Reaction Classification and Yield Prediction using the Differential Reaction Fingerprint DRFP

Daniel Probst,[*,†] Philippe Schwaller,[†,‡] and Jean-Louis Reymond[*,†]

[†]*Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

[‡]*IBM Research – Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*

E-mail: daniel.probst@dcb.unibe.ch; jean-louis.reymond@dcb.unibe.ch

Abstract

Predicting the nature and outcome of reactions using computational methods is a crucial tool to accelerate chemical research. The recent application of deep learning-based learned fingerprints to reaction classification and reaction yield prediction has shown an impressive increase in performance compared to previous methods such as DFT- and structure-based fingerprints. However, learned fingerprints require large training data sets, are inherently biased, and are based on complex deep learning architectures. Here we present the differential reaction fingerprint *DRFP*. The *DRFP* algorithm takes a reaction SMILES as an input and creates a binary fingerprint based on the symmetric difference of two sets containing the circular molecular n-grams generated from the molecules listed left and right from the reaction arrow, respectively, without the need for distinguishing between reactants and reagents. We show that *DRFP* outperforms DFT-based fingerprints in reaction yield prediction and other structure-based fingerprints in reaction classification, reaching the performance of state-of-the-art learned fingerprints in both tasks while being data-independent.

Introduction

Computational methods to predict the nature and outcome of reactions are important tools to accelerate chemical research.[1–11] The nature of a reaction is well-described by its name and class, where a reaction class is defined by the general reaction-type and the participating chemical entities.[12–14] Automating the classification of reactions provides a tool for chemists to search databases and to quickly evaluate and optimise a novel reaction based on the nature of similar reactions. An important outcome of a chemical reaction is its yield, the percentage of successfully converted reactants into the desired product. Computational methods for predicting such yields are highly valuable in synthesis-planning, where high yields are of paramount importance–especially in multi-step reactions. Earlier work used physics-based descriptors or structure-based molecular fingerprints to classify chemical reactions or predict reaction yields.[6,15,16] However, computational complexity and inherent biases have introduced seemingly insurmountable challenges to these approaches. Recently, with the availability of large data sets and the resurgence of artificial neural networks (ANN), deep learning-based learned fingerprints have been introduced as an alternative to earlier methods, outperforming them by considerable margins.[11] However, these approaches come with several drawbacks as well. Training a learned fingerprint requires large amounts of data of acceptable quality and must be retrained when new data becomes available, posing a challenge to accessibility and reproducibility. Due to the nature of the ANNs, learned fingerprints are challenging to interpret, as they, for example, require a careful analysis of attention weights.[11] Finally, the training and evaluation of the models require specialised hard- and software to become computationally tractable.

Here we present the differential reaction fingerprint (*DRFP*) for reaction search and categorization as well as yield prediction. The reaction fingerprint *DRFP* borrows the creation of circular substructures from a molecule and the subsequent hashing of their SMILES representations from the chemical fingerprints ECFP and MHFP, respectively (see Figure 1 and Molecular n-grams).[18,19] However, as reaction SMILES consist of multiple molecules in the form `REACTANTS>AGENTS>PRODUCTS`, three additional steps have to be introduced: (I) The agents are added to the reactants, resulting in the representation `REACTANTS+AGENTS>>PRODUCTS`; (II) molecules on each side of the reaction representation are processed individually, resulting in two sets of SMILES R and P; (III) the symmetric difference of the two sets $S = R \triangle P$ is taken, hashed using an arbitrary hash function with a sufficiently low collision probability (BLAKE2), and then further hashed into a fix-length binary vector using $h(k) = k \mod d$, where $k \in S$, and d is the desired dimensionality of the fingerprint. Compared to the approach introduced by Schneider et al.[16], *DRFP* does not apply weights based on atom-mapping to differentiate between reactants and agents, does not require the calculation of molecular properties for the agents, and does not apply arithmetic operations on individual molecular fingerprints, such as the atom pair fingerprint, to create a reaction fingerprint.

Given this conceptually simple fingerprint, we show that its performance, when applied to tasks mentioned above, rivals or even surpasses that of state-of-the-art methods while using minimal non-specialised computational resources and no specialised hard- or software (see Computational Resources). The fingerprint requires an unannotated, non-atom-mapped reaction SMILES as input and embeds this molecular representation from reaction SMILES space into an arbitrary low dimensional binary metric space through set operations and subsequent mod hashing. We show that a k-NN classifier trained with *DRFP* significantly outperforms those trained on existing, non-learned fingerprints and rivals or surpasses the performance of learned fingerprints without the need for supervised learning pre-classification. Furthermore, the fingerprint can act as an unbiased benchmark for new methods. Finally, we show that this method, based on a simple set operation and hashing scheme, can outperform both deep learning-based learned fingerprints and physics-based descriptors in yield prediction tasks. We make the fingerprint creation algorithm available as a pypi package (drfp). The source code and documentation are available on GitHub (`https://github.com/reymond-group/drfp`).

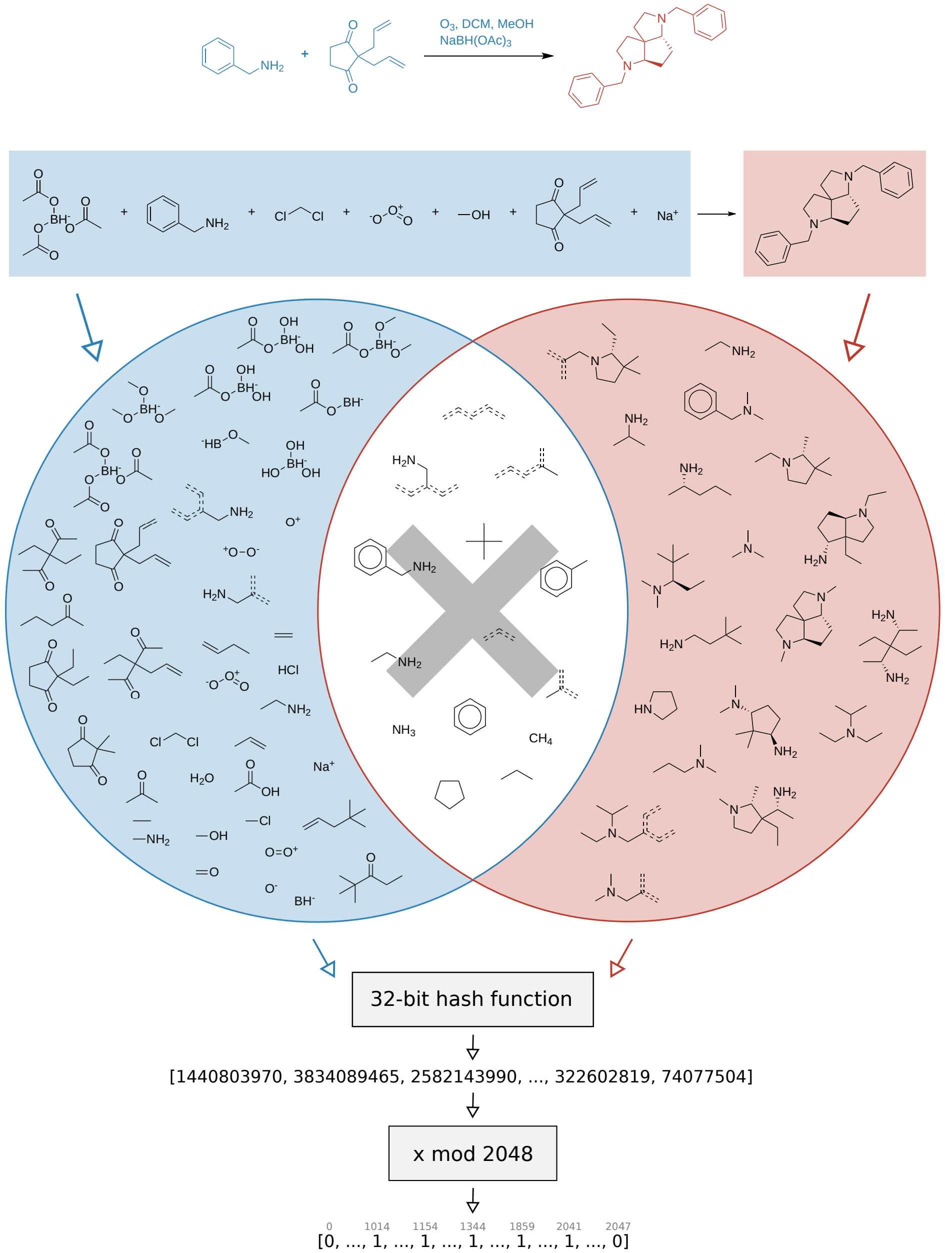


Figure 1: Encoding a reaction[17] without distinguishing between reactants and agents into an *DRFP* fingerprint is achieved by first extracting circular substructures of radius r ($r = 3$ in the above example) into two sets (blue and red circles for reactants and products, respectively). In a second step, the two sets' symmetric difference (blue and red shaded areas) is hashed using an arbitrary hash function. Finally, the resulting set is hashed into a binary vector using modular hashing.

Results and Discussion

Reaction Classification

The reaction classification was carried out using the k-nearest neighbor classifier based on faiss[20] as defined by Schwaller et al.[11]. Initially, DRFP was evaluated on the USPTO 1k TPL set using a number of different configurations, namely for radius $r \in \{2, 3, 4\}$ and dimensionality $d \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$. For all chosen radii, the accuracy increases strongly between $d = 16$ to $d = 128$, while only increasing slightly from $d = 256$ to $d = 2048$. The $r = 2$ variant performs significantly better than $r \in \{3, 4\}$ for $d \in \{16, 32\}$ (Figure 2a). This is due to fewer collisions during mod hashing resulting from fewer extracted sub-structures. Starting with $d = 256$, the $r = 3$ variant performs better than both the other variants.

Reducing the training set to 10 and 1% of its original size, aside from a general reduction in accuracy, also leads to a better relative performance of the $r = 2$ variant across all dimensions d (Figure 2b,c). These results suggest that choosing the $r = 2$ variant might be advantageous in low data settings, and there is no value in choosing $r = 4$ over $r = 2$ or $r = 3$, independent from d and the amount of available training data. However, as the $r = 3$ variant performed best in the case of the complete training set for high d, the $r = 3$ and $d = 2048$ variant is chosen for all further benchmarks, including reaction yield predictions.

Table 1: Reaction classification accuracy on the USPTO 1k TPL data set.

USPTO 1k TPL	Classifier	Accuracy	CEN	MCC
rxnfp	5-NN	**0.989**	0.006	0.989
AP3 256	5-NN	0.295	0.242	0.292
AP3 256	MLP	0.809	0.101	0.808
DRFP	5-NN	0.917	0.041	0.917
DRFP	MLP	0.977	0.011	0.977

Evaluating the k-nearest neighbour classification benchmark on the TPL data set, *DRFP* outperforms the structure-based fingerprint AP3 256 by a factor of 3.1 and reaches 93% of

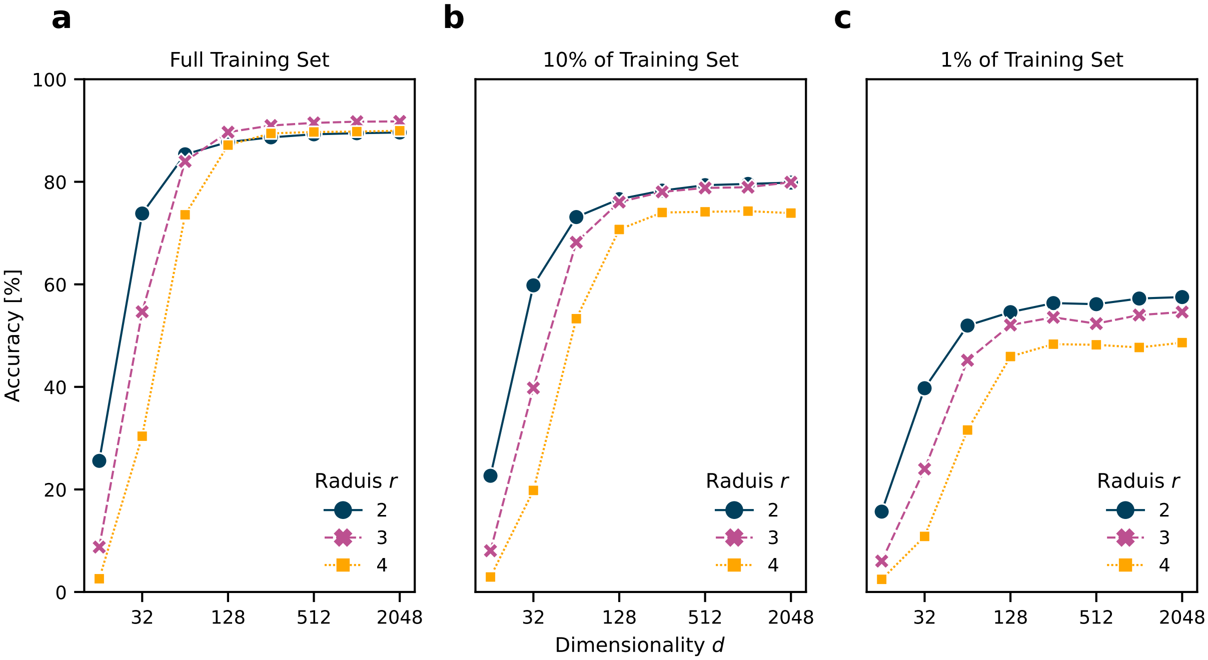


Figure 2: Accuracy of the k-nearest neighbor classification on (**a**) the entire TPL data set, (**b**) 10% of the data set, and (**c**) 1% of the data set using *DRFP* fingerprints for dimensionality $d \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$ and $r \in \{2, 3, 4\}$. The accuracy starts to plateau with $d = 128$ independently from the amount of training data. However, a lower r increases the accuracy in low data settings and when a low dimensionality d is chosen due to increased generality and fewer collisions, respectively.

the performance of the learned fingerprint rxnfp. Replacing the k-nearest neighbour classifier with a simple multilayer perceptron (MLP), *DRFP* reaches 99% of the performance of rxnfp. This result suggests that conceptual complexity, including learning, can be factored out of fingerprint creation, moving it instead to the classification task with a minor impact on classification performance. A non-learned fingerprint has the advantages of reducing bias and increasing the interpretability of results as each feature can be mapped to one or more molecular substructures.

Reaction Yield Prediction

Comparing the yield prediction performance of *DRFP* to that of learned and physical descriptor-based fingerprints shows that this simple fingerprint is competitive, as it demonstrates consistent performance on all test sets. Averaging the 11 tests shown in Table 2, *DRFP* outperforms Yield-BERT, an augmented version of Yield-BERT, as well as a DFT-based method, in a yield prediction task on a data set of Buchwald Hartwig reactions. It also outperforms rxnfp in yield prediction of USPTO reaction data and a data set of Suzuki Miyaura reactions (Tables 3 and 4).

Table 2: R^2 of yield prediction on Buchwald Hartwig reactions.

R^2	DFT[6]	Yield-BERT[10]	Yield-BERT (aug.)[21]	DRFP (xgboost)
rand 70/30	0.92	0.95 ± 0.005	**0.97 ± 0.003**	0.95 ± 0.005
rand 50/50	0.9	0.92 ± 0.01	**0.95 ± 0.01**	0.93 ± 0.01
rand 30/70	0.85	0.88 ± 0.01	**0.92 ± 0.01**	0.89 ± 0.01
rand 20/80	0.81	0.86 ± 0.01	**0.89 ± 0.01**	0.87 ± 0.01
rand 10/90	0.77	0.79 ± 0.02	**0.81 ± 0.02**	0.80 ± 0.02
rand 5/95	0.68	0.61 ± 0.04	**0.74 ± 0.03**	0.73 ± 0.02
rand 2.5/97.5	0.59	0.45 ± 0.05	**0.61 ± 0.04**	**0.61 ± 0.04**
test 1	0.8	**0.84 ± 0.01**	0.8 ± 0.01	0.81 ± 0.01
test 2	0.77	0.84 ± 0.03	**0.88 ± 0.02**	0.83 ± 0.003
test 3	0.64	**0.75 ± 0.04**	0.56 ± 0.08	0.71 ± 0.001
test 4	**0.54**	0.49 ± 0.05	0.43 ± 0.04	0.49 ± 0.004
avg. 1-4	0.69	**0.73**	0.58 ± 0.33	0.71 ± 0.16
avg. overall	0.75 ± 0.12	0.76 ± 0.17	0.778 ± 0.18	**0.784 ± 0.14**

In order to predict reaction yields using *DRFP*, gradient boosting with early stopping was chosen as a machine learning technique. 10% of each training split was set aside and used to evaluate for early stopping. Hyperparameter optimisation was performed on five random splits (70/30). The resulting performance (R^2) is then compared to the density functional theory (DFT) based fingerprint with a random forest regressor by Ahneman et al.[6], Yield-BERT, an extension of the learned rxnfp fingerprint with a regression layer, and an augmented variant of the latter (Table 2). The data set used is a collection of 3,955 Pd-catalysed Buchwald–Hartwig C-N cross-coupling reactions from a high throughput experiment by Ahneman et al.[6]. For this data set, 11 splits were defined; seven splits where the relative size of the training set was decreased from 70 to 2.5% and four out-of-sample splits based on isoxazole additives. *DRFP* performs better on the random splits than the DFT-based fingerprint with random forests and Yield-BERT but is outperformed by the augmented Yield-BERT by a narrow margin. In the out-of-sample splits, *DRFP* performs better than the augmented version of Yield-BERT and the DFT-based method, yet the non-augmented Yield-BERT performs slightly better. When averaging over all 11 tests, *DRFP* performs best.

Table 3: R^2 of yield prediction on Suzuki Miyaura reactions.

R^2	Yield-BERT	DRFP (gradient boost)
avg.	0.81 (± 0.01)	**0.85** (± 0.01)

The performance of *DRFP* was further tested on a data set containing Suzuki-Miyaura reactions and the USPTO reaction data set. In both cases, *DRFP* performed better than Yield-BERT. However, similar to the Buchwald-Hartwig reaction data, the difference between the two approaches is relatively small. Both methods perform better on reactions with a sub-gram scale yield.

Table 4: The R^2 of yield prediction on the USPTO data set that has been divided into gram scale and sub-gram scale yield subsets.

USPTO Random Split	rxnfp	*DRFP*
Gram Scale	0.117	**0.13**
Sub-Gram Scale	0.195	**0.197**

Overall, *DRFP* reaches a compelling performance in yield prediction using a gradient boosting regressor that does not require hyperparameter tuning between different sets.

Conclusion

We have introduced a reaction fingerprint encoding scheme, *DRFP*, based on a simple 4-step process comprised of extracting circular n-grams, XORing, hashing, and folding. *DRFP* is capable of reaching state-of-the-art performance without extending the use of machine learning models from classification or regression tasks to the fingerprint creation task. The fingerprint creation algorithm is available as a pypi package (drfp). Source code and documentation are available on GitHub (`https://github.com/reymond-group/drfp`).

Methods

Computational Resources

We ran all of the training runs as well as the evaluations of the models on a DELL XPS Laptop with 16 GB of main memory, no dedicated GPU, and an 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz CPU.

Molecular n-grams

Molecular n-grams are generated from SMILES using the RDKit library. Given a radius r, we iterate over the heavy atoms in an input molecule and extract sub-SMILES centred on each atom with radii 0 to r, where a radius of 0 is the single central atom. In addition, rings from the SSSR (smallest set of smallest rings) are extracted as well. Compared to the atom pair-based approach by Schneider et al.[16], the n-grams-based fingerprint also captures stereochemistry.

Gradient Boosting

For regression by gradient boosting, we used the Python library xgboost. Hyperparameter tuning was carried out on the rand 70/30 set of the Buchwald-Hartwig reaction data set.

Algorithm 1 Generating molecular n-grams

```
$shingling \leftarrow$ empty set
for $atom$ in $molecule$ do
    for $radius = 0, \ldots, r$ do
        Add substructure with $radius$ rooted at $atom$ to $shingling$ as SMILES
    end for
end for
for $ring$ in $sssr(molecule)$ do
    Add substructure of $ring$ to $shingling$ as SMILES
end for
```

We applied the same hyperparameter values (`n_estimators=999999`, `learning_rate=0.01`, `max_depth=15`, `min_child_weight=8`, `colsample_bytree=0.2125`, `subsample=1`) in all uses of xgboost. For each test, 10% of the training data were randomly selected as the validation set an removed from the training set. The validation data sets were used as evaluation sets for early stopping (20 rounds for all data sets with the exception of the USPTO, data for which 10 rounds were used to speed up the calculation).

k-Nearest Neighbours Classifier

The k-Nearest Neighbour classifier was implemented according to Schwaller et al.[11] using faiss with $k = 5$.

Multilayer Perceptron Classifier

In addition to *DRFP* + 5-NN classifier, *DRFP* + multilayer perceptron (MLP) classifier was applied to the USPTO 1k TPL data set. The MLP was implemented using Tensorflow 2.4.1 and consists of an input layer the size of the input vector (2,048), a dense hidden layer of size 1,664 and a tanh activation function, and a dense output layer with a softmax activation function. The loss function was sparse categorical cross-entropy. Adam was used as an optimiser. The model was trained over 10 epochs with a batch size of 64 on a CPU.

For the evaluation of AP3 256, the number of units in the hidden layer was changed to 1024, and the model was trained for 100 epochs.

Acknowledgement

This work was supported financially by the Swiss National Science Foundation, NCCR TransCure.

References

(1) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–76.

(2) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Sci. Rep.* **2017**, *7*, 3582.

(3) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **2017**, *3*, 1237–1245.

(4) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

(5) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.

(6) 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* **2018**, *360*, 186–190.

(7) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **2020**, *11*, 3601.

(8) 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. *React. Chem. Eng.* **2020**, *5*, 1963–1972.

(9) 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**, *7*, 2269–2277.

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

(11) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell* **2021**, *3*, 144–152.

(12) Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. Analysis of the reactions used for the preparation of drug candidate molecules. *ChemInform* **2006**, *37*.

(13) Roughley, S. D.; Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011**, *54*, 3451–3479.

(14) Ontologies, R. S. C. rxno.

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

(16) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39–53.

(17) Meier, K.; Arús-Pous, J.; Reymond, J.-L. A potent and selective Janus kinase inhibitor with a chiral 3D-shaped triquinazine ring system from chemical space. *Angew. Chem. Weinheim Bergstr. Ger.* **2021**, *133*, 2102–2105.

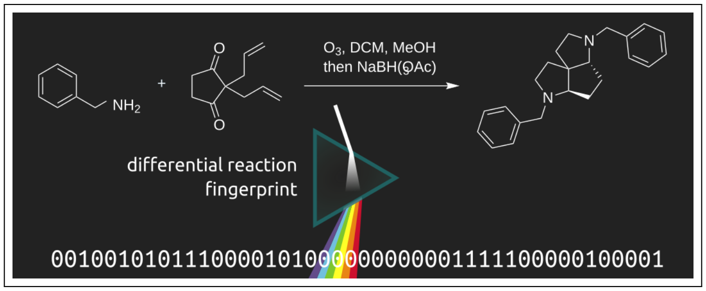
(18) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(19) Probst, D.; Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminform.* **2018**, *10*, 66.

(20) Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734* **2017**,

(21) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *ChemRxiv preprint, 10.26434/chemrxiv.13286741.v1* **2020**,

Graphical TOC Entry


NH_2
+
O
O
O_3, DCM, MeOH
then NaBH(OAc)$_3$
N
N
differential reaction fingerprint
0010010101110000101000000000001111100000100001