Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

COMMIT

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 1988

Commit: Reaction classification and yield prediction using the differential reaction fingerprint DRFP

Daniel Probst

In "Reaction classification and yield prediction using the differential reaction fingerprint DRFP", we introduced a chemical reaction fingerprint based on the symmetric difference $A\Delta B$ of two sets A and B. With DRFP, were present a reaction as the two sets R and P, where R contains the fragments of one or more reactants and P the fragments of one or more products. The SMILES strings of the fragments in the symmetric difference of fragments $R\Delta P$ are then hashed and folded into a binary vector. We evaluated DRFP-trained models on high through put experiment data where it performed at least as well as DFT-based and learned fingerprints. In this commit, we present the evaluation of DRFP-trained XGBoost and Random Forest regressors on a recently released set of electronic laboratory notebook-extracted Buchwald–Hartwig reactions where it performs better than other methods by a wide margin. This result underlines the status of DRFP as a strong baseline for reaction representation and yield prediction.

Received 6th March 2025
Accepted 23rd June 2025

DOI: 10.1039/d5dd00089k

rsc.li/digitaldiscovery

1 Introduction

In reaction classification and yield prediction using the differential reaction fingerprint DRFP,[1] we introduced a chemical reaction fingerprint based on the symmetric difference $A\Delta B$ of two sets A and B. With DRFP, we represent a reaction as the two sets R and P, where R contains the fragments of one or more reactants and P the fragments of one or more products. The SMILES strings of the fragments in the symmetric difference of fragments $R\Delta P$ are then hashed and folded into a binary vector.

We showed that gradient boosting models based on this conceptually simple reaction fingerprint can perform at least as well as DFT- and learned fingerprint-based approaches in reaction yield prediction on high-throughput experiment (HTE) data of palladium-catalysed Buchwald–Hartwig reactions.[2] In a reaction classification task on the USPTO 1k TPL data set,[3] our method outperformed the baseline set by another fingerprint-based approach and performed similar to a large language model Yield-BERT.[4] However, since the inception of DRFP, a more challenging data set of electronic laboratory notebook-extracted (ELN) Buchwald–Hartwig reactions with experimentally determined yields has been released by Saebi *et al.*[5]

Compared to HTE reactions, those in the ELN data set cover a much broader and diverse reaction space and, due to the nature of manual experiments, differ in regard to reaction conditions and operator.[5] While the HTE data set encompasses an exhaustive combinatorial space of 15 aryl and heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives resulting in 4608 reactions including controls, the ELN data set consists of 781 samples from a reaction space exceeding 450 000 000 possible combinations of 340 aryl halides, 260 amines, 24 ligands, 15 bases, and 15 solvents.[2,5] This difference in the size of the underlying reaction space makes yield predictions on the ELN data a significantly more challenging task than training and testing models on the HTE data.

2 Results & discussion

Benchmarking DRFP on the data released by Saebi *et al.*,[5] we show that XGBoost or Random Forest (RF) regressors trained on DRFP reaction fingerprints perform better than the large language model-based Yield-BERT, the graph neural network YieldGNN,[5] and our recently released MSR2-RXN, which is, to the best of our knowledge, the currently best performing model on the ELN data set.[6] The DRFP-trained XGBoost and RF regressors improve the mean absolute error (MAE) by 20% and 13% compared to Yield-BERT and YieldGNN, respectively. Compared to our recently published set-based MSR2-RXN model, the DRFP-trained models improve the MAE by 4.2%.

These results show that, given reaction data sampled from a large, diverse reaction space, architecturally simple machine learning methods, paired with a sample distribution-agnostic computational representation of the reactions, retain more of their predictive performance compared to deep learning-based methods, which learn reaction representations from the samples or pretraining data. While the HTE data set is larger ($n = 4608$) than the ELN set ($n = 781$), this size difference does not explain the lower performance as Yield-BERT, YieldGNN, and DRFP have been evaluated on as little as 115 training samples (a

Bioinformatics Group Wageningen University & Research Wageningen, The Netherlands. E-mail: daniel.probst@wur.nl

Table 1 Yield prediction on ELN-extracted Buchwald–Hartwig reactions. Yield-BERT, YieldGNN, MSR2-RXN, and DRFP-trained XGBoost and Random Forest (RF) models compared to a random basline where the ground truth values were shuffled. The best results per metric are printed in bold, the runners-up are underlined

Method	R^2 (↑)	MAE (↓)
Shuffle	−0.16 ± 0.060	0.25 ± 0.011
Yield-BERT	−0.01 ± 0.110	0.25 ± 0.010
YieldGNN	0.05 ± 0.007	0.23 ± 0.001
MSR2-RXN	0.13 ± 0.080	0.21 ± 0.012
DRFP (XGBoost)	0.21 ± 0.052	**0.20 ± 0.010**
DRFP (RF)	**0.24 ± 0.036**	**0.20 ± 0.007**

2.5% training and 97.5% test split) during ablation studies on the HTE data set.[1] However, unlike DRFP, Yield-BERT, YieldGNN, and MSR2-RXN increasingly suffer from the sparsity of the ELN data set, which covers only a small subset ($|T \subset S| = 781$) of the reaction space ($|S| = 450\,000\,000$); a known challenge for deep learning, and specifically deep representation learning, approaches that learn a lower-dimensional representation of the reactions from the input or pretraining data (Table 1).[7,8]

As Yield-BERT and YieldGNN fail to substantially improve on a random baseline (shuffled ground truth yield values), the improvements by the DRFP-trained models are still only of limited use in a laboratory setting. Nevertheless, we show that DRFP provides a strong baseline for yield prediction on ELN-extracted reaction data as well as HTE data, which has not been reached by recent large language models (Yield-BERT), graph neural networks (YieldGNN), or set representation-based methods (MSR2-RXN). Furthermore, beyond setting a baseline for accuracy in yield prediction in real-world settings, DRFP also readily integrates with explainable machine learning methodologies due to the deterministic nature of the fingerprint.[9] Finally, the DRFP-based models were again trained and evaluated on a laptop CPU (11th Gen Intel(R) Core(TM) i7-1165G7@2.80 GHz), highlighting the computational efficiency of the method compared to deep learning-based approaches.

A potential limitation of the approach is that both the HTE and ELN data sets contain small molecule reactions that are well-suited to the DRFP algorithm, which is based on extracting molecular substructures. Therefore, DRFP-based models suffer from the same limitations as substructure fingerprints, such as ECFP, namely, reduced performance on large or repetitive molecules, including lipids, carbohydrates, peptides, and polymers in general.[10] However, taking inspiration from more recent developments in molecular fingerprints, such as MAP4, which generalizes across diverse molecules, may improve DRFP-based models when applied to reaction data sets containing large molecules, as is often the case with natural products.[11]

3 Conclusion

As with the previously studied HTE reaction data sets, DRFP-trained models perform well on ELN-extracted reaction data compared to other state-of-the-art models. While DRFP managed to match the performance of Yield-BERT and YieldGNN on the HTE data, it performs substantially better on the ELN-extracted data set, showing an improvement of up to 20% in mean absolute error (MAE). We therefore believe that DRFP-based models provide an excellent baseline for learning on diverse, real-world reaction data, as they mitigate the negative effects of under-sampled or biased training data from these large and complex reaction spaces.

Data availability

The source code, data and processing scripts for this paper, including the scripts to generate the fingerprints and the models are available at **https://github.com/reymond-group/drfp**. A release associated with the commit has been uploaded to Zenodo under the record **https://zenodo.org/records/14991185** with **https://doi.org/10.5281/zenodo.5268143**.

Conflicts of interest

There are no conflicts of interest to declare.

References

1 D. Probst, P. Schwaller and J. L. Reymond, Reaction classification and yield prediction using the differential reaction fingerprint DRFP, *Digital Discovery*, 2022, **1**(2), 91–97, DOI: **10.1039/D1DD00006C**.
2 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning, *Science*, 2018, **360**(6385), 186–190.
3 P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, *et al.*, Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks, *Nat. Mach. Intell.*, 2021, **3**(2), 144–152.
4 P. Schwaller, A. C. Vaucher, T. Laino and J. L. Reymond, Prediction of chemical reaction yields using deep learning, *Mach. Learn.: Sci. Technol.*, 2021, **2**(1), 015016, DOI: **10.1088/2632-2153/abc81d**.
5 M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, *et al.*, On the use of real-world datasets for reaction yield prediction, *Chem. Sci.*, 2023, **14**(19), 4997–5005, DOI: **10.1039/D2SC06041H**.
6 M. Boulougouri, P. Vandergheynst and D. Probst, Molecular set representation learning, *Nat. Mach. Intell.*, 2024, **6**(7), 754–763, DOI: **10.1038/s42256-024-00856-0**.
7 R. Geirhos, J. H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, *et al.*, Shortcut Learning in Deep Neural Networks, *Nat. Mach. Intell.*, 2020, **2**(11), 665–673, available from, **https://www.nature.com/articles/s42256-020-00257-z**.
8 A. K. Lampinen, S. C. Y. Chan and K. Hermann, Learned Feature Representations Are Biased by Complexity, Learning Order, Position, and More, available from, **https://openreview.net/forum?id=aY2nsgE97a**.

9 D. Probst, An explainability framework for deep learning on chemical reactions exemplified by enzyme-catalysed reaction classification, *J. Cheminform.*, 2023, **15**(1), 113, DOI: **10.1186/s13321-023-00784-y**.
10 D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *J. Chem. Inf. Model.*, 2010, **50**(5), 742–754, DOI: **10.1021/ci100050t**.
11 A. Capecchi, D. Probst and J. L. Reymond, One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome, *J. Cheminform.*, 2020, **12**(1), 43, DOI: **10.1186/s13321-020-00445-4**.