Data-driven Reaction Template Fingerprints

Anubhab Chakraborty
IBM Research Europe - Zurich
Rueschlikon, Switzerland
anubhab.chakraborty@ibm.com

Amol Thakkar∗
IBM Research Europe - Zurich
Rueschlikon, Switzerland
tha@zurich.ibm.com

Alain C. Vaucher∗
IBM Research Europe - Zurich
Rueschlikon, Switzerland
ava@zurich.ibm.com

Teodoro Laino∗
IBM Research Europe - Zurich
Rueschlikon, Switzerland
teo@zurich.ibm.com

Abstract

Chemical reactions can be classified into distinct categories that encapsulate concepts for how one molecule is transformed into another. One can encode these concepts in rules specifying the set of atoms and bonds that change during a transformation, which is commonly known as a reaction template. While there exist multiple possibilities to represent a chemical reaction in a vector representation, or fingerprint, this is not the case for reaction templates. As a consequence, methods to navigate the space of reaction templates are limited. In this work, we introduce the first reaction template fingerprint. To this end, we follow a data-driven approach relying on a masked language modelling task on SMIRKS strings. We combine unsupervised pre-training with fine-tuning on the classification of templates according to the RXNO ontology, for which we achieve up to 98.4% classification accuracy. We highlight how the learned embeddings can be extracted and used in downstream applications.

1 Introduction

In recent years the application of Natural Language Processing (NLP) methods, particularly using transformer architectures, have shown to be effective in learning the language of chemistry [1]. This has given rise to chemical language modelling (CLM), which has been successfully applied to the tasks of de novo molecular generation [2], the prediction of chemical reactions [3, 4] and experimental procedures [5], and molecular property prediction [6]. Together these form the basis of hypothesis generation for design, make, test, and analyse cycles for accelerating scientific discovery.

Chemical reactions consist of a set of reacting molecules that are transformed into a set of product molecules. Each molecule can be represented by a text-based string using the SMILES (Simplified Molecular Input Line Entry System) language [7], and by extension reaction SMILES can be used for the textual representation of reactions (Figure 1). However, the reaction can be represented at a more granular level and described in terms of the set of atoms and bonds that change during a transformation, which is commonly known as a reaction template. Because the reaction templates encode the core chemical transformation, they are widely used for reaction classification and synthesis planning tasks [8, 9, 10, 11, 12]. The reaction template is specified using an extension of the SMILES language, known as SMIRKS, that encodes atom and bond primitives, connected via a series of logical expressions, with an internally consistent atom-mapping (Appendix A.1) [13]. However, the added complexity of SMIRKS in comparison to SMILES makes them difficult to create and interpret.

∗National Center for Competence in Research-Catalysis (NCCR-Catalysis), Switzerland

Dolfus et al. have recently developed a tool that aids interpretation through visualisation [14], and the automated creation of reaction templates has been addressed by cheminformatics packages that implement automated extraction of reaction templates from a set of reaction SMILES [12, 15, 16]. Herein, we address the challenge of vectorizing reaction templates, by learning a data-driven reaction template fingerprint. In contrast to reaction SMILES, for which a variety of vector representations already enable similarity searching, clustering, and modelling [17, 18, 19], no such method exists for reaction templates. We build upon the creation of reaction fingerprints from reaction SMILES [17], to create reaction template fingerprints. We show that transformer models can be trained on SMIRKS, to predict the reaction classes from reaction templates with up to 98.4% classification accuracy. Furthermore, we show how the learned embeddings can be extracted for potential downstream applications, such as template similarity search, development of new retrosynthesis schemes, and the comparison of chemical reaction datasets.

2 Methods

2.1 Dataset

The Pistachio dataset (version 2022Q1) from NextMove software [20] contains reactions in the SMILES notation. These were atom-mapped using RXNMapper [1] and filtered to a subset of 1.9M reactions using a preprocessing pipeline [21] to remove reactions failing to meet the following criteria: one product; between 2 and 10 reactants; up to 300 reactant tokens; up to 200 product tokens; successful extraction of a reaction template. Reaction templates were extracted from the reaction SMILES using RDChiral [12]. The reactions were classified using NameRXN’s hierarchical classification scheme [22, 23, 24], where the top level denotes the superclass, the middle level denotes the reaction category, and the final level is the named reaction. The database contained ~20% ‘unrecognized’ reaction classes denoted by class code 0.0. The data-set was split into training, validation, and test sets in a 90:5:5 ratio respectively and tokenized in two stages (Appendix A.1).

2.2 Model

Model training was done in two steps. In the first step, we carried out pre-training of the BERT [25] model on the reaction templates using masked language modelling. In the second step, we fine-tuned the model for the task of reaction class prediction.

The transformers library [26] was used to train all models in this study. The BERTForMaskedLM pre-defined model was used for pre-training with the following parameters: hidden size of 256, an
intermediate size of 512 and 4 attention heads. The remaining parameters were unchanged from the reference implementation [25]. During pre-training, tokens in the SMIRKS sequence were masked with a 0.15 probability using the special [MASK] token. Additionally, a [CLS] token was prepended to the beginning of the input sequence. Since the [CLS] is never masked, the embedding of the [CLS] token encodes a global representation of the entire template sequence [25]. The pre-training was done for 270,000 steps over 5 epochs with a batch size of 32 and a learning rate of $10^{-4}$ on two V100 GPUs for 11 hours. The pre-trained model, evaluated on the validation data-set, predicted 99.6% of the masked tokens correctly.

We fine-tuned the pre-trained model on the reaction classification task using the pre-defined BERTForSequenceClassification model for 22 hours over 10 epochs on one V100 GPU at a learning rate of $10^{-5}$.

3 Results and Discussion

3.1 Template Classification

We fine-tuned two models for each of the three levels classification hierarchy, one including the templates with ‘unrecognized’ (0.0) class labels and one without, for a total of six models. We tested the performance of each of the models at its target classification task. Additionally, the ability of the model trained in predicting the full classification hierarchy to generalize for the broader levels of classification was tested using the testing scheme described in A.2. The results are shown in Table 1.

Table 1: Classification results for the six trained models, at their trained hierarchy level. The model trained on the full classification hierarchy showed similar performance to the models trained on the higher levels of hierarchy. Additionally removing unclassified datapoints, improved the performance of the models at the higher levels of classification.

<table>
<thead>
<tr>
<th>Model trained on</th>
<th>Model tested on</th>
<th>Total samples</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full classification hierarchy</td>
<td>Full class</td>
<td>97,197</td>
<td>89.7</td>
</tr>
<tr>
<td></td>
<td>top 2 levels</td>
<td>97,197</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>top level</td>
<td>97,197</td>
<td>95.7</td>
</tr>
<tr>
<td>Full classification hierarchy excluding 0.0s</td>
<td>Full class</td>
<td>76,528</td>
<td>89.3</td>
</tr>
<tr>
<td>top 2 levels of reaction classes</td>
<td>top 2 levels</td>
<td>97,197</td>
<td>95.5</td>
</tr>
<tr>
<td>top 2 levels of reaction classes excluding 0.0s</td>
<td>top 2 levels</td>
<td>76,528</td>
<td>97.8</td>
</tr>
<tr>
<td>top level of reaction classes</td>
<td>top level</td>
<td>97,197</td>
<td>96.0</td>
</tr>
<tr>
<td>top level of reaction classes excluding 0.0s</td>
<td>top level</td>
<td>76,528</td>
<td>98.4</td>
</tr>
</tbody>
</table>

The models showed a general trend of higher accuracy for broader classification tasks. This is expected to a certain extent as reactions can be classified based on the reagents used. As a result, different reactions undergoing the same transformation but using different reagents may have identical reaction templates but different reaction classes. This is validated by our observation that the prediction accuracy of the model trained and tested on the full classification hierarchy was measured to be 89.7%. However, when tested at broader classification levels, the prediction accuracy of the same model jumped to above 95%, nearly matching the performance of the models trained and tested exclusively on the broad classification levels (Table 1).

To gauge the influence of the ‘unrecognized’ datapoints on the model training, we compared the weighted average of the accuracy per class (excluding the ‘unrecognized’ class) of the model, trained on the dataset with the ‘unrecognized’ classes to the overall accuracy of the model trained on the dataset without the ‘unrecognized’ class. The weighted average of the accuracy per class excluding the ‘unrecognized’ class for superclasses was measured to be 96.6% (See Table 2 in the Appendix). Interestingly, the model trained on the dataset with ‘unrecognized’ datapoints removed, saw a non-trivial increase in performance for the remaining classes resulting in an overall accuracy of 98.4% at the broadest level of classification. However this performance gain shrunk when moving to more granular classification levels, ultimately showing no practical improvement in accuracy when tested on the full classification level.
3.2 Mapping the chemical transformation space

To show that the model has learned to partition the reaction transform space for the task of reaction classification, we used TMAP [27] to create a visualisation using the learned embeddings shown in Figure 2. The figure shows that the model has learnt representations which clusters in the vector space based on the reaction types. The embeddings were extracted from the BERT model which uses the [CLS] token to encode a global representation of the template. Thus, we could obtain a data-driven fingerprint that is able to encode chemical transformations given as SMIRKS patterns representing reaction templates.

![Figure 2: Template fingerprint space visualised using a TMAP for dimensionality reduction. Each point in the TMAP represents a fingerprint of an unique chemical transformation generated using the fully trained model. The points are coloured by their ground truth superclass, to reveal the clustering of similar transformations.](image)

3.3 Applications

Having the possibility, for the first time, to represent reaction templates as vectors, opens several research directions, which we will be studying in further work. In addition to the possibility to plot and navigate reaction templates in two dimensions, as illustrated in the previous section, the new fingerprint provides a straightforward way to compare templates and calculate their similarity (bypassing the need for substructure matching algorithms). It will make comparisons of reaction datasets more pertinent and succinct, by automatically narrowing down redundant reactions to their core transformation. Furthermore, for retrosynthesis prediction the fingerprint will provide a method for curating the underlying template libraries, and enhance performance for single-step retrosynthesis models as shown by Baylon et. al. [28].

4 Conclusion

Using a novel tokenization scheme, we trained a BERT model to learn the language of reaction templates with up to 99.6% accuracy. We then fine-tuned the model on the reaction classification task with up to 98.4% accuracy. We use this result as a demonstration of BERT’s ability to generate meaningful vector representations or ‘fingerprints’ of reaction templates. These fingerprints, first of its kind, enable the efficient use of reaction templates in a number of downstream applications previously deemed impractical.
Acknowledgments and Disclosure of Funding

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References


A Appendix

A.1 SMIRKS tokenization

The overall structure of an example SMIRKS string is shown in Figure 3. Each square bracket \([N\ldots]\) specifies an individual atom along with its properties and atom mapping. The properties of the atom inside the square bracket is concatenated with conditional statements. The bonds are represented with -, = and # for single, double and triple bond respectively.

For meaningful tokenization, we grouped together letters which represented certain properties into one token. For example, D and 3 by themselves do not mean anything in the SMIRKS string, but when used together D3 meant that the atom had a degree of 3, that is, the atom is bonded to three other atoms excluding hydrogen.
Figure 3: Top: Structure of a SMIRKS string. Rather than tokenizing the string by individual letters, we chose to keep letters having independent meaning, together, as a single token. Bottom: Tokenized SMIRKS string.

A.2 Note on classification analysis

For testing the model at multiple levels of the hierarchy, we compared the trimmed reaction class string. For example, suppose that for a given template the ground truth reaction class is 9.1.1 and that the model predicts the reaction class to be 9.1.5, then we consider the model to have made the correct prediction for the category 9.1 and the superclass 9, even though it predicted the reaction class wrong.

A.3 Classification results per class

Table 2 shows that the model has an overall consistent behaviour across classes, except for a lower accuracy in heterocycle formation reactions.

Table 2: Accuracy of the classification model on reaction templates from different superfamilies. The samples were taken from the test split and inference was performed using the model trained on the task of superclass prediction.

<table>
<thead>
<tr>
<th>Superclass</th>
<th>Total samples</th>
<th>Fraction of data-set (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>'unrecognized'</td>
<td>20,525</td>
<td>21.1</td>
<td>93.1</td>
</tr>
<tr>
<td>Heteroatom alkylation and arylation</td>
<td>19,857</td>
<td>20.4</td>
<td>97.5</td>
</tr>
<tr>
<td>Acylation and related processed</td>
<td>16,823</td>
<td>17.3</td>
<td>97.0</td>
</tr>
<tr>
<td>C-C bond formation</td>
<td>9,224</td>
<td>9.5</td>
<td>96.6</td>
</tr>
<tr>
<td>Heterocycle formation</td>
<td>3,077</td>
<td>3.2</td>
<td>88.1</td>
</tr>
<tr>
<td>Protections</td>
<td>1,312</td>
<td>1.3</td>
<td>91.4</td>
</tr>
<tr>
<td>Deprotections</td>
<td>10,501</td>
<td>10.8</td>
<td>98.8</td>
</tr>
<tr>
<td>Reductions</td>
<td>4,418</td>
<td>4.5</td>
<td>98.5</td>
</tr>
<tr>
<td>Oxidations</td>
<td>1,974</td>
<td>2.0</td>
<td>97.0</td>
</tr>
<tr>
<td>Functional group interconversion</td>
<td>6,972</td>
<td>7.2</td>
<td>96.4</td>
</tr>
<tr>
<td>Functional group addition</td>
<td>2,323</td>
<td>2.4</td>
<td>97.0</td>
</tr>
<tr>
<td>Resolution</td>
<td>191</td>
<td>0.2</td>
<td>95.8</td>
</tr>
</tbody>
</table>