

Data-Driven Chemical Reaction Classification, Fingerprinting and Clustering using Attention-Based Neural Networks

Philippe Schwaller,^{*,†,‡} Daniel Probst,[‡] Alain C. Vaucher,[†] Vishnu H. Nair,[†]
Teodoro Laino,[†] and Jean-Louis Reymond[‡]

[†]*IBM Research – Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*

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

E-mail: phs@zurich.ibm.com

Abstract

Organic reactions are usually assigned to classes grouping reactions with similar reagents and mechanisms. The classification process is a tedious task, requiring first an accurate mapping of the reaction (atom mapping) followed by the identification of the corresponding reaction class template. In this work, we present two transformer-based models that infer reaction classes from the SMILES representation of chemical reactions. Our best model reaches a classification accuracy of 98.2%. We study the incorrect predictions of the models and show that they reveal different biases and mistakes in the underlying data set. Using the embeddings of our classification model, we introduce reaction fingerprints that do not require knowing the reaction center or distinguishing between reactants and reagents. This conversion from chemical reactions to feature vectors enables efficient clustering and similarity search in the reaction space.

We compare the reaction clustering for combinations of self-supervised, supervised, and molecular shingle-based reaction representations.

1 Introduction

Name reactions¹ 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 analyze and infer optimal reaction conditions. Today, several hundreds of name reactions exist in the RXNO ontology.¹ 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.

In the last decade, computer-based systems²⁻⁶ became an important asset available to chemists for reaction prediction tasks. The knowledge of the class of a predicted reaction has a great value for expert chemists to assess the quality of the prediction. For this reason, the demand for robust algorithms to categorize chemical reactions is high. The current state-of-the-art in reaction classification is represented by commercially available tools,^{7,8} which classify reactions based on a library of expert-written rules. These tools typically make use of SMIRKS,⁹ a language to describe transformations in the simplified molecular-input line-entry system (SMILES) format.^{10,11} 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. Among the few attempts made to infer name reactions using machine

¹For convenience, in this work, “name reaction” refers to reaction classes that have an established name in chemistry, and not only to reactions that carry the name of the discoverer(s).

learning methods, Schneider et al.¹² developed a reaction classifier based on traditional reaction fingerprints. Unfortunately, the limited set of reaction classes used (only 50 most important ones) makes it difficult to judge how this algorithm would perform on a more comprehensive set. Recent work by Ghiandoni et al.¹³ introduced an alternative hierarchical classification scheme and random forest classifier for reaction classification. However, the model requires reaction center information as input, which limits the applicability of the method.

Here, we use a labeled set of chemical reactions as ground truth to train a FastText classifier,¹⁴ a MinHash fingerprint (MHFP)¹⁵-based k -NN classifier, and two transformer-based deep learning models as architecture.^{16,17} The ground truth data is composed of chemical transformations represented as SMILES, and its labeling (classification) was taken from the Pistachio data set,¹⁸ which uses NameRXN for the reaction classification.⁷

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. We show that the transformer-based sequence-2-sequence (seq-2-seq) model¹⁶ was able to match the ground-truth classification with an accuracy of 95.2% and the Bidirectional Encoder Representations from Transformers (BERT) classifier¹⁷ with 98.2%. The mismatches are mainly related to unrecognized 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 analyze 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. We then show that the class embeddings learned by the BERT model can be used as reaction fingerprints. Traditionally, reaction fingerprints were hand-crafted using the reaction center or a combination of the

reactant, reagent and product fingerprints. ChemAxon,¹⁹ for instance, provides eight types of such reaction fingerprints. One of the most frequently used hand-crafted fingerprint is the difference fingerprint developed by Schneider et al.¹² However, their fingerprint requires to know the reactant-reagent split, as reactants and reagents are weighted differently. The difference fingerprint¹² has successfully been applied to predict reaction conditions,²⁰ where the reagents were not taken into account for the reaction description. Based on the differentiable molecule fingerprint by Duvenaud et al.,²¹ the first example of a learned reaction fingerprint was presented by Wei et al.²² 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. The reaction fingerprints we introduce in this work, enable efficient similarity searches and clustering in the chemical reaction space without the requirement of knowing the reaction center or the reactant-reagent split.

2 Data & Models

The data consisted of 2.6M reactions extracted from the Pistachio database¹⁸ (version 191118), where we removed duplicates and filtered invalid reactions using RDKit.²³ 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,⁷ a rule-based software that classifies roughly 1000 different name reactions. The classification is organized in superclasses,²⁴ reaction categories and name reactions according to the RXNO ontology.¹ For more detail on name reactions and their categories, we refer the reader to the work of Schneider et al.²⁵ As commonly done, we represent the chemical reactions with reaction SMILES.^{10,11} We tokenize the reaction SMILES as in Schwaller et al.⁵ without enforcing any distinction between reactants and reagents. Therefore, our method is universally applicable, including those reactions where the reactant-reagent distinction is

subtle.²⁶

To have a baseline, we first trained a supervised classifier using FastText^{14,27} as well as an approximate k -nearest neighbor classifier for an MHFP-based reaction fingerprint.¹⁵ The fingerprint, which has been shown to perform better than comparable molecular fingerprints like the Extended-connectivity fingerprint (ECFP)²⁸ in ligand-based virtual screening benchmarks, encodes a reaction SMILES as a 512-dimensional MinHash vector by minhashing a shingling consisting of SMILES-encoded circular substructures of the molecule. The fingerprint has been adapted to work on reactions by creating a shingling of the symmetric difference between the SMILES-encoded circular substructures of the products and those of the precursors of the reaction SMILES. An implementation of the local sensitive hashing (LSH) forest algorithm facilitates fast k -NN searches with $k = 3$, $k_c = 1000$, and the number of trees $l = 32$.²⁹ The k -NN algorithm employs a distance-weighted class-wise count of nearest neighbors as a scoring function, where the score of a class is defined as $s_y = \sum_{n \in N_y} \frac{1}{d(q,n)^2}$, where q is the query and N_y is the subset of the k -nearest neighbors with class label y .

We then 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.¹⁶ 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,^{5,30} which is state-of-the-art in chemical reaction prediction.

One of the major recent advancement in natural language processing is BERT,¹⁷ 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. 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 tokenized reaction SMILES. The [CLS] token was never masked during this self-supervised training. In contrast to the original BERT pretraining,¹⁷ 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. 31, 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. 31. 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 pytorch³² models, which matched the Huggingface³³ interface, as it facilitated further analysis.

3 Results and Discussion

3.1 Reaction classification

A summary of the results can be found in Table 1. We observe that a simple N-gram sentence classification model¹⁴ cannot capture the details of the reactions and is only able to correctly match the ground truth 77.5 % of the time. The *k*-NN classifier using the MHFP-based fingerprints as input achieved an accuracy of 86.1%. The transformer enc2-dec1 model matched the ground truth classification with 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. We analyze the BERT classifier in more detail and present an elaborate comparison between the two transformer-based models.

First, we identified different types of incorrect predictions by the transformer BERT classifier model, which are summarized in Table 2. Most errors are related to the "Unrecognized"

Table 1: Classification results

Model	Test Accuracy [%]
FastText (autotuned)	77.5
MHFP-based k -NN classifier	86.1
transformer enc2-decl	95.2
BERT classifier	98.2

class of the RXNO ontology. The most frequent error type is the prediction of a reaction class for a reaction classified as “Unrecognized” (47.9% of all incorrect predictions), and the second most frequent error type is predicting “Unrecognized” 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 “Unrecognized”	1111	0.84%
Model predicts “Unrecognized” 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 half of these reaction classes (12 out of 15), the most common error source is the failure to assign a reaction class, thus predicting “Unrecognized”. Among the other most common failures, there is the “Bouveault-Blanc reduction”, which is a reaction 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.

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 Unrecognized
3.9.41 Decarboxylative coupling	72.1	0.0 Unrecognized
9.7.140 Defluorination	75.6	0.0 Unrecognized
7.4.2 Bouveault-Blanc reduction	76.4	7.4.1 Ester to alcohol reduction
11.1 Chiral separation	83.6	0.0 Unrecognized
8.8.11 Hydroxylation	83.7	0.0 Unrecognized
4.3.11 Thiazoline synthesis	85.7	0.0 Unrecognized
3.9.12 Olefin metathesis	85.8	0.0 Unrecognized
2.5.5 Nitrile + amine reaction	86.0	0.0 Unrecognized
9.7.42 Chloro to fluoro	86.4	0.0 Unrecognized
10.4.2 Methylation	88.9	0.0 Unrecognized
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 Unrecognized
1.1.6 Chloro Menshutkin reaction	90.6	0.0 Unrecognized
5.1.2 N-Cbz protection	90.9	2.1.1 Amide Schotten-Baumann

Although the large number of “Unrecognized” reactions in Pistachio makes an extensive analysis difficult, the inspection of a few dozen cases provides interesting insights. Part of the “Unrecognized” 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 “Unrecognized” 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 examples 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 recognize 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 in the predictions are related to the "Unrecognized" class. 3511 out of 5108 reactions that were correctly predicted by the BERT classifier but not the seq-2-seq model belong to the "Unrecognized" 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 "Unrecognized", which were classified as recognized name reactions by the BERT model. Besides the "Unrecognized" 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.

3.2 Visualization of Attention Weights

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.

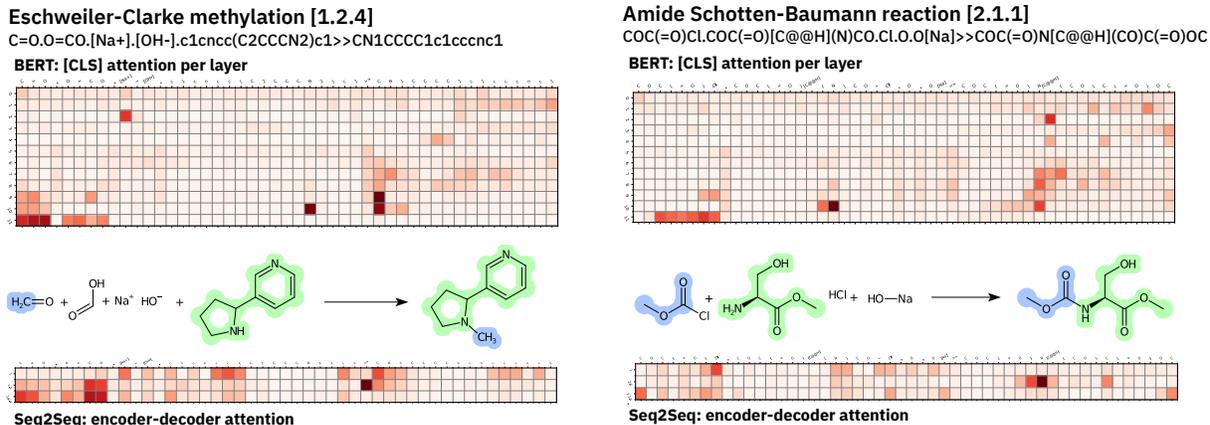


Figure 1: Layer-wise [CLS] token attention for the BERT classifier and encoder-decoder attention for the enc2-dec1 transformer model. The horizontal axis contains the SMILES tokens of the input reaction. The darker the token the more attention a specific token had in that particular layer or output step. The coloring on the reaction depictions made with CDK depict³⁴ shows the mapping from precursors to product in the ground truth.

3.3 Reaction Fingerprints and Maps

In this work, we present reaction fingerprints which can be applied to any reaction data set, as they do not require a reactant-reagent split or a fixed number of precursors. As a baseline, we suggest a reaction fingerprint variant of SECFP,¹⁵ a fingerprint from the MHFP family that is based on the hashing of molecular shinglings and supports being folded into a binary vector. However, exhaustively adding SMILES of circular substructures of products and precursors to a molecular shingling showed relatively poor performance. The best results were achieved using the symmetric difference between the shinglings of the products and the shinglings of the precursors; shinglings were created with radius $r \in \{0, 1, 2\}$ in both cases. The SECFP

hashes of the shinglings were subsequently folded into a 2048-dimensional binary vector and minhashed using a weighted hashing scheme resulting in a 4096-dimensional MinHash vector that can be readily merged with other MinHash vectors that have the same length and were encoded with the same hashing scheme. One of the advantages of MinHash-based fingerprints is that they can be arbitrarily extended with additional data. Hence, we can for example combine shingle-based fingerprints with learned fingerprints as shown below.

Our BERT model did not only perform best on the classification task but also allows to generate a vectorial representation of a chemical reaction. 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. For the supervised fine-tuning, the embeddings of the [CLS] token are then taken as input for the classification head and further refined. In our case, the [CLS] token embedding is a vector of size 256. Before the fine-tuning the [CLS] token embeddings are learned by self-supervision. For 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.

In Figure 2 the different fingerprints of the test set reactions are visualized as reaction atlases using a TMAP algorithm²⁹ and the Faerun visualization library.³⁵ The colors correspond to the 12 superclasses found in the data set. a) shows the reaction atlas using the fingerprints computed from a pretrained reaction BERT model without classification supervision. Surprisingly, applying a purely unsupervised masked language modelling training the model is already able to extract features relevant for reaction classification and some cluster-

ing can be observed in the figure. The BERT embeddings were minhashed using a weighted hashing scheme to make them compatible with the LSH forest. b) shows the reaction atlas made from the shingle-based SECFP fingerprint.¹⁵ The reactions of the same superclass tend to be clustered in the same branches. Although it is still a bit noisy the overall clustering seems to be better than the one from the pretrained model. c) shows the reaction atlas made from the fingerprints computed by the BERT model after fine tuning on the classification task. The individual classes are almost perfectly clustered. d) shows the reaction atlas created by merging the pretrained BERT fingerprint found in a) with the shingle-based reaction fingerprint found in b). Compared to a) the separation of "Heteroatom alkylation and arylation", "Acylation and related processes" and "C-C bond formation" seem to have improved, while keeping a good clustering for "Oxidations", "Functional group interconversions" and "Functional group additions". e) shows the reaction atlas created by merging the BERT classifier fingerprint in c) with the shingle-based reaction fingerprint in b). The reaction atlas is similar to the one found in c) and a more detailed analysis should be performed to uncover the differences.

In Figure 3, we show an annotated version of a reaction atlas made with the embeddings of a BERT classifier fine-tuned for three epochs. 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.

3.4 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 model⁵ or a

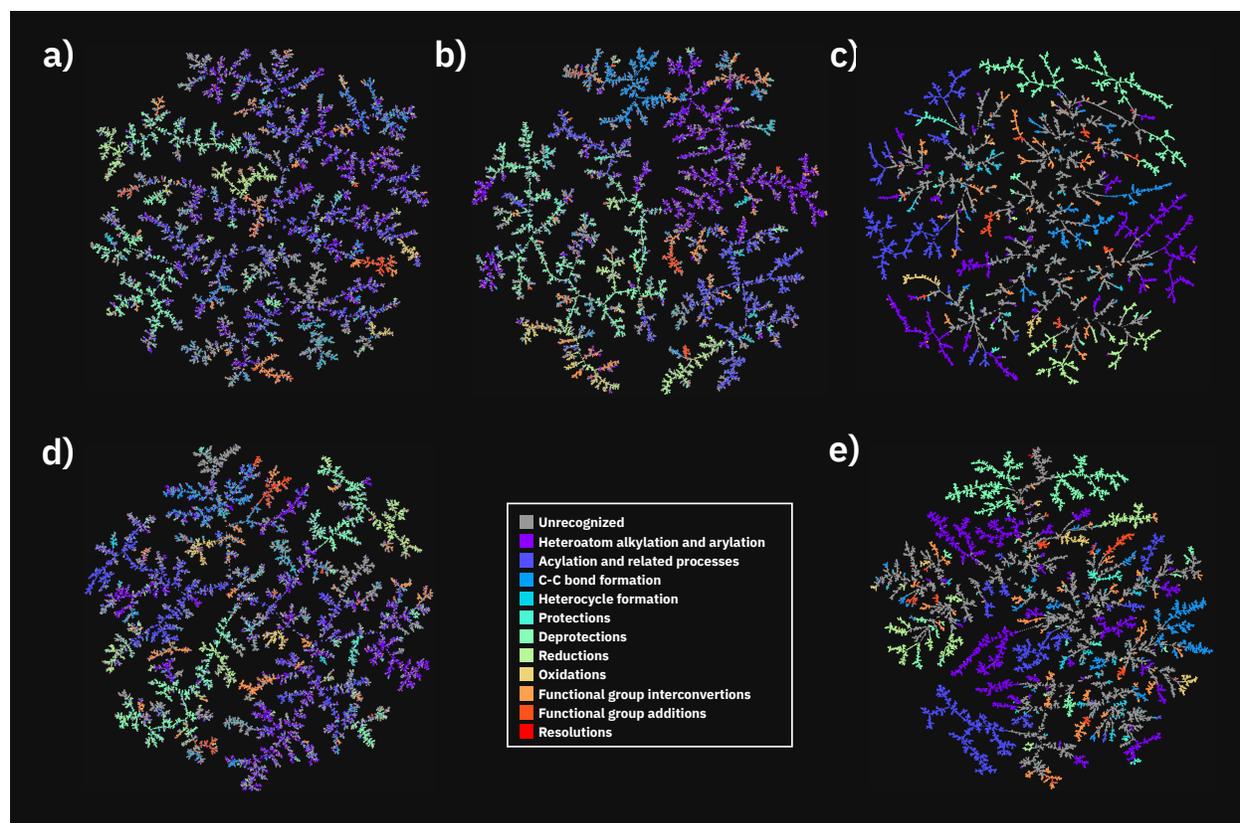


Figure 2: Different reaction atlases made with TMAP²⁹ and Faerun:³⁵ a) Embeddings extracted from the pretrained BERT model (unsupervised), b) SECFP fingerprints (hand-crafted), c) Embeddings extracted from the fine-tuned BERT classifier (supervised), d) SECFP merged with pretrained BERT embeddings, and e) SECFP merged with the fine-tuned BERT embeddings.

retrosynthesis model³⁶ 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 4 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 LSH forest from the TMAP module developed by Probst and Reymond,²⁹ the search on the training set containing 2.4M reactions was performed within milliseconds using unoptimized 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

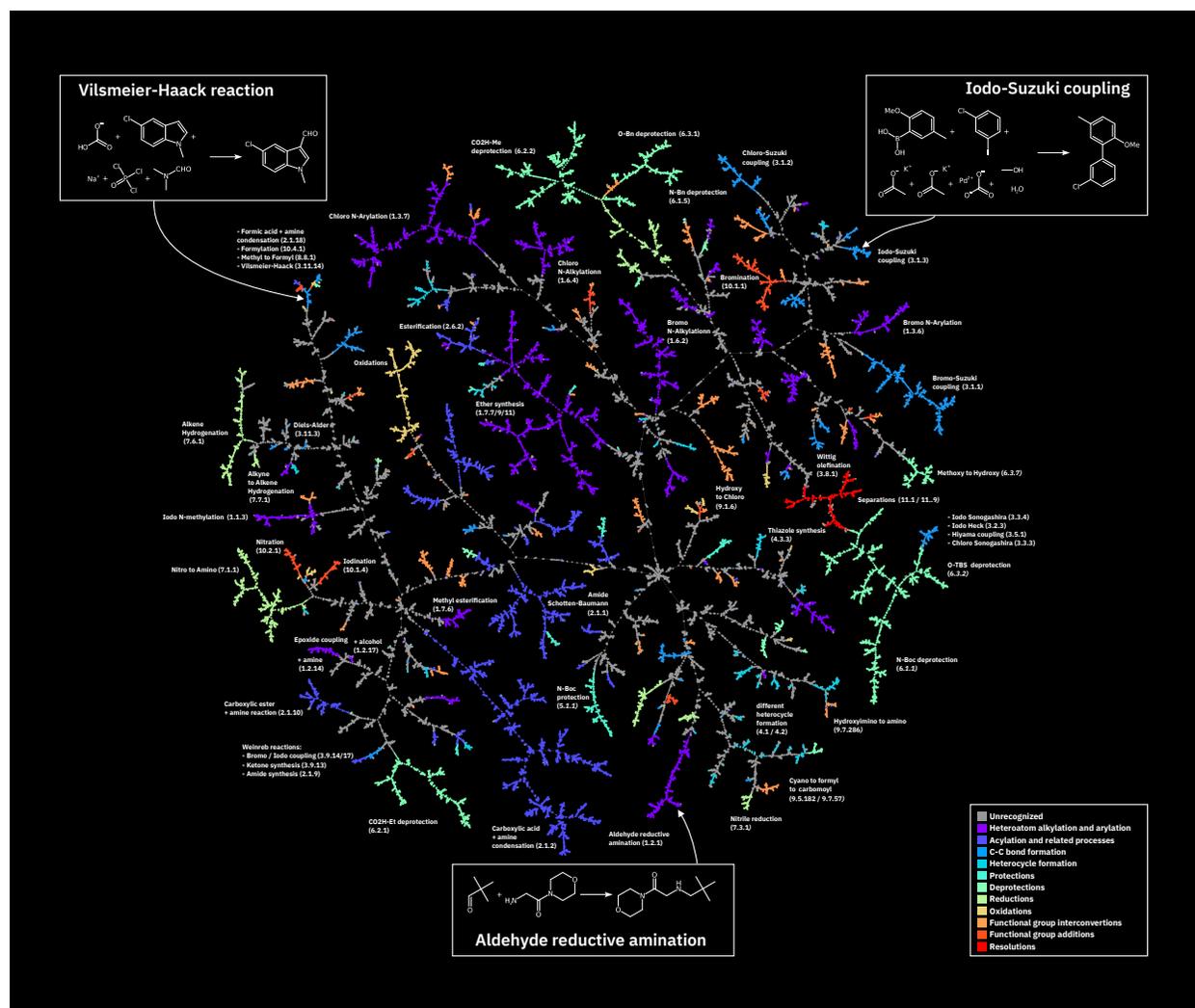


Figure 3: Annotated reaction atlas made from the embeddings of the BERT classifier after 3 epochs of fine-tuning.

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.

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

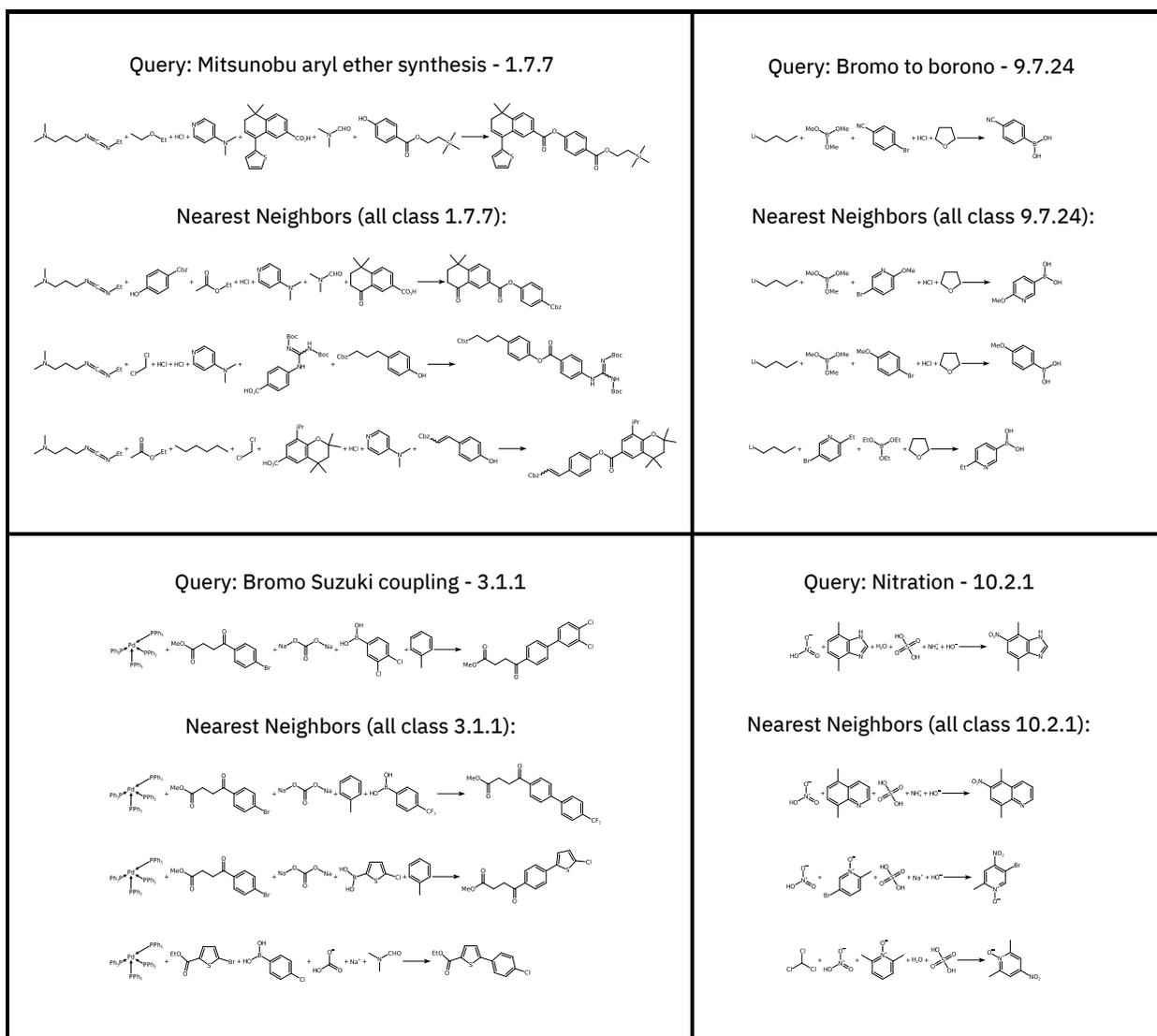


Figure 4: Four examples of reaction SMILES queries, retrieving the three nearest neighbors in the LSHforest²⁹ of the training set containing 2.4M reactions. All the retrieved reactions belong to the same reaction class as the query reaction and show similar precursors.

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 “Unrecognized”, 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 "Unrecognized". Functional group interconversions seem to be more difficult to recover.

4 Conclusion

In this work, we focused on the data-driven classification of chemical reactions with natural language processing methods and on the use of side product information to design a reaction fingerprint. 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%. Out of the 1.8% of incorrect predictions, 1.2% are linked to the "Unrecognized" class of the underlying database. Our model is 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.

To the best of our knowledge, this is the first work that applied a BERT-like pretraining¹⁷ to chemical reactions. We showed that the embeddings computed by our BERT classifier could be used as reaction fingerprints. Those BERT 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.

5 Acknowledgement

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

References

- (1) RSC's RXNO Ontology. <http://www.rsc.org/ontologies/RXNO/index.asp>, (Accessed Sep 13, 2019).
- (2) Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E. The 'wired' universe of organic chemistry. *Nature Chemistry* **2009**, *1*, 31–36.
- (3) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.
- (4) 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.
- (5) 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*.
- (6) IBM RXN for Chemistry. <https://rxn.res.ibm.com>, (Accessed Sep 13, 2019).
- (7) Nextmove Software NameRXN. <http://www.nextmovesoftware.com/namerxn.html>, (Accessed Jul 29, 2019).
- (8) 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.

- (9) Daylight Theory Manual, Chapter 5. <http://www.daylight.com/dayhtml/doc/theory/theory.smirks.htm> (accessed May 25, 2014).
- (10) 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.
- (11) 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.
- (12) 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.
- (13) 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.
- (14) Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017; pp 427–431.
- (15) Probst, D.; Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *Journal of cheminformatics* **2018**, *10*, 66.
- (16) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems. 2017; pp 5998–6008.

- (17) 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.
- (18) Nextmove Software Pistachio. <http://www.nextmovesoftware.com/pistachio.html>, (Accessed Jul 29, 2019).
- (19) ChemAxon. <https://docs.chemaxon.com/display/ltsargon/Reaction+fingerprint+RF>, (Accessed Dec 21, 2019).
- (20) 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.
- (21) 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.
- (22) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS central science* **2016**, *2*, 725–732.
- (23) Landrum, G. et al. rdkit/rdkit: 2019_03_4 (Q1 2019) Release. 2019; <https://doi.org/10.5281/zenodo.3366468>.
- (24) 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.
- (25) 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.

- (26) 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.
- (27) Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* **2016**,
- (28) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.
- (29) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *arXiv preprint arXiv:1908.10410* **2019**,
- (30) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proc. ACL. 2017.
- (31) BERT code. <https://github.com/google-research/bert#sentence-and-sentence-pair-classification-tasks>, (Accessed Oct 15, 2019).
- (32) Paszke, A. et al. *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc., 2019; pp 8024–8035.
- (33) 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*.
- (34) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, 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.
- (35) Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2017**, *34*, 1433–1435.

- (36) 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**,