* **2023**, *64*, 42–56.

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

(8) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360*, 186–190.

(9) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2021**, *2*, 015016.

(10) Liu, Z.; Moroz, Y. S.; Isayev, O. The challenge of balancing model sensitivity and robustness in predicting yields: a benchmarking study of amide coupling reactions. *Chemical Science* **2023**, *14*, 10835–10846.

(11) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine learning for chemical reactivity: the importance of failed experiments. *Angewandte Chemie International Edition* **2022**, *61*, e202204647.

(12) Raghavan, P.; Rago, A. J.; Verma, P.; Hassan, M. M.; Goshu, G. M.; Dombrowski, A. W.; Pandey, A.; Coley, C. W.; Wang, Y. Incorporating Synthetic Accessibility in Drug Design: Predicting Reaction Yields of Suzuki Cross-Couplings by Leveraging AbbVie's 15-Year Parallel Library Data Set. *Journal of the American Chemical Society* **2024**, *146*, 15070–15084, PMID: 38768950.

(13) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Accounts of Chemical Research* **2017**, *50*, 2976–2985, PMID: 29172435.

(14) Fu, Z.; Li, X.; Wang, Z.; Li, Z.; Liu, X.; Wu, X.; Zhao, J.; Ding, X.; Wan, X.; Zhong, F.; others Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki–Miyaura cross-coupling reaction. *Organic Chemistry Frontiers* **2020**, *7*, 2269–2277.

(15) Götz, J.; Jackl, M. K.; Jindakun, C.; Marziale, A. N.; André, J.; Gosling, D. J.; Springer, C.; Palmieri, M.; Reck, M.; Luneau, A.; others High-throughput synthesis

provides data for predicting molecular properties and reaction success. *Science advances* **2023**, *9*, eadj2314.

(16) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. Machine learning C–N couplings: Obstacles for a general-purpose reaction yield prediction. *ACS omega* **2023**, *8*, 3017–3025.

(17) Xu, Y.; Gao, Y.; Su, L.; Wu, H.; Tian, H.; Zeng, M.; Xu, C.; Zhu, X.; Liao, K. High-Throughput Experimentation and Machine Learning-Assisted Optimization of Iridium-Catalyzed Cross-Dimerization of Sulfoxonium Ylides. *Angewandte Chemie International Edition* **2023**, *62*, e202313638.

(18) Qiu, J.; Xu, Y.; Su, S.; Gao, Y.; Yu, P.; Ruan, Z.; Liao, K. Auto Machine Learning Assisted Preparation of Carboxylic Acid by TEMPO-Catalyzed Primary Alcohol Oxidation. *Chinese Journal of Chemistry* **2023**, *41*, 143–150.

(19) Xu, Y.; Ren, F.; Su, L.; Xiong, Z.; Zhu, X.; Lin, X.; Qiao, N.; Tian, H.; Tian, C.; Liao, K. HTE and machine learning-assisted development of iridium (I)-catalyzed selective O–H bond insertion reactions toward carboxymethyl ketones. *Organic Chemistry Frontiers* **2023**, *10*, 1153–1159.

(20) Yu, Z.; Kong, Y.; Li, B.; Su, S.; Rao, J.; Gao, Y.; Tu, T.; Chen, H.; Liao, K. HTE- and AI-assisted development of DHP-catalyzed decarboxylative selenation. *Chemical Communications* **2023**, *59*, 2935–2938.

(21) Qiu, J.; Xie, J.; Su, S.; Gao, Y.; Meng, H.; Yang, Y.; Liao, K. Selective functionalization of hindered meta-C–H bond of o-alkylaryl ketones promoted by automation and deep learning. *Chem* **2022**, *8*, 3275–3287.

(22) Li, B.; Su, S.; Zhu, C.; Lin, J.; Hu, X.; Su, L.; Yu, Z.; Liao, K.; Chen, H. A deep learning framework for accurate reaction prediction and its application on high-throughput experimentation data. *Journal of Cheminformatics* **2023**, *15*, 72.

(23) Xu, Y.; Ren, F.; Su, L.; Xiong, Z.; Zhu, X.; Lin, X.; Qiao, N.; Tian, H.; Tian, C.; Liao, K. HTE and machine learning-assisted development of iridium (I)-catalyzed selective O–H bond insertion reactions toward carboxymethyl ketones. *Organic Chemistry Frontiers* **2023**, *10*, 1153–1159.

(24) Lowe, D. Chemical reactions from US patents (1976-Sep2016). `https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873`, 2017; Accessed: 12 June 2024.

(25) Daylight Chemical Information Systems, I. SMiles ARbitrary Target Specification (SMARTS). `http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`, 2023; Accessed: 2024-06-12.

(26) RDKit RDKit: cheminformatics and machine learning software. 2023; `https://www.rdkit.org/`.

(27) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

(28) Jolliffe, I. T. Mathematical and statistical properties of sample principal components. *Principal component analysis* **2002**, 29–61.

(29) Thakkar, A.; Kogej, T.; Reymond, J.-L.; others Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science* **2019**, *10*, 10302–10313.

(30) Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(31) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.

(32) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.

(33) Erickson, N.; Mueller, J.; Gupta, S. R.; et al. AutoGluon-Tabular: Robust and accurate AutoML for structured data. *arXiv preprint arXiv:2003.06505* **2020**,

(34) Lu, J.; Zhang, Y. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *Journal of Chemical Information and Modeling* **2022**, *62*, 1376–1387, PMID: 35266390.

(35) AutoGluon AutoML for text, image, and tabular data. 2023; `https://auto.gluon.ai/`.

(36) Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* **2018**, *abs/1810.04805*.

(37) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* **2019**, *abs/1910.10683*.

(38) Stuyver, T.; Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *The Journal of Chemical Physics* **2022**, *156*.

(39) Elsevier Reaxys. `https://www.reaxys.com`, Accessed: 2024-06-28.

TOC Graphic