Uncertainty-Aware Yield Prediction with Multimodal Molecular Features

Jiayuan Chen[1], Kehan Guo[2], Zhen Liu[3], Olexandr Isayev[3], Xiangliang Zhang[2]*

[1]The Ohio State University
[2] Department of Computer Science and Engineering, University of Notre Dame
[3]Department of Chemistry, Carnegie Mellon University
chen.12930@osu.edu, kguo2@nd.edu, liu5@andrew.cmu.edu, olexandr@olexandrisayev.com, xzhang33@nd.edu

Abstract

Predicting chemical reaction yields is pivotal for efficient chemical synthesis, an area that focuses on the creation of novel compounds for diverse uses. Yield prediction demands accurate representations of reactions for forecasting practical transformation rates. Yet, the uncertainty issues broadcasting in real-world situations prohibit current models to excel in this task owing to the high sensitivity of yield activities and the uncertainty in yield measurements. Existing models often utilize single-modal feature representations, such as molecular fingerprints, SMILES sequences, or molecular graphs, which is not sufficient to capture the complex interactions and dynamic behavior of molecules in reactions. In this paper, we present an advanced Uncertainty-Aware Multimodal model (UAM) to tackle these challenges. Our approach seamlessly integrates data sources from multiple modalities by encompassing sequence representations, molecular graphs, and expert-defined chemical reaction features for a comprehensive representation of reactions. Additionally, we address both the model and data-based uncertainty, refining the model's predictive capability. Extensive experiments on three datasets, including two high throughput experiment (HTE) datasets and one chemist-constructed Amide coupling reaction dataset, demonstrate that UAM outperforms the state-of-the-art methods. The code and used datasets are available at https://github.com/jychen229/Multimodal-reaction-yield-prediction.

Introduction

Computer-Assisted Synthesis Prediction (CASP) has emerged as a key area of focus in the intersection of artificial intelligence in scientific domains. The goal of CASP revolves around tackling a diverse array of chemical challenges, including the prediction of reaction products (Coley et al. 2017) and the intricacies of retro-synthesis (Ishida et al. 2019). Yield prediction, among the spectrum of CASP tasks, is particularly crucial. The target of yield prediction is to accurately estimate the practical conversion rates in chemical reactions, illustrating the transition from reactants to products. In this context, yield prediction lays the foundation for reaction-related predictions, thereby supporting the advancements in CASP (Ahneman et al. 2018).

When conceptualized as a machine learning problem, yield prediction is essentially a regression task. The development of an effective yield prediction model depends critically on obtaining high-quality representations of the reactants and products involved in chemical reactions. Early, molecular fingerprints were employed to depict chemical structures, yet their efficacy in handling complex structures was limited. Deep learning-based methods can automatically learn intricate patterns and features from data. For instance, (Schwaller et al. 2020) employ BERT (Devlin et al. 2018), a bidirectional transformer language model, for learning the representation of molecules involved in chemical reactions based on their sequential SMILES expressions. This learned representation is then utilized in a subsequent regression model to predict yields. Similarly, (Kwon et al. 2022) employ molecular graphs to represent molecules within chemical reactions and utilize graph neural networks to learn useful features for yield prediction. These current yield prediction models exhibit strong performance on specially curated reaction datasets, such as the High-Throughput (HTE) datasets (Ahneman et al. 2018; Perera et al. 2018). However, when applied to real-world tasks, their efficacy diminishes significantly (Saebi et al. 2023). One primary reason for this decline is the pervasive issue of uncertainty in real-world yield prediction datasets, manifesting in two major aspects.

High sensitivity of yield. In chemical reactions, structural isomers—compounds with identical molecular formulas but different arrangements of atoms—can significantly impact the yield. Even minor structural variations within the reactants themselves can lead to pronounced discrepancies in the resulting yields. For example, the addition of a methoxy group that is far from the reaction center can lower the reaction center by as much as 55% (Schierle et al. 2020). This highlights how real-world reactions can be extremely sensitive to slight variations in the reactants and products involved. Existing models, as referenced by (Schwaller et al. 2021), primarily utilize single-modal data such as graphs or sequences, and thus may not adequately capture the subtle structural variations in molecules. These subtle yet critical variations include minor differences in stereochemistry and the presence of specific functional groups, both of which can have a significant impact on reaction pathways and yields.

Uncertainty in the yield measurement. The yield from

[*] The corresponding author.

the reaction process depends on many factors in the reaction cycle, including the properties of the molecules, the environmental condition, and human operations. As a result, the same reaction can exhibit significant yield variations. For example, (Liu, Moroz, and Isayev 2023) pointed out that the yield standardized deviation can be as large as 23.7% when the same reaction was reported by different research groups. Although (Kwon et al. 2022) considered yield prediction uncertainty and introduced an uncertainty-related loss for training the prediction model, the inherent intricacies of data uncertainty hinder a precise prediction.

To address the aforementioned challenges, we propose an advanced Uncertainty-Aware Multimodal model (UAM) for yield prediction by taking into account multi-modal features to combat the prediction uncertainty. Specifically, we introduce a multi-modal feature extractor that integrates sequence features, graph structural features, and human-defined reaction condition features to acquire a more comprehensive representation of reactants and products. Moreover, aided by cross-modal contrastive learning, we facilitate modal fusion to capture the shared information and distinctive features across modalities to alleviate discrepancies induced by the high sensitivity of yield. Additionally, we incorporate a Mixture-of-Experts (MoE) module to enhance model expressiveness without additional computational costs. This facilitates a dynamic equilibrium between the model's sensitivity to variations and its ability to discern reaction types. Last, we introduce an uncertainty quantification module, which mitigates the inherent training uncertainty of the model while focusing on quantifying the uncertainty presented in the data itself, thereby enhancing predictive accuracy.

Our contributions in this work are summarized as follows:

- We study the reaction yield prediction problem and proposed a novel model called UAM to tackle the uncertainty issue by fusing multi-modal molecular features;
- We explore an innovative and effective way to utilize cross-modal contrastive learning and an additional MoE module is added to enhance the reaction representation;
- Experimental results on three real-world datasets demonstrate the effectiveness of UAM in comparison to the state-of-the-art approaches.

Related Work

Molecular Representation Learning

Molecular representation learning is a crucial link between machine learning and chemistry and is gaining rising awareness in computational chemistry. Early techniques manually compute chemical descriptors like Morgan fingerprints (Pattanaik and Coley 2020; Sandfort et al. 2020) or Density functional theory (DFT) descriptors (Hu et al. 2003) to obtain numerical vector representations of molecules. Lately, deep learning is gaining attention with two main categories: sequence-based and graph-based methods. The first category builds upon the practice that molecules are often represented as SMILES string (Weininger, Weininger, and Weininger 1989). These methods leverage sequence deep neural network models such as Recurrent Neural Network (Segler et al. 2018) and Transformer (Schwaller et al. 2019, 2021) to effectively encode molecular information. The second category, graph-based methods, concentrates on the atom-atom connection patterns within molecules (Guo et al. 2023c). This approach stems from the understanding that a molecule's activity and properties are often closely linked to its structural information. Although SMILES string captures sequential details, they can lose global context in cases of lengthy SMILES sequences. In contrast, graph-based molecular representation (Hu et al. 2019; Guo et al. 2021; Wang et al. 2021; Li, Zhao, and Zeng 2022) preserves structural information by naturally mapping molecules into graphs with atoms as nodes and bonds as edges. However, molecular representations that rely on a single modality have inherent limitations. Graph-based models may not inherently represent the stereochemistry of molecules, such as the R/S configuration in chiral centers or E/Z configuration in double bonds. SMILES, however, can be extended to include stereochemical information by using or symbols. While human-defined features incorporate abundant domain knowledge, they require complex pre-computation and may not produce the most task-relevant and generalizable molecular features. In this paper, we propose a multi-modal molecular representation encoding followed by a late fusion, so it effectively captures the inherent characteristics of chemical reactions.

Reaction Yield Prediction

Chemical reaction yield prediction is a crucial application in machine learning for chemical synthesis. The reaction yield is typically a certain percentage of the theoretical chemical conversion. Therefore, in evaluating the reaction yield, the representation learning of both reactants and products plays an important role. Earlier, (Ahneman et al. 2018) utilizes molecular descriptors with off-the-shelf machine learning models such as Random Forest to predict cross-coupling reactions. However, such methods are limited to specific reaction categories and require expert intuition to select the appropriate chemical fingerprints. Deep learning has enabled the utilization of sequence-based and graph-based models for general reaction yield prediction (Guo et al. 2021, 2023a,b). For instance, YieldBert (Schwaller et al. 2020, 2021) employs transformers to encode reaction SMILES for context-dependent molecular information. Meanwhile, other approaches (Gilmer et al. 2017; Kwon et al. 2022) leverage GNNs to predict yields using graph-based molecular representations. However, due to the inherent limitations of learning representations from single-model data, these models exhibit suboptimal performance on real-world datasets. They fail to account for the uncertainty arising from factors such as reaction conditions (temperature, time), side reactions, reactant degradation, and other influences. (Kwon et al. 2022) is the most related work to ours for considering uncertainty in yield prediction. However, it merely predicted additional variance for auxiliary training without conducting an intricate and comprehensive analysis of uncertainty inherent in chemical reactions. In this paper, we analyze the sources of uncertainty and employ uncertainty quantification techniques to enhance the performance of yield predictions.

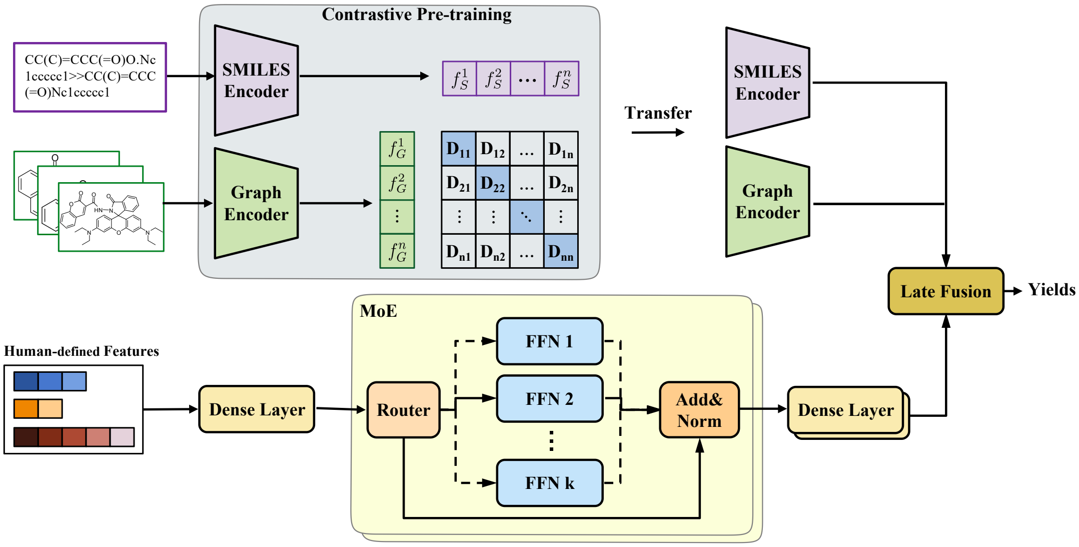

Figure 1: The framework of our approach UAM, which consists of three encoders: graph encoder, SMILES encoder, and human-defined feature encoder. The top part shows the contrastive pre-training for combining the representation from SMILES and graph encoder. The lower part depicts the encoding process for human-defined features. This process is structured with a densely-connected layer, followed by the Mixture of Experts (MoE) module, and then another series of dense layers. The late fusion module is designed by either voting fusion, feature concatenation, or self-attention weighted fusion for predicting the yields. The SMILES and graph encoders are initially pre-trained through contrastive learning, and then, along with the dense layers, the MoE and fusion modules, they undergo an end-to-end fine-tuning.

Methodology

In this section, we first define the multi-modal yield prediction problem and then present the details of our model.

Problem Definition Let $R = \{R_1, ..., R_N\}$ be a set of chemical reactions and $Y = \{y_1, ..., y_N\}$ be the reaction yields representing the percentage conversion of reactants into products, where N is the number of reactions. Given a reaction $R_i \in R$, our model's input comprises molecular graphs $\{G^i_{r_1}, ..., G^i_{r_n}, G^i_{p_1}, ..., G^i_{p_m}\}$, SMILES sequence S^i, and human-defined features H^i (e.g., molecular fingerprints, reaction conditions), where r denotes reactants, p represents products, and n and m are their respective quantities. Typically, most of the reactions involve n=2 reactants and m=1 or 2 products. The yield of a reaction y_i is a real value between 0 and 1. The goal of yield prediction is to develop a mapping function, $f_\Theta : R \rightarrow Y$. This function involves encoding R_i into representation vectors and subsequently associating these vectors with the prediction target y_i.

Model Architecture The architecture of our approach is shown in Figure 1. The model consists of four components: graph encoder, SMILES encoder, human-defined feature encoder, and multi-modal fusion. The SMILES and graph encoders are pre-trained with a contrastive learning strategy. Subsequently, these encoders, in conjunction with the dense layers, MoE and fusion modules, are subjected to end-to-end fine-tuning. The embedding vectors for the reactant-product SMILES sequences are represented as f_S, while those for the reactant-product molecular graphs are denoted as f_G. The human-defined reaction features, after being processed through a mixture-of-experts feature encoder, are represented as low-dimensional features f_H. These features, derived from the three modalities, are then fed into a perceiver for late fusion. Finally, we introduce an uncertainty quantification module to enhance the model's performance. The following sections detail each component of the model.

Graph Encoder. For reaction R_i, the graph encoder encodes the reactants and products separately, and concatenates them as the output embedding f^i_G:

$$f^i_G = \textbf{Concat}\left[Enc(G^i_{r_1}), \dots, Enc(G^i_{p_m})\right]. \tag{1}$$

As shown in Figure 2, the graph encoder includes a node information propagation module and a graph-level global pooling module. The node information propagation module has two components: feature mapping for nodes and edges, and feature aggregation. Considering the atom heterogeneity and bonding affinity in molecules, we designed a high-frequency information capture layer to enrich the features of the nodes. The graph-level pooling part can be a simple permutation invariant function such as Max and Mean, or a more sophisticated algorithm like GlobalAttention.

SMILES Encoder. Similar to YieldBert (Schwaller et al. 2020, 2021), the SMILES encoder is constructed by stacking multiple transformer encoders (Vaswani et al. 2017). It can capture long-range dependencies of elements in reactions and obtain the embedding vector of reaction SMILES sequence:

$$f^i_S = Enc(S^i) \tag{2}$$

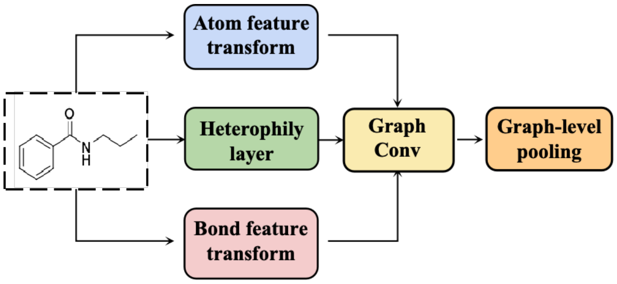

Figure 2: Graph Encoder, including atom and bond feature propagation, as well as graph-level pooling.

For a detailed introduction to the encoders, please refer to the implementation at https://github.com/jychen229/Multimodal-reaction-yield-prediction.

Multi-Modal Contrastive Learning To integrate the long-range dependencies identified in SMILES sequences with the spatial and structural information derived from molecular graphs, we employ a multi-modal contrastive learning strategy. Our approach is built based on the idea that the encoding vectors derived from SMILES sequences and those from molecular graphs should be similar if they correspond to the same reaction, and distinct if they refer to different reactions. Specially, we consider (f_S^j, f_G^j) as a positive pair, as they represent the same reaction R_i through both molecular graph and sequence modalities. Conversely, pairs such as (f_S^j, f_G^k) and (f_S^k, f_G^j), where $k \neq j$, are considered negative pairs, since these SMILES sequences and molecular graphs correspond to different reactions. To ensure that positive pairs have closely aligned encoding vectors and negative pairs have divergent ones, we minimize the following contrastive training loss, with learnable temperature $\tau \in \mathbb{R}^+$:

$$\mathcal{L}_c = -\frac{1}{2}\log\frac{e^{\langle f_G^j, f_S^j\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle f_G^j, f_S^k\rangle/\tau}} - \frac{1}{2}\log\frac{e^{\langle f_G^j, f_S^j\rangle/\tau}}{\sum_{k=1}^{N} e^{\langle f_G^k, f_S^j\rangle/\tau}},$$

where $e^{\langle\cdot,\cdot\rangle}$ ensures dimension flexibility by transforming the multi-modal encoded vectors through a nonlinear projection to fixed-dimensional vectors for contrastive learning (Zhang et al. 2022). In the pre-training stage, the SMILES encoder and graph encoder are trained using this contrastive learning loss on the input dataset. These pre-trained encoders will be fine-tuned later with other modules.

Mixture-of-Experts Feature Encoder The human-defined features include Morgan fingerprints, Mordred features, and QM descriptors (Liu, Moroz, and Isayev 2023). Due to the complexity of reactions, these features are often represented as high-dimensional sparse vectors. In order to extract and compress the most relevant information from these high-dimensional inputs, we employ a sparse MoE model, which is designed to uncover the shared subspaces common to subsets of reactions. Each expert can specialize in different aspects found within the high-dimensional data, and characterize the common features shared by specific subsets of reactions. The router automates expert assignment for each reaction's feature extraction. The nature that only a subset of experts is activated per input significantly reduces computational load.

Specifically, for the input features H, we first process them through a dense layer and then feed the obtained x_H into the MoE layers. The router, a gate function with trainable weights: $\mathcal{G}(x_H) = \mathrm{Softmax}\left(W_g \cdot x_H\right)$, assigns each input reaction to t out of k experts, $E = \{E_1, ..., E_k\}$. Each experts E_i is a feed-forward network (FFN). One MoE layer presents the output:

$$\mathrm{MoE}(x_H) = \sum_{i=1}^{t} \mathcal{G}(x_H)_i \cdot E_i(x_H) \tag{3}$$

which is a linear combination of the outputs from t FFNs. If required, $\mathrm{MoE}(x_H)$ can be passed through another MoE layer that possesses the same functional design. Following (Shazeer et al. 2017), we introduce an auxiliary loss $\mathcal{L}_a$ to encourage balanced routing to all experts. The output of MoE is transformed as f_H by another dense layers to get integration with f_G and f_S.

Late Fusion and Prediction The multi-modal reaction representation f_G, f_S and f_H can be incorporated with various strategies such as voting fusion, feature concatenation, or self-attention weighted fusion, all aimed at effectively predicting the corresponding yield. The final prediction is denoted as $\hat{y}$. We next introduce our prediction loss with uncertainty quantification.

Uncertainty Quantification Uncertainty is commonly categorized into aleatoric uncertainty and epistemic uncertainty. In reaction yield prediction, we further attribute uncertainty to *model uncertainty* and *data uncertainty*. Our model aims to minimize model uncertainty while employing the Bayesian learning framework (Kendall and Gal 2017) to model data uncertainty to enhance prediction performance and assist users in better evaluation reactions.

Molecules in chemical reactions often contain conformers of differing energy levels, which could result in different yields being reported for the same reaction. Therefore, we consider the reaction yield $\hat{y}$ as a random variable to account for the *data uncertainty*. By learning a probability distribution with the features $\boldsymbol{x} = \{f_G, f_S, f_H\}$, we sample from the distribution to obtain the final yield prediction. Taking the normal distribution as an example, we learn the mean $\mu(\boldsymbol{x})$ and variance $\sigma(\boldsymbol{x})$ of the distribution, and obtain the final prediction through the reparameterization trick (Kingma and Welling 2013):

$$\hat{y} = \mu(\boldsymbol{x}) + \epsilon * \sigma(\boldsymbol{x}) \tag{4}$$

where ϵ is an input independent variable, and $p(\epsilon) \sim \mathcal{N}(0, 1)$. The introduction of reparameterization enables models to consider uncertainty while maintaining differentiability, ensuring end-to-end training.

Based on the above uncertainty quantification, the prediction loss function is defined as follows:

$$\mathcal{L}_u = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\sigma\left(\boldsymbol{x}_i\right)^2}\left\|y_i - \mu\left(\boldsymbol{x}_i\right)\right\|^2 + \log\sigma\left(\boldsymbol{x}_i\right)^2\right]. \tag{5}$$

To reduce the *model uncertainty*, we employ the regularization method proposed in (Wu et al. 2021), where an additional KL-divergence loss $\mathcal{L}_r$ is introduced.

During the end-to-end training process, the overall loss function $\mathcal{L}$ is defined by combining the prediction loss with uncertainty quantification $\mathcal{L}_u$, the aforementioned auxiliary loss $\mathcal{L}_a$ for MoE, and the regularized dropout loss $\mathcal{L}_r$:

$$\mathcal{L} = \alpha\mathcal{L}_u + \beta\mathcal{L}_a + \gamma\mathcal{L}_r \tag{6}$$

where α, β and γ are hyper-parameters. More details of the loss functions and the implementation code are available at https://github.com/jychen229/Multimodal-reaction-yield-prediction.

Experiment

Experimental Setup

Datasets We use three evaluation datasets (see Table 1), two of which are popularly employed High-throughput experiment (HTE) datasets and the third one is constructed from patent literature by expert chemists.

- **High-throughput (HTE) datasets**. We used Buchwald-Hartwig dataset (Ahneman et al. 2018) and Suzuki-Miyaura dataset (Perera et al. 2018), which respectively involve high-throughput experiments on the class of Pd-catalyzed Buchwald-Hartwig C-N cross-coupling reactions and Suzuki-Miyaura cross-coupling reactions.
- **Amide coupling reaction (ACR) dataset**[1]. This is a recently launched large literature dataset, containing 41,239 amide coupling reactions extracted from Reaxys (Reaxys 2020). It is considerably more complex than the two HTE datasets. In addition to the SMILES representations of reactants and products, it furnishes contextual information about the reactions, including time, temperature, reagents, conditions, and solvent, which are important for yield prediction.

Baselines We evaluated the proposed method against three types of baselines: sequence models, graph-based models, and multi-modal models:

- **One-hot** (Chuang and Keiser 2018) represents the chemical reaction as one-hot vectors of reactants and products, indicating the presence or absence of each component.
- **YieldBert** (Schwaller et al. 2020, 2021) takes reaction SMILES as input and applies the large-scale sequence model BERT for yield prediction and is fine-tuned on the dataset based on the rxnfp pre-trained model.
- **MPNN** (Kwon et al. 2022), a graph-based model, represents reaction as a set of molecular graphs and utilizes graph neural networks for prediction.
- **YieldGNN** (Saebi et al. 2023) conducts prediction by combining molecular graphs and chemical features such as Morgan substructure fingerprints calculated by Rdkit (Landrum et al. 2019) and canonical MDS using Tanimoto similarity metric.

Dataset	No. reactions
Buchwald-Hartwig reaction	3,955
Suzuki-Miyaura reaction	5,760
Amide coupling reaction	41,239

Table 1: The statistics of experimental datasets.

Model	MAE ↓	RMSE ↓	R^2 ↑
Mordred	15.99 ± 0.14	21.08 ± 0.16	0.168 ± 0.010
YieldBert	16.52 ± 0.20	21.12 ± 0.13	0.172 ± 0.016
YieldGNN	15.27 ± 0.18	19.82 ± 0.08	0.216 ± 0.013
MPNN	16.31 ± 0.22	20.86 ± 0.27	0.188 ± 0.021
Ours	**14.76 ± 0.15**	**19.33 ± 0.10**	**0.262 ± 0.009**

Table 2: Results on the Amide coupling reaction dataset.

Implementation Details Our model is implemented by Pytorch and optimized with Adam optimizer and cosine learning rate scheduler with warming up. For the graph-level pooling module, the model utilizes a transformer decoder. The expert assignment in MoE is configured with t=1 and k=6. For the HTE datasets, we adopted the experimental settings from the (Kwon et al. 2022) to ensure a fair comparison. In the experiments on the ACR dataset, the late fusion module is designed with feature concatenation, and the MoE is structured with two stacked layers. We adopted a train/-valid/test split of $6/2/2$ and employed early-stopping for avoid overfitting. Regarding the baseline models, for YieldBert, we utilized the model with augmented data. As for YieldGNN, the human-defined features utilized as inputs are identical to those employed in our model. To ensure the robustness of evaluation, we perform 10 random shuffles of each dataset, and we subsequently report both the mean and the standard deviation of these results. All experiments are executed on a single NVIDIA RTX3090 GPU. Additional details of the model architecture and specific experimental settings can be found at the shared GitHub link.

Results on the ACR Dataset

The performance of UAM and baselines on the ACR dataset are reported in Table 2, where the best results are highlighted in **bold** and the second best baseline scores are underlined. It is observed that UAM achieved the best performance compared to all baselines. Other observations are as follows: Notably, we observe that all models exhibit suboptimal predictive performance on this dataset, with R^2 consistently below 0.5. This phenomenon stems from the inherent complexity of the ACR dataset and the presence of numerous incongruous reaction yields. On the contrary, our UAM results significantly surpass those of the baseline models in terms of three key metrics: R^2, mean absolute error (MAE), and root mean squared error (RMSE). In comparison to the baseline model, our approach has achieved an improvement of nearly 25% in terms of R^2 performance. This underscores the substantial efficacy of our model's enhancements in addressing uncertainty in real-world datasets. It is indeed the uncer-

[1] Available at https://github.com/isayevlab/amide_reaction_data

Model	MAE ↓	RMSE ↓	R^2 ↑
One-hot	6.08 ± 0.08	9.02 ± 0.16	0.890 ± 0.005
YieldBert	3.09 ± 0.12	4.80 ± 0.26	0.969 ± 0.004
YiledGNN	3.89 ± 0.14	6.01 ± 0.21	0.953 ± 0.003
MPNN	2.92 ± 0.06	4.43 ± 0.09	0.974 ± 0.001
Ours	**2.89** ± **0.06**	**4.36** ± **0.10**	**0.976** ± **0.001**

Table 3: Results on the Buchwald–Hartwig reactions dataset.

Model	MAE ↓	RMSE ↓	R^2 ↑
One-hot	8.55 ± 0.08	12.27 ± 0.15	0.809 ± 0.023
YieldBert	6.60 ± 0.27	10.52 ± 0.48	0.859 ± 0.012
YiledGNN	6.96 ± 0.25	11.00 ± 0.37	0.845 ± 0.011
MPNN	6.12 ± 0.22	9.47 ± 0.46	0.886 ± 0.010
Ours	**6.04** ± **0.18**	**9.23** ± **0.40**	**0.888** ± **0.009**

Table 4: Results on the Suzuki–Miyaura reactions dataset.

tainty within the dataset that hinders the accurate predictions of baselines. Furthermore, UAM not only demonstrates the highest predictive accuracy but also exhibits smaller standard deviations, showcasing the model's stability. We can also find that YieldGNN outperforms MPNN on the ACR dataset. This can be attributed to YieldGNN's incorporation of human-defined features, enabling more accurate predictions than MPNN. However, YieldBert and MPNN, which solely utilize sequence or graph structural information, yield less favorable results. And our model not only leverages information from three modalities but also employs enhanced feature extractors, resulting in superior performance on the large-scale real-world dataset.

Results on Two HTE Datasets

The performance of UAM and baseline models on the two HTE datasets are reported in Table 3 and 4. The results of the baseline models are reported from (Kwon et al. 2022). One can observe that most of the models have achieved R^2 values exceeding 0.95 or 0.85 on these two datasets. This can be attributed to the relatively homogeneous reaction types within the HTE datasets, rendering the intrinsic features of reactions easier to extract. Building upon this foundation, our model has achieved noticeable enhancements, affirming the superiority of our model's encoders. Furthermore, while YieldGNN, MPNN, and our model all incorporate GNN modules, YieldGNN's performance lags slightly behind. This discrepancy arises due to the adoption of the encoder-decoder pooling architecture in both our model and MPNN, which inherently outperforms the graph convolution utilized in YieldGNN.

Notably, one can observe that our model's performance improvement on the ACR dataset surpasses that on the HTE dataset by a significant margin. This phenomenon can be attributed to the characteristic of the HTE dataset, which consists of reactions carefully curated by chemists, resulting in a relatively straightforward linkage between yields and reactions. Consequently, nearly all baseline models achieve R^2 values above 0.95 or 0.85. In contrast, the ACR dataset represents a large-scale real-world dataset, as we mentioned earlier, and the inherent uncertainty within the dataset poses challenges for baseline models to make accurate predictions. The model design of the UAM effectively addresses these challenges, leading to substantial performance enhancements.

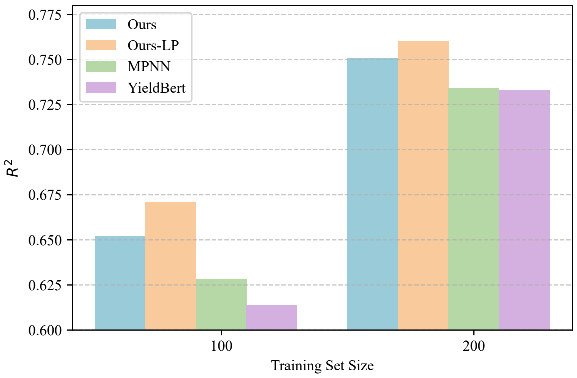

Figure 3: Label efficient learning performance on the Buchwald–Hartwig reactions dataset.

Performance of Label Efficient Learning

We conducted further analysis of the model's performance within the context of Label Efficient Learning. Here, we additionally implemented a variant of our model with Linear Probe (*Ours-LP*). In this setting, the parameters of both the graph encoder and the SMILES encoder are held constant, while the human-defined feature encoder is omitted from the configuration. Training is exclusively conducted for the regressor component of the model. The results in Figure 3 show that our models demonstrate superior performance compared to the baseline models when trained on a limited number of samples (2.5% and 5% of the original training set). Particularly, *Ours-LP* attains optimal performance. This achievement can be attributed to the benefits of contrastive learning pretraining, which effectively captures the shared and complementary information among different modalities. This underscores the substantial potential of our model in scenarios where limited literature-recorded data are available for specific reaction categories.

Ablation Studies

In this section, we study the influence of different components in our model, including the uncertainty quantification loss function $\mathcal{L}_u$, the regularized dropout loss $\mathcal{L}_r$, features from the three modalities, and the MoE module. We report the main results in Table 5.

Impact of the Uncertainty Quantification Loss $\mathcal{L}_u$ To study the impact of the uncertainty quantification loss $\mathcal{L}_u$, we switched the loss function back to the normal L2 loss. The experimental results demonstrated a noticeable decrease in accuracy. This highlights the crucial role of uncertainty

Model	MAE ↓	RMSE ↓	R^2 ↑
Ours	14.76 ± 0.15	19.33 ± 0.10	0.262 ± 0.009
w/o UQ	15.08 ± 0.13	19.63 ± 0.09	0.249 ± 0.009
w/o $\mathcal{L}_r$	14.80 ± 0.16	19.51 ± 0.10	0.261 ± 0.010
w/o MoE	15.12 ± 0.18	20.03 ± 0.13	0.230 ± 0.012
w/o Seq.	14.97 ± 0.16	19.55 ± 0.11	0.261 ± 0.010
w/o Graph	15.06 ± 0.15	19.59 ± 0.10	0.260 ± 0.009
w/o H.	15.83 ± 0.20	20.46 ± 0.18	0.212 ± 0.016

Table 5: Results of ablation study on the ACR dataset. UQ represents the Uncertainty Quantification, $\mathcal{L}_r$ is the regularized dropout loss, Seq represents the SMILES sequence, H. denotes the human-defined features, and w/o stands for the ablated model variant *without* a specific design element.

assessment in real-world datasets. Meanwhile, there was no significant difference in the standard deviations of the results when changing the loss functions. This suggests that the uncertainty quantification does not adversely affect the robustness of the model.

Impact of the Regularized Dropout Loss $\mathcal{L}_r$ We conducted ablation experiments regarding the regularized dropout loss $\mathcal{L}_r$, for evaluating its effectiveness on mitigating the model's intrinsic uncertainty. The results without $\mathcal{L}_r$ indicate that the model's training-time uncertainty does indeed impact its performance to a certain extent.

Impact of Mixture-of-Experts Another key design of UAM is to introduce Mixture-of-Experts layers. The MoE module allocates reactions to specific experts, enabling each FFN to handle particular reaction types. In the ablation study, we substituted the MoE module with an equally layered FFN. From Table 5, we observe that the model without MoE exhibited a performance decrease of approximately 10%. This highlights the effectiveness of MoE on extracting and compressing human-defined features compared to FFN.

To gain a deeper insight into the expert selection process, we have visualized the distribution of expert selections in both the first and second MoE layers during the testing phase of experiments on the ACR dataset, as shown in Fig. 4. On the left side of the figure, it is evident that in the first layer, each expert is assigned a varying number of reactions. In contrast, the distribution of expert selections in the second layer is considerably more balanced compared to the first. This allocation in the MoE layers significantly boosts the model's ability to expressively handle high-dimensional yet low-rank molecular descriptors and reaction condition information for predictive analysis. Moreover, this data allocation partitions the overall dataset uncertainty into submodules, leading to heightened prediction stability.

Impact of Multi-Modal Features We also investigated the importance of multi-modal features for prediction. From the results in Table 5, it can be observed that both sequence and graph representations have an impact on yield prediction but are not significant. In comparison, human-defined features play a vital role in the prediction outcome. This phenomenon can be attributed to two reasons: firstly, the human-defined features include molecular descriptors like fingerprints, which cover partial sequence and graph structural information. Secondly, by incorporating the rich reaction context such as temperature, time, reagents, and conditions, these features provide a crucial supplement for yield prediction. Additionally, removing sequence and graph data has a limited impact on model performance, validating the partial redundancy in the information contained within SMILES and graph representations. It is worth mentioning that while the contribution of each modality varies with specific datasets, it is evident that the integration of multi-modal features positively enhances prediction performance.

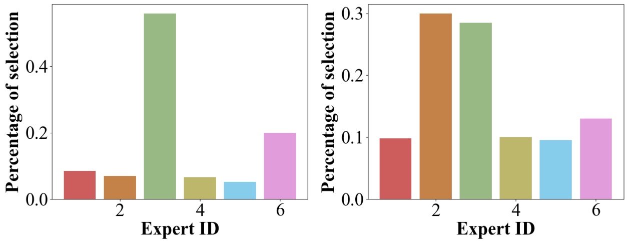

Figure 4: The distribution of expert selection in the first (left) and second (right) MoE layer.

Conclusion and Broader Impact

In this paper, we address the uncertainty inherent in predicting yields within real-world chemical reaction datasets. We introduce an uncertainty-aware multi-modal yield prediction model that synthesizes multi-modal molecular representation and incorporates a dedicated uncertainty quantification loss, thereby elevating predictive accuracy. Our experimental results reveal notable performance enhancements relative to existing yield prediction models. While our model has achieved significant improvement over baselines on the ACR dataset, there is still room for further enhancement. A promising direction could be the incorporation of additional modality, particularly those designed to handle 3D graph data (Schütt et al. 2017; Liu et al. 2021, 2022). This integration could potentially increase the model's performance by providing a more comprehensive understanding of molecular structures. As our model consists of multiple integrated modules, another future work will delve into the relationships between these components with the aim of refining model interpretability.

Acknowledgments

This work was supported by the National Science Foundation (CHE–2202693) through the NSF Center for Computer Assisted Synthesis (C-CAS, https://ccas.nd.edu/). O.I acknowledges CHE200122 allocation award from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; and Doyle, A. G. 2018. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385): 186–190.

Chuang, K. V.; and Keiser, M. J. 2018. Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science*, 362.

Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; and Jensen, K. F. 2017. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5): 434–443.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *ICML*, 1263–1272.

Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2023a. What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. In *NeurIPS*.

Guo, T.; Ma, C.; Chen, X.; Nan, B.; Guo, K.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2023b. Modeling non-uniform uncertainty in Reaction Prediction via Boosting and Dropout. *arXiv preprint arXiv:2310.04674*.

Guo, Z.; Guo, K.; Nan, B.; Tian, Y.; Iyer, R. G.; Ma, Y.; Wiest, O.; Zhang, X.; Wang, W.; Zhang, C.; and Chawla, N. V. 2023c. Graph-based Molecular Representation Learning. In *IJCAI*, 6638–6646.

Guo, Z.; Zhang, C.; Yu, W.; Herr, J.; Wiest, O.; Jiang, M.; and Chawla, N. V. 2021. Few-shot graph learning for molecular property prediction. In *Proceedings of the Web Conference*, 2559–2567.

Hu, L.; Wang, X.; Wong, L.; and Chen, G. 2003. Combined first-principles calculation and neural-network correction approach for heat of formation. *The Journal of Chemical Physics*, 119(22): 11501–11507.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; and Okuno, Y. 2019. Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks. *Journal of chemical information and modeling*, 59(12): 5026–5033.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? *NeurIPS*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kwon, Y.; Lee, D.; Choi, Y.-S.; and Kang, S. 2022. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *Journal of Cheminformatics*, 14: 2.

Landrum, G.; Tosco, P.; Kelley, B.; sriniker; gedeck; NadineSchneider; Vianello, R.; Dalke, A.; Ric; Cole, B.; AlexanderSavelyev; Turk, S.; Swain, M.; Vaucher, A.; N, D.; Wójcikowski, M.; Pahl, A.; JP; Berenger, F.; strets123; JLVarjo; O’Boyle, N.; Cosgrove, D.; Fuller, P.; Jensen, J. H.; Sforna, G.; DoliathGavid; Leswing, K.; Leung, S.; and van Santen, J. 2019. rdkit/rdkit: 2019_03_4 (Q1 2019) Release.

Li, H.; Zhao, D.; and Zeng, J. 2022. KPGT: knowledge-guided pre-training of graph transformer for molecular property prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 857–867.

Liu, Y.; Wang, L.; Liu, M.; Zhang, X.; Oztekin, B.; and Ji, S. 2021. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*.

Liu, Z.; Moroz, Y. S.; and Isayev, O. 2023. The challenge of balancing model sensitivity and robustness in predicting yields: a benchmarking study of amide coupling reactions. *Chemical Science*, 14(39): 10835–10846.

Liu, Z.; Zubatiuk, T.; Roitberg, A.; and Isayev, O. 2022. Auto3d: Automatic generation of the low-energy 3d structures with ANI neural network potentials. *Journal of Chemical Information and Modeling*, 62(22): 5373–5382.

Pattanaik, L.; and Coley, C. W. 2020. Molecular Representation: Going Long on Fingerprints. *Chem*, 6(6): 1204–1207.

Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; and Sach, N. W. 2018. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374): 429–434.

Reaxys. 2020. Reaxys Database. Accessed: Feb 10, 2020.

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.; et al. 2023. On the use of real-world datasets for reaction yield prediction. *Chemical Science*, 14(19): 4997–5005.

Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; and Glorius, F. 2020. A structure-based platform for predicting chemical reactivity. *Chem*, 6(6): 1379–1390.

Schierle, S.; Helmstädter, M.; Schmidt, J.; Hartmann, M.; Horz, M.; Kaiser, A.; Weizel, L.; Heitel, P.; Proschak, A.; Hernandez-Olmos, V.; et al. 2020. Dual farnesoid X receptor/soluble epoxide hydrolase modulators derived from Zafirlukast. *ChemMedChem*, 15(1): 50–67.

Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. In *NeurIPS*, 992–1002.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9): 1572–1583.

Schwaller, P.; Vaucher, A. C.; Laino, T.; and Reymond, J.-L. 2020. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *Proceedings of NeurIPS 2020 Machine Learning for Molecules Workshop*.

Schwaller, P.; Vaucher, A. C.; Laino, T.; and Reymond, J.-L. 2021. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1): 015016.

Segler, M. H.; Kogej, T.; Tyrchan, C.; and Waller, M. P. 2018. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1): 120–131.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. *NeurIPS*.

Wang, H.; Li, W.; Jin, X.; Cho, K.; Ji, H.; Han, J.; and Burke, M. D. 2021. Chemical-reaction-aware molecule representation learning. *arXiv preprint arXiv:2109.09888*.

Weininger, D.; Weininger, A.; and Weininger, J. L. 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences*, 29(2): 97–101.

Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.-Y.; et al. 2021. R-drop: Regularized dropout for neural networks. *NeurIPS*, 10890–10905.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25.