Unsupervised Attention-Guided Atom-Mapping

Philippe Schwaller 1,2  Benjamin Hoover 3  Jean-Louis Reymond 2  Hendrik Strobelt 3  Teodoro Laino 1

Abstract

Knowing how atoms rearrange during a chemical transformation is fundamental to numerous applications aiming to accelerate organic synthesis and molecular discovery. This labelling is known as atom-mapping and is an NP-hard problem. Current solutions use a combination of graph-theoretical approaches, heuristics, and rule-based systems. Unfortunately, the existing mappings and algorithms are often prone to errors and quality issues, which limit the effectiveness of supervised approaches. Self-supervised Transformers, on the other hand, have recently shown tremendous potential when applied to textual representations of different domain-specific data, such as chemical reactions. We used an interactive and visual tool to explore the attention weights learned by a Transformer trained on unlabeled chemical reaction data, and discovered that it encoded atom rearrangement information between the products and reactants. We use this information to build a chemically agnostic attention-guided reaction mapper that shows a remarkable performance in terms of accuracy and speed, even for strongly imbalanced reactions. Our work suggests that unannotated collections of chemical reactions contain all the relevant information needed to construct coherent sets of reaction rules and emphasizes the importance of interpretability methods for learned weights and representations. The finding provides the missing link between data-driven and rule-based approaches for numerous chemical reaction tasks.

1. Introduction

The principle of mass conservation states that mass is conserved within an isolated system. In low energy regimes like chemical reactions, this means that every atom in the products has a unique counterpart in the reactants. This match is called atom-mapping and is crucial for numerous tasks like template-based reaction prediction (Coley et al., 2017; Segler & Waller, 2017) and retrosynthesis planning methods (Segler et al., 2018; Thakkar et al., 2020; Fortunato et al., 2020), reaction graph neural network algorithms (Jin et al., 2017; Coley et al., 2019), reactant-reagent role assignments (Schneider et al., 2016), reaction rules extraction (Coley et al., 2017; Segler et al., 2018), identification of metabolic pathways (Rahman et al., 2014), and knowledge extraction from reaction databases (Chen et al., 2002). The better the atom-mapping, the better the downstream models that depend on it.

Because of the impracticality of manually assigning atom-mapping, automatic algorithms to approximate solutions for the underlying NP-hard problem have been developed since the 1970s (Chen et al., 2013; Gonzalez et al., 2017). Most of the available approaches are either structure-based (Lynch & Willett, 1978; Moock et al., 1988; Vleduts, 1977; McGregor & Willett, 1981; Funatsu et al., 1988; Körner & Apostolakis, 2008; Apostolakis et al., 2008) or optimisation-based (Jochum et al., 1980; Akutsu, 2004; Crabtree & Mehta, 2009; First et al., 2012; Latendresse et al., 2012). The current state-of-the-art is a combination of heuristics, a set of expert-curated rules that precompute candidates for complex reactions, and a graph-theoretical algorithm to generate the final mapping as developed by Jaworski et al. (2019). Nonetheless, complex preprocessing steps, computationally intensive strategies, and the need for expert-curated rules hinder its wider adoption. Applications requiring properly mapped reactions currently rely on more popular alternatives based on expert-curated rule-based methods (NameRxn; Indigo). Combined with expert annotations, these tools could be used to supervise the training of a data-driven model, but the performance of this model would be inherently limited to the quality of the labels used. Therefore, developing methodologies to extract hidden atom-mapping information from unsupervised models trained on unlabeled data remains a key objective.

1IBM Research Europe, CH-8803 Rüschlikon, Switzerland
2Department of Chemistry and Biochemistry, University of Bern, Switzerland
3IBM Research – Cambridge / MIT-IBM Lab, United States. Correspondence to: Philippe Schwaller <phs@zurich.ibm.com>.

ML Interpretability for Scientific Discovery Workshop of the 37th International Conference on Machine Learning, Vienna, Austria, 2020. Copyright 2020 by the authors.
Neural networks, and in particular natural language processing (NLP) models (Oztürk et al., 2020), have recently had a significant impact on synthetic chemistry (Almeida et al., 2019). NLP models encode latent knowledge from a training set of molecules and reactions encoded as text (SMILES (Weininger, 1988)) without needing to embed chemical rules. Molecular transformer models, for instance, are the state-of-the-art for forward reaction prediction tasks, achieving an accuracy higher than 90% (Vaswani et al., 2017; Schwaller et al., 2019a; Schwaller & Laino, 2019). This impressive performance is likely due to learned representations in the model’s architecture that capture characteristic reaction data patterns. Unfortunately, extracting scientifically meaningful insights and explaining predictions of these models is difficult and remains largely unexplored.

Using a visual exploration tool, we discovered evidence that atom-mapping is learned as a key signal in Transformer models trained on unmapped reactions on the self-supervised task of predicting the randomly masked parts in a reaction sequence (Devlin et al., 2018; Lan et al., 2020). We also show that Transformer architectures can learn the underlying structure of chemical reactions without any human labelling or supervision, solely based on atom-wise tokenisation of a large data set of reaction SMILES (Weininger, 1988). After establishing an attention-guided atom-mapper and introducing a neighbour attention multiplier, we were able to achieve 99.4% of correct full atom-mappings on a test set of 49k strongly unbalanced patent reactions (Schneider et al., 2016). We are making available the reaction mapper (RXNMapper), which can handle stereochemistry and unbalanced reactions, and the public data set of Lowe (2017) annotated with RXNMapper, hoping that both contributions will have an impact on all applications that build on top of atom-mapping. This completely unsupervised approach to atom-mapping links data-driven approaches to traditional rule-based systems, demonstrating how a consistent set of atom-mapping rules is a latent component within large data sets of chemical reactions.

2. Attention-guided chemical reaction mapping

Self-attention is the major component of algorithms that are setting new records on NLP benchmarks (e.g., BERT, ALBERT, and GPT-2) (Devlin et al., 2018; Lan et al., 2020; Radford et al., 2019), and even creating breakthroughs in the chemical domain (Schwaller et al., 2019b). Transformer models use self-attention across multiple layers to learn a contextual representation of each token (e.g., each atom and bond in a reaction SMILES) from all the tokens in the same input. Each layer may consist of multiple self-attention modules, called heads, each learning to attend to the inputs independently. When applied to chemical reactions, Transformers use attention mechanisms to focus on atoms relevant to understanding the molecular structure, describing the chemical transformation, and gathering latent information. These context-dependent atom representations have a high potential to encode much more information than could be manually done by a human expert. Fortunately, the internal attention mechanisms are intuitive to visualise and interpret using interactive tools (Hoover et al., 2019; Wiegrefe & Pinter, 2019; Vig, 2019). Figure 1 shows an example of the attention weights connecting an input sequence of SMILES tokens to itself. Visual analysis revealed the ability of some Transformer heads to learn distinct chemical features, where one specific head (Figure 1, Head 6) learned how to connect product atoms to reactant atoms, the process defined above as atom-mapping.

Throughout this work, our Transformer architecture of choice is ALBERT (Lan et al., 2020). ALBERT’s primary advantage over its predecessor BERT (Devlin et al., 2018) is that ALBERT shares network weights across layers during training. Not only does the weight sharing make the model smaller, but it also keeps the functionality learned by a particular head the same across layers and consistent across different inputs, as shown in Figure 2. Learned functions such as forward scanning and backward scanning of the sequence, focusing on non-atomic tokens (ring openings/closures), and atom-mapping all perform similarly, irrespective of the input. We performed a model parameter randomisation sanity check (Adebayo et al., 2018) by taking a freshly initialised model and analysing the attention weights. We could not observe any of the functionalities and
therefore, confirmed that the model learned the functionalities during the self-supervised training. Moreover, every transformer model we trained on chemical reactions showed the atom-mapping signal in at least one head.

![Figure 2. Comparison of the atom-mapping signal in a BERT and ALBERT model.](image)

#### 2.1. Interpreting Transformers

Transformers are a class of deep neural network architectures that rely on multiple and sequential applications of self-attention layers (Vaswani et al., 2017). These layers are composed of one or more heads, each of which learns a square attention matrix $A \in \mathbb{R}^{N \times N}$ of weights that connect each token’s embedding $Y_i$ in an input sequence $Y$ of length $N$ to every other token’s embedding $Y_j$. Thus, each element $A_{ij}$ is the attention weight connecting $Y_i$ to $Y_j$. This formulation makes the attention weights in the Transformer architecture amenable to visualisations as the curves connecting an input sequence to itself, shown in Figure 1, where a thicker, darker line indicates a higher attention value.

The calculation of the attention matrix of each head can be easily interpreted as a probabilistic hashmap or lookup table over all other elements $Y_j$. Each head in a self-attention layer will first convert the vector representation of every token $Y_i$ into a key, query, and value vector using the following operations:

$$
K_i = W_k Y_i \quad Q_i = W_q Y_i \quad V_i = W_v Y_i \tag{1}
$$

where $W_k \in \mathbb{R}^{d_k \times d_e}$, $W_q \in \mathbb{R}^{d_q \times d_e}$, and $W_v \in \mathbb{R}^{d_v \times d_e}$ are learnable parameters. $A_i$, or the vector of attention out of token $Y_i$, is then a discrete probability distribution over the other input tokens, and it is calculated by taking a dot product over that token’s query vector and every other token’s key vector followed by a softmax to convert the information into probabilities:

$$
A_i = \text{softmax}\left(\frac{Q_i (W_k Y_i^\top)}{\sqrt{d_k}}\right). \tag{2}
$$

Note that one can define input sequence $Y$ as an $N \times d_e$ matrix and matrix $W_k$ as a $d_k \times d_e$ matrix, where $d_e$ is the embedding dimension of each token and $d_k$ is the embedding dimension shared by the query and the key.

Each head must learn a unique function to accomplish the masked language modeling task, and some of these functions are inherently interpretable to the domain of the data. For example, in Natural Language Processing (NLP), it has been shown that certain heads learn dependency and part of speech relationships between words (Vig & Belinkov, 2019; Clark et al., 2019). Interactive visual tools can help exploring and interpreting these learned functions (Hoover et al., 2019). Such a visual inspection lead to the formulation of our hypothesis that an atom mapping signal might be present in some attention heads.

#### 2.2. From raw attention to atom-mapping

To benchmark the performance of the learned atom-mapping signal, we needed to extract a single atom-map from the bidirectional attention signal of an atom-mapping head into a product-to-reactant atom-map. This qualification ensures that each atom in the products corresponds to an atom in the reactants. It is an important definition given that the most sizable open-source reaction data sets (Lowe, 2012; 2017) report only major products and show reactions that have fewer product atoms than reactant atoms.

The product atoms are mapped to reactant atoms one at a time, starting with product atoms that have the largest attention to an identical atom in the reactants. At each step, we introduce a neighbour attention multiplier that increases the attention connection from adjacent atoms of the newly mapped product atom to adjacent atoms of the newly mapped reactant atom, boosting the likelihood of an atom having the same adjacent atoms in reactants and products. This process continues until all product atoms are mapped to corresponding reactant atoms.

We selected the best performing modellayer/head combination after evaluating them on a curated set of 1k patent reactions originally mapped with the rule-based NameRXN tool (Schneider et al., 2016; NameRXN). We consider the atom maps in NameRXN (NameRXN) to be of high quality because they are a side product of matched reaction rules designed by human experts. We will refer to the best ALBERT model configuration (8 heads, layer 11, head 6 and multiplier 90) as RXNMapper.

#### 2.3. Atom-mapping evaluation

The predominant use case for atom-mapping algorithms is to map heavily imbalanced reactions, such as those in patent reaction data sets (Lowe, 2012; 2017) or those predicted by data-driven reaction prediction models (Schwaller et al., 2019a). After training the RXNMapper model on unmapped reactions (Lowe, 2017), we investigated the chemical knowl-
edge our model had extracted by comparing our predicted atom maps to a set of 49k patent reactions by Schneider et al. (2016) with high-quality atom-maps. Impressively, the majority (96.8%) of the atom-mappings matched, including methylene transfers, epoxidations and Diels-Alder reactions (Figure 3). We manually annotated the remaining discrepancies to discover edge cases where RXNMapper seemingly failed. Out of the 1551 non-matching reactions, we only found 284 incorrect predictions by our model. In 415 reactions, our atom-mapping was equivalent to the original atom maps (e.g., tautomers), and in 436, the atom-mapping generated by RXNMapper was even better. In 369 cases, the original reaction was questionable and likely wrongly extracted from patents. For 47 reactions, the key reagents to determine the reaction mechanisms were missing.

Among the most frequent failures of RXNMapper, we find examples of wrong atom ordering in rings and azide compounds (Figure 3, (d)). In other failure cases the only difference is one oxygen atom, like in reductions where the model predicts the wrong oxygen atom to leave (Figure 3, (e)), or in Mitsunobu reactions (Figure 3, (f)), where the phenolic oxygen should become part of the product, but the model maps the primary or secondary alcohol instead. We also observed counterexamples of Mitsunobu reactions (Figure 3, (c)) for which our model correctly mapped the reacting oxygen while the rule-based maps contained the wrong mapping as a result of the reaction not matching the Mitsunobu reaction rule. Human-made rules are inflexible and therefore extremely brittle. Using RXNMapper, we were able to identify important limitations in the rules-based annotated ground truth. RXNMapper correctly assigned primary alcohols to be part of the major product for esterification reactions (Figure 3, (a)) like Fischer-Speier and Steglich esterifications, our model correctly favoured anhydrides (Figure 3, (b)) and peroxides as reactants in acylation and oxidation reactions. We have extracted this attention information to develop an attention-guided reaction mapper that exhibits remarkable performance in both speed and accuracy across a wide distribution of reaction classes, especially for strongly imbalanced reactions that are otherwise difficult to handle using existing methods. In contrast to earlier work, our purely data-driven approach can create a state-of-the-art atom-mapping tool within two days of training without the need for tedious and potentially biased expert encoding or curation. Because the approach is completely unsupervised, adding specific reaction datasets can further improve the atom-mapping performance for corner cases. Finally, our work provides the first evidence that unannotated collections of chemical reactions contain all the relevant information necessary to construct a coherent set of atom-mapping rules.

### References


### Table 1. Comparison of RXNMapper (ours), Indigo Toolkit (Indigo), and Mappet (Jaworski et al., 2019).

<table>
<thead>
<tr>
<th></th>
<th>RXNMapper</th>
<th>Indigo</th>
<th>Mappet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg time (short)</td>
<td>6.4 ms</td>
<td>17.0 ms</td>
<td>Slower than Indigo</td>
</tr>
<tr>
<td>Avg time (strongly unbalanced)</td>
<td>7.7 ms</td>
<td>&gt;4000 ms</td>
<td>Not handled</td>
</tr>
<tr>
<td>Quality on complex reactions</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Quality on strongly unbalanced reactions</td>
<td>High</td>
<td>Low</td>
<td>–</td>
</tr>
<tr>
<td>Open Source code?</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

3. Discussion

Our work demonstrates the importance of using visual tools to explore how models understand their training data. We have shown that self-supervised attention-based language models can learn atom rearrangements between products and reactants/reagents. We have extracted this attention information to develop an attention-guided reaction mapper that exhibits remarkable performance in both speed and accuracy across a wide distribution of reaction classes, especially for strongly imbalanced reactions that are otherwise difficult to handle using existing methods. In contrast to earlier work, our purely data-driven approach can create a state-of-the-art atom-mapping tool within two days of training without the need for tedious and potentially biased expert encoding or curation. Because the approach is completely unsupervised, adding specific reaction datasets can further improve the atom-mapping performance for corner cases. Finally, our work provides the first evidence that unannotated collections of chemical reactions contain all the relevant information necessary to construct a coherent set of atom-mapping rules.
Figure 3. Top: Analysis of the 49k Schneider test set. Visualizing the results on the whole set and with focus on the mis-matched atom-mappings (1.5k matches for context) using reaction TMAPs (Probst & Reymond, 2020; Schwaller et al., 2019b). Bottom: Examples of atom-mappings generated by RXNMapper, where there was a mis-match. Reactants and reagents were not separated in the inputs.

**Interdisciplinary Reviews: Computational Molecular Science, 3(6):560–593, 2013.**


Moock, T., Nourse, J., Grier, D., and Hounshell, W. The implementation of aam and related reaction features in the reaction access system (reaccs), 1988.


