Prediction of Chemical Reaction Yields using Deep Learning

Philippe Schwaller

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

E-mail: phs@zurich.ibm.com

Alain C. Vaucher, Teodoro Laino

IBM Research – Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

Jean-Louis Reymond

Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland

Abstract. Artificial intelligence is driving one of the most important revolutions in organic chemistry. Multiple platforms, including tools for reaction prediction and synthesis planning based on machine learning, successfully became part of the organic chemists' daily laboratory, assisting in domain-specific synthetic problems. Unlike reaction prediction and retrosynthetic models, the prediction of reaction yields has received less attention in spite of the enormous potential of accurately predicting reaction conversion rates. Reaction yields models, describing the percentage of the reactants converted to the desired products, could guide chemists and help them select high-yielding reactions and score synthesis routes, reducing the number of attempts. So far, yield predictions have been predominantly performed for high-throughput experiments using a categorical (one-hot) encoding of reactants, concatenated molecular fingerprints, or computed chemical descriptors. Here, we extend the application of natural language processing architectures to predict reaction properties given a text-based representation of the reaction, using an encoder transformer model combined with a regression layer. We demonstrate outstanding prediction performance on two high-throughput experiment reactions sets. An analysis of the yields reported in the open-source USPTO data set shows that their distribution differs depending on the mass scale, limiting the dataset applicability in reaction yields predictions.

1. Introduction

Chemical reactions in organic chemistry are described by writing the structural formula of reactants and products separated by an arrow, representing the chemical transformation by specifying how the atoms rearrange between one or several reactant

molecules and one or several product molecules [1]. Economic, logistic, and energetic considerations drive chemists to prefer chemical transformations capable of converting all reactant molecules into products with the highest yield possible. However, side-reactions, degradation of reactants, reagents or products in the course of the reaction, equilibrium processes with incomplete conversion to a product, or simply by product isolation and purification undermine the quantitative conversion of reactants into products, rarely reaching optimal performance.

Reaction yields are usually reported as a percentage of the theoretical chemical conversion, i.e., the percentage of the reactant molecules successfully converted to the desired product compared to the theoretical value. It is not uncommon for chemists to synthesise a molecule in a dozen or more reaction steps. Hence, low-yield reactions may have a disastrous effect on the overall route yield because of the individual steps' multiplicative effect. Therefore, it is not surprising that designing new reactions with yields higher than existing ones attracts much effort in organic chemistry research.

In practice, specific chemical reaction classes are characterised by lower or higher yields, with the actual value depending on the reaction conditions (temperature, concentrations, etc.) and on the specific substrates.

Estimating the reaction yield can be a game-changing asset for synthesis planning. It provides chemists with the ability to evaluate the overall yield of complex reaction paths, addressing possible shortcomings well ahead of investing hours and materials in wet-lab experiments. Computational models predicting reaction yields could support synthetic chemists in choosing an appropriate synthesis route among many predicted by data-driven algorithms. Moreover, reaction yields prediction models could also be employed as scoring functions in computer-assisted retrosynthesis route planning tools [2, 3, 4, 5], to complement forward prediction models [6, 4] and in-scope filters [2].

Most of the existing efforts in constructing models for the prediction of reactivity or of reaction yields focused on a particular reaction class: oxidative dehydrogenations of ethylbenzene with tin oxide catalysts [7], reactions of vanadium selenites [8], Buchwald–Hartwig aminations [9, 10, 11], and Suzuki–Miyaura cross-coupling reactions [12, 13, 14]. To the best of our knowledge, there was only one attempt to design a general-purpose prediction model for reactivity and yields, without applicability constraints to a specific reaction class [15]. In this work, the authors design a model predicting whether the reaction yield is above or below a threshold value and conclude that the models and descriptors they consider cannot deliver satisfactory results.

Here, we build on our legacy of treating organic chemistry as a language to introduce a new model that predicts reaction yields starting from reaction SMILES [16]. More specifically, we fine-tune the rxnfp models by Schwaller et al. [17] based on a BERT-encoder [18] by extending it with a regression layer to predict reaction yields. BERT encoders belong to the transformer model family, which has revolutionised natural language processing [19, 18]. These models take sequences of tokens as input to compute contextualised representations of all the input tokens, and can be applied to reactions represented in the SMILES [20] format. In this work, we demonstrate for the first

time, that these natural language architectures are very useful not only when working with language tokens, but also to provide descriptors of high quality to predict reaction properties such as reaction yields.

It is possible to train our approach both on data specific to a given reaction class or on data representing different reaction types. Thus, we initially trained the model on two high-throughput experimentation (HTE) data sets. Among the few HTE reaction data sets published in recent years, we selected the data sets for palladium-catalysed Buchwald–Hartwig reactions provided by Ahneman et al. [9] and for Suzuki–Miyaura coupling reactions provided by Perera et al. [21]. Finally, we trained our model on patent data available in the USPTO data set [22, 23].

HTE and Patent data sets are very different in terms of content and quality. HTE data sets typically cover a very narrow region in the chemical reaction space, with chemical reaction data related to one or a few reaction templates applied to large combinations of selected precursors (reactants, solvents, bases, catalysts, etc.). In contrast, patent reactions cover a much wider reaction space. In terms of quality, HTE data sets report reactions represented uniformly and with yields measured using the same analytical equipment, thus providing a consistent and high quality collection of knowledge. In comparison, the yields from patents were measured by different scientists using different equipments. Incomplete information in the original documents, such as unreported reagents or reaction conditions, and the extensive limitation in text mining technologies makes the entire set of patent reactions quite noisy and sparse. An extensive analysis of the USPTO data set revealed that the experimental conditions and reaction parameters, such as scale of the reaction, concentrations, temperature, pressure, or reaction duration, may have a significant effect on the measured reaction yields. The functional dependency of the yields from the reaction conditions poses additional constraints, as the model presented in this work does not consider those values explicitly in the reaction descriptor. The basic assumption is that every reaction yield reported in the data set is optimised for the reaction parameters.

Our best performing model reached an R^2 score of 0.956 on a random split of the Buchwald-Hartwig data set while the highest R^2 score on the smoothed USPTO data was 0.388. These numbers reflect how the intrinsic data set limitations increase the complexity of training a sufficiently good performing model on the patent data, resulting into a more difficult challenge than training a model for the HTE data set.

2. Models & experimental pipeline

We base our models directly on the reaction fingerprint (rxnfp) models by Schwaller et al. [17]. We use a fixed size encoder model size, tuning only the hyperparameter for dropout rate and learning rate, thus avoiding often encountered difficulties of neural networks with numerous hyperparameters. During our experiments, we observed good performances for a wide range of dropout rates (from 0.1 to 0.8) and conclude that the initial learning rate is the most important hyperparameter to tune. To facilitate

the training, our work uses simpletransformers [24], huggingface transformer [25] and PyTorch framework [26]. The overall pipeline is shown in Figure 1.

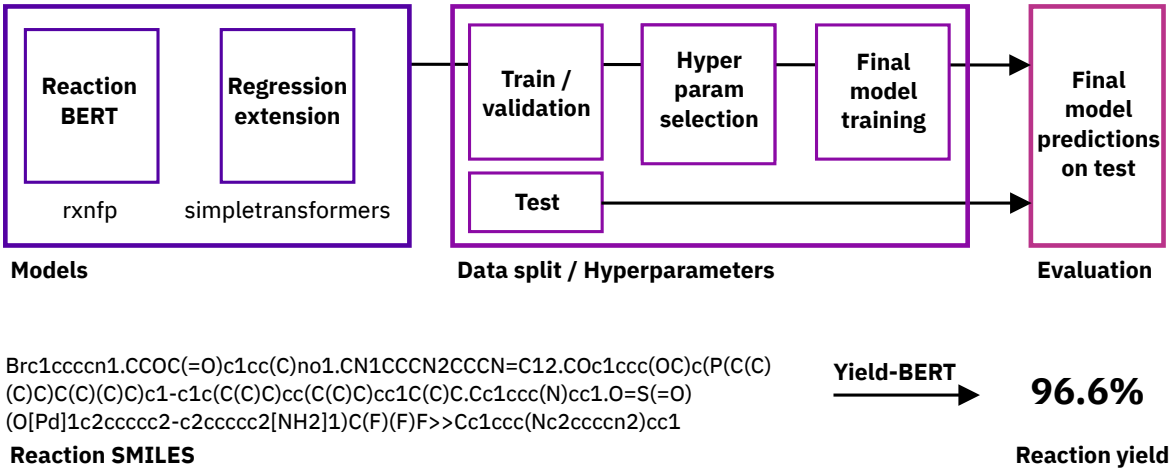

Figure 1. Training/evaluation pipeline and task description.

To provide an input compatible with the rxnfp model we use the same RDKit [27] reaction canonicalisation and SMILES tokenization [6] as in the rxnfp work [17].

3. High-throughput experiment yield predictions

3.1. Buchwald–Hartwig reactions

Ahneman et al. [9] performed high-throughput experiments on Pd-catalysed Buchwald–Hartwig C-N cross coupling reactions, measuring the yields for each reaction. For the experiments, they used three 1536-well plates spanning a matrix of 15 aryl and heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives resulting in 3955 reactions. As inputs for their models, Ahneman et al. [9] computed 120 molecular, atomic and vibrational properties with density functional theory using Spartan for every halide, ligand, base and additive combination. The descriptors included HOMO and LUMO energy, dipole moment, electronegativity, electrostatic charge and NMR shifts for atoms shared by the reagents. Compared to reaction SMILES that can vary in length, the input in the work of Ahneman et al. [9] was a fixed-size vector. They investigated numerous methods, including linear models, k-Nearest-Neighbours, support vector machines, Bayes generalised linear models, artificial neural networks and random forests. Eventually, they selected their random forest model as the best performing. The work of Ahneman et al. [9] was challenged by Chuang and Keiser [10], who pointed out several issues. First, by replacing the computed chemical features with random features of the same length or one-hot encoded vectors Chuang and Keiser got similar performance than the original paper with the chemical features. Therefore, they weakened the original claim that additive features were the most important for the predictions. However, the additive features were on average still estimated to

be the most important features by the random forest model when the yields were shuffled [10]. Recently, Sandfort et al. [11] used a concatenation of multiple molecular fingerprints as an alternative reaction representation to demonstrate superior yield prediction performance compared to one-hot encoding.

Table 1. Comparing methods on the Buchwald-Hartwig data set. All results shown in this table used the rxnfp pretrained model as base encoder.

R^2	DFT [9]	one-hot [10, 11]	MFF [11]	Yield-BERT
rand 70/30	0.92	0.89	0.927 ± 0.007	**0.951 ± 0.005**
rand 50/50	0.9			0.92 ± 0.01
rand 30/70	0.85			0.88 ± 0.01
rand 20/80	0.81			0.86 ± 0.01
rand 10/90	0.77			0.79 ± 0.02
rand 5/95	0.68			0.61 ± 0.04
rand 2.5/97.5	0.59			0.45 ± 0.05
test 1	0.8	0.69	0.85	0.84 ± 0.01
test 2	0.77	0.67	0.71	0.84 ± 0.03
test 3	0.64	0.49	0.64	0.75 ± 0.04
test 4	0.54	0.49	0.18	0.49 ± 0.05
avg. 1-4	0.69	0.59	0.60	0.73

Unlike previous work, we directly use the reaction SMILES as input to a BERT-based reaction encoder [17] enriched with a regression layer (Yield-BERT). To investigate the suggested method, we used the same splits as Sandfort et al. [11]. In contrast, to their work, we used 1/7 of the training set from the first random split as a validation set to select optimal values for the two hyperparameters, namely, learning rate and dropout probability. Once selected, we kept the hyperparameters identical for all the subsequent experiments.

The results are shown in Table 1. Using solely a reaction SMILES representation, our method achieves an average R^2 of 0.951 on the random splits and outperforms not only the MFF by Sandfort et al. [11], but also the chemical descriptors computed with DFT by Ahneman et al. [9]. Moreover, for the out-of-sample tests where the isoxazole additives define the splits our method performs on average better than MFF and one-hot descriptors and comparable to the chemical descriptors. As in the work of Sandfort et al. [11], the test 3 split resulted in the worst model performance. For the rest of the out-of-sample, our method performs better than the others. We also reduced the training set to 5% (197 reactions), 10% (395 reactions) and 20% (791 reactions) and observed that the model learned to reasonably predict yields despite the significantly smaller training set.

3.2. Suzuki–Miyaura reactions

Perera et al. [21] used HTE technologies to the class of the Suzuki–Miyaura reactions. They considered 15 pairs of electrophiles and nucleophiles, each leading to a different product. For each pair, they varied the ligands (12 in total), bases (8), and solvents (4), resulting in a total of 5760 measured yields. The same data set was also investigated in the work of Granda et al. [12].

Here, we first trained our yield prediction models with the same hyperparameters as for the Buchwald–Hartwig reaction experiment above, achieving an R^2 score of 0.79±0.01. Second, we tuned the dropout probability and learning rate, similarly to the previous experiment, using a split of the training set of the first random split. The resulting hyperparameters were then used for all the splits. The hyperparameter tuning did not lead to better performance compared to the parameters used for the Buchwald–Hartwig reactions. This shows that the models have a stable performance for a wide range of parameters and that they are transferable from one data set to another related data set.

Table 2. Summary of the average R^2 scores on the Suzuki–Miyaura reactions data set using a Yield-BERT with different base encoders. We used 10 different random folds (70/30).

Base encoder rxnfp [17] Hyperparameters	pretrained same as 3.1	pretrained tuned	ft same as 3.1	ft tuned
random 70/30	0.79 ± 0.01	0.79 ± 0.02	**0.81 ± 0.02**	**0.81 ± 0.01**

We also compared two different base encoder models that are available from the rxnfp library [17], namely the BERT model pretrained with a masked language modelling task, and the BERT model subsequently fine-tuned on a reaction class prediction task. The results are displayed in Table 2. In contrast to the Buchwald–Hartwig data set, where no difference between the two base encoders was observed, the ft model achieves an R^2 score of 0.81 ± 0.01, outperforming the pretrained base encoder on the Suzuki–Miyaura reactions.

3.3. Discovery of high yielding reactions with reduced training sets

Granda et al. [12] proposed to train on a random (10%) portion of the original data set to evaluate the rest of the reactions with the purpose of selecting the next reactions to test. Similarly, we trained our models on different fractions of the training set and used them to evaluate the yields of the remaining reactions. The aim here is to evaluate how well the models are at selecting high-yielding reactions after having seen a small fraction of randomly chosen reactions.

As can be seen from Figure 2, training on only 5% of the reactions already enables a chemist to select some of the highest yielding reactions for the next round of the

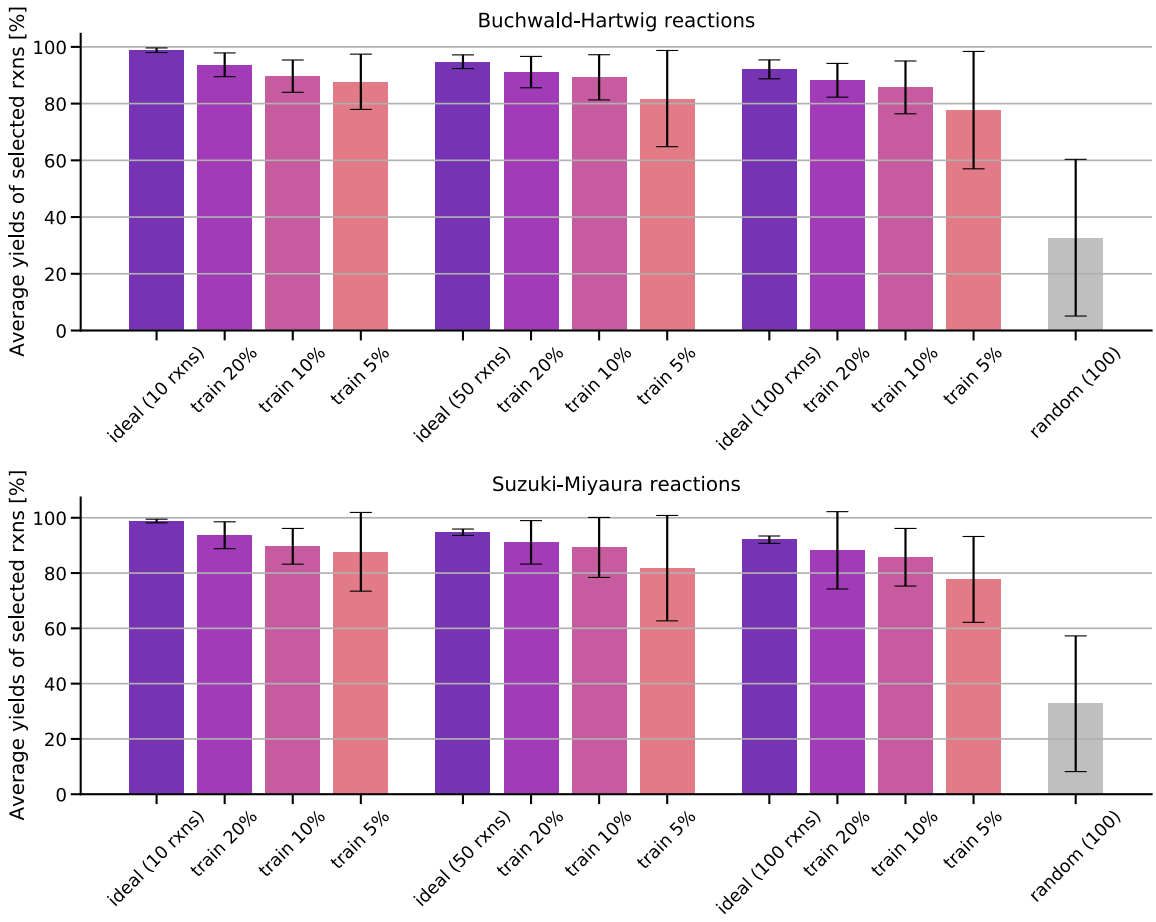

Figure 2. Average and standard deviation of the yields for the 10, 50, and 100 reactions predicted to have the highest yields after training on a fraction of the data set (5%, 10%, 20%). The ideal reaction selection and a random selection are plotted for comparison.

experiments. With a training set of 10% the yields of the selected reactions are close to the best possible selection marked with "ideal" in the Figure. For the Buchwald–Hartwig reaction, using a model trained on 10% of the data set, the 10 reactions from the remaining unseen data set predicted to have the highest yields, have an average yield of 90 ± 6 %, compared to the ideal selection of 98.7 ± 0.9 %. In contrast, a random selection of 10 reactions would have let to yields of 34 ± 27 %. The selection works similarly for the Suzuki–Miyaura reactions.

We performed a purely greedy selection, as we aimed to find highest yielding reactions after one training round. A wider chemical reaction space exploration with a reaction selection using more elaborate uncertainty estimates and an active learning strategy was investigated by Eyke et al. [14].

4. Patent yield predictions

In this section, we analyse USPTO data set [22, 23] yields. We started from the same set as in our previous work [28], keeping only reactions for which yields and product mass were reported. In contrast to HTE, where reactions are typically performed in sub-gram scale, the patent data contains reactions spanning a wider range, from grams to sub-grams scales.

4.1. Gram versus sub-gram scale

When investigating the yields for different mass scales, we observed that gram and sub-gram scales had statistically different yield distributions, as shown in Figure 3. One reason could be that the reaction sub-gram scale reactions are generally less optimised than gram-scale. In sub-gram scale, the primary goal is to show that the desired product is present. To be able to synthesise a specific compound on a larger scale, reactions are optimised and predominantly high yielding reactions are employed. Therefore, we split the USPTO reactions into two data sets according to the product mass. If for the same canonical reaction SMILES multiple yields were reported in the same mass scale, we took the average of those yields.

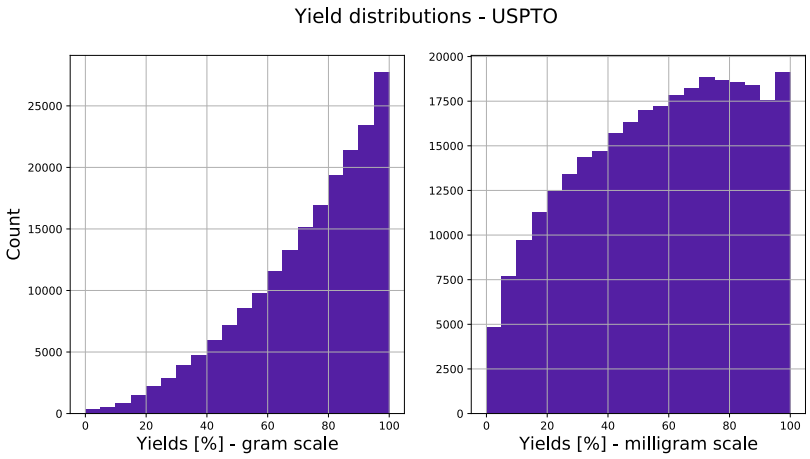

Figure 3. USPTO yields historgrams separated in gram and sub-gram scale

We performed various experiments summarised in Table 3. The R^2 scores for the randomly train-test splits with 0.117 for gram scale and 0.195 low. As expected, the tasks become even more difficult when the time split is used. In our experiment, we took all reactions first published in 2012 and before as training/validation set and the reactions published after 2012 as test set. To show that the model was still able to learn, we performed a sanity check by randomising the yields across the training reactions. The resulting performance on the test set was a R^2 score of 0.

Unfortunately, the yields from the USPTO data set could not be accurately predicted. To better understand why, we further inspected the USPTO reaction yields

with a visual analysis using reaction atlases built using TMAP [29], faerun [30] and our reaction fingerprints [17]. Figure 4 reveals that globally reaction classes tend to have similar yields. However, if a local neighbourhood is analysed the nearest neighbours often have extremely diverse reaction yields. Those diverse yields make it challenging for the model to learn anything but yield averages for similar reactions and hence, explain the low performance on the patent reactions. This analysis opens up relevant questions on the quality of the reported information (relative to the mass scale) and its extraction accuracy from text, which could severely hamper the development of reaction yield predictive models. The need of cleaned and consistent reaction yields data set is even more important than for other reaction prediction tasks.

Table 3. Summary of the R^2 scores on the different USPTO reaction sets.

scale	gram	sub-gram
random split	0.117	0.195
time split	0.095	0.142
random split (smoothed)	0.277	0.388
randomized yields	0.0	0.0

In Table 3, the "random split (smoothed)" row shows an experiment inspired from the observations above. As some of the yields values are probably incorrect in the data set, we smoothed the yields by computing the average of the three nearest neighbour yields plus twice the own yield of the reaction. The nearest neighbours were estimated using the *rxnfp ft* [17] and faiss [31]. On the smoothed data sets, the performance of our models more than triples in the gram scale and doubles on the sub-gram scale, achieving R^2 scores of 0.277 and 0.388, respectively. The removal of noisy reactions [32] or reaction data augmentation techniques [33] could potentially lead to further improvements.

5. Conclusion

In this work, we combined a reaction SMILES encoder with a reaction regression task to design a reaction yield predictive model. We analysed two HTE reaction data sets, showing excellent results. On the Buchwald–Hartwig reaction data set, our models outperform previous work on random splits and perform similar to models trained on chemical descriptors computed with DFT on test sets where specific additives were held out from the training set. Compared to random forest models, the feature importance can not directly be obtained. Future work could (visually) investigate the attention weights to find out what tokens and molecules contribute the most to the predictions [34, 35].

We analysed the yields in the public patent data and show that the distribution of reported yields strongly differs depending on the reaction scale. Because of the

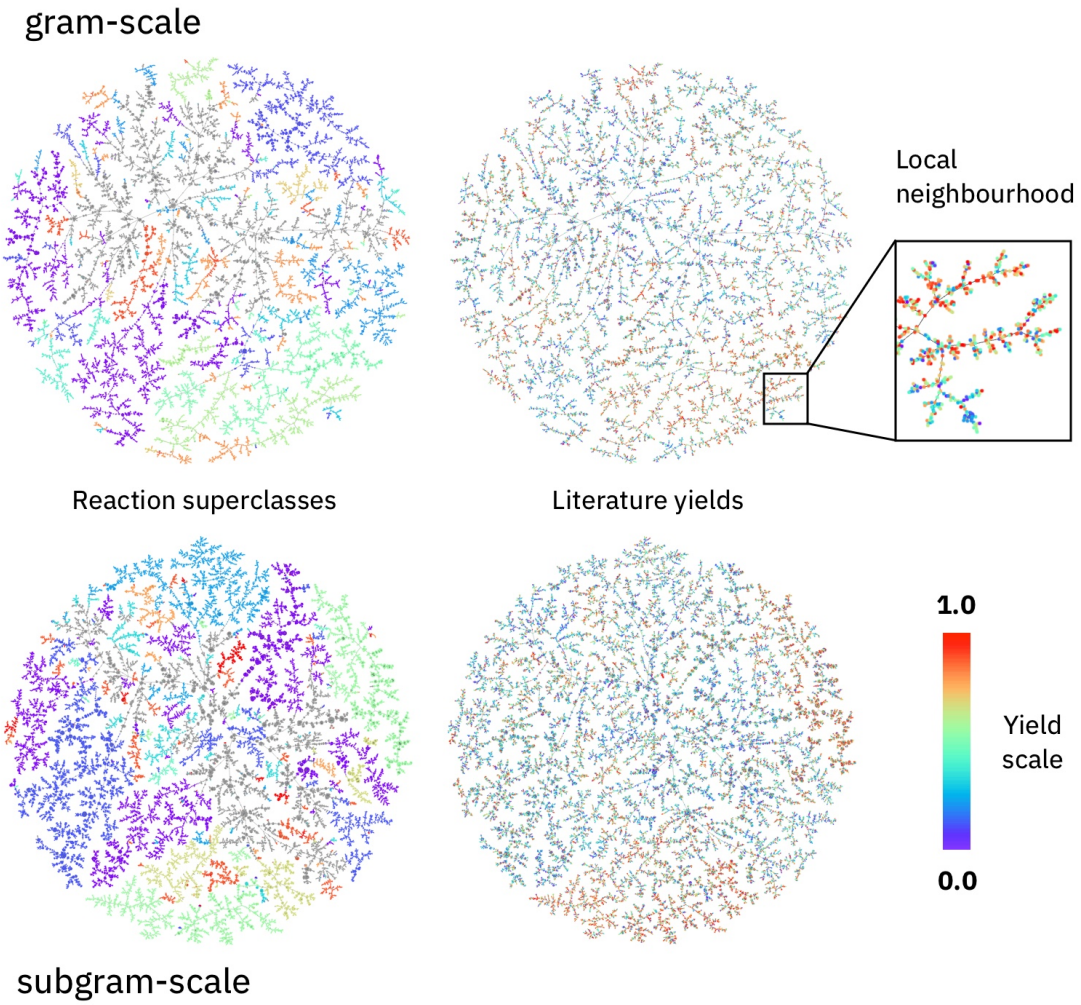

Figure 4. Reaction Atlases. Top: gram scale. Bottom: sub-gram scale. Left: Reaction superclass distribution, reactions belonging to the same superclass have the same colour. Right: Corresponding reaction yields.

intrinsic lack of consistency and quality in the patent data, our proposed method fails to predict patent reaction yields accurately. While we cannot rule out the existence of any other architecture potentially performing better than the one presented in this manuscript, we raise the need for a more consistent and better quality public data set for the development of reaction yields prediction models. The suspect that the patent data yields are inconsistently reported is substantiated by the large variability of methods used to purify and report yields by the different reaction mass scales and the different optimisation in each reported reaction. Our reaction atlases [30, 29, 17] reveal globally higher yielding reaction classes. However, nearest neighbours often have significantly scattered yields. We show that better results can be achieved by smoothing the patent data yields using the nearest neighbours.

Our approach to yield predictions can be extended to any reaction regression task, for example, for predicting reaction activation energies [36, 37, 38], and is expected to

have a broad impact in the field of organic chemistry.

The code and data are available on `https://rxn4chemistry.github.io/rxn_yields/`.

Acknowledgements

We acknowledge the RXN for Chemistry team for insightful discussion.

References

[1] 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).

[2] Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).

[3] Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science* **365**, eaax1566 (2019).

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

[5] Genheden, S. *et al.* AiZynthFinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. *ChemRxiv preprint* doi:10.26434/chemrxiv.12465371.v1 (2020).

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

[7] Kite, S., Hattori, T. & Murakami, Y. Estimation of catalytic performance by neural network — product distribution in oxidative dehydrogenation of ethylbenzene. *Appl. Catal., A* **114**, L173 – L178 (1994).

[8] Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).

[9] 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).

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

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

[12] Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).

[13] Fu, Z. *et al.* Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki–Miyaura cross-coupling reaction. *Org. Chem. Front.* (2020).

[14] 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).

[15] Skoraczyński, G. *et al.* Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Sci. Rep.* **7**, 3582 (2017).

[16] 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).

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

[18] 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).

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

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

[22] Lowe, D. M. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge (2012).

[23] Lowe, D. Chemical reactions from US patents (1976-Sep2016) (2017).

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

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

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

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

[28] 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).

[29] Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminf.* **12**, 1–13 (2020).

[30] Probst, D. & Reymond, J.-L. Fun: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **34**, 1433–1435 (2017).

[31] Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734* (2017).

[32] 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).

[33] Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. Augmented transformer achieves 97% and 85% for Top5 prediction of direct and classical retro-synthesis. *arXiv preprint arXiv:2003.02804* (2020).

[34] Hoover, B., Strobelt, H. & Gehrmann, S. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276* (2019).

[35] Vig, J. & Belinkov, Y. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 63–76 (2019).

[36] Grambow, C. A., Pattanaik, L. & Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **7**, 1–8 (2020).

[37] von Rudorff, G. F., Heinen, S., Bragato, M. & von Lilienfeld, A. Thousands of reactants and transition states for competing E2 and SN2 reactions. *Machine Learning: Science and Technology* (2020).

[38] Jorner, K., Brinck, T., Norrby, P.-O. & Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *ChemRxiv preprint* doi:10.26434/chemrxiv.12758498.v1 (2020).

Supplementary Information: Prediction of Chemical Reaction Yields using Deep Learning

Philippe Schwaller

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

E-mail: phs@zurich.ibm.com

Alain C Vaucher, Teodoro Laino

IBM Research – Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

Jean-Louis Reymond

Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland

Contents

1 Detailed results on Buchwald Hartwig reactions 1

2 Detailed results on Suzuki-Miyaura reactions 9

3 Detailed analysis of USPTO yields data 14

4 Hyperparameter tuning 15

1. Detailed results on Buchwald Hartwig reactions

Figure S1-S14 show the correlation between the measured yields and the predicted yields for the different splits published by Sandfort et al. [1]. Moreover, the root mean squared error (RMSE) and the mean average error (MAE) are shown in the figures.

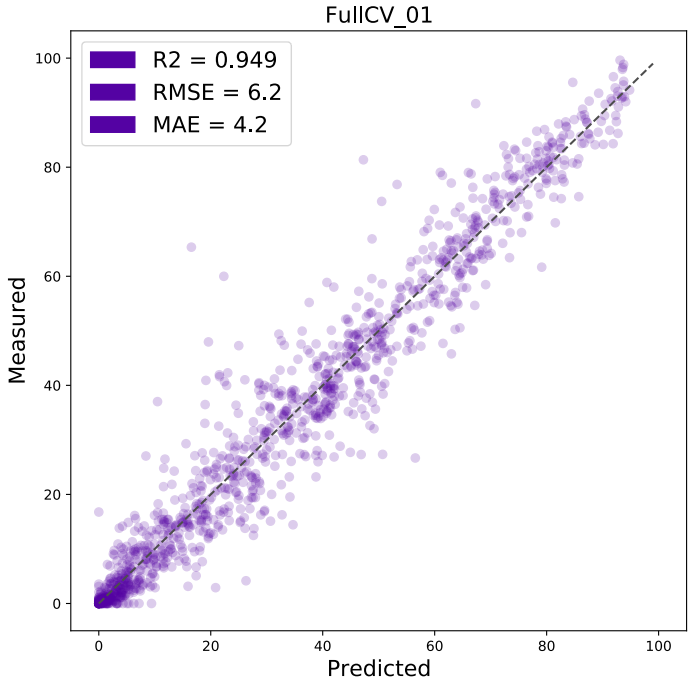

Figure S1. Measured vs predicted yields [%] - FullCV_01

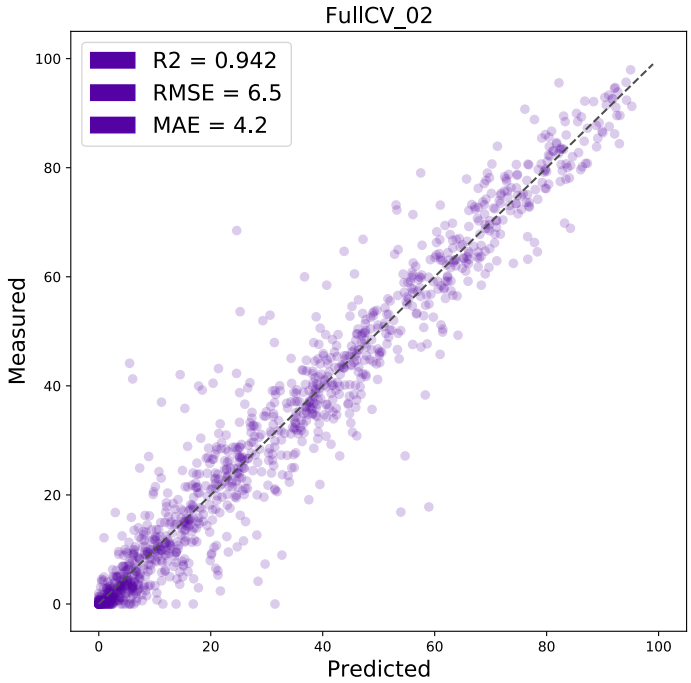

Figure S2. Measured vs predicted yields [%] - FullCV_02

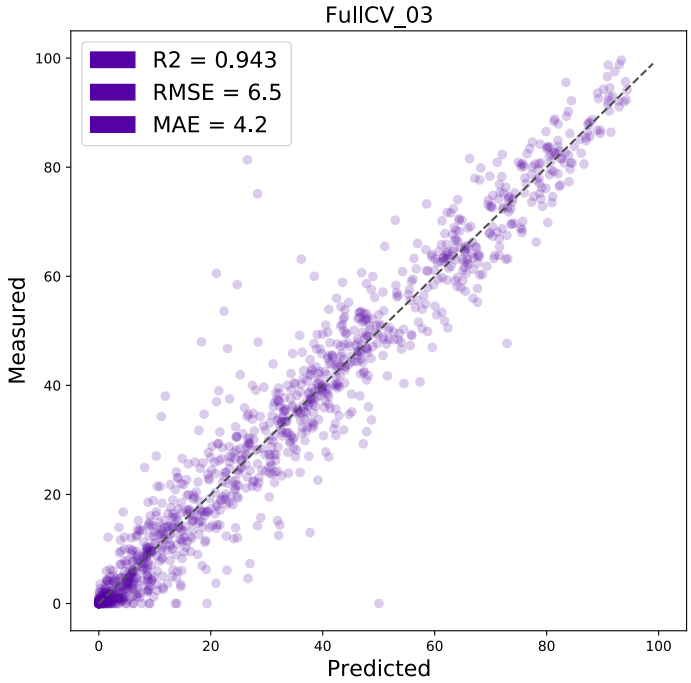

Figure S3. Measured vs predicted yields [%] - FullCV_03

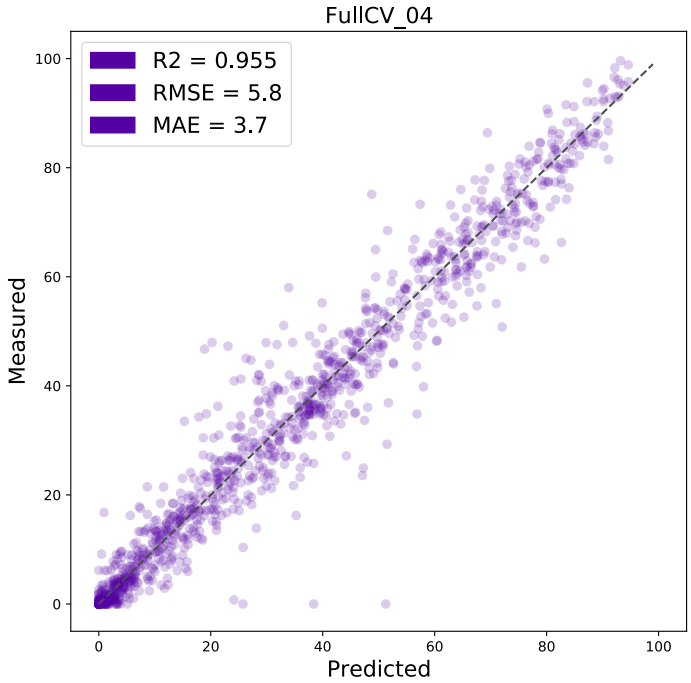

Figure S4. Measured vs predicted yields [%] - FullCV_04

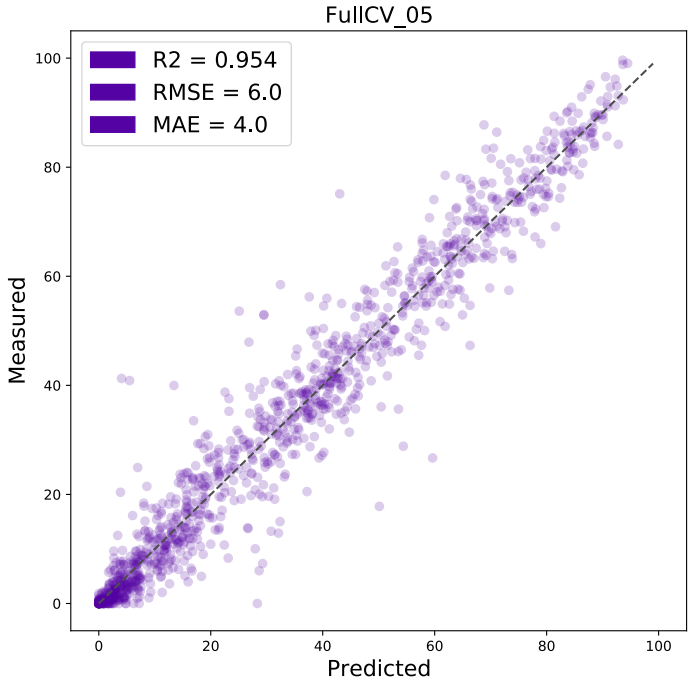

Figure S5. Measured vs predicted yields [%] - FullCV_05

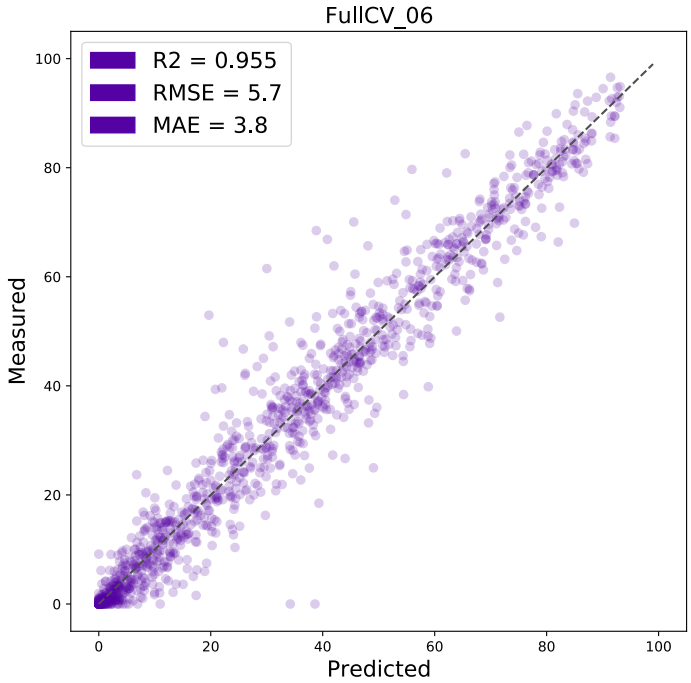

Figure S6. Measured vs predicted yields [%] - FullCV_06

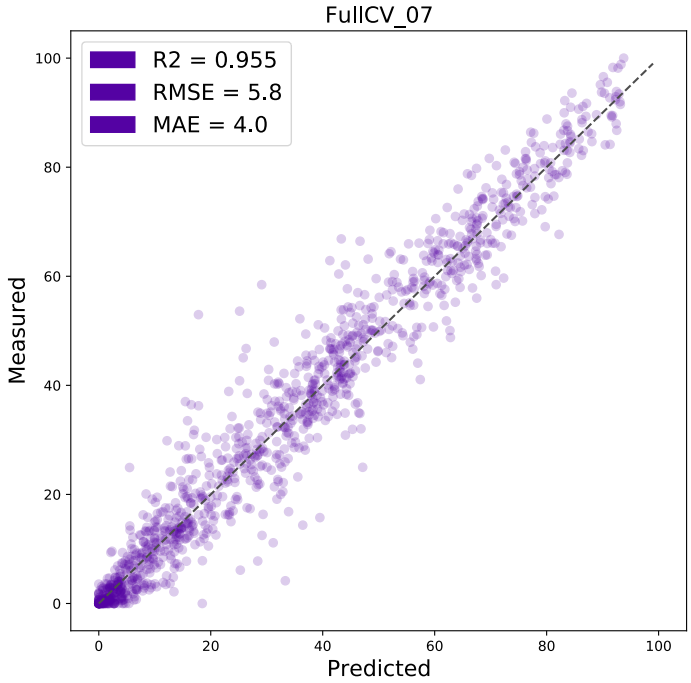

Figure S7. Measured vs predicted yields [%] - FullCV_07

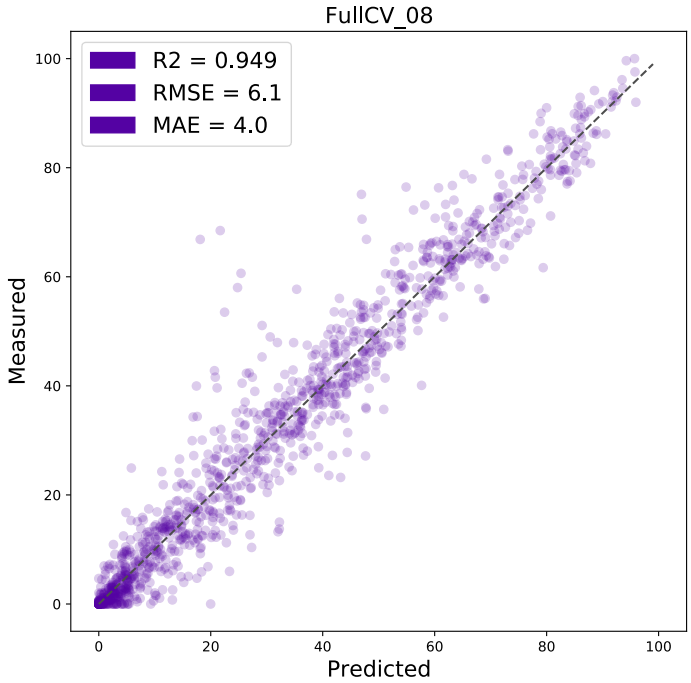

Figure S8. Measured vs predicted yields [%] - FullCV_08

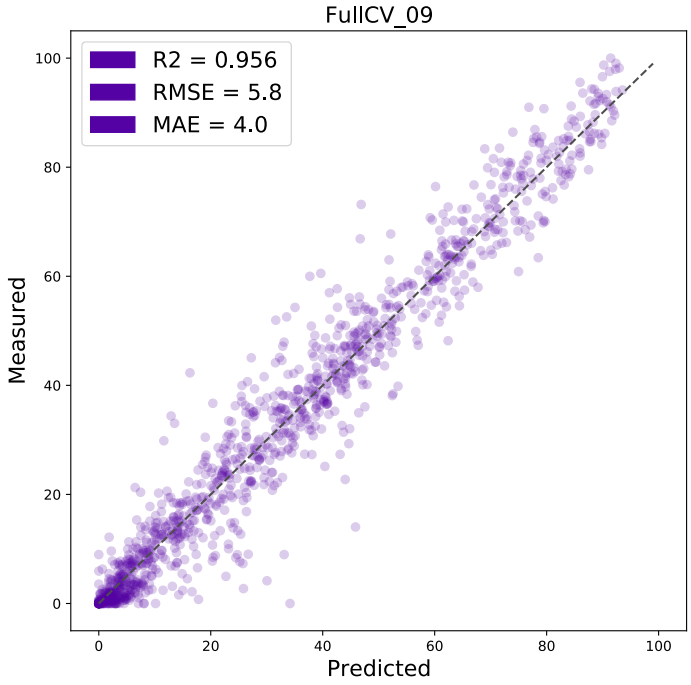

Figure S9. Measured vs predicted yields [%] - FullCV_09

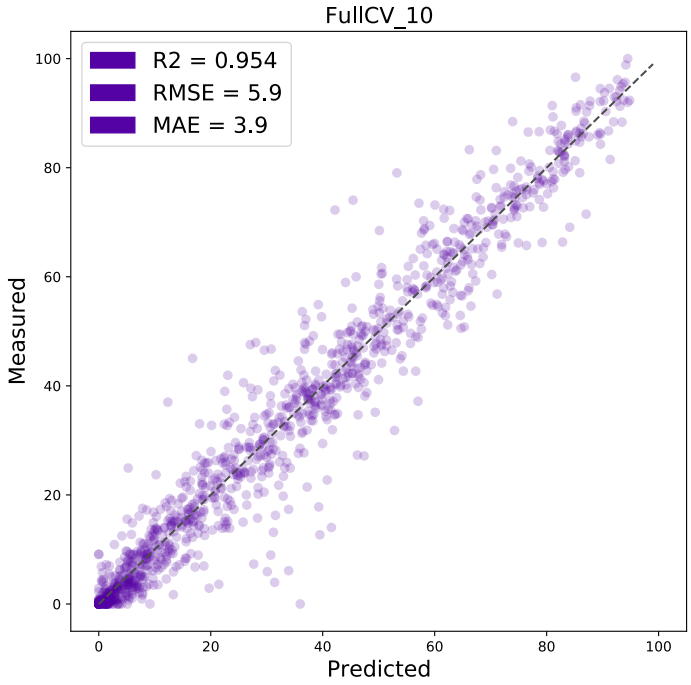

Figure S10. Measured vs predicted yields [%] - FullCV_10

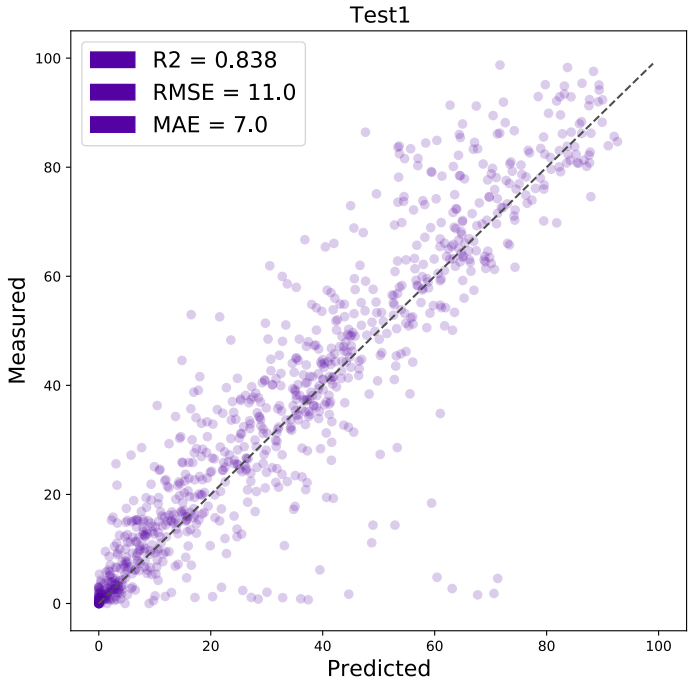

Figure S11. Measured vs predicted yields [%] - Test1

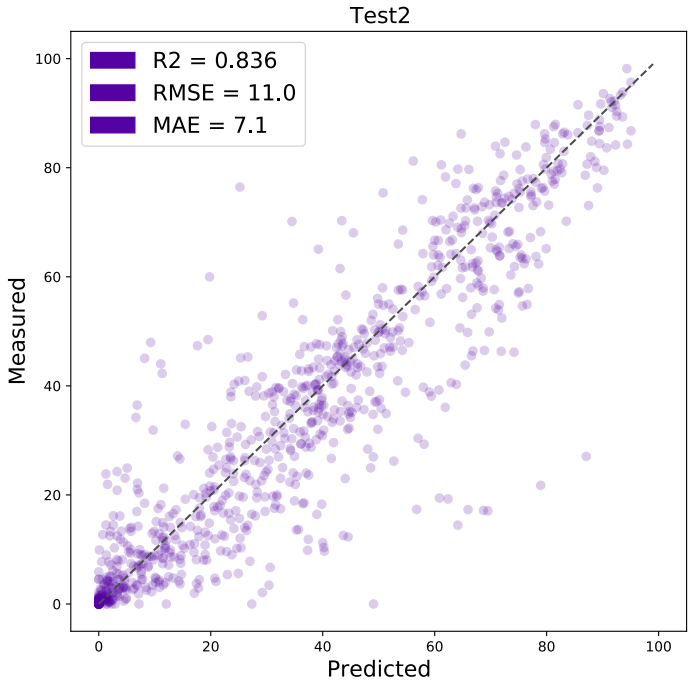

Figure S12. Measured vs predicted yields [%] - Test2

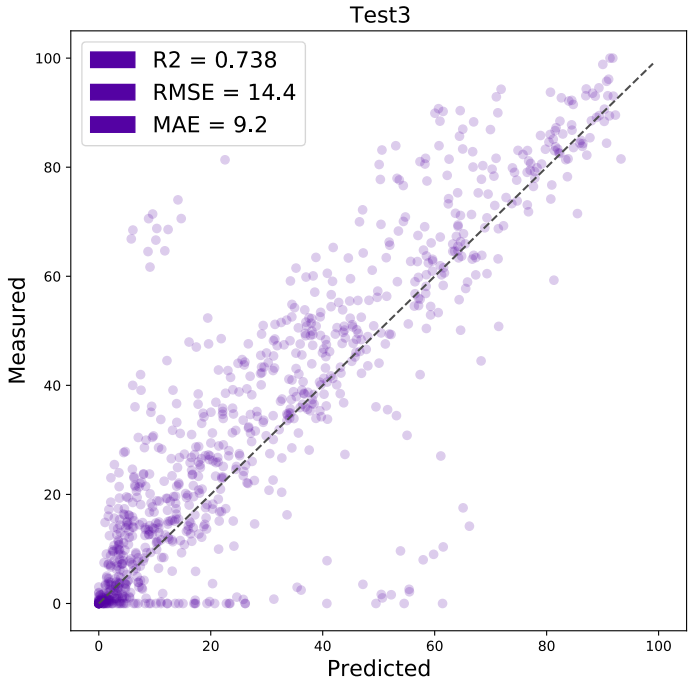

Figure S13. Measured vs predicted yields [%] - Test3

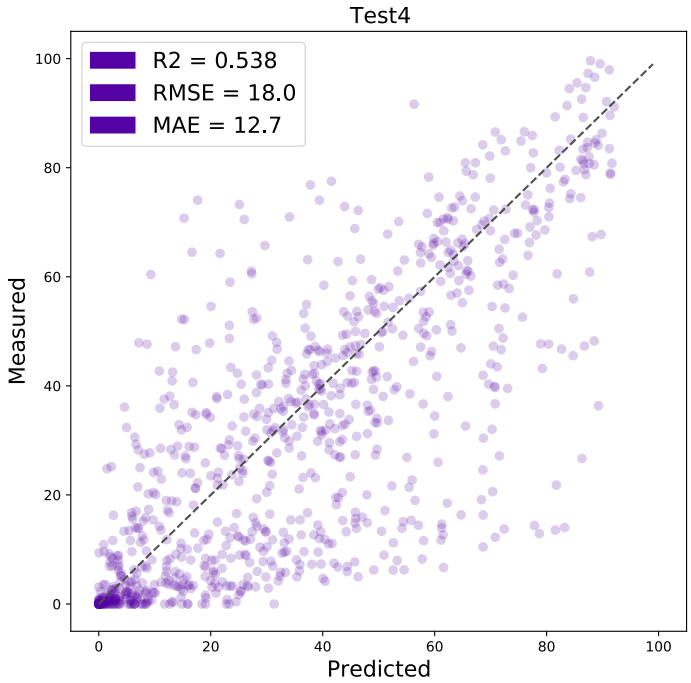

Figure S14. Measured vs predicted yields [%] - Test4

2. Detailed results on Suzuki-Miyaura reactions

Figure S15–S24 show the correlation between the measured yields and the predicted yields for model with the *rxnfp ft* base encoder on the 10 random splits.

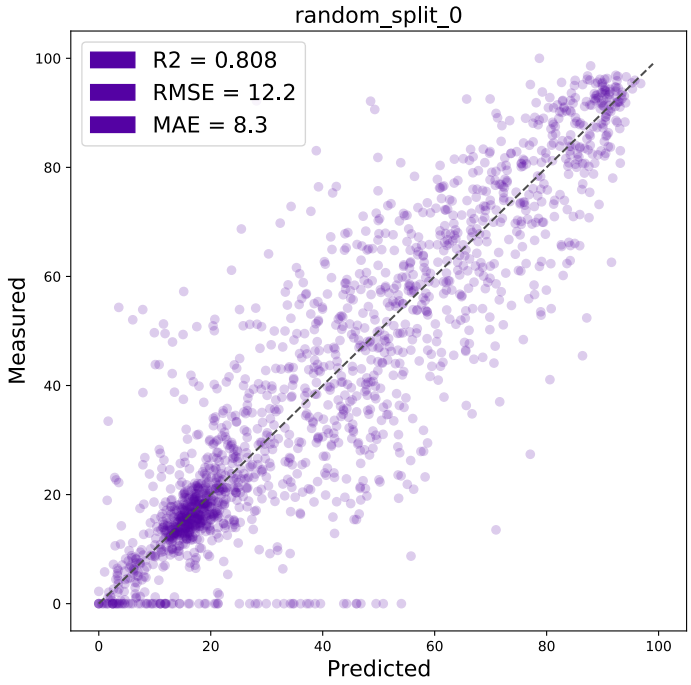

Figure S15. Measured vs predicted yields [%] - random_split_0

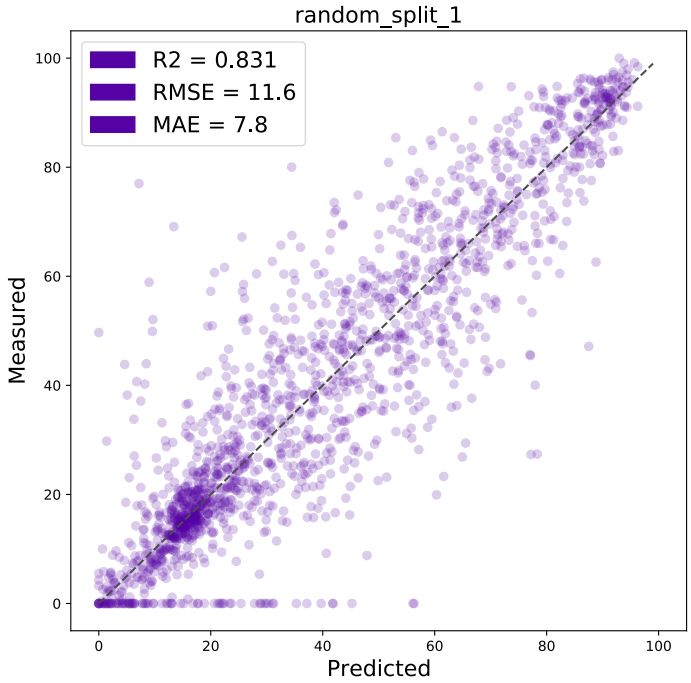

Figure S16. Measured vs predicted yields [%] - random_split_1

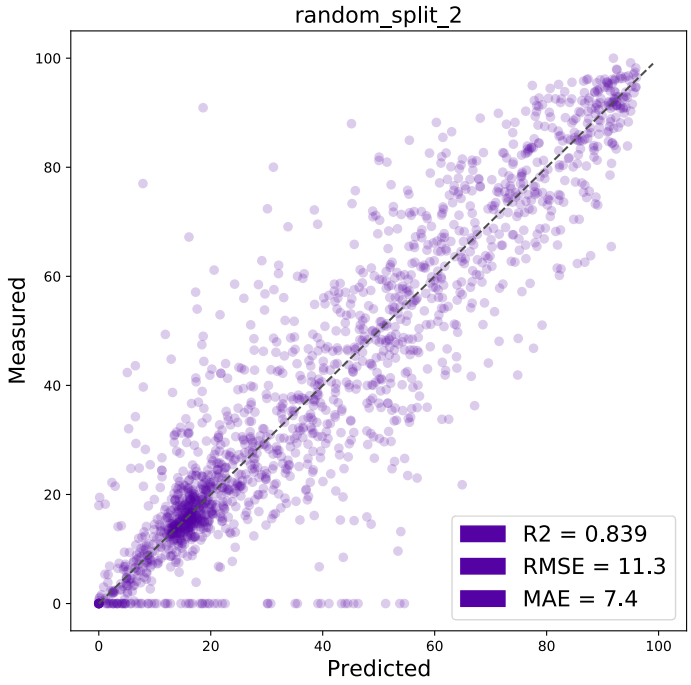

Figure S17. Measured vs predicted yields [%] - random_split_2

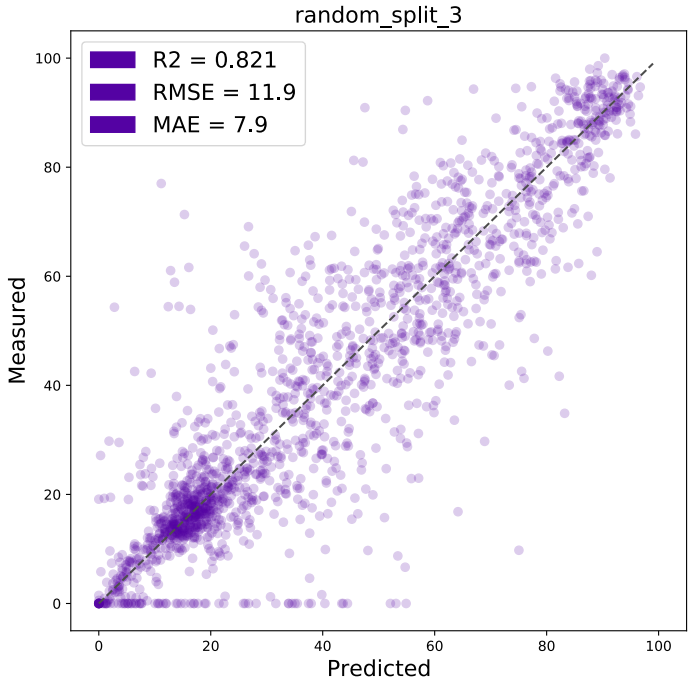

Figure S18. Measured vs predicted yields [%] - random_split_3

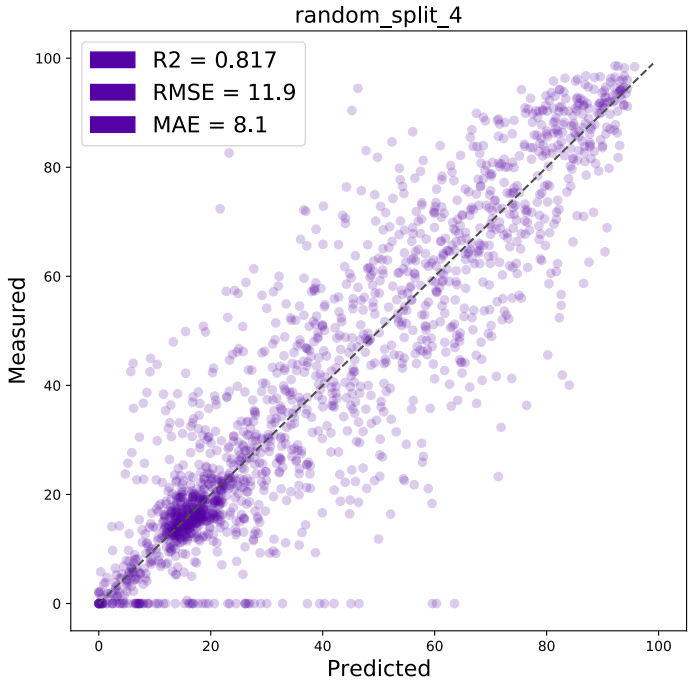

Figure S19. Measured vs predicted yields [%] - random_split_4

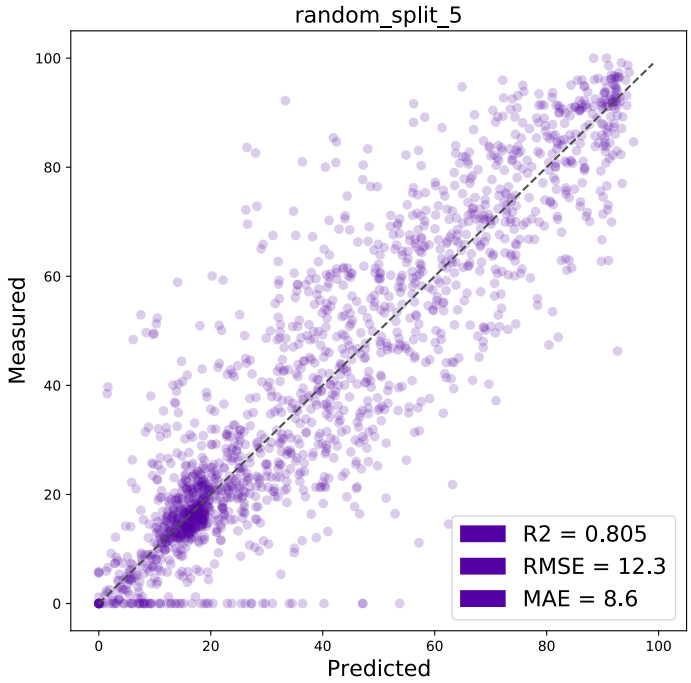

Figure S20. Measured vs predicted yields [%] - random_split_5

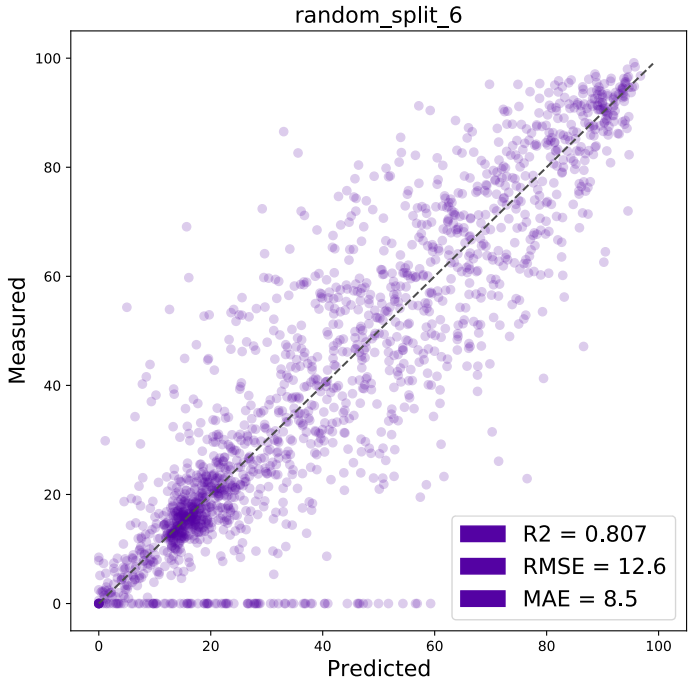

Figure S21. Measured vs predicted yields [%] - random_split_6

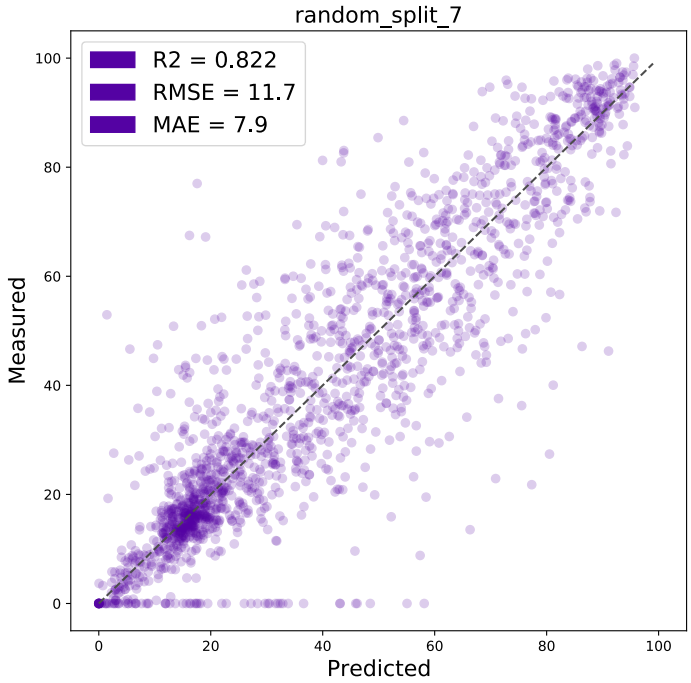

Figure S22. Measured vs predicted yields [%] - random_split_7

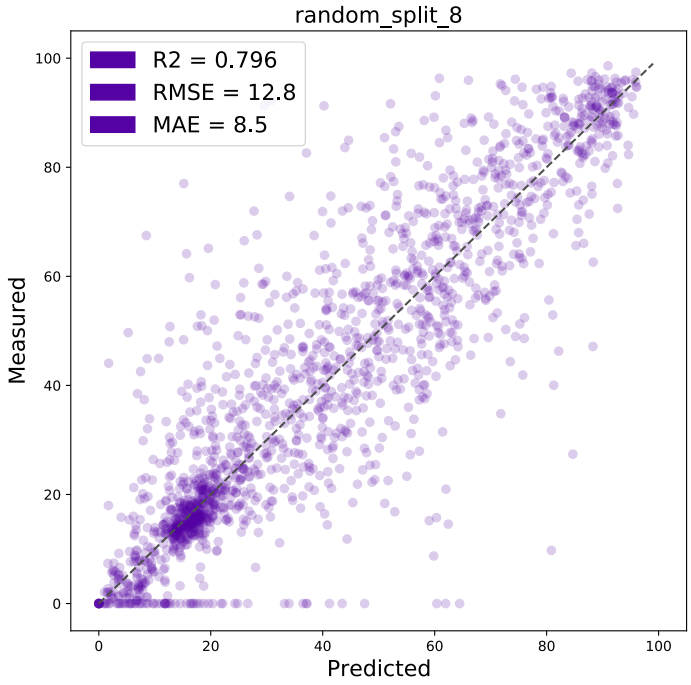

Figure S23. Measured vs predicted yields [%] - random_split_8

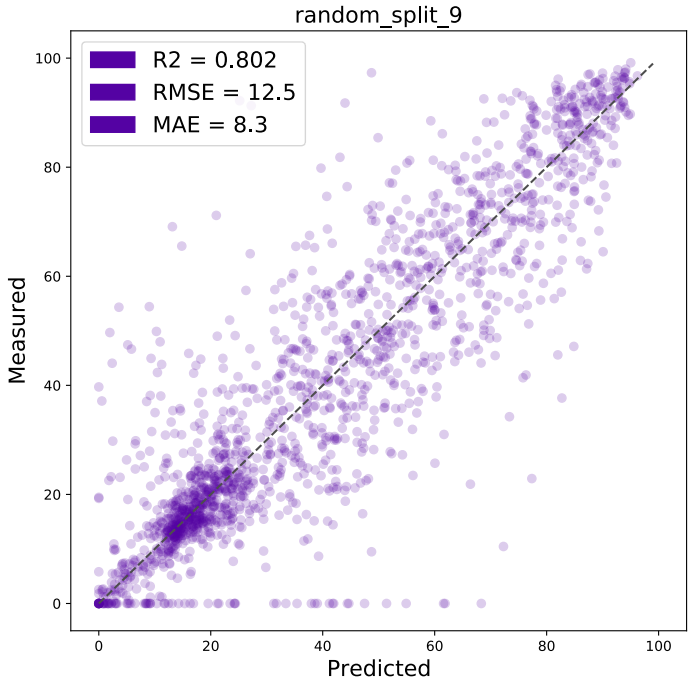

Figure S24. Measured vs predicted yields [%] - random_split_9

3. Detailed analysis of USPTO yields data

Table S1 show global statistics on the gram scale and sub-gram scale USPTO yields data sets.

Table S1. USPTO yield statistics

	gram scale	subgram scale
count	197619	302040
mean	73.2	56.8
std	20.9	26.6
min	0.0	0.0
25%	60.2	35.5
50%	78.0	58.9
75%	90.3	79.5
max	100.0	100.0

Tables S2 and S3 show the yields average in the random split test set for the different reaction superclasses.

Figure S25 shows the distributions of the smoothed yields. To smooth the yields of the USPTO data set [2, 3] we calculated the average of the 3 nearest-neighbours of the reaction, computed using the *rxnfp ft* [4] and faiss [5], and twice the own reaction yield.

Table S2. Test set sub-gram scale. Average and standard deviation per class.

Class	Name	Mean [%]	Std	Count
0	Unrecognised	52.1	26.8	12359
1	Heteroatom alkylation and arylation	53.3	25.8	12995
2	Acylation and related processes	54.8	25.6	10583
3	C-C bond formation	53.2	25.6	5111
4	Heterocycle formation	48.0	25.1	2043
5	Protections	69.8	22.3	527
6	Deprotections	68.7	25.2	8542
7	Reductions	67.5	26.1	3528
8	Oxidations	63.4	25.3	1078
9	Functional group interconversion (FGI)	62.3	25.2	2779
10	Functional group addition (FGA)	56.2	25.1	863

Table S3. Test set gram scale. Average and standard deviation per class.

Class	Name	Mean [%]	Std	Count
0	Unrecognised	69.4	22.0	10327
1	Heteroatom alkylation and arylation	71.9	20.9	7912
2	Acylation and related processes	74.5	19.7	4745
3	C-C bond formation	70.7	20.0	2547
4	Heterocycle formation	67.1	22.9	1417
5	Protections	79.9	18.5	1154
6	Deprotections	82.2	16.9	3332
7	Reductions	81.2	18.2	3105
8	Oxidations	76.0	18.8	742
9	Functional group interconversion (FGI)	74.9	20.1	2751
10	Functional group addition (FGA)	71.7	21.7	1491

4. Hyperparameter tuning

The two hyperparameters we tuned were dropout rate (between 0.05 and 0.8) and learning rate (between 1e-6 and 1e-4). For the *rxnfp pretrained* model on the Buchwald-Hartwig reactions a learning rate of 9.659e-05 and dropout probability of 0.7987 led to the highest validation R^2 score. We observe high R^2 scores for a wide range of dropout probabilities. The hyperparameter tuning was performed on a single *Nvidia RTX 2070 super* GPU and the optimal hyperparameters were found in less than 12 hours. A typical training run (10 epochs) on the same hardware takes 4 minutes and 30 seconds. We trained the final models for 15 epochs.

On the Suzuki-Miyaura reactions, we selected a learning rate of 5.812e-05 and

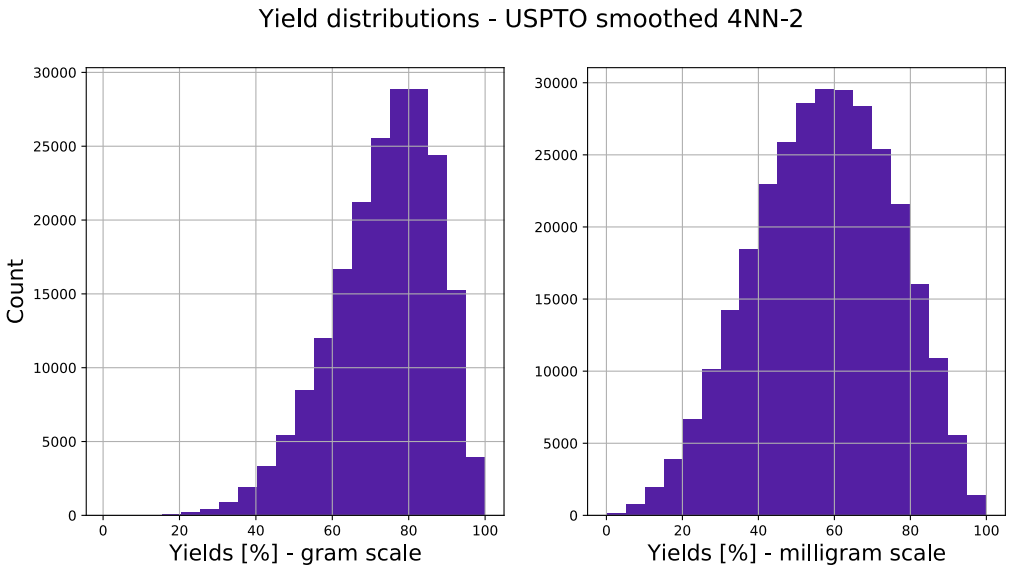

Figure S25. Smoothed USPTO yields distribution separated in gram and sub-gram scale

dropout probability of 0.5848 for the *rxnfp pretrained* base encoder and a learning rate of 9.116e-05 and dropout probability 0.7542 for the *rxnfp ft* base encoder model.

On the USPTO data we performed a hyperparameter search using a reduced training set of 50k reactions and only 3 epochs. We selected a learning rate of 1.562e-05 and dropout probability of 0.5237 for the gram scale and 2.958e-05 and 0.5826 respectively, for the sub-gram scale. The final models were trained for 2 epochs on the complete training data, as an evaluation showed signs of over-fitting from the third epochs on.

Figure S26 – S30 show the hyperparameters with the corresponding R^2 values on the validation set. The validation was made on subsplit of the training set of the first random split for all three data sets. Overall, the learning rate seemed to be more important to tune than the dropout probability.

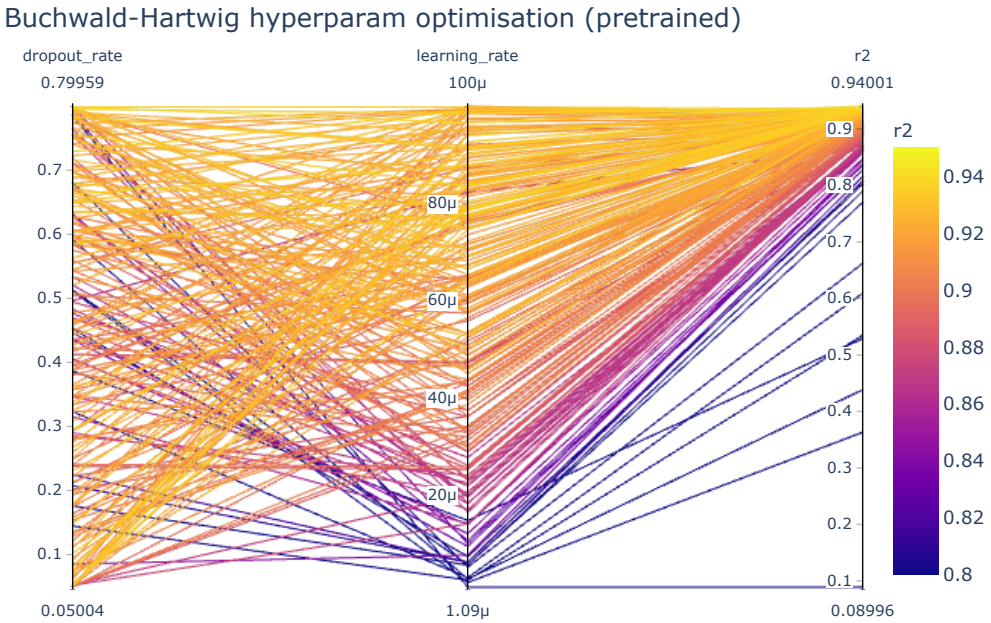

Figure S26. Hyperparameter optimisation on Buchwald-Hartwig data set (pretrained base encoder)

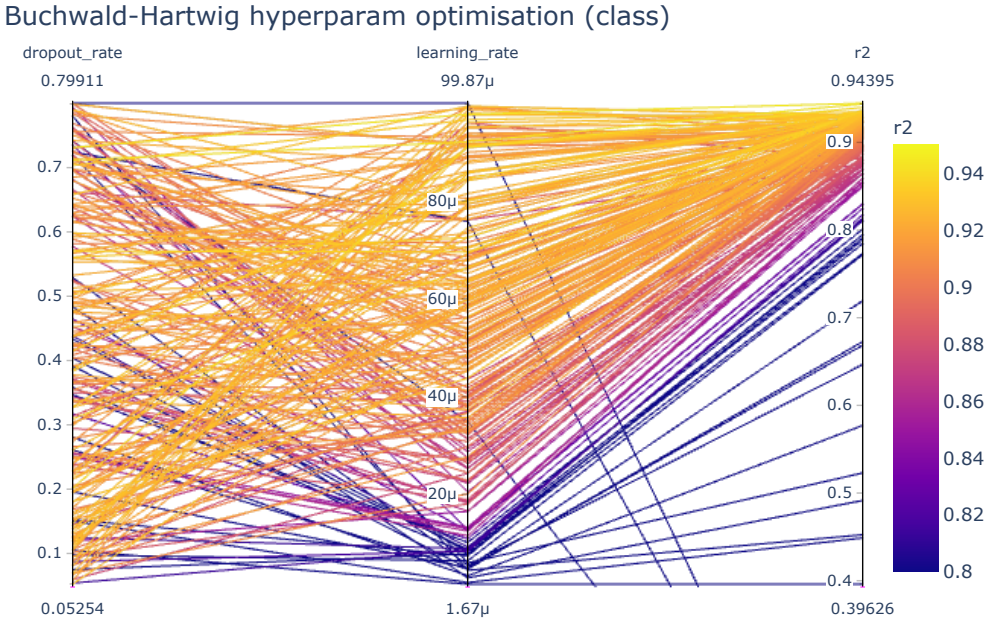

Figure S27. Hyperparameter optimisation on Buchwald-Hartwig data set (class base encoder)

References

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

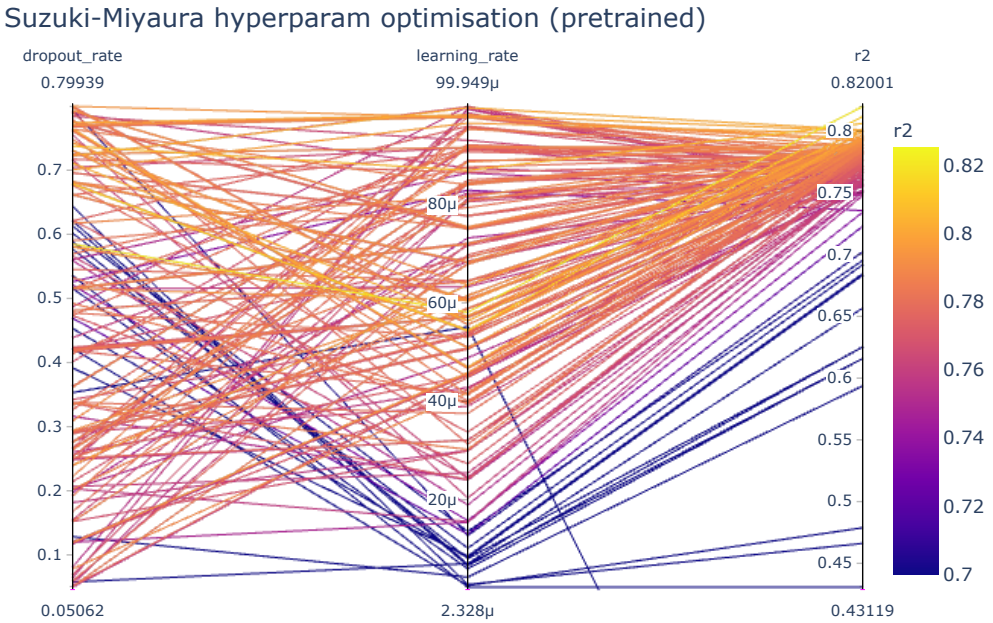

Figure S28. Hyperparameter optimisation on Suzuki-Miyaura data set (pretrained base encoder)

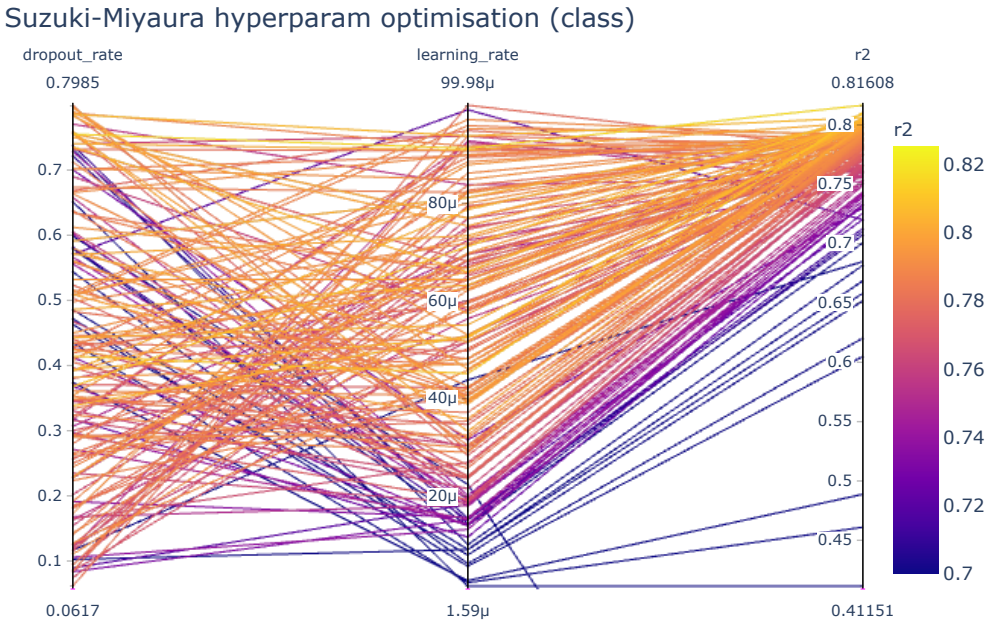

Figure S29. Hyperparameter optimisation on Suzuki-Miyaura data set (class base encoder)

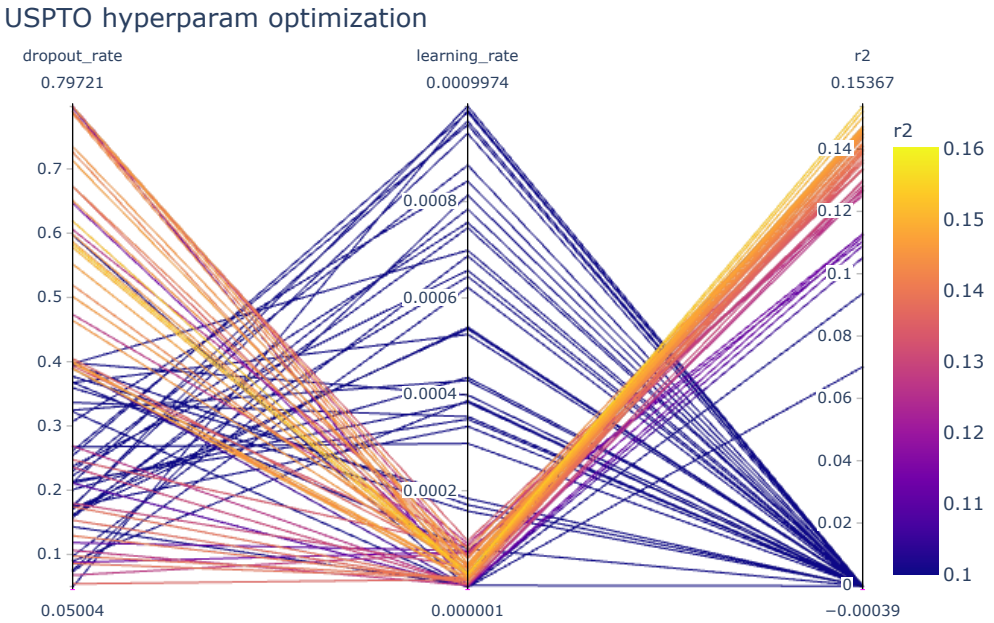

Figure S30. Hyperparameter optimisation on USPTO subgram data set (pretrained base encoder)

[2] Lowe, D. M. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge (2012).

[3] Lowe, D. Chemical reactions from US patents (1976-Sep2016) (2017).

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

[5] Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734* (2017).