Graph Neural Networks for Predicting Chemical Reaction Performance

MANDANA SAEBI, BOZHAO NAN, JOHN HERR, JESSICA WAHLERS, OLAF WIEST, and NITESH V. CHAWLA, University of Notre Dame

Chemical reactions are a complex process, as they involve interaction between several molecular compounds. As a result, predicting the success of a reaction is a non-trivial task, which often requires running several experiments in the lab. This process is is expensive, time consuming, and inefficient. As a result, in recent years, researchers have explored the use of machine learning algorithms to predict reaction success. These methods mainly rely on chemical properties of the molecules involved in the reactions. Despite their promising success, none of existing methods explored the use of structural properties of molecules in predicting reaction success. In this work, we develop an Attributed Graph Neural Network model that integrates both structural properties as well as chemical properties of molecules for predicting reaction success. Our model shows remarkable performance on two hand-crafted datasets obtained from high-throughput experiments, as well as one real-world dataset.

1 INTRODUCTION

Predicting the Performance of chemical reactions is a fundamental problem in organic chemistry. The ability to predict whether a reaction will be successful or not can save significant time and effort of organic chemists and expedite the process of generating chemical compounds. Existing methods still depend on handcrafted reaction rules [2, 6] or heuristically extracted reaction templates [3, 10], and therefore, are not well generalizable to unseen reactions. Another major challenge is the availability of data on both failed and successful experiments. Existing research in the literature has mainly focused on successful experiments (i.e., reactions with a high yield), thus making it hard for a machine learning model to infer what makes a reaction successful. Recently, machine learning models have been proposed to predict the reaction performance based on molecular features [8, 11]. Ahneman et al. [1] advanced the field by proposing a regression model based on the synthetic reaction data obtained from high-throughput experiments [1, 7**?**]. However, these methods only consider features such as molecular, atomic, and vibrational properties and do not use any information about the complex structure of molecular graphs. We argue that in order to solve this problem effectively, an intelligent AI system should have two key capabilities: (1) Understanding the molecular graph structure of the input reactants to identify complex interactions between reaction components, and (2) Incorporating domain knowledge of organic chemists in the form of molecular, atomic, and vibrations characteristics of reactants to learn the rules that organic chemists use for predicting reaction success. To this end, we propose an framework which combines Attributed Graph Neural Network (AGNN) and chemical properties about reaction compounds to predict reaction performance. The chemical properties are incorporated into the model both directly (via the domain module) and indirectly (via attributed graphs of molecules).

2 MODEL DESIGN

2.1 Problem formulation

We define a chemical reaction as a combination of molecular graphs $(G_r, G_l, G_s, G_b, G_p)$, where G_r, G_l, G_s, G_b, and G_p represent the reactant, ligand, solvent, base, and the products [5] graph,

Authors' address: Mandana Saebi, msaebi@nd.edu; Bozhao Nan, bnan@nd.edu; John Herr, ; Jessica Wahlers, jwahlers@nd.edu; Olaf Wiest, owiest@nd.edu; Nitesh V. Chawla, nchawla@nd.edu, University of Notre Dame, Notre Dame, Notre Dame, IN, 46556.

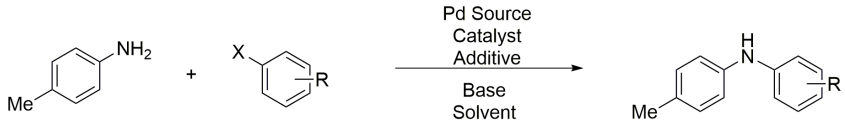

Fig. 1. Pd-catalysed Buchwald–Hartwig C-N cross coupling reaction

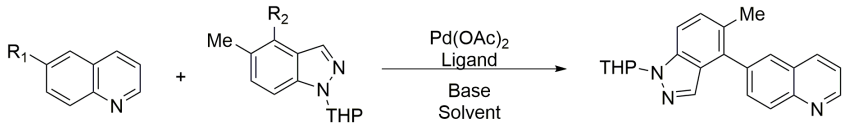

Fig. 2. Suzuki–Miyaura cross-coupling reaction

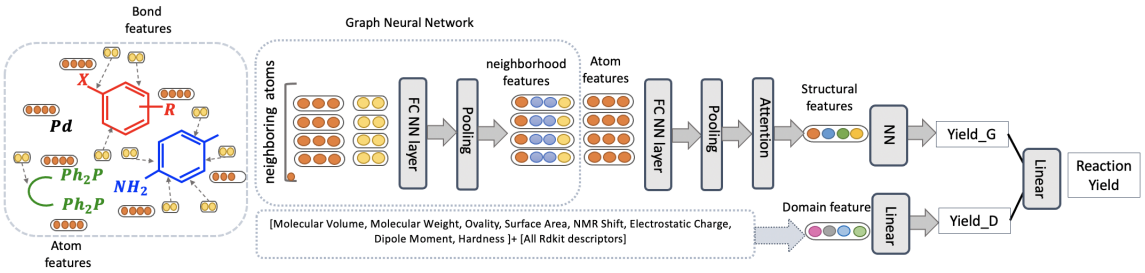

Fig. 3. **Model Overview. (A) The Pd-catalyzed Buchwald-Hartwig reaction which is used as a model reaction for the yield prediction task. The bond changes in the highlighted components under the Pd catalyst results in the generation of the final product. (B) For each reaction, we extract structural features by aggregating atom and bond features over the neighborhoods. We concatenate structural features to the domain-features to obtain the final reaction features. We obtain two yield scores by feeding structural features and domain-based features. We average these two scores to generate the reaction yield predictions.**

respectively. A molecular graph is described as $G = (V, E)$, where $V = a_1, a_2, \cdots, a_n$ is the set of atoms and $E = b_1, b_2, \cdots, b_m$ is the set of associated bonds of varying types (single, double, aromatic, etc.). Note that G_r normally consists of two or more connected components since reactants contain multiple molecules. Given a reaction $(G_r, G_l, G_s, G_b, G_p)$, our objective is to predict the reaction yield (i.e., reaction performance) based on molecular properties and interactions between the molecular graphs of reactants. We treat this problem as a regression task. For this work, we focus on Pd-catalyzed Buchwald-Hartwig reaction [1] (as shown in Figure 1 and Suzuki–Miyaura reactions (as shown in Figure 2) because of their broad value in pharmaceutical synthesis. Below we detail our model design.

2.2 Model Architecture

Our model incorporates the domain knowledge about molecular properties and the complex interaction between molecular graphs using an attributed graph neural network (AGNN). An overview of the model is shown in Figure 3. The top module represents the AGNN which learns the structural features and the bottom module captures the chemical properties. We detail the process of feature selection in subsection 4.1.

Molecular Graph Neural Networks. We propose AGNN for capturing the complex interactions between molecular components. GNNs have been shown to be very successful in capturing the higher-order interactions between neighboring components of a graph [12]. During a reaction, particular components of two molecules interact with each other, resulting in changes in atoms and bonds. Under favorable experimental conditions (temperature, pressure) these interactions transfer the reactants into the final product. Predicting the efficiency of these complex interactions relies on understating the substructures that are likely to interact.

We use Weisfeiler-Lehman Network (WLN) **??** for the GNN component to capture these substructural features. We then apply attention mechanism to capture the global context and improve model performance.

For each reaction, we extract structural features by aggregating atom and bond features over the higher-order neighborhoods by passing the molecules through the GNN and attention layer. We concatenate structural features to the domain-features to obtain the final reaction features. We obtain two yield scores from the structural features and domain-based features, and then feed the two scores to a linear layer to generate the final reaction yield predictions. Our model is in part inspired by [4]. However, in this work we focus on prediction reaction yield by combining both structural graph-based features as well as chemical properties.

3 DATA

3.1 Buchwald–Hartwig reactions

We use two datasets that contain a collection of Buchwald–Hartwig reactions:

3.1.1 B-H. reactions from Ahneman et al.[1]. Obtained from high-throughput experiments (HTE) on Pd-catalysed Buchwald– Hartwig C-N cross coupling reactions. This dataset contains 4140 reactions covering 15 aryl and heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives.

3.1.2 B-H. reactions from AstraZeneca. The AstraZeneca dataset used in this work is an unpublished dataset containing 1000 Buchwald-Hartwig reactions generated by AstraZeneca from the Electronic Laboratory Notebooks (ELN). Compared with the well-designed HTE datasets, the AstraZeneca dataset from industrial ELN can be regarded as a real-world dataset that requires extensive curation. The reaction components has been extracted from the original data table and then classified into categories such as aryl halides, amines, ligands, and metals. After the data curation, the dataset consists of 757 reactions with 535 amines, aryl and heteroaryl halides, 24 ligands, 15 bases, and 15 solvent system.

3.2 Suzuki–Miyaura reactions

The Suzuki-Miyaura reaction dataset [7] include 5760 Suzuki-Miyaura reactions which was generated using both nanomole-scale reaction screening and micromole-scale synthesis in a highly automated flow system [7]. In this dataset a wide range of reaction variables in a Suzuki-Miyaura reaction was tested, including 12 ligands, 11 substrates, 8 base system and 4 solvent systems.

4 EXPERIMENTS

4.1 Feature Selection

For the chemical properties, we collect features from two main sources. The first source are the set of descriptors available in the RDKit library. The second source are the features from DFT calculation using Gaussian 16 with B3LYP(6-31G*,6-311G*) basis set. The rest of features include the

surface area generate from pymol, pKaH of the base, solvent dielectric component from compound database. The following set shows the chemical features used for model training:

Molecular features: molecular volume, surface area, ovality, molecular weight, HOMO/LUMO Energy, electronegativity, hardness, and dipole moment.
Atomic features: Electrostatic charge and NMR shift
Reaction features: Temperature, Reaction scale and volume for some of reactions

We combine the features from both sources and train the random forest model to select the features that contribute to model performance. This model serves as a baseline and also helps us reduce the number of the parameters used in our deep learning model. Note that, we do not perform feature engineering on the structural features and they are automatically generated by the GNN model.

4.2 Baselines

We compare our model with several state-of-the-art machine learning models. Below we provide a brief description of each baseline:

(1) **Ahneman et al.[1]** Trains several conventional machine learning models on a selected set of features that are only based on chemical properties of the reaction compounds. Since they found random forests to perform the best among several other models (such as SVM, neural networks, linear regression), we use this model as our first baseline.
(2) **Schwaller et al. [9]** This work treats reaction smiles as text and fine-tunes a pre-trained language model to predict the reaction yield. For our experimental results, we directly quote the performance of this model from [9].
(3) **Random forest model**. We also use a random forest model with all available chemical features for feature selection. This model is different from [1] as it includes a much larger subset of chemical properties.

We also isolate the GNN module in our model to measure the power of structural features for predicting reaction yield.

For all the above models we report R^2 as a measure of the regression performance.

4.3 Parameter selection

We performed a grid-search for each hyperparamter and to tuned them for each dataset separately. For both datasets, batch size and initial learning rate are set to 40 and 0.01, respectively. Dropout is also set to 0.04 for both datasets. For [1] data, we decay the learning rate with a 0.9 ratio if the loss plateaus. We use a 1-hop neighborhood and set the size of all hidden layers to 100. For [7] data, we decay the learning rate with a 0.5 ratio upon loss plateau. We use a 2-hop neighborhood and set the size of hidden layers to 200. We also clip the gradient with a 0.8 ratio to avoid the exploding gradient problem.

4.4 Results

4.5 Yield prediction

Our experimental results are described inTable 1. Starting with the random forest models, we notice that Random forest-2 provides significantly better performance for Suzuki-Miyaura reactions while Random forest-1 shows slightly better performance for Buchwald-Hartwig [1] data, although it contains a much smaller feature set. This indicates that for Buchwald-Hartwig [1] data, feature selection using [1] method plays a more important role in achieving a good final performance using a random forest model.

Data Set	Suzuki-Miyaura [7]	Buchwald-Hartwig [1]	AstraZeneca
Random forest-1	0.797 ± 0.008	0.912 ± 0.008	0.273 ± 0.0372
Random forest-2	0.832 ± 0.004	0.892 ± 0.0098	0.295 ± 0.0371
Schwaller et al. [9]	0.81 ± 0.01	0.951 ± 0.005	—
YieldGNN - domain	0.899 ± 0.067	0.90 ± 0.0401	0.871 ± 0.0084
YieldGNN + domain	**0.962 ± 0.013**	**0.970** ± 0.021	**0.925 ± 0.0096**

Table 1. **Experimental results on the three reaction datasets.**

Interestingly, for Suzuki-Miyaura reactions, the Random forest-2 outperforms the Schwaller et al. [9] model, while for Buchwald-Hartwig [1] data Schwaller et al. [9] model provides better performance. For both Suzuki-Miyaura reactions and Buchwald-Hartwig data [1], YieldGNN + domain provides the best R^2. However, YieldGNN - domain still provides a relatively good performance compared to other baselines, despite having little knowledge about molecular properties. This indicates that the molecular structure provides important information for predicting reaction yield.

We should also note that we were able to make incremental improvements in the GNN models (with or without chemical properties). The first improvement happened as a result of adding the attention layer to the GNN. These results improved for both datasets after adding the chemical properties (GNN+domain). The final major improvement –especially with Suzuki-Miyaura reactions– happened after adding the solvent and base molecules to the pool of reaction molecules. Initially, we only included the molecules of reactants, and that resulted in an R^2 value very close to the current value in Table 1. However, our model was struggling with the Suzuki-Miyaura data. After adding solvent and base molecules, we were able to achieve an R^2 value of 0.962 ± 0.01.

4.5.1 Performance on AstraZeneca data. Although this dataset also contain Buchwald-Hartwig reactions, it is quite different from the other two dataset, as it is collected from ELNs. As a result, there is a much larger variation in the reaction space for this data. Furthermore, reactions in this dataset run at different volumes and temperatures, which adds several additional complexities to the yield prediction problem for this data. Given this large reaction space, we have a lot fewer training examples to learn how different factors are contributing to the reaction yield. Despite all these factors, GNN+domain provides a significantly better results compared to the random forest model.

REFERENCES

[1] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. 2018. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* 360, 6385 (2018), 186–190.

[2] Jonathan H Chen and Pierre Baldi. 2009. No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms. *Journal of chemical information and modeling* 49, 9 (2009), 2034–2043.

[3] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. 2017. Prediction of organic reaction outcomes using machine learning. *ACS central science* 3, 5 (2017), 434–443.

[4] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* 10, 2 (2019), 370–377.

[5] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. 2017. Predicting organic reaction outcomes with weisfeiler-lehman network. In *Advances in Neural Information Processing Systems*. 2607–2616.

[6] Matthew A Kayala and Pierre F Baldi. 2011. A machine learning approach to predict chemical reactions. In *Advances in Neural Information Processing Systems*. 747–755.

[7] Damith Perera, Joseph W Tucker, Shalini Brahmbhatt, Christopher J Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W Sach. 2018. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* 359, 6374 (2018), 429–434.

[8] Paul Raccuglia, Katherine C Elbert, Philip DF Adler, Casey Falk, Malia B Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A Friedler, Joshua Schrier, and Alexander J Norquist. 2016. Machine-learning-assisted materials discovery using failed experiments. *Nature* 533, 7601 (2016), 73–76.

[9] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. 2021. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* 2, 1 (2021), 015016.

[10] Marwin HS Segler and Mark P Waller. 2017. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal* 23, 25 (2017), 5966–5971.

[11] G Skoraczyński, P Dittwald, B Miasojedow, S Szymkuć, EP Gajewska, Bartosz A Grzybowski, and A Gambin. 2017. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Scientific reports* 7, 1 (2017), 1–9.

[12] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894* (2019).