

# Mapping the Space of Chemical Reactions using Attention-Based Neural Networks

Philippe Schwaller,<sup>\*,†,‡</sup> Daniel Probst,<sup>‡</sup> Alain C. Vaucher,<sup>†</sup> Vishnu H. Nair,<sup>†</sup>  
David Kreutter,<sup>‡</sup> Teodoro Laino,<sup>†</sup> and Jean-Louis Reymond<sup>‡</sup>

<sup>†</sup>*IBM Research – Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*

<sup>‡</sup>*Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012  
Bern, Switzerland*

E-mail: [phs@zurich.ibm.com](mailto:phs@zurich.ibm.com)

## Abstract

Organic reactions are usually assigned to classes grouping reactions with similar reagents and mechanisms. Reaction classes facilitate communication of complex concepts and efficient navigation through chemical reaction space. However, the classification process is a tedious task, requiring the identification of the corresponding reaction class template via annotation of the number of molecules in the reactions, the reaction center and the distinction between reactants and reagents. In this work, we show that transformer-based models can infer reaction classes from non-annotated, simple text-based representations of chemical reactions. Our best model reaches a classification accuracy of 98.2%. We also show that the learned representations can be used as reaction fingerprints which capture fine-grained differences between reaction classes better than traditional reaction fingerprints. The unprecedented insights into chemical reaction space enabled by our learned fingerprints is illustrated by an interactive reaction atlas providing visual clustering and similarity searching.

In the last decade, computer-based systems<sup>[1-3]</sup> became an important asset available to chemists, with deep learning methods standing out, not only for reaction prediction tasks,<sup>[4-6]</sup> but also for synthesis route planning<sup>[7-9]</sup> and synthesis procedures to actions conversions.<sup>[10]</sup> Among the few approaches, natural language processing methods<sup>[11,12]</sup> applied to Simplified molecular-input line-entry system (SMILES)<sup>[13,14]</sup> and other text-based representation of molecules and reactions are particularly effective in the chemical domain. Recently, Schwaller et al.<sup>[15]</sup> demonstrated that neural networks were able to capture the atom rearrangements in chemical reactions without supervision.

Name reactions play a crucial role in the language of organic chemists. They represent an efficient way to communicate what a chemical reaction does or how it works in terms of atomic rearrangements. For this reason, those name reactions are currently used to navigate large databases of reactions, to retrieve similar members of the same reaction class to help chemists to analyse and infer optimal reaction conditions. Today, several hundreds of name reactions exist in the RXNO ontology.<sup>[16]</sup> Often their name honors the persons who discovered that chemical reaction or who refined an already known transformation, substantially raising its popularity. An example is the Friedel-Crafts reaction, named after Charles Friedel and James Mason Crafts, who discovered the catalytic effect of aluminum chloride in electrophilic substitutions. Name reactions can also be named after the reaction type, using the initials or referring to structural features.

The demand for robust algorithms to categorise chemical reactions is high because knowledge of the class of a reaction has a great value for expert chemists, for example to assess the quality of the reaction prediction.<sup>[17]</sup> The current state-of-the-art in reaction classification is represented by commercially available tools,<sup>[18,19]</sup> which classify reactions based on a library of expert-written rules. These tools typically make use of SMIRKS,<sup>[20]</sup> a language to describe transformations in the SMILES format.<sup>[14,21]</sup> On the contrary, classifiers based on machine learning have the potential to increase the robustness to noise in the reaction equations and to avoid the explicit formulation of rules.

Early work in the 90s used self-organising neural networks to map organic reactions and investigate similarities between them.<sup>[22-24]</sup> More recently, Schneider et al.<sup>[25]</sup> developed a reaction classifiers based on traditional reaction fingerprints. Their best performing fingerprint combines a products-reactants difference fingerprint with molecular features calculated on the reagents, tested on a limited set of 50 reaction classes. The difference fingerprint developed by Schneider et al.<sup>[25]</sup> is currently one of the most frequently used hand-crafted fingerprint. It has been for example successfully applied to reaction conditions predictions,<sup>[26]</sup> where the reagents were not taken into account for the reaction description. Ghiandoni et al.<sup>[27]</sup> introduced an alternative hierarchical classification scheme and random forest classifier for reaction classification. Their algorithm outputs a confidence score by means of conformal prediction. The fingerprints of Schneider et al.<sup>[25]</sup> and Ghiandoni et al.<sup>[27]</sup> both require a reactants-reagents role separation,<sup>[28]</sup> which is often ambiguous and thus limits their applicability.

Traditionally, reaction fingerprints were hand-crafted using the reaction center or a combination of the reactant, reagent and product fingerprints. ChemAxon,<sup>[29]</sup> for instance, provides eight types of such reaction fingerprints. Based on the differentiable molecule fingerprint by Duvenaud et al.,<sup>[30]</sup> the first example of a learned reaction fingerprint was presented by Wei et al.<sup>[31]</sup> and used to predict chemical reactions. Unfortunately, their fingerprint was restricted to a fixed reaction scheme consisting of two reactants and one reagent, and hence, only working for reactions conform with that scheme. Similarly, the multiple fingerprint features by Sandfort et al.<sup>[32]</sup> are made by concatenating multiple fingerprints for a fixed number of molecules.

In the first part of our work, we predict chemical reaction classes using attention-based neural networks belonging to the family of transformers.<sup>[11,12]</sup> Instead of relying on the formulation of specific rules and on the need to have every reaction properly atom-mapped, our deep learning models learn the atomic motifs that differentiate reactions belonging to different classes from raw reaction SMILES without reactant-reagent role annotations. The

transformer-based sequence-2-sequence (seq-2-seq) model<sup>[11]</sup> matched the ground-truth classification with an accuracy of 95.2% and the Bidirectional Encoder Representations from Transformers (BERT) classifier<sup>[12]</sup> with 98.2%. The mismatches are mainly related to unrecognised reactions, some of which are correctly classified by our model. Moreover, both architectures show very high robustness towards errors in the SMILES representation. We report cases where, despite an error in the converted molecules, our model was able to classify correctly the reaction that was originally described by chemists in the patent procedure text. We analyse the encoder-decoder attention of the seq-2-seq model and the self-attention of the BERT model and observe that atoms involved in the reaction center, as well as reagents specific to the reaction class, have larger attention weights.

In the second part, we demonstrate that the representations learned by the BERT models, unsupervised and supervised, can be used as reaction fingerprints. The reaction fingerprints we introduce are independent of the number of molecules taking part in a reaction. The BERT models trained on chemical reactions convert any reaction SMILES into a vector without requiring atom-mapping or a reactant-reagent separation. Therefore our reaction fingerprints are universally applicable to any reaction database. Based on those reaction fingerprints and TMAP,<sup>[33]</sup> a method to visualise high-dimensional spaces as tree-like graphs, we were able to map the chemical reaction space and show in our reaction atlases nearly perfect clustering according to the reaction classes. Moreover, our fingerprints enable efficient similarity searches in the chemical reaction space. On a imbalanced data set our fingerprints and classifiers reach an overall accuracy of more than 98%, compared to 41 % when using a traditional reaction fingerprint. The ability to accurately classify chemical reactions and represent them as fingerprints, which can capture fine-grained differences between chemical reactions improves how chemical reactions can be accessed by machines and humans alike. Hence, our work has the potential to unlock new insights in the field of organic synthesis.

# Results and Discussion

## Reaction classification

### Classification results

We used a labeled set of chemical reactions as ground truth to train two transformer-based deep learning models as architecture.<sup>[11][12]</sup> The ground truth data is composed of chemical transformations represented as SMILES, and its labeling (classification) was taken from the strongly imbalanced Pistachio data set,<sup>[34]</sup> which uses NameRXN for the reaction classification.<sup>[18]</sup> We analysed the classification performance of our models on the test set, which contained 132k reactions belonging to 792 different classes. A summary of the results can be found in Table [1](#). The transformer enc2-dec1 model matched the ground truth classification with an accuracy of 95.2%. The Reaction BERT classifier predicted the correct name reaction with an accuracy of 98.2%, therefore achieving significantly better results than with the seq-2-seq approach. As a comparison to previous work,<sup>[25]</sup> we computed transformation fingerprint AP3 (folded) + featureFP on the Pistachio data and used a 5-NearestNeighbour (5-NN) classifier<sup>[35]</sup> to classify the test set reactions. Even though for this fingerprint we separated the reactants and reagents using RDKit,<sup>[36]</sup> the classifier only achieved an overall accuracy of 41.0%. The traditional fingerprint was not be able to represent the fine-grained differences between the reaction classes. The “Unrecognised”, “Carboxylic acid + amine condensation”, “Amide Schotten-Baumann” and “N-Boc deprotection” classes contained the most false positives.

In contrast, our BERT classifier without reactant-reagent separation was also the best performing model, when looking at the confusion entropy of a confusion matrix (CEN)<sup>[37]</sup> and overall Matthews correlation coefficient (MCC).<sup>[38][39]</sup>

To show that the worse performance of the traditional reaction fingerprint did not stem from the choice of the 5-NN classifier, we took the embeddings of the pretrained (*rxnfp (pretrained)*) and finetuned BERT (*rxnfp*) as inputs to the 5-NN classifier, classified the test

Table 1: Classification results. The lower the confusion entropy of a confusion matrix and the higher the Matthews correlation coefficient the better. The traditional fingerprint is a AP3 256 (folded) + agents features developed by Schneider et al.<sup>[25]</sup>

Model	Accuracy	CEN	MCC
Traditional fp <sup>[25]</sup> + 5-NN classifier	0.410	0.365	0.305
Transformer enc2-dec1	0.952	0.039	0.946
BERT classifier	<b>0.982</b>	<b>0.014</b>	<b>0.980</b>
<i>rxnfp</i> ( <i>pretrained</i> ) + 5-NN classifier	0.819	0.121	0.797
<i>rxnfp</i> + 5-NN classifier	<b>0.989</b>	<b>0.010</b>	<b>0.988</b>

set reactions, and computed the scores. As expected, the results for *rxnfp*, which corresponds to the input of the classifier layer in the BERT classifier, perfectly matched the scores of the BERT classifier. An elaborate description of both rxnfps is presented in the section on data-driven reaction fingerprints below. A comparison of our data-driven approach to traditional fingerprints on a balanced data set of 50k reactions can be found in the supplementary information. Even using as little as 10k training reactions from 50 different classes the fine-tuned embeddings are able to outperform traditional fingerprints by increasing precision, recall and F1-score from 0.97 to 0.99.

### Analysis of incorrect predictions

We analysed the BERT classifier in more detail and compared it to the seq-2-seq transformer model. First, we identified different types of incorrect predictions by the transformer BERT classifier model, which are summarised in Table 2. Most errors are related to the “Unrecognised” class of the RXNO ontology. The most frequent error type is the prediction of a reaction class for a reaction classified as “Unrecognised” (47.9% of all incorrect predictions), and the second most frequent error type is predicting “Unrecognised” when a class should be predicted (22.8%). The third most frequent error is predicting the incorrect name reaction (third number of the class string, 17.5%). The remaining errors are predicting an incorrect superclass (first number of the class string, 8.3%) and predicting an incorrect category (second number of the class string, 3.5%).

Table 2: Types of incorrect predictions of the BERT model on the test set consisting of a total of 132213 reactions.

	Count	Percentage
Correctly predicted	129892	98.24%
Model predicts name reaction instead of “Unrecognised”	1111	0.84%
Model predicts “Unrecognised” instead of name reaction	529	0.40%
Incorrect name rxn	407	0.31%
Incorrect superclass	193	0.15%
Incorrect category	81	0.06%

In Table 3, we show the reaction classes for which our model makes incorrect predictions most frequently. Due to statistical sampling, we restricted this analysis to reactions with at least 20 occurrences in the test set. For 12 out of 15 of these reaction classes, the most common error source is the failure to assign a reaction class, thus predicting “Unrecognised”. Among the other most common failures, there is the “Bouveault-Blanc reduction”, where an ester is reduced to a primary alcohol. Hence, it is very similar to the Ester to alcohol reduction class, with which it is most mistaken. The difference lies in the specific precursors used in the “Bouveault-Blanc reduction”, such as sodium and ethanol or methanol. The “1,3-Dioxane synthesis” reaction class has an overall accuracy of 88.9%. However, there are some reactions mistaken for “Dioxolane synthesis”, for which the newly formed heterocycle in the product has an additional carbon atom.

Although the large number of “Unrecognised” reactions in Pistachio makes an extensive analysis difficult, the inspection of a few dozen cases provides interesting insights. Part of the “Unrecognised” reactions should actually belong to a name reaction. The data-driven approach can be more robust than rule-based models and assign the correct reaction class. For example, in contrast to rule-based models, data-driven ones are often able to capture the reaction class despite changes in the tautomeric state between precursors and product. Another part of those “Unrecognised” reactions belongs to the category for which multiple transformations occur simultaneously. In this case, the reaction cannot be classified into a single name reaction, and our model predicts one of the corresponding reactions. Such exam-

Table 3: Worst-predicted reaction classes with more than 20 occurrences in the test set for the BERT classifier.

Reaction class	Accuracy [%]	Most frequent incorrectly predicted class
1.1.2 Menshutkin reaction	62.1	0.0 Unrecognised
3.9.41 Decarboxylative coupling	72.1	0.0 Unrecognised
9.7.140 Defluorination	75.6	0.0 Unrecognised
7.4.2 Bouveault-Blanc reduction	76.4	7.4.1 Ester to alcohol reduction
11.1 Chiral separation	83.6	0.0 Unrecognised
8.8.11 Hydroxylation	83.7	0.0 Unrecognised
4.3.11 Thiazoline synthesis	85.7	0.0 Unrecognised
3.9.12 Olefin metathesis	85.8	0.0 Unrecognised
2.5.5 Nitrile + amine reaction	86.0	0.0 Unrecognised
9.7.42 Chloro to fluoro	86.4	0.0 Unrecognised
10.4.2 Methylation	88.9	0.0 Unrecognised
4.2.39 1,3-Dioxane synthesis	88.9	4.2.20 Dioxolane synthesis
4.1.53 1,2,4-Triazole synthesis	90.0	0.0 Unrecognised
1.1.6 Chloro Menshutkin reaction	90.6	0.0 Unrecognised
5.1.2 N-Cbz protection	90.9	2.1.1 Amide Schotten-Baumann

ples can be found in deprotection reactions where more than one distinct functional group is removed. Another interesting aspect comes from molecules that are incorrectly parsed in Pistachio. If the SMILES string of a molecule involved in the reaction was incorrectly derived from the name, rule-based approaches fail to recognise the atomic rearrangements and thus to classify the reaction. For minor parsing errors, our model shows its potential, recognizing the correct transformation in several instances.

The accuracy of the enc2-dec1 seq-2-seq model was 3% worse than the one of the BERT classifier. When comparing the predictions of the two models, we observe that most of the differences are related to the "Unrecognised" class. 3511 out of 5108 reactions that were correctly predicted by the BERT classifier but not the seq-2-seq model belong to the "Unrecognised" class. Moreover, the three classes containing the most examples of reaction classes predicted correctly by the BERT classifier but not by the seq-2-seq model were "Carboxylic acid + amine condensation" (2.1.2), "Methylation" (10.4.2) and "Williamson ether synthesis" (1.7.9) reactions with 90, 61 and 37 examples respectively. In contrast, the

seq-2-seq model was able to classify 474 reactions as “Unrecognised”, which were classified as recognised name reactions by the BERT model. Besides the “Unrecognised” reactions, the three reaction types with the most examples that were correctly predicted by the seq-2-seq model but not by the BERT classifier were “Bouveault-Blanc reduction” (7.4.2), “Ester to alcohol reduction” (7.4.1) reactions with 33 and 15 examples respectively. The seq-2-seq seems to capture the subtle difference between the two distinct “Ester to alcohol” (7.4) classes better.

## Visualisation of Attention Weights

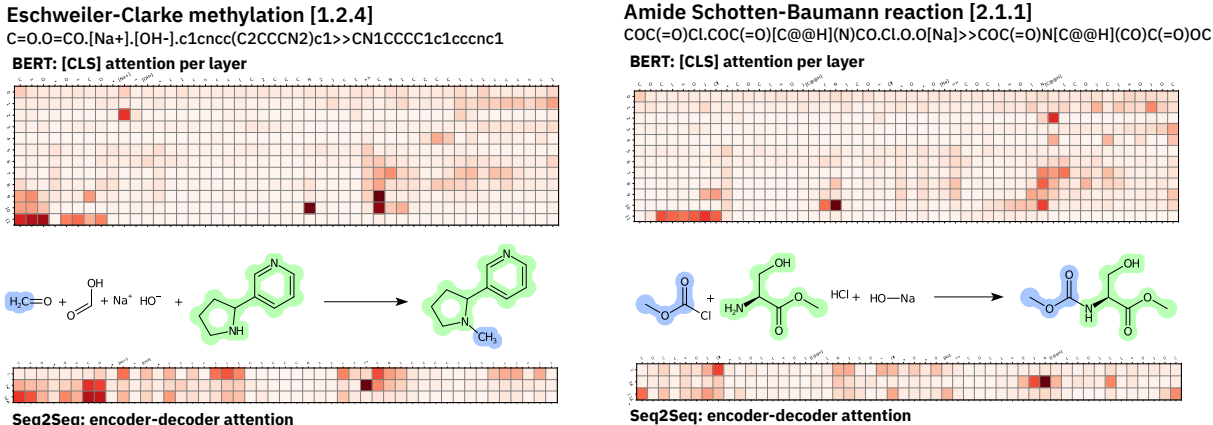


Figure 1: Layer-wise [CLS] token attention for the BERT classifier and encoder-decoder attention for the enc2-dec1 transformer model. The horizontal axis contains the SMILES tokens of the input reaction. The darker the token the more attention a specific token had in that particular layer or output step. The coloring on the reaction depictions made with CDK depict<sup>40</sup> shows the mapping from precursors to product in the ground truth.

Figure 1 shows the layer-wise [CLS] token attention of the BERT classifier (above the reaction) and the encoder-decoder attention of the seq-2-seq model (below the reaction) for two different chemical transformations. We note that the larger weights are associated with the atoms that are part of the reaction center or precursors specific to the reaction class. Just like a human expects to see a certain group of atoms based on the name reaction, for the seq-2-seq model, the decoder learned to focus on the atoms involved in the rearrangement to classify reactions. For the BERT classifier, the initial layers have weak attention on all

the reaction tokens, middle layers tend to attend either the product or on the precursors, and the last layers focus on the reaction center and the precursors that are important for the classification.

## Mapping Chemical Reaction Space

### Data-driven Reaction Fingerprints

Molecular fingerprints are widely used to screen molecules with similar properties or map chemical space.<sup>[41]</sup> Our reaction BERT models does not only perform best on the classification task but also allows chemists to generate vectorial representations of chemical reactions. Here we introduce reaction fingerprints based on the embeddings computed by BERT<sup>[12]</sup> models, which can be applied to any reaction data set, as they do not require a reactant-reagent split or a fixed number of precursors. The pretraining of the BERT model works by masking and predicting individual tokens in the reaction SMILES. As the prepended [CLS] token is never masked, the model is always able to attend the representation of this token to recover the masked tokens. The intuition is that the model uses the [CLS] token to embed a global description of the reaction. Before the fine-tuning the [CLS] token embeddings are learned purely by self-supervision. We refer to this fingerprint as *rxnfp* (*pretrained*). For the supervised fine-tuning, the embeddings of the [CLS] token are then taken as input for a one layer classification head and further refined. We refer to the fingerprint fine-tuned on the Pistachio training set as *rxnfp*. In our case, the [CLS] token embedding is a vector of size 256, corresponding to the hidden size of the BERT model. During the supervised classification task, the model has to focus on the reaction center and certain precursors that are specific to the individual name reactions. For instance, the Eschweiler-Clarke methylation (1.2.4) is a methylation reaction that can be distinguished from other methylation reactions as its precursors contain formaldehyde and formic acid (see Figure [1](#)). Another example are Suzuki-type coupling reactions, where the “-type” suffix means that the metal catalyst is missing but the described reaction would correspond to a Suzuki coupling reaction.

## Reaction Atlases

In Figure 2, we show an annotated version of a reaction atlas made with the embeddings of a BERT classifier fine-tuned for three epochs. The colors correspond to the 12 superclasses found in the data set. The individual classes are almost perfectly clustered. It is worth noting that the sub-trees in the TMAP group closely related reaction classes. For instance, in the upper left, one sub-tree contains all "Formylation"-related reactions, Weinreb reactions are clustered in a branch in the lower left and Suzuki-type reactions are sharing the same branch as the corresponding Suzuki reactions. The unannotated reaction atlas was made using the fingerprints computed from a pretrained reaction BERT model without classification fine-tuning. Surprisingly, applying a purely unsupervised masked language modeling training the model was already able to extract features relevant for reaction classification and some clustering can be observed in the figure.

An interactive reaction TMAP<sup>33</sup> visualising the public Schneider 50k<sup>25</sup> data set using the *rxnfp* (10k) embeddings and highlighting different precursor and product properties can be found on [https://rxn4chemistry.github.io/rxnfp//tmaps/tmap\\_ft\\_10k.html](https://rxn4chemistry.github.io/rxnfp//tmaps/tmap_ft_10k.html).

## Reaction search

One of the primary use cases of reaction fingerprints is the search for similar reactions in a database. An atom-mapping independent reaction fingerprint is extremely powerful, as it unlocks the possibility of reaction retrieval without the need of knowing the reaction center. For instance, when a black box model like a forward reaction prediction model<sup>6</sup> or a retrosynthesis model<sup>9</sup> predict a reaction, the most similar reactions from the training set of those models could be retrieved. Such retrieval of similar reactions does not only increase the explainability of deep learning models but allows chemists to access the metadata (including yield and reaction conditions) of the closest reactions if available.

In Figure 3 the three approximate nearest neighbors of the BERT classifier fingerprint can be found for four test set reactions from four distinct reaction classes. Based on the

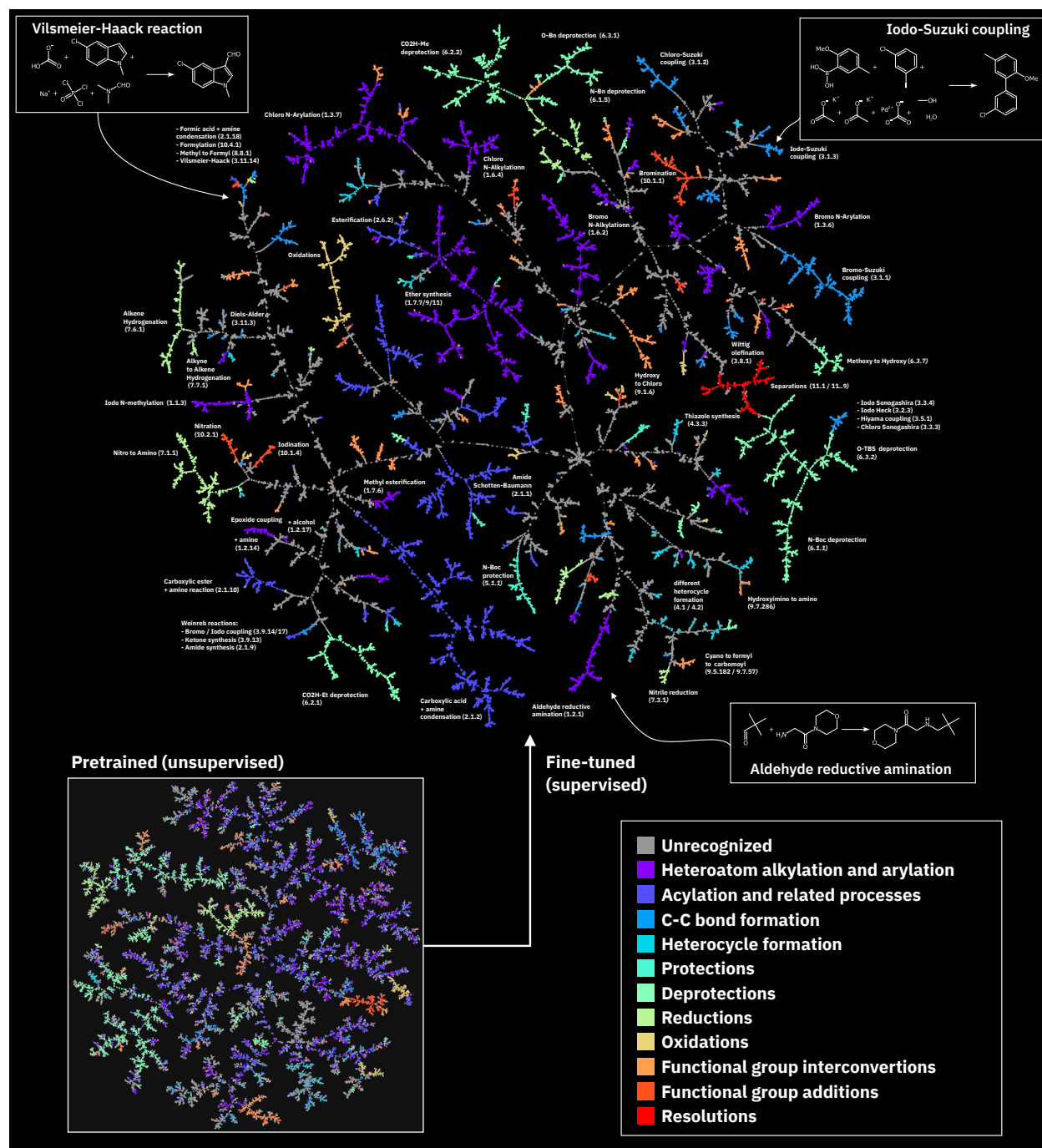


Figure 2: Top: Annotated reaction atlas made from *rxnfp*. Bottom: reaction atlas made from *rxnfp* (pretrained). The different fingerprints of the test set reactions are visualised using a TMAP algorithm<sup>[33]</sup> and the Faerun visualisation library.<sup>[42]</sup> The minhashed using a weighted hashing scheme to make them compatible with the LSH forest.

LSH forest from the TMAP module developed by Probst and Reymond,<sup>[33]</sup> the search on the training set containing 2.4M reactions was performed within milliseconds using unoptimised

python code on a MacBook Pro (Processor: 2.7 GHz Intel Core i7, Memory: 16 GB 2133 MHz LPDD). In all searches, the nearest neighbors corresponded to the same class as the query reaction. The similarities between the query reaction and the retrieved nearest neighbors are clearly visible even for non-experts. The reactions share similar if not the same precursors and the products show similar features. One of the great advantages of this reaction search method is that it only requires a reaction smiles as input.

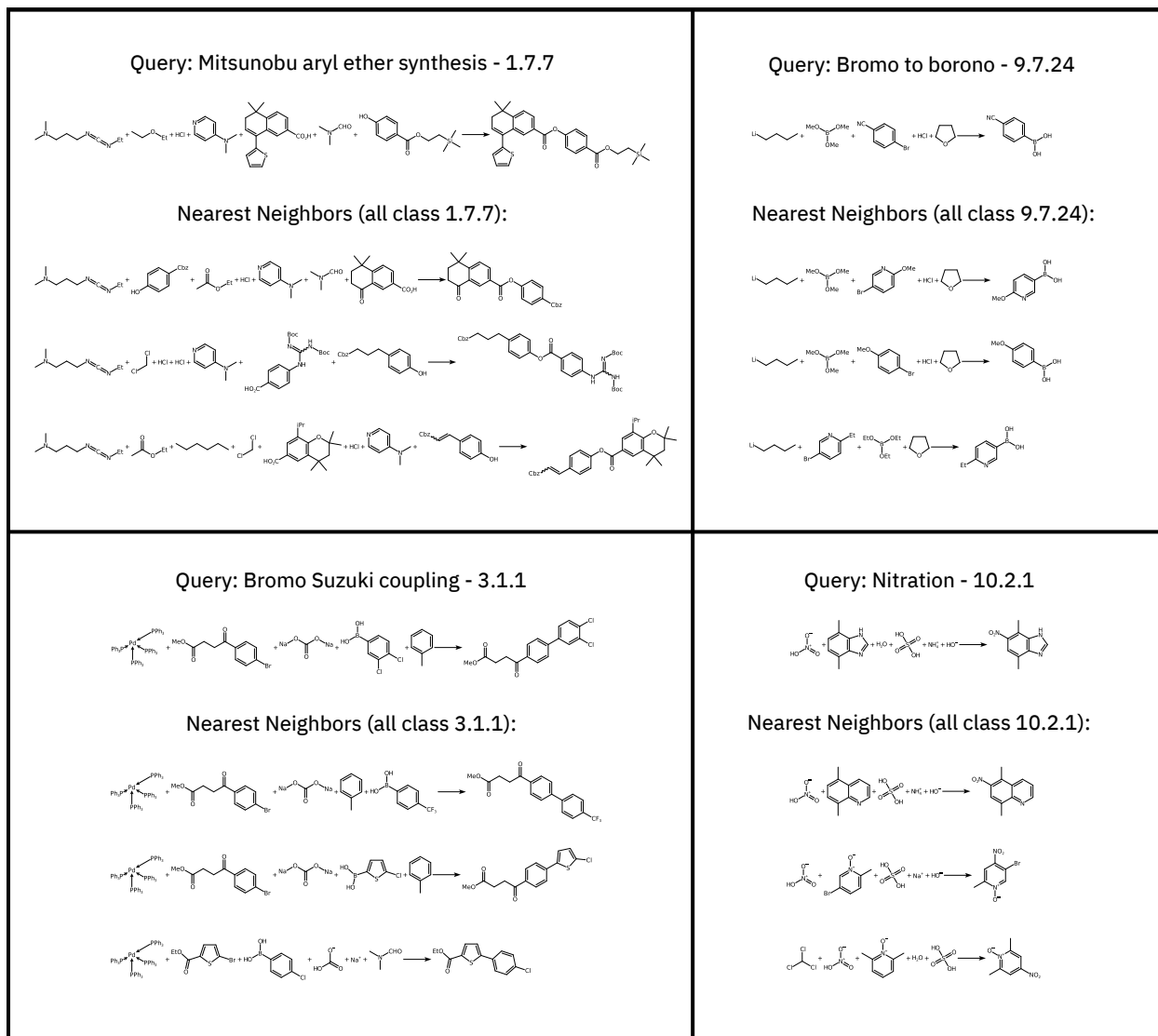


Figure 3: Four examples of reaction SMILES queries, retrieving the three nearest neighbors in the LSHforest<sup>33</sup> of the training set containing 2.4M reactions. All the retrieved reactions belong to the same reaction class as the query reaction and show similar precursors.

To investigate the robustness of our BERT classifier embeddings we removed three classes

from the fine-tuning training set (Number of removed reactions: ‘1.6.4 - Chloro N-alkylation’: 24109, ‘3.9.17 - Weinreb Iodo coupling’: 225, ‘9.7.73 - Hydroxy to azido’: 1526) and fine-tuned another BERT classifier. After 5 epochs, we generated the embeddings for the test set reactions belonging to the three removed classes. While for the “Chloro N-alkylation” and the “Hydroxy to azido” class the most common prediction was “Unrecognised”, all the predictions of the BERT model trained without the removed classes for the “Weinreb Iodo coupling” were “Weinreb bromo coupling” that differs just by the type of the reacting halogen atom. Interesting is also the retrieval of nearest neighbors from the original training set for the embeddings generated by the BERT model trained without the removed classes. Out of the 1370 “Chloro N-alkylation” reactions in the test set, for 1078 reactions the nearest neighbor in the initial training set (including all the reaction classes) was a “Chloro N-alkylation” reaction. For the 10 “Weinreb Iodo coupling” reactions, the nearest neighbors in the original training set were four “Weinreb Bromo coupling” and other four “Bromo Grignard + nitrile ketone synthesis” reactions, which are both closely related reaction types. There was no clear dominating reaction class in the nearest neighbors with 44 out of 76 reactions being “Unrecognised”.

## Conclusion

In this work, we focused on the data-driven classification of chemical reactions with natural language processing methods and on the use of their embedded information to design reaction fingerprints. Our transformer-based models could learn the classification schemes using a broad set of chemical reactions as ground-truth labeled with the use of commercially available reaction classification tool. With the BERT classifier, we match the rule-based classification with an accuracy of 98.2%, compared to 41% for a traditional fingerprint plus 5-nearest neighbour classifier. Our models are able to learn the atomic environment characteristic of each class and provides a rationale easily interpretable by expert chemists. The possibility

to understand the reasoning behind each classification may help the end-user chemists along the adoption process of these technologies.

We showed that the representation learned by our BERT models could be used as reaction fingerprints. Those data-driven reaction fingerprints do not only unlock the possibility to map the reaction space without knowing the reaction centers or the reactant-reagent split but also to perform nearest neighbor searches efficiently on reaction data sets containing millions of reactions.

## Methods

### Data

The data consisted of 2.6M reactions extracted from the Pistachio database<sup>[34]</sup> (version 191118), where we removed duplicates and filtered invalid reactions using RDKit.<sup>[36]</sup> The data set was split into train, validation and test sets (90% / 5% / 5%), keeping reactions with identical products in the same set. The reaction data in Pistachio was classified with NameRXN,<sup>[18]</sup> a rule-based software that classifies roughly 1000 different name reactions. The classification is organised in superclasses,<sup>[43]</sup> reaction categories and name reactions according to the RXNO ontology.<sup>[16]</sup> For more detail on name reactions and their categories, we refer the reader to the work of Schneider et al.<sup>[44]</sup> As commonly done, we represent the chemical reactions with reaction SMILES.<sup>[14][21]</sup> We tokenise the reaction SMILES as in Schwaller et al.<sup>[6]</sup> without enforcing any distinction between reactants and reagents. Therefore, our method is universally applicable, including those reactions where the reactant-reagent distinction is subtle.<sup>[28]</sup> To compare with previous work, we used the reaction data set published by Schneider et al.<sup>[25]</sup> containing 50k reactions belonging to 50 different reaction classes.

## Models

We trained two different types of deep learning models inspired by recent progress in Natural Language Processing. The first model is an autoregressive encoder-decoder transformer model.<sup>[11]</sup> We constructed the model with 2 layers and 1 decoder layer. For the target, we split the class prediction into superclass, category and name reaction prediction. This means, for example, that the target string for the name reaction “1.2.3” would be “1 1.2 1.2.3”. As the source and target are dissimilar, we did not share encoder and decoder embeddings. For the remaining hyperparameters, we used the same as were used for the training of the Molecular Transformer,<sup>[6,45]</sup> which is state-of-the-art in chemical reaction prediction.

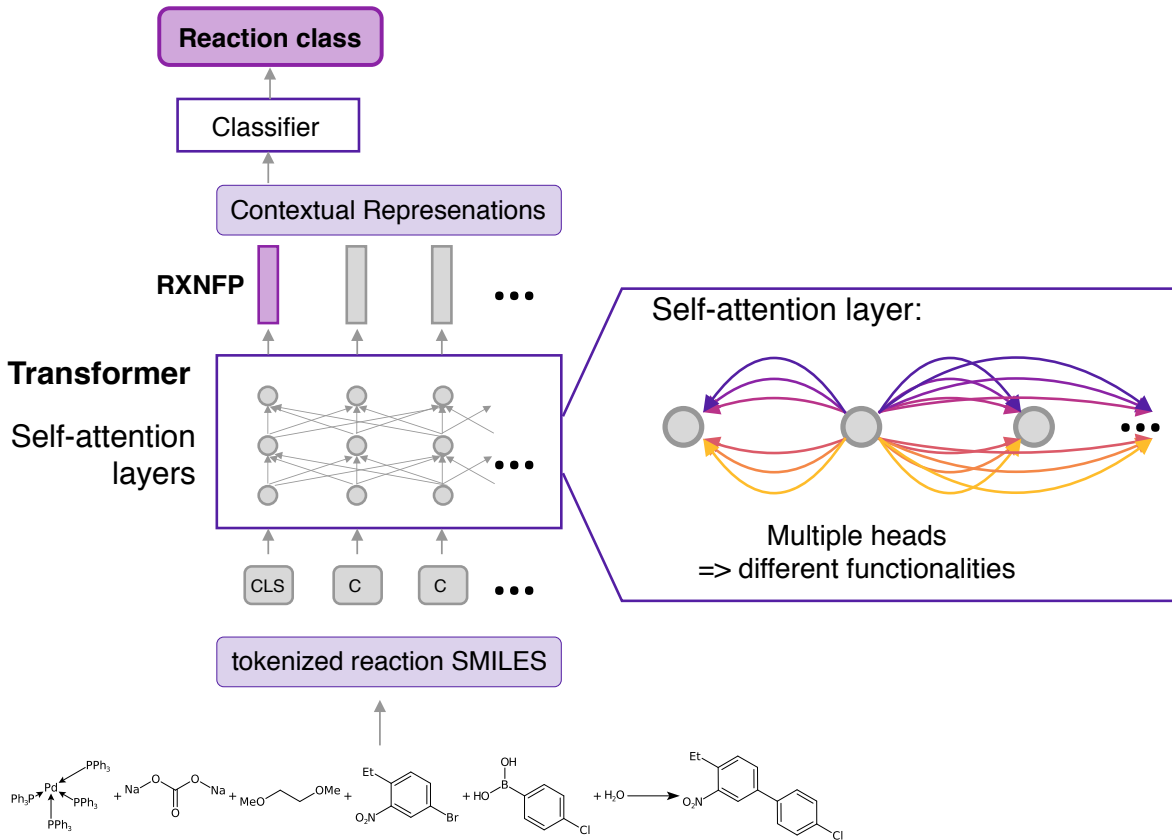


Figure 4: BERT model with classification layer applied to reaction SMILES.

One of the major recent advancement in natural language processing is BERT,<sup>[12]</sup> which compared to the seq-2-seq architecture only consists of a transformer encoder with specific

heads that can be fine-tuned for different tasks such as multi-class prediction. The model is visualised in Figure 4. We pretrained a BERT model using masked language modeling loss on the chemical reactions. The task of the model in masked language modeling consists of predicting individual tokens of the input sequence that have been masked with a probability of 0.15. Same as in the BERT training, a special class token [CLS] was prepended to the tokenised reaction SMILES. The [CLS] token was never masked during this self-supervised training. In contrast to the original BERT pretraining,<sup>12</sup> we did not use the next sentence prediction task. We then fine-tuned the pretrained model with a classifier head on the name reaction classes. The embeddings of the [CLS] token were taken as input to the classifier head. Compared with the hyperparameters of the BERT-Base model in Ref. 46, we decreased the hidden size to 256, the intermediate size to 512, and the number of attention heads to 4. For the pretraining, we set 820k steps with a learning rate of 1e-4 and a maximum sequence length of 512, the rest of the parameters were kept as suggested in Ref. 46. For the classification fine-tuning, we only changed the learning rate to 2e-5, kept the maximum sequence length of 512 and fine-tuned for 5 epochs. After training, we converted the models to pytorch<sup>47</sup> models, which matched the Huggingface<sup>48</sup> interface, as it facilitated further analysis.

## k-Nearest Neighbor Classifier

The k-nearest neighbor classifier used to assess the quality of the proposed reaction representations is based on the FAISS framework developed by Facebook research.<sup>35</sup> As FAISS provides an efficient implementation of brute-force k-nearest neighbour searches that can be applied on relatively large data sets, possible biases introduced through approximation methods were avoided. The number of nearest neighbours  $k = 5$  and the Euclidean metric (L2) are chosen for all tests. The predicted class of the query is assumed to be the one that is most represented within the result set. Ties are broken using the distance between the query and one or more neighbours.

## TMAP

TMAP<sup>[33]</sup> is an dimensionality reduction algorithm capable of handling millions of data points. The advantage of TMAP compared to other dimensionality reduction algorithms is the 2-dimensional tree-like output, which preserves both local and global structure, with a focus of local structure. The algorithm consists of four steps: 1) LSH Forest-based indexing, 2) k-nearest neighbour graph generation, 3) minimum spanning tree calculation using Kruskal’s algorithm and 4) creating the tree-like layout. The resulting layout is then displayed using the interactive data visualisation framework Faerun.<sup>[42]</sup>

TMAP<sup>[33]</sup> and Faerun<sup>[42]</sup> were originally developed to visualise large molecular data sets, but have also been shown to be applicable to a wide range of other data. Here, we extended the framework with a customised version of SmilesDrawer<sup>[49]</sup> which has been extended to allow for the display of chemical reactions.

## Evaluation metrics

To compare the results on the imbalanced classification test set using the confusion entropy of the confusion matrix (CEN)<sup>[37]</sup> calculated as follows,

$$P_{i,j}^j = \frac{Matrix(i,j)}{\sum_{k=1}^{|C|} \left( Matrix(j,k) + Matrix(k,j) \right)}, \quad P_{i,j}^i = \frac{Matrix(i,j)}{\sum_{k=1}^{|C|} \left( Matrix(i,k) + Matrix(k,i) \right)}$$
$$CEN_j = - \sum_{k=1, k \neq j}^{|C|} \left( P_{j,k}^j \log_{2(|C|-1)} \left( P_{j,k}^j \right) + P_{k,j}^j \log_{2(|C|-1)} \left( P_{k,j}^j \right) \right)$$
$$P_j = \frac{\sum_{k=1}^{|C|} \left( Matrix(j,k) + Matrix(k,j) \right)}{2 \sum_{k,l=1}^{|C|} Matrix(k,l)}$$
$$CEN = \sum_{j=1}^{|C|} P_j CEN_j$$

where Matrix is the confusion matrix, and the overall Matthews Correlation Coefficient (MCC),<sup>[38,39]</sup>

$$\begin{aligned} cov(X, Y) &= \sum_{i,j,k=1}^{|C|} \left( Matrix(i, i) Matrix(k, j) - Matrix(j, i) Matrix(i, k) \right) \\ cov(X, X) &= \sum_{i=1}^{|C|} \left[ \left( \sum_{j=1}^{|C|} Matrix(j, i) \right) \left( \sum_{k,l=1, k \neq i}^{|C|} Matrix(l, k) \right) \right] \\ cov(Y, Y) &= \sum_{i=1}^{|C|} \left[ \left( \sum_{j=1}^{|C|} Matrix(i, j) \right) \left( \sum_{k,l=1, k \neq i}^{|C|} Matrix(k, l) \right) \right] \\ MCC &= \frac{cov(X, Y)}{\sqrt{cov(X, X) \times cov(Y, Y)}}. \end{aligned}$$

Both are recommended metrics for imbalanced multi-class classification problems. We computed the scores using PyCM.<sup>[50]</sup> For the comparison on the balanced data set, we used the average recall, precision and F1 score, as those metrics were used by Schneider et al.<sup>[25]</sup> The recall, precision and F1 score values for the individual classes are shown in the supplementary material.

## Data availability

The Schneider 50k data set is publicly available.<sup>[25]</sup> The commercial Pistachio data set can be obtained from NextMove Software.<sup>[34]</sup>

## Code availability

The rxnfp code and the experiments on the public data set, as well as an interactive TMAP, can be found on <https://rxn4chemistry.github.io/rxnfp>.

## Acknowledgement

DP and JLR acknowledge financial support by the Swiss National Science Foundation (NCCR TransCure).

## References

- (1) Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E. The 'wired' universe of organic chemistry. *Nature Chemistry* **2009**, *1*, 31–36.
- (2) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **2017**, *3*, 1237–1245.
- (3) IBM RXN for Chemistry. <https://rxn.res.ibm.com>, (Accessed Sep 13, 2019).
- (4) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* **2017**, *3*, 434–443.
- (5) 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* **2018**, *9*, 6091–6098.
- (6) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *in press*.
- (7) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.
- (8) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical science* **2020**, *11*, 154–168.

- (9) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy. *arXiv preprint arXiv:1910.08036* **2019**,
- (10) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. **2020**,
- (11) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*. 2017; pp 5998–6008.
- (12) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019; pp 4171–4186.
- (13) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28*, 31–36.
- (14) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences* **1989**, *29*, 97–101.
- (15) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Unsupervised Attention-Guided Atom-Mapping. **2020**,
- (16) RSC’s RXNO Ontology. <http://www.rsc.org/ontologies/RXNO/index.asp>, (Accessed Sep 13, 2019).

- (17) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted Noise-Reduction of Chemical Reactions Data Sets. **2020**,
- (18) Nextmove Software NameRXN. <http://www.nextmovesoftware.com/namerxn.html>, (Accessed Jul 29, 2019).
- (19) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for reaction classification. *Journal of chemical information and modeling* **2013**, *53*, 2884–2895.
- (20) Daylight Theory Manual, Chapter 5. <http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html> (accessed Dec 10, 2019).
- (21) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.
- (22) Chen, L.; Gasteiger, J. Organic Reactions Classified by Neural Networks: Michael Additions, Friedel–Crafts Alkylations by Alkenes, and Related Reactions. *Angewandte Chemie International Edition in English* **1996**, *35*, 763–765.
- (23) Chen, L.; Gasteiger, J. Knowledge discovery in reaction databases: Landscaping organic reactions by a self-organizing neural network. *Journal of the American Chemical Society* **1997**, *119*, 4033–4042.
- (24) Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. Classification of organic reactions: similarity of reactions based on changes in the electronic features of oxygen atoms at the reaction sites. *Journal of chemical information and computer sciences* **1998**, *38*, 210–219.
- (25) 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. *Journal of chemical information and modeling* **2015**, *55*, 39–53.
- (26) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS central science* **2018**, *4*, 1465–1476.
- (27) Ghiandoni, G. M.; Bodkin, M. J.; Chen, B.; Hristozov, D.; Wallace, J. E.; Webster, J.; Gillet, V. J. Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature. *Journal of chemical information and modeling* **2019**, *59*, 4167–4187.
- (28) Schneider, N.; Stiefl, N.; Landrum, G. A. What’s what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling* **2016**, *56*, 2336–2346.
- (29) ChemAxon. <https://docs.chemaxon.com/display/ltsargon/Reaction+fingerprint+RF>, (Accessed Dec 21, 2019).
- (30) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*. 2015; pp 2224–2232.
- (31) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS central science* **2016**, *2*, 725–732.
- (32) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **2020**,
- (33) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *arXiv preprint arXiv:1908.10410* **2019**,

- (34) Nextmove Software Pistachio. <http://www.nextmovesoftware.com/pistachio.html>, (Accessed Jul 29, 2019).
- (35) Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. 2017.
- (36) Landrum, G. et al. rdkit/rdkit: 2019\_03\_4 (Q1 2019) Release. 2019; <https://doi.org/10.5281/zenodo.3366468>.
- (37) Wei, J.-M.; Yuan, X.-J.; Hu, Q.-H.; Wang, S.-Q. A novel measure for evaluating classifiers. *Expert Systems with Applications* **2010**, *37*, 3799–3809.
- (38) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* **1975**, *405*, 442–451.
- (39) Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational biology and chemistry* **2004**, *28*, 367–374.
- (40) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliaskova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O., et al. The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics* **2017**, *9*, 33.
- (41) Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics* **2020**, *12*, 1–15.
- (42) Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2017**, *34*, 1433–1435.
- (43) Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. Analysis of the reactions used for the preparation of drug candidate molecules. *Organic & biomolecular chemistry* **2006**, *4*, 2337–2347.

- (44) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists’ bread and butter. *Journal of medicinal chemistry* **2016**, *59*, 4385–4402.
- (45) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proc. ACL. 2017.
- (46) BERT code. <https://github.com/google-research/bert#sentence-and-sentence-pair-classification-tasks>, (Accessed Oct 15, 2019).
- (47) Paszke, A. et al. *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc., 2019; pp 8024–8035.
- (48) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Brew, J. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv* **2019**, *abs/1910.03771*.
- (49) Probst, D.; Reymond, J.-L. Smilesdrawer: parsing and drawing SMILES-encoded molecular structures using client-side javascript. *Journal of chemical information and modeling* **2018**, *58*, 1–7.
- (50) Haghighi, S.; Jasemi, M.; Hessabi, S.; Zolanvari, A. PyCM: Multiclass confusion matrix library in Python. *Journal of Open Source Software* **2018**, *3*, 729.

# Supplementary Information: Mapping the Space of Chemical Reactions using Attention-Based Neural Networks

Philippe Schwaller,<sup>\*,†,‡</sup> Daniel Probst,<sup>‡</sup> Alain C. Vaucher,<sup>†</sup> Vishnu H. Nair,<sup>†</sup>  
David Kreutter,<sup>‡</sup> Teodoro Laino,<sup>†</sup> and Jean-Louis Reymond<sup>‡</sup>

<sup>†</sup>*IBM Research – Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*

<sup>‡</sup>*Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012  
Bern, Switzerland*

E-mail: [phs@zurich.ibm.com](mailto:phs@zurich.ibm.com)

## 1 Reaction elements and property maps

Figure [1](#) shows the chemical reaction found in the 50k set by Schneider et al. [1](#) visualised with TMAP [2](#) using the *rxnfp* (10k). The BERT model, which generated this reaction fingerprint was trained on the 10k training reactions. The reaction maps are made of the 10k training reactions plus 40k unseen reactions. The reactions corresponding to same reaction classes are well clustered together. We highlight reactions that contain specific elements in the precursors and observe that they found in the same branches of the map. Moreover, we visualize product properties and also observe defined clustering.

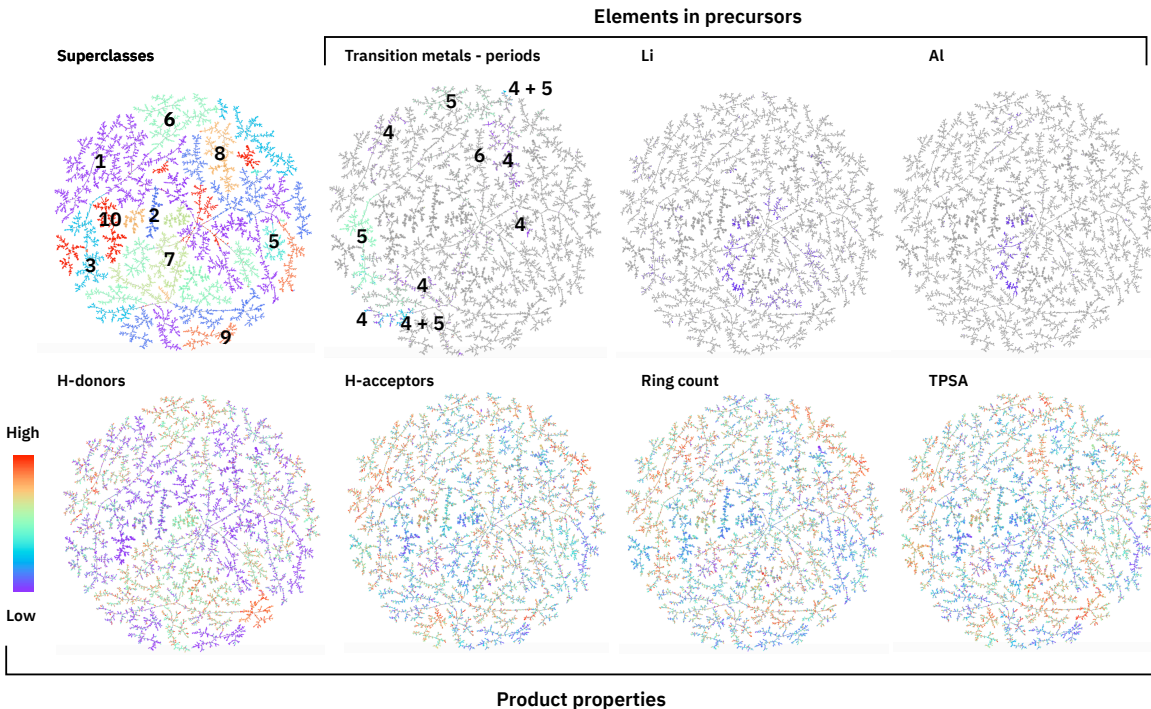


Figure 1: TMAP<sup>[2]</sup> of the Schneider 50k set using the *rxnfp* (10k) embeddings. The superclasses, as well as specific metallic elements in the precursors and product properties are highlighted in the different maps. An interactive version of this map is also available as a separate file.

## 2 Detailed results on balanced data set with 50k

Schneider et al.<sup>[1]</sup> evaluated their reaction fingerprints by analysing how well it could classify chemical reactions using a logistic regression classifier.<sup>[3]</sup> For a given reaction input, they trained their classifier to predict 1 out of 50 named reaction classes using 200 training/validation and 800 testing examples per class. To be able to directly compare to the results of Ref. [1], we investigated our learned fingerprints on their data sets, pretrained and fine-tuned on the same 10k training reactions resulting in *rxnfp* (10k). A summary where we report recall, precision and F-score averaged over the 50 classes can be found in Table [1]. While the *rxnfp* (pretrained) does not suffice to match the performance of the handcrafted fingerprint on this balanced data set, *rxnfp* (10k), generated after fine-tuning the model on as little as the 10k reactions, is able to reach scores of 0.99 compared to 0.97 for the

hand-crafted fingerprint.

Table 1: Comparing fingerprints on the 50k reactions classification benchmark by Schneider et al.<sup>[1]</sup> (50 classes, 1000 reactions per class, 200 for training/validation and 800 for testing)

Fingerprint	recall	precision	F-score	
AP3 256 (folded) <sup>[1]</sup> + Agent features	0.97	0.97	0.97	handcrafted, reactants-reagents separation
<i>rxnfp</i> (pretrained)	0.90	0.90	0.90	after pretraining
<i>rxnfp</i> (10k)	0.99	0.99	0.99	fine-tuning on 10k reactions training set <sup>[1]</sup>

Table 2 and Figure 2 show the detailed results for *rxnfp* (10k). Table 3 and Figure 3 show the results of for *rxnfp* (pretrained) computed by the model never fine-tuned on reaction classification.

For both data-driven fingerprints the methylation class seems to be the hardest to predict correctly. Using the pretrained fingerprint it is hard to distinguish between reaction classes that differ only by one atom, like “CO2H-Et deprotection” and “CO2H-Me deprotection”. “Carboxylic acid + amine condensation” are confused with “Amide Schotten-Baumann” reactions and “Mitsunobu aryl ether synthesis” with “Williamson ether synthesis” reactions. It is likely that in future unsupervised reaction fingerprints will be developed that capture this fine-grained information better.

## References

- (1) 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. *Journal of chemical information and modeling* **2015**, *55*, 39–53.
- (2) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *arXiv preprint arXiv:1908.10410* **2019**,

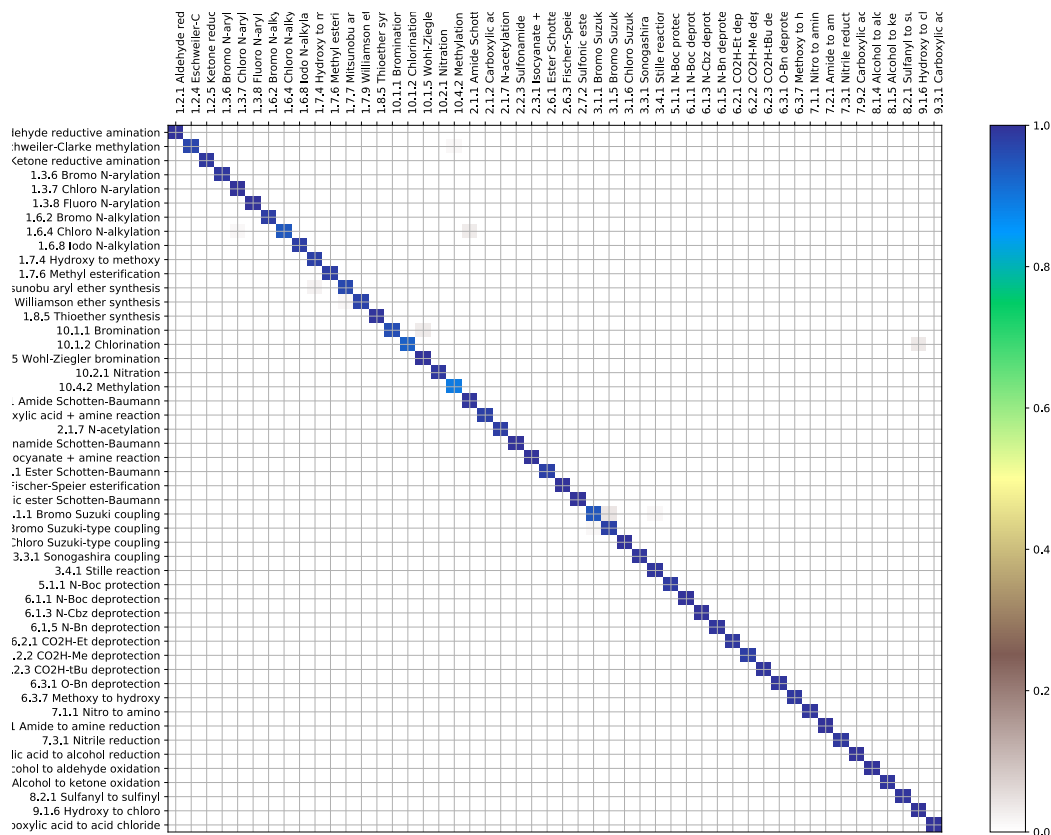


Figure 2: Confusion matrix for *rxnfp* (10k) train

- (3) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, 12, 2825–2830.

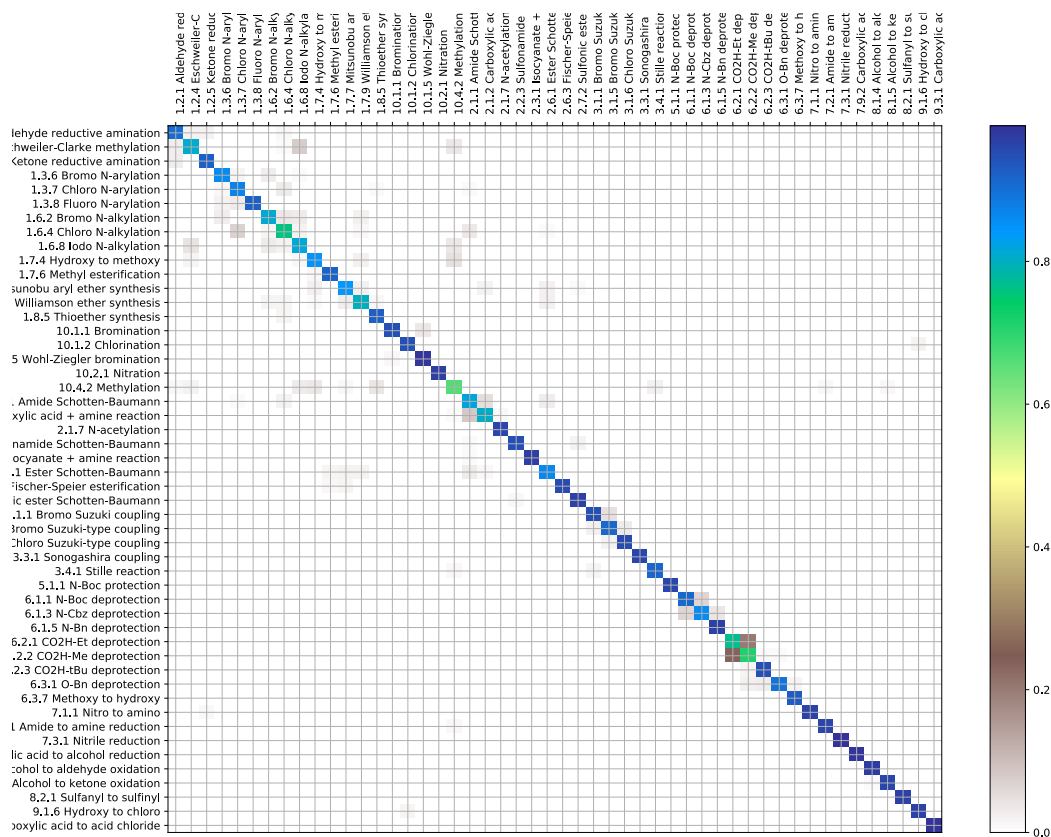


Figure 3: Confusion matrix for *rxnfp* (pretrained)

Table 2: *rxnfp* (10k) train: 50k reactions classification benchmark by Schneider et al.<sup>[1]</sup>

	recall	prec	F-score	reaction class	
0	0.9988	0.9901	0.9944	Aldehyde reductive amination	1.2.1
1	0.9712	0.9848	0.9780	Eschweiler-Clarke methylation	1.2.4
2	0.9888	0.9950	0.9918	Ketone reductive amination	1.2.5
3	0.9912	0.9863	0.9888	Bromo N-arylation	1.3.6
4	0.9962	0.9827	0.9894	Chloro N-arylation	1.3.7
5	0.9975	0.9876	0.9925	Fluoro N-arylation	1.3.8
6	0.9825	0.9788	0.9807	Bromo N-alkylation	1.6.2
7	0.9437	0.9921	0.9673	Chloro N-alkylation	1.6.4
8	0.9838	0.9825	0.9831	Iodo N-alkylation	1.6.8
9	0.9775	0.9678	0.9726	Hydroxy to methoxy	1.7.4
10	0.9838	0.9838	0.9838	Methyl esterification	1.7.6
11	0.9675	0.9639	0.9657	Mitsunobu aryl ether synthesis	1.7.7
12	0.9750	0.9665	0.9708	Williamson ether synthesis	1.7.9
13	0.9938	0.9938	0.9938	Thioether synthesis	1.8.5
14	0.9575	0.9935	0.9752	Bromination	10.1.1
15	0.9313	0.9868	0.9582	Chlorination	10.1.2
16	0.9988	0.9685	0.9834	Wohl-Ziegler bromination	10.1.5
17	0.9888	0.9987	0.9937	Nitration	10.2.1
18	0.8938	0.9483	0.9202	Methylation	10.4.2
19	0.9950	0.9522	0.9731	Amide Schotten-Baumann	2.1.1
20	0.9788	0.9899	0.9843	Carboxylic acid + amine reaction	2.1.2
21	0.9838	0.9975	0.9906	N-acetylation	2.1.7
22	0.9975	0.9975	0.9975	Sulfonamide Schotten-Baumann	2.2.3
23	1.0000	0.9950	0.9975	Isocyanate + amine reaction	2.3.1
24	0.9775	0.9726	0.9751	Ester Schotten-Baumann	2.6.1
25	0.9962	0.9815	0.9888	Fischer-Speier esterification	2.6.3
26	1.0000	1.0000	1.0000	Sulfonic ester Schotten-Baumann	2.7.2
27	0.9463	0.9818	0.9637	Bromo Suzuki coupling	3.1.1
28	0.9800	0.9596	0.9697	Bromo Suzuki-type coupling	3.1.5
29	1.0000	0.9950	0.9975	Chloro Suzuki-type coupling	3.1.6
30	0.9925	0.9937	0.9931	Sonogashira coupling	3.3.1
31	0.9925	0.9778	0.9851	Stille reaction	3.4.1
32	0.9850	0.9975	0.9912	N-Boc protection	5.1.1
33	1.0000	0.9780	0.9889	N-Boc deprotection	6.1.1
34	0.9975	1.0000	0.9987	N-Cbz deprotection	6.1.3
35	0.9950	0.9925	0.9938	N-Bn deprotection	6.1.5
36	0.9888	0.9875	0.9881	CO <sub>2</sub> H-Et deprotection	6.2.1
37	0.9825	0.9800	0.9813	CO <sub>2</sub> H-Me deprotection	6.2.2
38	0.9950	0.9925	0.9938	CO <sub>2</sub> H-tBu deprotection	6.2.3
39	0.9950	0.9925	0.9938	O-Bn deprotection	6.3.1
40	0.9888	0.9900	0.9894	Methoxy to hydroxy	6.3.7
41	0.9938	0.9925	0.9931	Nitro to amino	7.1.1
42	0.9975	0.9803	0.9888	Amide to amine reduction	7.2.1
43	0.9912	0.9925	0.9919	Nitrile reduction	7.3.1
44	0.9988	0.9938	0.9963	Carboxylic acid to alcohol reduction	7.9.2
45	1.0000	0.9963	0.9981	Alcohol to aldehyde oxidation	8.1.4
46	0.9950	0.9987	0.9969	Alcohol to ketone oxidation	8.1.5
47	0.9950	0.9962	0.9956	Sulfanyl to sulfinyl	8.2.1
48	0.9962	0.9614	0.9785	Hydroxy to chloro	9.1.6
49	0.9975	0.9888	0.9932	Carboxylic acid to acid chloride	9.3.1
	0.99	0.99	0.99	Average	

Table 3: *rxnfp* (pretrained): 50k reactions classification benchmark by Schneider et al.<sup>[1]</sup>

	recall	prec	F-score	reaction class	
0	0.9012	0.8990	0.9001	Aldehyde reductive amination	1.2.1
1	0.8063	0.8323	0.8190	Eschweiler-Clarke methylation	1.2.4
2	0.9213	0.9213	0.9213	Ketone reductive amination	1.2.5
3	0.8600	0.8632	0.8616	Bromo N-arylation	1.3.6
4	0.8712	0.7938	0.8308	Chloro N-arylation	1.3.7
5	0.9225	0.9498	0.9360	Fluoro N-arylation	1.3.8
6	0.8113	0.8353	0.8231	Bromo N-alkylation	1.6.2
7	0.7600	0.7696	0.7648	Chloro N-alkylation	1.6.4
8	0.8125	0.7908	0.8015	Iodo N-alkylation	1.6.8
9	0.8500	0.8662	0.8580	Hydroxy to methoxy	1.7.4
10	0.9200	0.9258	0.9229	Methyl esterification	1.7.6
11	0.8413	0.8519	0.8465	Mitsunobu aryl ether synthesis	1.7.7
12	0.8000	0.7960	0.7980	Williamson ether synthesis	1.7.9
13	0.9225	0.8902	0.9061	Thioether synthesis	1.8.5
14	0.9437	0.9461	0.9449	Bromination	10.1.1
15	0.9463	0.9232	0.9346	Chlorination	10.1.2
16	0.9838	0.9633	0.9734	Wohl-Ziegler bromination	10.1.5
17	0.9738	0.9725	0.9731	Nitration	10.2.1
18	0.6625	0.7172	0.6888	Methylation	10.4.2
19	0.8175	0.7861	0.8015	Amide Schotten-Baumann	2.1.1
20	0.8013	0.8250	0.8129	Carboxylic acid + amine reaction	2.1.2
21	0.9600	0.9588	0.9594	N-acetylation	2.1.7
22	0.9450	0.9345	0.9397	Sulfonamide Schotten-Baumann	2.2.3
23	0.9725	0.9569	0.9647	Isocyanate + amine reaction	2.3.1
24	0.8625	0.8582	0.8603	Ester Schotten-Baumann	2.6.1
25	0.9525	0.9658	0.9591	Fischer-Speier esterification	2.6.3
26	0.9700	0.9395	0.9545	Sulfonic ester Schotten-Baumann	2.7.2
27	0.9437	0.9333	0.9385	Bromo Suzuki coupling	3.1.1
28	0.9113	0.9045	0.9078	Bromo Suzuki-type coupling	3.1.5
29	0.9550	0.9340	0.9444	Chloro Suzuki-type coupling	3.1.6
30	0.9625	0.9686	0.9655	Sonogashira coupling	3.3.1
31	0.9150	0.9150	0.9150	Stille reaction	3.4.1
32	0.9613	0.9661	0.9637	N-Boc protection	5.1.1
33	0.9100	0.9089	0.9094	N-Boc deprotection	6.1.1
34	0.8600	0.9005	0.8798	N-Cbz deprotection	6.1.3
35	0.9700	0.9293	0.9492	N-Bn deprotection	6.1.5
36	0.7688	0.7437	0.7560	CO <sub>2</sub> H-Et deprotection	6.2.1
37	0.7150	0.7259	0.7204	CO <sub>2</sub> H-Me deprotection	6.2.2
38	0.9450	0.9486	0.9468	CO <sub>2</sub> H-tBu deprotection	6.2.3
39	0.8962	0.9459	0.9204	O-Bn deprotection	6.3.1
40	0.9313	0.9418	0.9365	Methoxy to hydroxy	6.3.7
41	0.9663	0.9898	0.9779	Nitro to amino	7.1.1
42	0.9613	0.9470	0.9541	Amide to amine reduction	7.2.1
43	0.9900	0.9888	0.9894	Nitrile reduction	7.3.1
44	0.9838	0.9887	0.9862	Carboxylic acid to alcohol reduction	7.9.2
45	0.9750	0.9750	0.9750	Alcohol to aldehyde oxidation	8.1.4
46	0.9600	0.9540	0.9570	Alcohol to ketone oxidation	8.1.5
47	0.9700	0.9898	0.9798	Sulfanyl to sulfinyl	8.2.1
48	0.9663	0.9748	0.9705	Hydroxy to chloro	9.1.6
49	0.9875	0.9925	0.9900	Carboxylic acid to acid chloride	9.3.1
	0.90	0.90	0.90	Average	