Data augmentation strategies to improve reaction yield predictions and estimate uncertainty

Philippe Schwaller[1,2]
phs@zurich.ibm.com

Alain C. Vaucher[1]
ava@zurich.ibm.com

Teodoro Laino[1]
teo@zurich.ibm.com

Jean-Louis Reymond[2]
jean-louis.reymond@dcb.unibe.ch

[1]*IBM Research – Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*
[2]*Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

Abstract

Chemical reactions describe how precursor molecules react together and transform into products. The reaction yield describes the percentage of the precursors successfully transformed into products relative to the theoretical maximum. The prediction of reaction yields can help chemists navigate reaction space and accelerate the design of more effective routes. Here, we investigate the best-studied high-throughput experiment data set and show how data augmentation on chemical reactions can improve yield predictions' accuracy, even when only small data sets are available. Previous work used molecular fingerprints, physics-based or categorical descriptors of the precursors. In this manuscript, we fine-tune natural language processing-inspired reaction transformer models on different augmented data sets to predict yields solely using a text-based representation of chemical reactions. When the random training sets contain 2.5% or more of the data, our models outperform previous models, including those using physics-based descriptors as inputs. Moreover, we demonstrate the use of test-time augmentation to generate uncertainty estimates, which correlate with the prediction errors.

1 Introduction

The synthesis of new chemicals affects numerous aspects of our life, ranging from food and medicine to novel materials for technological applications. The current machine learning revolution in automated synthesis can significantly accelerate novel materials and molecules' development. In the last years, natural language processing methods emerged as robust and effective approaches in the field of organic chemistry, showing promising results in reaction prediction (1; 2; 3; 4), retrosynthesis planning (5; 6; 7; 8), data curation (9) and synthesis action generation (10; 11). In those studies the encoder-decoder transformer models introduced by Vaswani et al. (12) excel among all other neural network architectures. More recently, the use of encoder-only transformers such as BERT (13; 14) led to advances in reaction classification and fingerprints (15), as well as in unsupervised reaction atom-to-atom mapping (16) and reaction yield predictions (17).

Reaction yields describe the percentage of the reactant molecules converted into the desired product molecule during a chemical reaction. The prediction of reaction yields can guide chemists in selecting the next experiments to perform, and retrosynthetic planning tools in aiming for routes that maximize the overall yield, thus minimizing waste. Extensive chemical reaction yield data sets exist for high-throughput experiments (HTE). Examples are the Suzuki–Miyaura coupling reactions by Perera et

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

al. (18) and the palladium-catalyzed Buchwald–Hartwig reactions by Ahneman et al. (19), to date the best-studied HTE yield data set. In this work, we study reactions yield prediction using the latter data set (19), containing a total of 3955 Buchwald–Hartwig reactions with measured yields. Figure 1 a) provides an overview of the data set.

In a recent manuscript, Schwaller et al. (17) introduced a BERT (13) model with a regression head to predict reactions' yields given as input a reaction SMILES (20; 21), a text-based molecule and reaction representation. We show in Figure 1 a) and c) the task description, together with an example of a reaction SMILES. Here, we investigate how different data augmentation techniques (Figure 1 b), molecule permutations and SMILES randomizations (22; 23; 24; 8)) improve the performance of the yield prediction models. Moreover, we demonstrate the use of test-time augmentation (Figure 1 d)) to provide uncertainty estimates (25) on the reaction yields, that correlate with the predictions' errors.

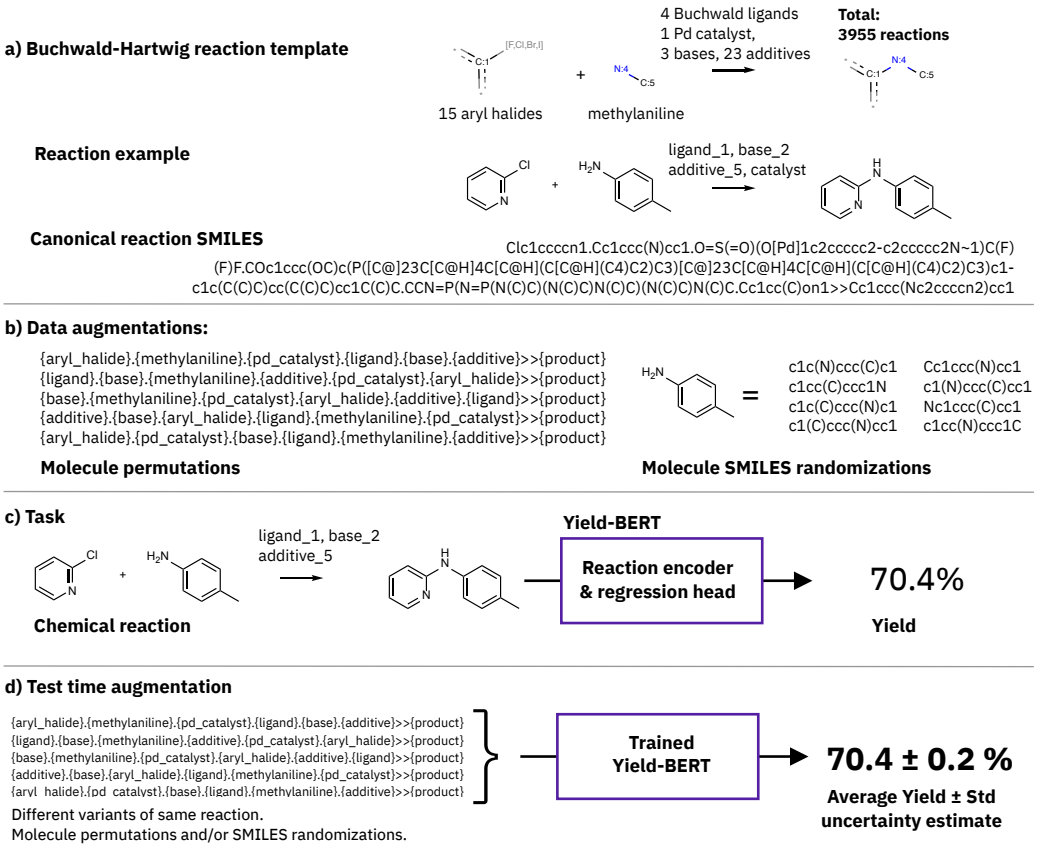

Figure 1: Training/evaluation pipeline and task description.

2 Results & Discussion

Our models were trained using Simpletransformers (26), huggingface transformers (27), PyTorch (28) and scripts adapted from the RXN yields github repository (17; 29). Canonicalizations and augmentations were done using RDKit (30). As described in the work of Schwaller et al. (17), fine-tuning a pretrained reaction BERT model (15) for a specific task provides the advantage of having most of the hyperparameters already optimized and fixed. Schwaller et al. (17) tuned only the dropout probability and the learning rate on the training data of the first random split, further split into a smaller training and validation set. Here, we initialized the dropout and learning rate using the values reported in (17) and we determined the optimal numbers of data augmentations using the same training/validation set. We investigated the two data augmentation techniques: molecule permutations, where we randomly shuffle the order of the precursors, SMILES randomizations, where

we generated multiple randomized SMILES for a given molecule (24), and the combination of the two. Examples of augmented reactions and molecules are shown in Figure 1 b).

2.1 Yield prediction

Most of the results in the literature were published on 70%/30% (training/testing) random splits. In Table 1, we compared the results of the canonical order, the permuted precursors, the randomized SMILES and the combination of both permutation plus randomization to previous studies (19; 31; 32; 17). While the use of the canonical order SMILES representation plus BERT with a regression head (17) already outperforms one-hot encodings (31), physics-based descriptors (19) and multi-fingerprint features (32) plus a random forest regressor, here we significantly improve the R^2 score using randomization. The same number of training augmentations, as stated in Table 1, was used throughout this work.

Table 1: Random splits 70/30, averaged over 10 splits

R^2	# samples/augmentations per rxn	mean	std
canonical	1	0.951	0.005
permuted	5	0.964	0.003
randomized	15	**0.970**	0.003
permuted, randomized	15	**0.970**	0.003
MFF + RF (32)		0.927	0.007
DFT + RF (19)		0.92	
one-hot + RF (31)		0.89	

Moreover, we investigated the prediction performance on reduced training sets (Table 2), an experiment also performed by Ahneman et al. (19). We observed that using SMILES randomization, we outperformed all other approaches, using only 2.5% (or 98 data points). Although deep learning models are typically criticized as being data-hungry, our combination of a pretrained base-encoder (15) and data augmentation leads to accurate predictions in the small data regime.

Table 2: Reduced training sets, averaged over 10 splits

R^2	2.5/97.5	5/95	10/90	20/80	30/70	50/50
can	0.45 ± 0.05	0.61 ± 0.04	0.79 ± 0.02	0.86 ± 0.01	0.88 ± 0.01	0.92 ± 0.01
permuted	0.47 ± 0.13	0.70 ± 0.06	0.81 ± 0.02	0.87 ± 0.02	0.90 ± 0.01	0.94 ± 0.01
randomized	$\mathbf{0.61 \pm 0.04}$	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.95 \pm 0.01}$
perm&rand	0.57 ± 0.08	0.71 ± 0.04	0.81 ± 0.02	0.89 ± 0.01	0.91 ± 0.01	0.95 ± 0.01
DFT+RF (19)	0.59	0.68	0.77	0.81	0.85	0.9

The data set of Ahneman et al. (19) also contains four out-of-sample splits, for which certain additives are only present in the test set. The results in Table 3 show that the models trained on canonical reaction SMILES without data augmentation perform best. For Test 4, the additives of the training set are the least representative of the ones in the test data. Therefore, the model trained on randomized SMILES, which better captures the patterns in the training data, unsurprisingly performs worse on that set.

2.2 Uncertainty estimation

We introduce test-time augmentation to provide an uncertainty estimation on our yield predictions. We input several data augmented versions of the same reaction and output the predicted yield as the average of the predicted yields using their standard deviation as the uncertainty estimate. Doing so does not significantly change the R^2 score. We measure the quality of the uncertainty estimates by computing the spearman's rank correlation coefficient (ρ) between absolute error and standard deviation of predicted yields, similar to the work by Hirschfeld et al. (33) on uncertainty quantification for molecular property predictions. The coefficient ranges between -1 and 1 and

Table 3: Out-of-sample test splits, averaged over 5 random seeds

R^2	Test 1	Test 2	Test 3	Test 4	Avg.
canonical	**0.84 ± 0.01**	0.84 ± 0.03	**0.75 ± 0.04**	0.49 ± 0.05	**0.73 ± 0.15**
permuted	0.82 ± 0.01	**0.90 ± 0.01**	0.63 ± 0.05	0.43 ± 0.07	0.69 ± 0.19
randomized	0.80 ± 0.01	0.88 ± 0.02	0.56 ± 0.08	0.07 ± 0.04	0.58 ± 0.33
perm&rand	0.79 ± 0.09	**0.90 ± 0.01**	0.55 ± 0.05	0.27 ± 0.14	0.63 ± 0.26
MFF + RF (32)	**0.85**	0.71	0.64	0.18	0.60
DFT + RF (19)	0.8	0.77	0.64	**0.54**	0.69
one-hot + RF (31)	0.69	0.67	0.49	0.49	0.59

measures the monotonic relation between errors and uncertainty estimates. Figure 2 a) shows that ρ increases for all augmentation methods with the number of test-time augmentations and converges to values above 0.4. For the example plots in Figure 2 b) and Figure 2 c), we used the models trained on randomized SMILES and applied 10 test-time augmentations. In Figure 2 b), we show how the predicted values get more certain and precise when increasing the data set from 2.5% to 70%. The out-of-sample test set plots in Figure 2 c) show that the uncertainty estimate correlates well with the error. Points with a larger error are generally more uncertain. Moreover, the models consistently predict a high yield for the reaction with the highest experimental yield independently of the split.

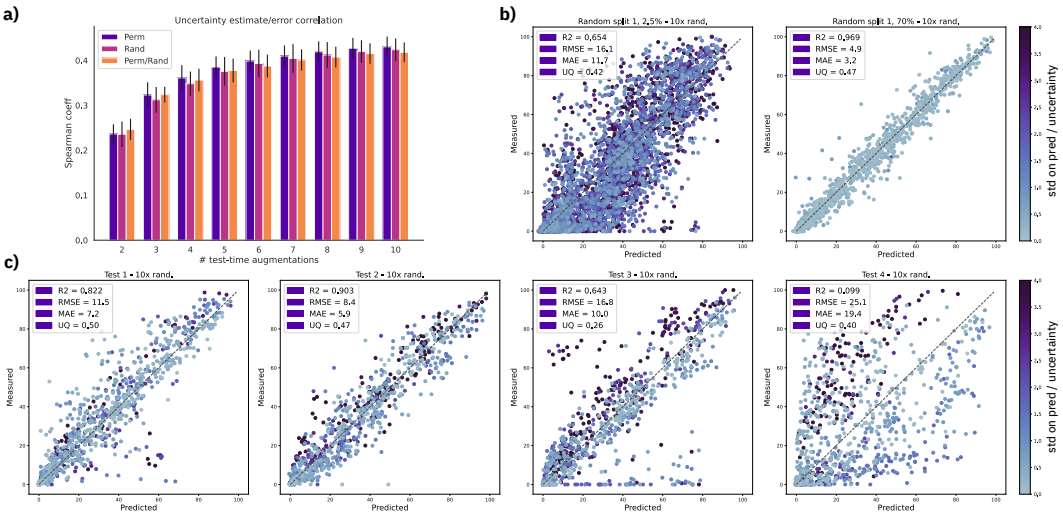

Figure 2: a) Spearman's rank correlation coefficient with increasing number of test time augmentations. b) Predictions and uncertainty on random split 01 with 2.5% and 70% training data using a fixed molecule order and 10 SMILES randomizations (randomized). c) Out-of-sample test set predictions using a fixed molecule order and 10 SMILES randomizations (randomized). Uncertainty scale was kept the same for all plots and capped at 4.0. MAE = mean average error, RMSE = root mean squared error, UQ = spearman's coefficient ρ.

3 Conclusion

In this manuscript, we presented augmentation strategies to increase reaction yield prediction using as input solely a text-based representation of chemical reactions. Even in a small data regime, a reaction BERT with regression head fine-tuned on randomized molecule representations was able to outperform physics-based descriptors plus random forest (19). Although data augmentations result in worse performance for strongly dissimilar out-of-sample test reactions, we show that test-time data augmentations can provide uncertainty estimates without the need of model retraining. The uncertainty estimates correlate with the error of the predictions and could be used to guide the chemical

space exploration (34; 35; 36; 37). The code and 400 trained models to produce the results described in this work are available for download (`https://github.com/rxn4chemistry/rxn_yields`).

References

[1] Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. *arXiv preprint arXiv:1612.09529* (2016).

[2] Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science* **9**, 6091–6098 (2018).

[3] Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).

[4] Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 1–8 (2020).

[5] Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).

[6] Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).

[7] Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chemical Science* **11**, 3355–3364 (2020).

[8] Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications* **11**, 1–11 (2020).

[9] Toniato, A., Schwaller, P., Cardinale, A., Geluykens, J. & Laino, T. Unassisted Noise-Reduction of Chemical Reactions Data Sets. *ChemRxiv preprint* doi:10.26434/chemrxiv.12395120.v1 (2020).

[10] Vaucher, A. C. *et al.* Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).

[11] Vaucher, A. C. *et al.* Inferring Experimental Procedures from Text-Based Representations of Chemical Reactions. *ChemRxiv preprint* doi:10.26434/chemrxiv.13118423.v1 (2020).

[12] Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).

[13] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Lan, Z. *et al.* Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations* (2020).

[15] Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *ChemRxiv preprint* doi:10.26434/chemrxiv.9897365 (2019).

[16] Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Unsupervised Attention-Guided Atom-Mapping. *ChemRxiv preprint* doi:10.26434/chemrxiv.12298559.v1 (2020).

[17] Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *ChemRxiv preprint* doi:10.26434/chemrxiv.12758474 (2020).

[18] Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018).

[19] 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* **360**, 186–190 (2018).

[20] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).

[21] Weininger, D., Weininger, A. & Weininger, J. L. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences* **29**, 97–101 (1989).

[22] Bjerrum, E. J. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076* (2017).

[23] Arús-Pous, J. *et al.* Randomized smiles strings improve the quality of molecular generative models. *J. Cheminf.* **11**, 1–13 (2019).

[24] Lambard, G. & Gracheva, E. Smiles-x: autonomous molecular compounds characterization for small datasets without descriptors. *Mach. Learn.: Sci. Technol.* **1**, 025004 (2020).

[25] Wang, G. *et al.* Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019).

[26] Simpletransformers (2020). URL `https://simpletransformers.ai`. (Accessed Oct 02, 2020).

[27] Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *ArXiv* **abs/1910.03771** (2019).

[28] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037 (2019).

[29] Rxn yields repo (2020). URL `https://rxn4chemistry.github.io/rxn_yields/`. (Accessed Oct 02, 2020).

[30] Landrum, G. *et al.* rdkit/rdkit: 2019_03_4 (q1 2019) release (2019). URL `https://doi.org/10.5281/zenodo.3366468`.

[31] Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **362** (2018).

[32] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* (2020).

[33] Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. Uncertainty quantification using neural networks for molecular property prediction. *arXiv preprint arXiv:2005.10036* (2020).

[34] Eyke, N. S., Green, W. H. & Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering* (2020).

[35] Thawani, A. R. *et al.* The photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry. *arXiv preprint arXiv:2008.03226* (2020).

[36] Felton, K., Rittig, J. & Lapkin, A. Summit: Benchmarking Machine Learning Methods for Reaction Optimisation. *ChemRxiv preprint* doi:10.26434/chemrxiv.12939806.v1 (2020).

[37] Häse, F. *et al.* Olympus: a benchmarking framework for noisy optimization and experiment planning. *arXiv preprint arXiv:2010.04153* (2020).