Unsupervised attention-guided atom-mapping

Philippe Schwaller,[*,†,‡] Benjamin Hoover,[¶] Jean-Louis Reymond,[‡] Hendrik Strobelt,[¶] and Teodoro Laino[†]

†*IBM Research Europe, CH-8803 Rüschlikon, Switzerland*

‡*Department of Chemistry and Biochemistry, University of Bern, Switzerland*

¶*IBM Research – Cambridge / MIT-IBM Lab, United States*

E-mail: phs@zurich.ibm.com

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 neural networks called Transformers, on the other hand, have recently shown tremendous potential when applied to textual representations of different domain-specific data, such as chemical reactions. Here we demonstrate that attention weights learned by a Transformer, without supervision or human labelling, encode atom rearrangement information between products and reactants. We 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 to construct coherent sets of reaction rules. This finding provides the missing link between data-driven and

rule-based approaches and will stimulate machine-assisted discovery in the chemical domain.

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[1,2] and retrosynthesis planning methods,[3–5] reaction graph neural network algorithms,[6,7] reactant-reagent role assignments,[8] reaction rules extraction,[1,3] identification of metabolic pathways,[9] and knowledge extraction from reaction databases.[10] 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.[11,12] Most of the available approaches are either structure-based[13–19] or optimisation-based.[20–24] 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.[25]. 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.[26,27] The renewed interest in data-driven algorithms and the use of a specific ground truth could lead to models trained explicitly for atom-mapping tasks. This approach would inherently rely on having either experts or rule-based systems annotating large data sets with the potential to ideally achieve the same annotation quality. Therefore, a key objective is to develop methodologies to extract hidden atom-mapping information from unlabelled data.

Neural networks, and in particular natural language processing (NLP) models,[28] have recently had a significant impact on synthetic chemistry.[29] NLP models encode latent knowledge from a training set of molecules and reactions encoded as text (SMILES[30]) without needing to embed chemical rules. Molecular Transformer models, a recent addition to the NLP family, are the state-of-the-art for forward reaction prediction tasks, achieving an accuracy higher than 90%.[31–33] This impressive performance is likely due to learned representations in the model's architecture that capture characteristic reaction data patterns. Unfortunately, the lack of an explicit declarative knowledge representation makes it incredibly difficult to explain the predictions.

Here, we report 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.[34,35] 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.[30] 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.[8] We are making available the reaction mapper (RXNMapper), which can handle stereochemistry and unbalanced reactions, and the public data set of Lowe[36] 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.

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),[34,35,37] and even creating breakthroughs in the chemical domain.[38] 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.[39–41] 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.[35] ALBERT's primary advantage over its predecessor BERT[34] 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. Learned functions such as forward scanning and backward scanning of the sequence, focusing on non-atomic tokens (ring openings/closures), and atom-mapping (Figure 1 c) all perform similarly, irrespective of the input.

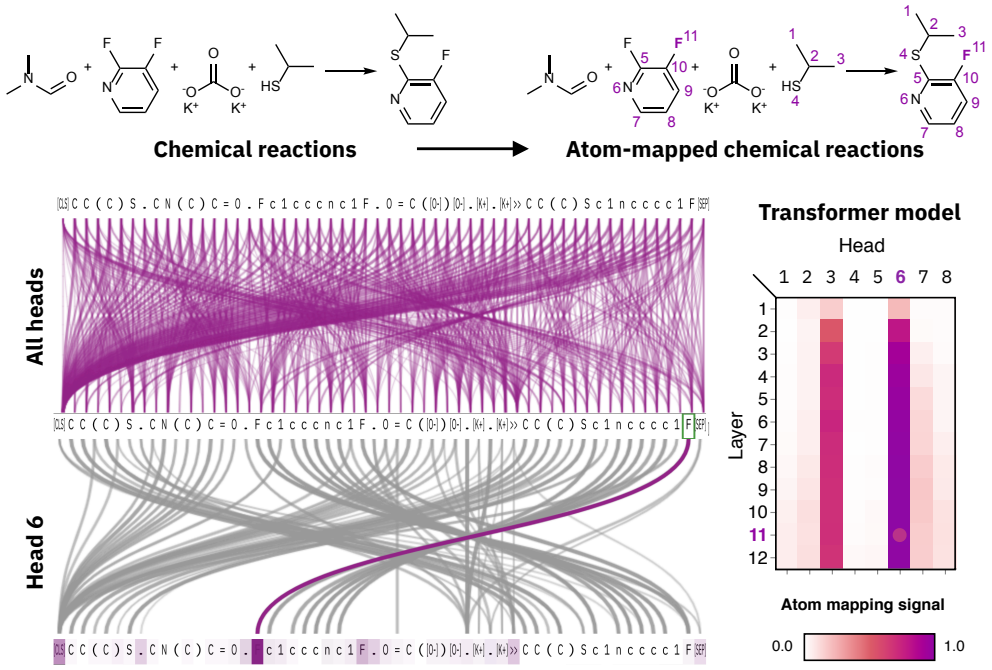


Figure 1: Visualisation of the inner workings of a transformer model for a given chemical reaction displayed on top. Superimposed curves connecting the input to itself represent the attention weights for a given reaction input for all heads (Layer 11) in the middle. The attention of a head that learned the atom-mapping mapping signal (Layer 11, Head 6) on the bottom. The atom-mapping signal measured per layer and head is displayed on the right.

From raw attention to atom-mapping

To quantify our observation, we developed an attention-guided algorithm that converts the bidirectional attention signal of an atom-mapping head into a product-to-reactant atom-mapping. 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[36,42] 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. Interestingly, the constraint of mapping only to equivalent atoms led to negligible improvements in terms of atom-mapping correctness, indicating that the model had already learned this rule in its atom-mapping function.

We selected the best performing model/layer/head combination after evaluating them on a curated set of 1k patent reactions originally mapped with the rule-based NameRXN tool.[8,26] We consider the atom maps in NameRXN[26] 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.

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[36,42] or those predicted by data-driven reaction prediction models.[32] After training the RXNMapper model on unmapped reactions,[36] we investigated the chemical knowledge our model had extracted by comparing our predicted atom maps to a set of 49k patent reactions by Schneider et al.[8] with high-quality atom-maps. Impressively, the majority (96.8%) of the atom-mappings matched, including methylene transfers, epoxidations and Diels-Alder reactions (Figure 2). 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 the RXNMapper, we find examples of wrong atom

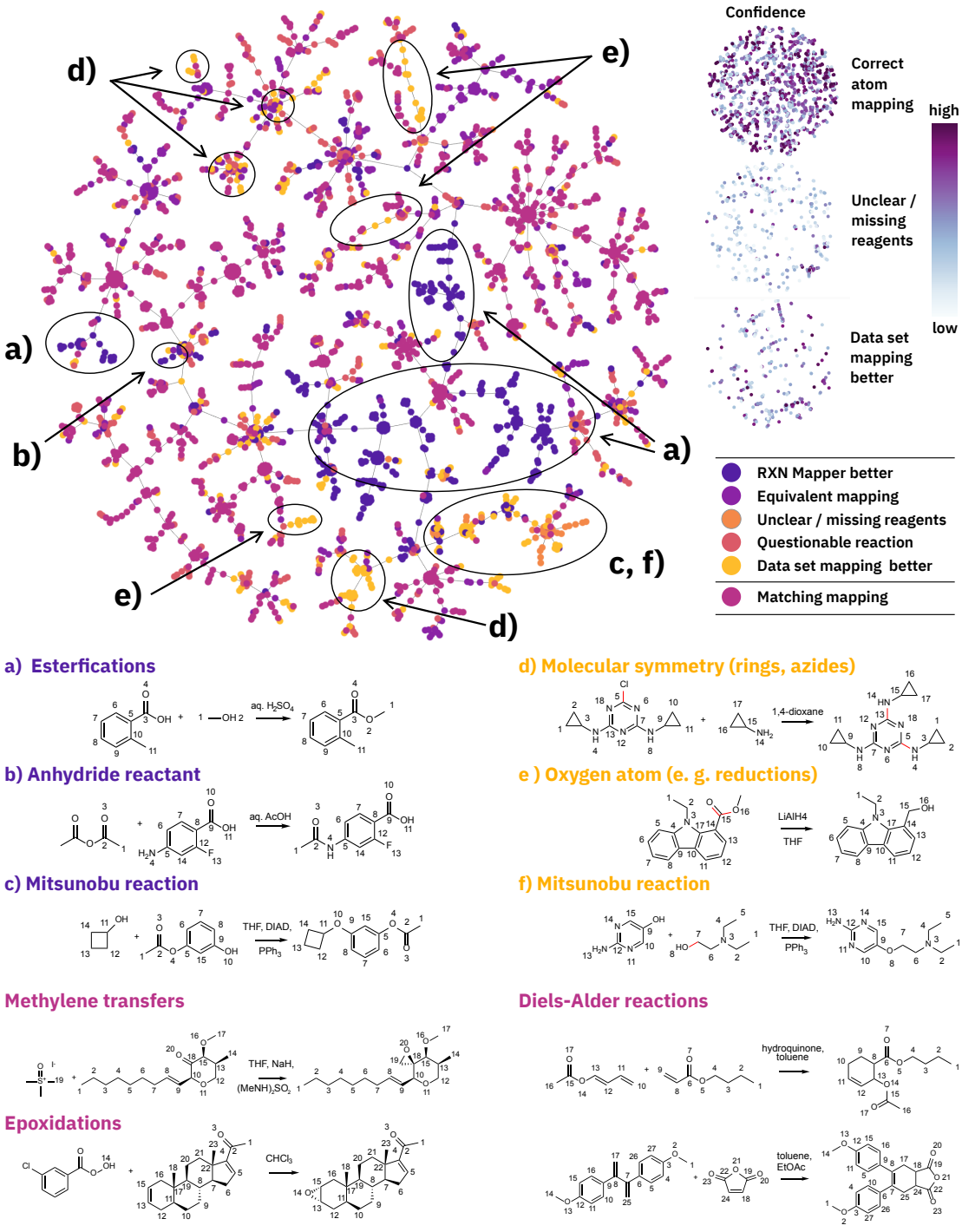


Figure 2: Reaction TreeMap[38,43] of the 1551 reactions, for which the predicted atom maps did not match the ground truth. Of the discrepancies, we found 851 reactions for which RXNMapper had generated an equal or better atom-mapping and only 284 reactions where the atom-mapping was incorrect. 1000 randomly selected matching atom-mappings place these discrepant reactions into a broader context.

ordering in rings and azide compounds (Figure 2, (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 2, (e)), or in Mitsunobu reactions (Figure 2, (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 2, (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 2, (a)) like Fischer-Speier and Steglich esterifications as opposed to the annotated ground truth. We also observed more subtle mistakes in the rules. For instance, our model correctly favoured anhydrides (Figure 2, (b)) and peroxides as reactants in acylation and oxidation reactions where the ground truth favoured formic acid and water. Moreover, it selected iodomethane over methanol (solvent) as the methylating reactant in Sandmeyer reactions with explicit copper-catalyst. The visualisations of the confidence scores for different categories of reactions in Figure 2 shows that wrongly predicted atom-mappings and those for unclear reactions are accompanied by lower model confidences than correctly predicted atom-mappings.

Table 1 provides results on the 49k patent test set. Overall, the generated atom-maps exactly match the original atom-maps in 96.9% of the cases. After removing questionable reactions from the statistics and counting the equivalent mappings as correct, the overall correctness increased to 99.4%. Table 1 shows the atom-mapping correctness divided into the different superclasses, where heterocycle formations were the most challenging superclass with 94.7% correctness.

Similar to Jaworski et al.[25], we analysed the atom-mapping in USPTO patent reactions according to the number of bond changes. RXNMapper performs better than Jaworski

Table 1: Results on the 49k patent test set

Reaction class	Total (curated)	Matching [%]	Correct [%]
Heteroatom alkylation and arylation	14836 (14698)	96.8	99.2
Acylation and related processes	11670 (11593)	95.7	99.8
C-C bond formation	5550 (5502)	98.0	99.4
Heterocycle formation	889 (881)	90.6	94.7
Protections	655 (652)	97.4	98.6
Deprotections	8055 (7983)	98.1	99.9
Reductions	4499 (4466)	97.6	99.1
Oxidations	809 (805)	98.0	99.9
Functional group interconversion (FGI)	1809 (1775)	96.2	99.8
Functional group addition (FGA)	228 (228)	89.0	99.1
All	49000 (48583)	96.8	99.4

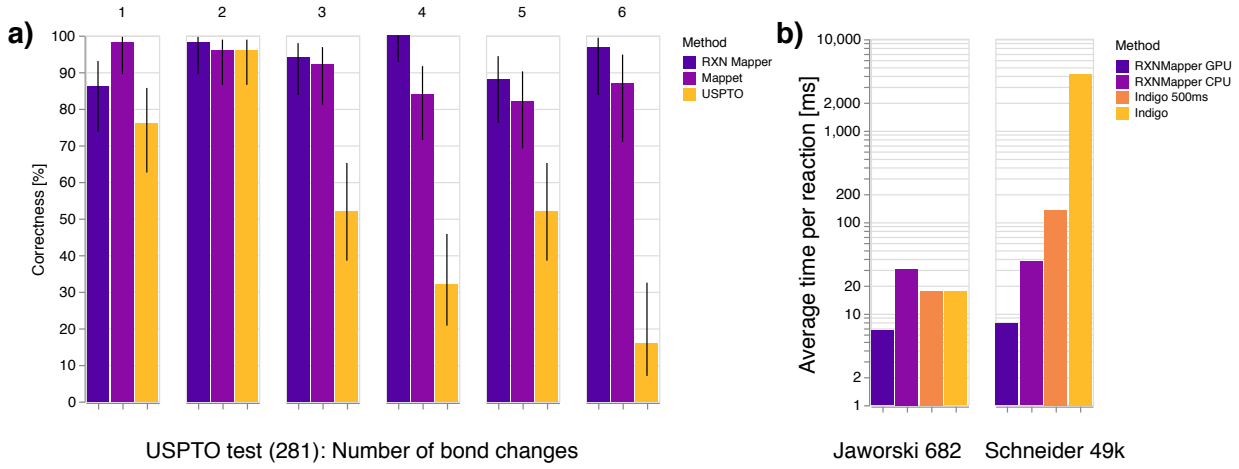


Figure 3: a) Atom-mapping correctness for different multipliers on the 1k validation set. b) Comparison of our RXNMapper, Mappet,[25] and a mapping from the USPTO data set (281 reactions). c) Mapping speed comparison between RXNMapper and Indigo,[27] which is faster than Mappet.[25] For Indigo 500ms, we set a time out of 500 ms, after which the tool would return an incomplete mapping. We averaged the timing on the imbalanced reactions for Indigo without timeout on 1000 reactions only.

et al.[25] on all reactions except for those involving only one bond change. With an average time to solution of 7.7 ms/reaction on GPU accelerators and 36.4 ms/reaction on CPU, RXNMapper's speed is similar to the Indigo toolkit[27] on balanced reactions and far exceeds Indigo on unbalanced ones. More information on the comparison and performance are available in the SI.

Discussion

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 from a Transformer model to develop an attention-guided reaction mapper that exhibits a remarkable performance in both speed and accuracy across a wide distribution of reaction classes. 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 entire approach is completely unsupervised, the use of specific reaction datasets can improve the atom-mapping performance on corner cases. Our approach is significantly faster and more effective, especially for strongly imbalanced reactions that are otherwise difficult to handle using existing methods. 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.

Outlook

The use of symbolic representations and the means to learn autonomously from rich chemical data led to the design of valuable assistants in chemical synthesis.[29] A strengthened trust between human and interpretable data-driven assistants will spark the next revolutions in chemistry, where domain patterns and knowledge can be easily extracted and explained from the inner architectures of trained models.

Data availability

All our generated atom-mappings, including those for the largest open-source patent data set,[36] the unmapped training, validation, and test set reactions, can be found in the following

repository `https://github.com/rxn4chemistry/rxnmapper`.

Code availability

The code is available at `https://github.com/rxn4chemistry/rxnmapper` and a demo at `http://rxnmapper.ai`.

Methods

Transformers

Transformers are a class of deep neural network architectures that relies on multiple and sequential applications of *self-attention* layers.[31] These layers are composed of one or more *heads*, each of which learns a square attention matrix $\mathbf{A} \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}}$ 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 $\mathbf{A_{ij}}$ 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:

$$K_i = \mathbf{W}_k Y_i \qquad Q_i = \mathbf{W}_q Y_i \qquad V_i = \mathbf{W}_v Y_i \tag{1}$$

where $W_k \in \mathbb{R}^{d_k \times d_e}$, $W_q \in \mathbb{R}^{d_k \times d_e}$, and $W_v \in \mathbb{R}^{d_v \times d_e}$ are learnable parameters. $\mathbf{A_i}$, 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:

$$\mathbf{A}_i = \text{softmax}\Big(\frac{Q_i(\mathbf{W}_k Y^\top)}{\sqrt{d_k}}\Big). \tag{2}$$

Note that one can define input sequence Y as an $N \times d_e$ matrix and matrix $\mathbf{W}_k$ 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.[44,45] Using visual tools can make exploring these learned functions easier.[39]

Model details

For our experiments, we used PyTorch (v1.3.1)[46] and huggingface transformers (v2.5.0).[47] The ALBERT model was trained for 48 hours on a single Nvidia P100 GPU with the hyperparameters stated in the supplementary information. Schwaller et al.[32] developed the tokenisation regex used to tokenise the SMILES. We expect further performance improvements when using more extensive data sets (e.g., commercially available ones). The RXN-Mapper model uses 12 layers, 8 heads, a hidden size of 256, an embedding size of 128, and an intermediate size of 512. In contrast to ALBERT base[35] with 12M parameters, our model is small and contains only 770k trainable parameters.

Data

The work by Lowe[36] provides the data sets used for training, composed of chemical reactions extracted from both grants and patent applications. We removed the original atom-mapping

from this dataset, canonicalised the reactions with RDKit,[48] and removed any duplicate reactions. The data set includes reactions with fragment information twice, once with and once without fragment bonds, as defined in the work of Schwaller et al.[49]. The final training set for the masked language modeling task contained a total of 2.8M reactions. For the evaluation and the model selection, we sampled 996 random reactions from the Schneider et al.[8] data set.

To test our models, we first used the remaining 49k reactions from the Schneider50k patents data set.[8] We do not distinguish between reactants and reagents in the inputs of our models. We also used the human-curated test sets that were introduced by Jaworski et al.[25] to compare our approach to previous methods. Table 2 shows an overview of the test sets. Note that patent reactions differ from the reactions in Jaworski et al.[25] because the latter removes most reactants and reagents in an attempt to balance the reactions.

Table 2: Data sets used for testing

Test set	Number of reactions	Avg. number of reactant atoms	Avg. number of product atoms
USPTO bond changes[25]	281	26.0	23.7
Schneider50k test[8]	49000	43.3	26.1

Attention-guided atom-mapping algorithm

The attention-guided algorithm relies on the construction of the attention matrix for a selected layer and head, where we sum the product-to-reactant and the corresponding reactant-to-product atom attentions. Algorithm 1 provides the exact atom-mapping algorithm. By default, after matching a product-reactant pair, the attentions to those atoms are zeroed. Optionally, atoms in product and reactants can have multiple corresponding atoms. We always mask out attention to atoms of different types.

Algorithm 1: Attention-guided atom-mapping algorithm

Data: Reaction SMILES S, multiplier W, model M
Result: Product → reactant atom-mapping P
begin
  $A \longleftarrow$ M(S) `// compute attention matrix`
  for $i \in range(len(P))$ `// iterate through product atoms`
  do
    Mask invalid atoms (not same type; optionally, already mapped)
    Select i, j pair with highest attention A_{ij}
    if $A_{ij} \neq 0$ **then**
      $P_i \longleftarrow$ j `// Map product atom` i `to reactant atom` j
      multiply attention of adjacent atoms of i to adjacent atoms of j by W
        `// Increase neighbour attentions`
    else
      $P_i \longleftarrow -1$ `// No corresponding reactant atom`
      break

Atom-mapping curation

Chemically equivalent atoms exist in many chemical reactions. Most of the chemically equivalent atoms could be matched after canonicalising the atom-mapped reaction using RDKit.[48,50] Exceptions were atoms of the same type connected to another atom with different bond types, which would form a resonance structure with delocalised electrons. We manually curated these exceptions and added them as alternative maps in the USPTO bond changes test set.[25]

References

(1) 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.

(2) Segler, M. H.; Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal* **2017**, *23*, 5966–5971.

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

(4) 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.

(5) Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning. **2020**,

(6) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. Advances in Neural Information Processing Systems. 2017; pp 2607–2616.

(7) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **2019**, *10*, 370–377.

(8) 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.

(9) Rahman, S. A.; Cuesta, S. M.; Furnham, N.; Holliday, G. L.; Thornton, J. M. EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nature methods* **2014**, *11*, 171.

(10) Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L. Over 20 years of reaction access systems from MDL: a novel reaction substructure search algorithm. *Journal of chemical information and computer sciences* **2002**, *42*, 1296–1310.

(11) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3*, 560–593.

(12) Gonzalez, G. A. P.; El Assal, L. R.; Noronha, A.; Thiele, I.; Haraldsdóttir, H. S.; Fleming, R. M. Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to Recon 3D. *Journal of cheminformatics* **2017**, *9*, 39.

(13) Lynch, M. F.; Willett, P. The automatic detection of chemical reaction sites. *Journal of Chemical Information and Computer Sciences* **1978**, *18*, 154–159.

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

(15) Vleduts, G. Development of a combined WLN/CTR multilevel approach to the algorithmic analysis of chemical reactions in view of their automatic indexing. *British Library, Research and Development Report No. 5399* **1977**,

(16) McGregor, J. J.; Willett, P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *Journal of Chemical Information and Computer Sciences* **1981**, *21*, 137–140.

(17) Funatsu, K.; Endo, T.; Kotera, N.; Sasaki, S.-I. Automatic recognition of reaction site in organic chemical reactions. *Tetrahedron Computer Methodology* **1988**, *1*, 53–69.

(18) Körner, R.; Apostolakis, J. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *Journal of chemical information and modeling* **2008**, *48*, 1181–1189.

(19) Apostolakis, J.; Sacher, O.; Körner, R.; Gasteiger, J. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *Journal of chemical information and modeling* **2008**, *48*, 1190–1198.

(20) Jochum, C.; Gasteiger, J.; Ugi, I. The principle of minimum chemical distance (PMCD). *Angewandte Chemie International Edition in English* **1980**, *19*, 495–505.

(21) Akutsu, T. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *Journal of Computational Biology* **2004**, *11*, 449–462.

(22) Crabtree, J. D.; Mehta, D. P. Automated reaction mapping. *Journal of Experimental Algorithmics (JEA)* **2009**, *13*, 1–15.

(23) First, E. L.; Gounaris, C. E.; Floudas, C. A. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *Journal of chemical information and modeling* **2012**, *52*, 84–92.

(24) Latendresse, M.; Malerich, J. P.; Travers, M.; Karp, P. D. Accurate atom-mapping computation for biochemical reactions. *Journal of chemical information and modeling* **2012**, *52*, 2970–2982.

(25) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic mapping of atoms across both simple and complex chemical reactions. *Nature communications* **2019**, *10*, 1–11.

(26) Nextmove Software NameRXN. `http://www.nextmovesoftware.com/namerxn.html`, (Accessed Apr 02, 2020).

(27) Indigo Toolkit. `https://lifescience.opensource.epam.com/indigo/`, (Accessed Apr 02, 2020).

(28) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today* **2020**,

(29) Almeida, A.; Moreira, R.; Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry* **2019**, *3*.

(30) 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.

(31) 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.

(32) 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**, *5*, 1572–1583.

(33) Schwaller, P.; Laino, T. Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches. *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*. 2019; pp 61–79.

(34) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

(35) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations. 2020.

(36) Lowe, D. Chemical reactions from US patents (1976-Sep2016). 2017; `https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873`.

(37) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. **2019**,

(38) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Laino, T.; Reymond, J.-L. Data-Driven Chemical Reaction Classification, Fingerprinting and Clustering using Attention-Based Neural Networks. *ChemRxiv preprint: 10.26434/chemrxiv.9897365.v2* **2019**,

(39) Hoover, B.; Strobelt, H.; Gehrmann, S. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276* **2019**,

(40) Wiegreffe, S.; Pinter, Y. Attention is not not Explanation. **2019**,

(41) Vig, J. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714* **2019**,

(42) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge, 2012.

(43) Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **2020**, *12*, 1–13.

(44) Vig, J.; Belinkov, Y. Analyzing the Structure of Attention in a Transformer Language Model. *CoRR* **2019**, *abs/1906.04284*.

(45) Clark, K.; Khandelwal, U.; Levy, O.; Manning, C. D. What Does BERT Look At? An Analysis of BERT's Attention. *CoRR* **2019**, *abs/1906.04341*.

(46) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems. 2019; pp 8024–8035.

(47) 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*.

(48) Landrum, G. et al. rdkit/rdkit: 2019_03_4 (Q1 2019) Release. 2019; `https://doi.org/10.5281/zenodo.3366468`.

(49) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **2020**, *11*, 3316–3325.

(50) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order - An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *Journal of chemical information and modeling* **2015**, *55*, 2111–2120.

Supplementary Information

Unsupervised attention-guided atom-mapping

Philippe Schwaller,[*,†,‡] Benjamin Hoover,[¶] Jean-Louis Reymond,[‡] Hendrik Strobelt,[¶] and Teodoro Laino[†]

†*IBM Research Europe, CH-8803 Rüschlikon, Switzerland*

‡*Department of Chemistry and Biochemistry, University of Bern, Switzerland*

¶*IBM Research – Cambridge / MIT-IBM Lab, United States*

E-mail: phs@zurich.ibm.com

http://rxnmapper.ai

Contents

A Comparison with atom-mapping tools

Recently, Jaworski et al.[1] developed an atom-mapper based on graph-theoretical approach augmented with human-expert written rules. They compared their tool called Mappet[1] to other methods. We performed the same tests using our RXNMapper. Figure S1 shows the correctness on three different test sets of our attention-based RXNMapper, Mappet,[1] Marvin JS (version 16.4.18),[2] ReactionMap,[3] ChemDraw Prime (version 16.0.0.82), and Indigo (version 1.3.0 beta).[4] The simple reactions set consists of 100 reactions from total syntheses reported in Org. Lett., J. Am. Chem. Soc., and J. Org. Chem., whereas the typical reactions set consists of 100 almost, but not fully, balanced patent reactions. RXNMapper achieves correctness scores similar to Mappet on both these sets. On the complex reaction set, which consists of 201 mechanistically complex reactions from recent literature, we perform slightly worse than Mappet but better than other reported methods. Still, the results are impressive as RXNMapper was not tuned specifically for any of these test sets. An overview of the test sets can be found in Table S1.

Table S1: Data sets for the comparison with other tools.

Test set	Number of reactions	Avg. number of reactant atoms	Avg. number of product atoms
Simple reactions[1]	100	27.1	27.1
Typical reactions[1]	100	19.9	19.6
Complex reactions[1]	201	25.7	24.8

RXNMapper performs remarkably well on reactions involving rearrangements of the carbon skeleton where correct atom mapping requires an understanding of the reaction mechanism. Striking examples include an intramolecular Claisen rearrangement used to construct

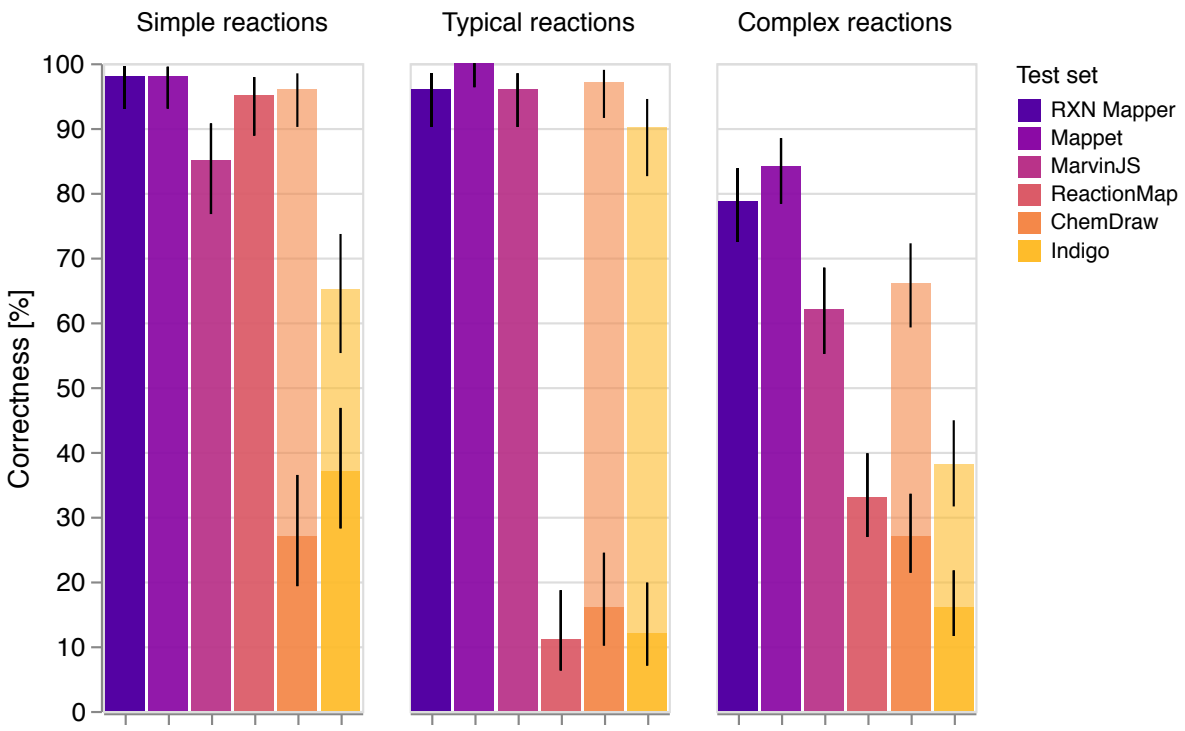


Figure S1: Tool comparison, test originally published by Jaworski et al.[1]. The error bars show the Wilson confidence interval.[5]

fused 7-8 membered ring in the synthesis of the natural product micrandilactone A (Figure S2 a)[6,7]), and the tandem Palladium-catalyzed semipinacol rearrangement / direct arylation used for a stereoselective synthesis of benzodiquinanes from cyclobutanols (Figure S2 b)[8]). In both cases, RXNMapper completes the correct atom mapping despite the entirely rearranged carbon skeletons resulting in different ring sizes and connections. By comparison, all other automated tools tested here, which comprised ReactionMap, Marvin, ChemDraw and Indigo, failed at the task. RXNMapper also succeeds in atom mapping for the ring rearrangement metathesis of a norbornene to form a bicyclic enone under catalysis by Grubbs-(I) catalyst (Figure S2 c)[9]). In this case, atom mapping also succeeded using the ChemDraw mapping tool, while the other tools failed. Furthermore, RXNMapper also performs well with multicomponent reactions such as the Ugi 4-component condensation of isonitriles, aldehydes, amines and carboxylic acids to form acylated aminoacid amides (Figure S2 d),[10]). In

this case, RXNmapper maps all atoms correctly except for the carbonyl oxygen atom of the isonitrile derived carboxamide. RXNMapper assigns this oxygen atom to the oxygen atom of the carbonyl group of the aldehyde reagent, though this atom actually comes from the hydroxyl group of the carboxylic acid reagent. Although correct mapping may seem less remarkable in this case, note that all other tested tools failed except for Mappet.

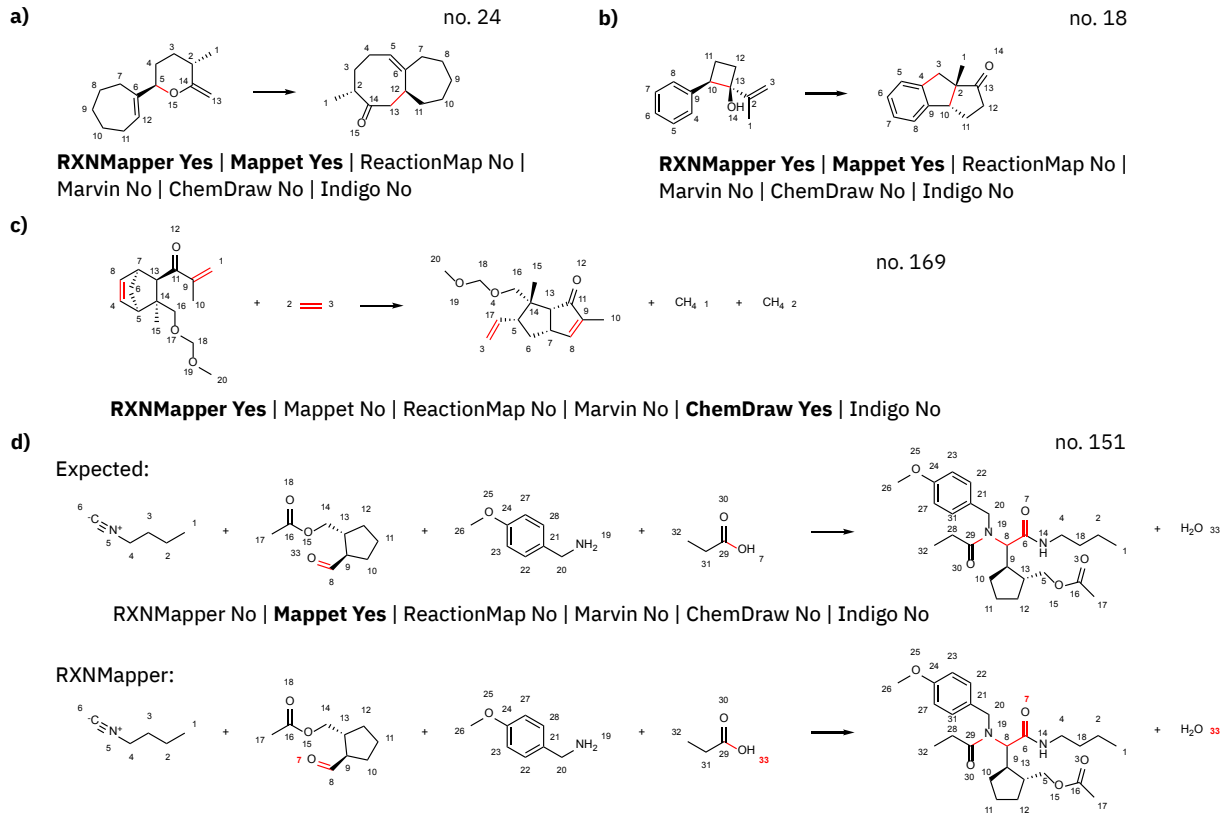


Figure S2: Examples from the complex reactions data set.[1] a) Bu3Al-promoted Claisen rearrangement[6,7] b) Palladium-Catalyzed Semipinacol Rearrangement and Direct Arylation.[8] c) Grubbs-catalyzed ring rearrangement metathesis reaction[9] d) Ugi reaction[10]

B Computational performance

In contrast to previous methods, RXNMapper does not require balanced or almost balanced reactions. It can compute atom-mapping for both patent reactions and reactions predicted by template-free reaction prediction models. RXNMapper maps the 682 balanced reactions from

the work of Jaworski et al.[1] at 33.3 reactions per second (30 ms/reaction) on a MacBook Pro: 2.7 GHz Intel Core i7, 16 GB 2133 MHz LPDD and reaches 156.2 reactions per second (6.4 ms/reaction), when the attention model inference is accelerated using a GPU (Nvidia RTX 2070 super). The computational performance is nearly the same when mapping reactions from the 49k patent reaction data set, which are mapped at a speed of 27.5 reactions per second (36.4 ms/reaction) on CPU only and 130 reactions per second (7.7 ms/reaction) using a GPU. In terms of speed RXNMapper performs similar to Indigo toolkit[4] on the balanced reactions, RXNMapper significantly outperforms Indigo on the patent reactions that contain many more reactants. The computational performance makes it feasible to apply RXNMapper to large reaction data sets in a reasonable time. We remapped the largest open-source reaction data set[11] at an average speed of 7.37 ms/reaction and made it available at `https://github.com/rxn4chemistry/rxnmapper`.

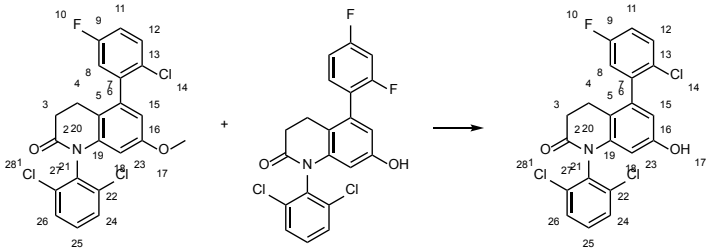
C Reactions examples

Questionable and unclear reactions

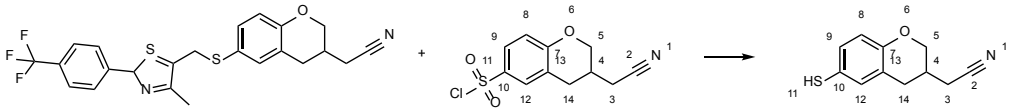
While analyzing the discrepancies in the atom-mapping generated on the 49k patents test set, we labelled 369 as questionable and 47 as unclear. Questionable reactions typically contain multiple products similar to reactants, as in Figure S3 a). The reason could be a wrong extraction from patents. Unclear reactions, on the other hand, have correct reactants but miss reagents, which are crucial to determine the reaction mechanism. The example shown in Figure S3 b) looks like a Mitsunobu reaction but the DEAD or DIAD reagents are not present. Despite the missing reagents, RXNMapper would have correctly mapped the phenolic alcohol.

a) questionable reactions

18646



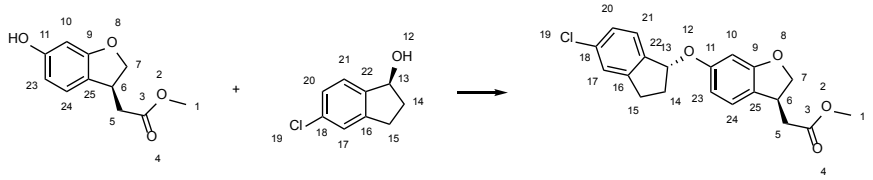
12534



b) unclear, missing reagents

Data set

44755



RXNMapper

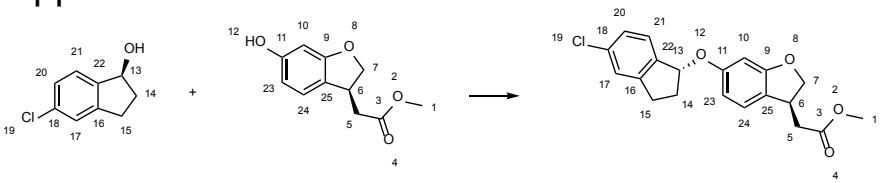


Figure S3: Examples of a) reactions that were classified as questionable. b) a reaction for which the correct atom-mapping is unclear as critical reagents are missing

Equivalent reactions

Figure S4 shows reactions that were counted as correct even though the atom-mapping was not identical with the one in the data set. Such reactions typically have two equivalent atoms or symmetry operations that make the atom maps equivalent. If there was twice the same molecule on the product side, the atom-mappings in the original data set pointed for both molecules to the same atoms in the reactants. In contrast, our algorithm in the default configuration mapped different atoms in the reactants.

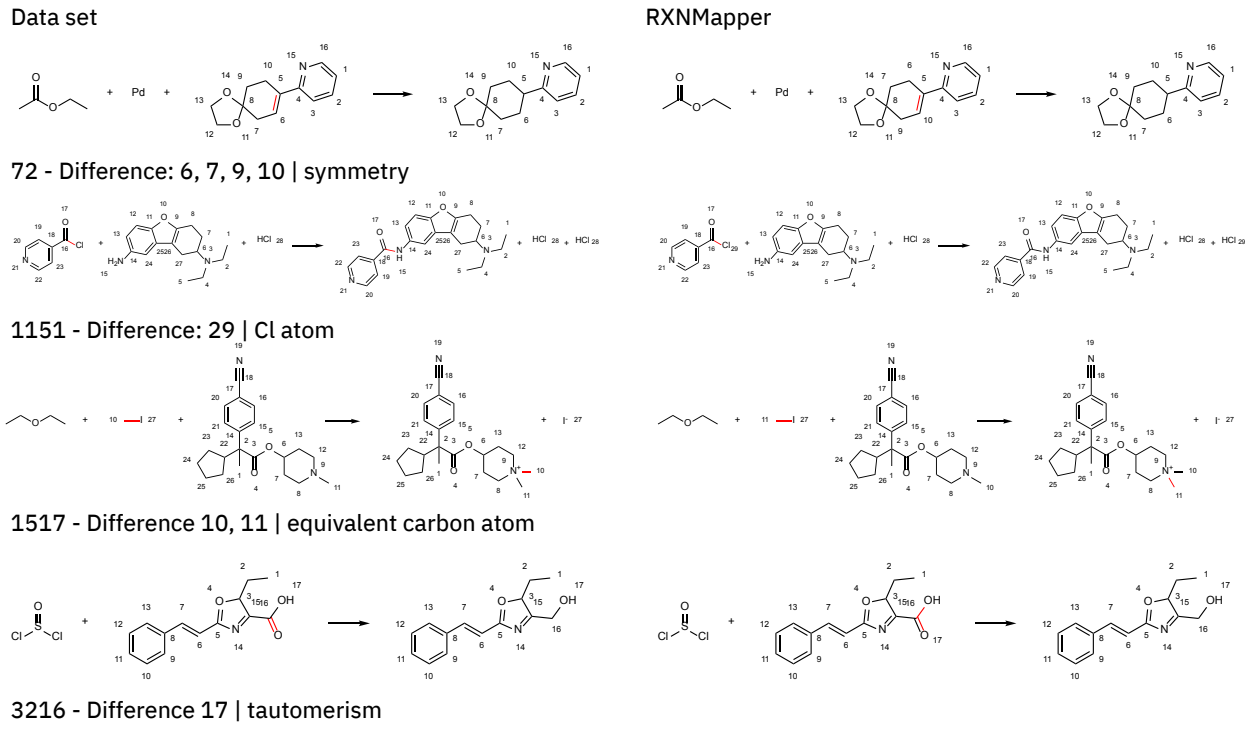


Figure S4: Examples of atom-mappings that differed from the data set but were counted as equally correct.

D Confidence score

The confidence score for atom-mapping is computed by multiplying the selected attention scores for all the mapped product atoms. As seen in Figure S5, correctly generated atom-mappings have, on average, a higher confidence score than those that contain mistakes. Questionable reactions (e.g., where the reaction was wrongly extracted from patents) contain the lowest confidence scores.

E Hyperparameters and model selection

Hyperparameters

We trained the models for 48 hours on a single Nvidia P100 GPU with a masked language masking probability of 0.15. We used the training scripts from huggingface[12] adapted to

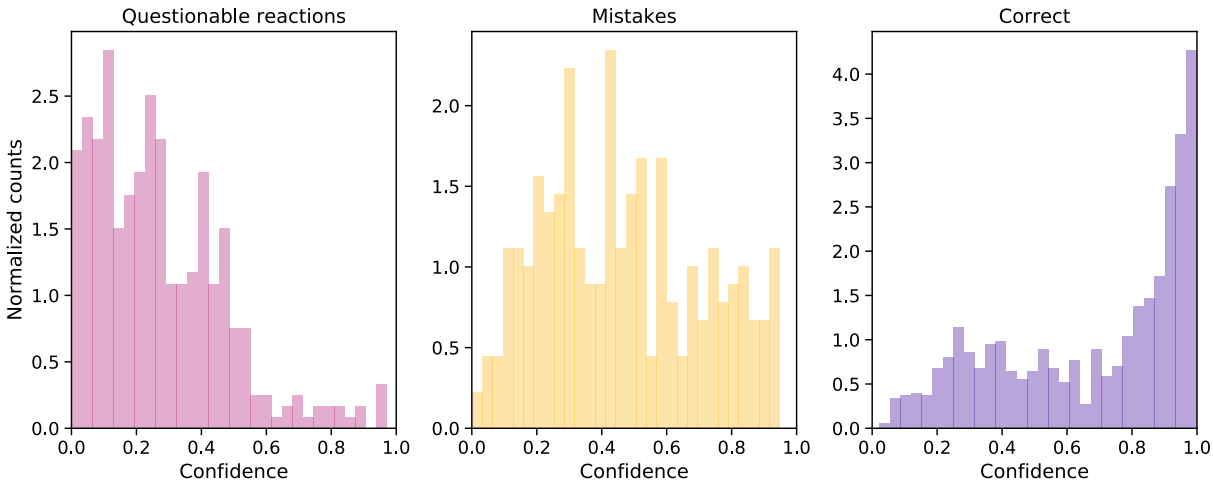


Figure S5: Normalized histograms of confidence scores on three categories of atom-mappings: atom-mappings on questionable reactions, wrongly generated atom-mappings and correct atom mappings.

work with a SmilesTokenizer, which we made available. For the ALBERT models, we fixed the number of layers to 12, the activation function to GELU, the dropout probability for 0.1, the embedding size to 128, the intermediate size to 512. We varied both the hidden size and the number of heads. The model with 8 heads uses a hidden size of 256, the model with 10 heads uses a hidden size of 320, and the model with 12 heads uses a hidden size of 384. We experimented with larger models, but the differences in atom mapping correctness were marginal. Our final model has only 770k trainable parameters, which is small compared to BERT base[13] with 108M and ALBERT base[14] with 12M parameters.

Model selection

The improvement of the atom-mapping correctness may increase up to 30% when changing the neighbour attention multiplier from 1 (basic algorithm) to a value of 20. Figure S6 shows the atom-mapping correctness on the validation reactions for all the heads and layers of different models. For the ALBERT pre-trained model, at least one head learned atom-mapping, and the position and role of the heads remained constant across all layers. The atom-mapping correctness increased in the first layers and is more or less constant from

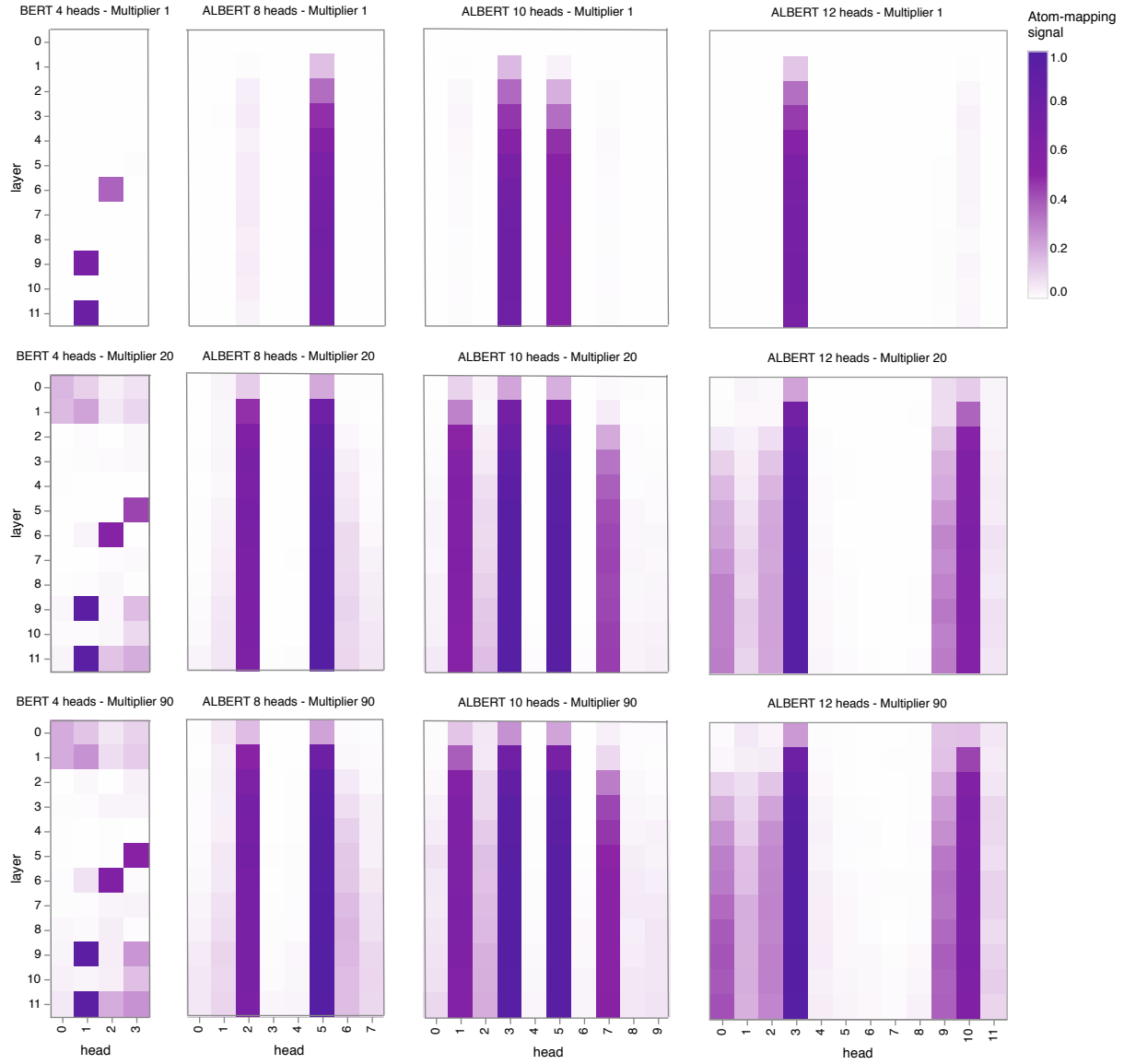


Figure S6: Atom-mapping performance of all layers and heads of one BERT and 3 ALBERT models on the patent validation set with multipliers of 1, 20 and 90.

layer 7 to 11. In contrast, for the BERT model does not share weights across layers and only particular heads in particular layers had learned an atom-mapping signal.

As shown in Figure S7, the atom-mapping correctness steeply increases in the first 100k training steps then continues to increase more slowly. We observed this behaviour for all models we trained. Moreover, models with more heads seemed to learn the atom-mapping signal faster, but the models with fewer heads quickly beat the performance of the larger

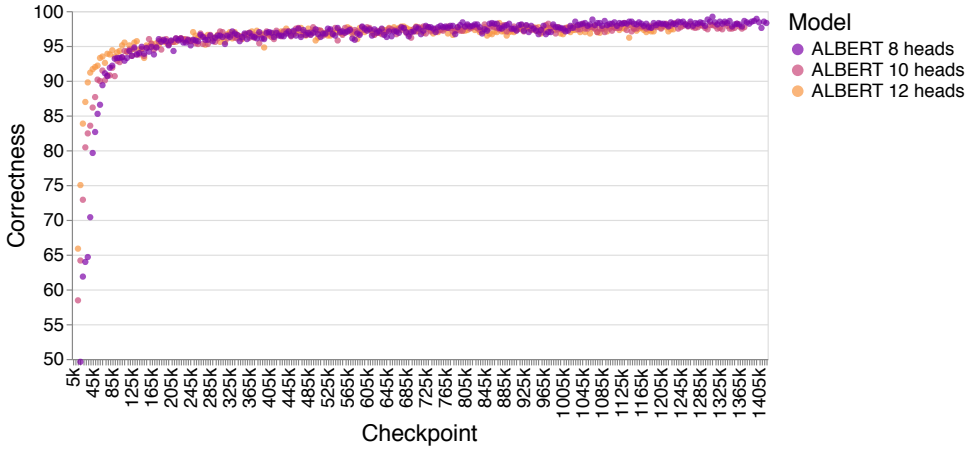


Figure S7: Evaluation atom-mapping correctness for checkpoints every 5k training steps on the validation set for ALBERT models with 8, 10 and 12 heads. The layer was fixed to 10, the multiplier to 90 and the head with the largest atom-mapping signal was selected.

models.

The top-20 model combinations are shown in Table S2. We selected checkpoint 1310k (layer 10, head 5) as the best performing model on the 1k patent validation set. We used this model to perform all experiments in the main paper.

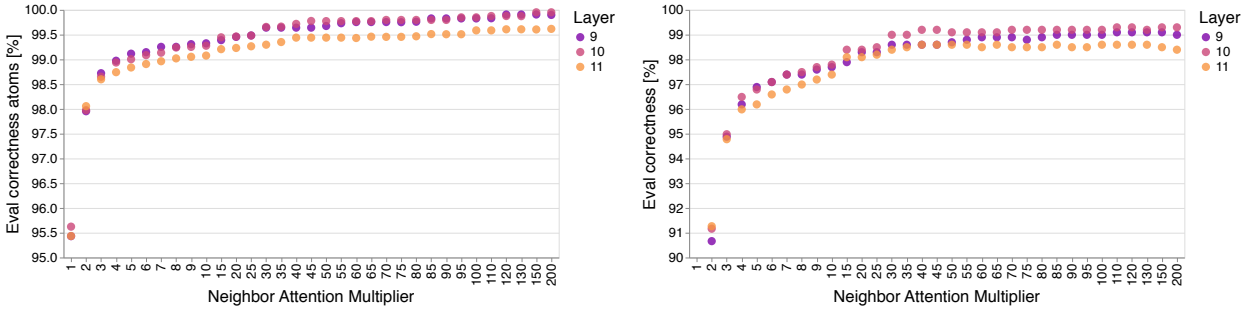


Figure S8: Evaluation atom-mapping correctness per atom (left) and per reaction (right) for different multiplier.

As shown in Figure S8, increasing the nearest neighbour multiplier increases the atom-wise and full reaction atom-mapping correctness.

Table S2: Top-20 model/layer/head combinations by correctness on the validation set for a multiplier of 90.

	name	checkpoint	layer	head	Atom correctness [%]	Correctness [%]
11740	ALBERT 8 heads	1310k	10	5	99.8	99.2
12009	ALBERT 8 heads	1400k	9	5	99.9	99.1
11739	ALBERT 8 heads	1310k	9	5	99.8	99.0
12010	ALBERT 8 heads	1400k	10	5	99.7	98.9
11709	ALBERT 8 heads	1300k	9	5	99.7	98.9
11710	ALBERT 8 heads	1300k	10	5	99.7	98.8
11005	ALBERT 8 heads	1065k	10	5	99.6	98.8
11291	ALBERT 8 heads	1160k	11	5	99.8	98.7
11604	ALBERT 8 heads	1265k	9	5	99.8	98.6
11845	ALBERT 8 heads	1345k	10	5	99.8	98.6
11995	ALBERT 8 heads	1395k	10	5	99.7	98.6
11996	ALBERT 8 heads	1395k	11	5	99.7	98.6
11006	ALBERT 8 heads	1065k	11	5	99.7	98.6
11935	ALBERT 8 heads	1375k	10	5	99.6	98.6
11679	ALBERT 8 heads	1290k	9	5	99.6	98.6
11381	ALBERT 8 heads	1190k	11	5	99.5	98.6
11080	ALBERT 8 heads	1090k	10	5	99.4	98.6
11289	ALBERT 8 heads	1160k	9	5	99.8	98.5
11560	ALBERT 8 heads	1250k	10	5	99.7	98.5
11725	ALBERT 8 heads	1305k	10	5	99.7	98.5

F Visualisation of self-attention

Visual inspection of the attention weights enabled the initial discovery that molecular Transformer models learned atom-mapping as a key signal. We release a tool called RXNMapper-Vis that allows others to explore the attentions of the ALBERT model behind RXNMapper interactively and make new hypotheses. RXNMapper-Vis maps the attentions from the tokenised SMILES onto a 2D skeletal structure to ease interpretation. The tool has been made available at `https://rxnmapper.ai`.

RXNMapper-Vis was inspired by previous work to visualise the attentions of Transformer models in the natural language processing (NLP).[15–17] These tools can reveal learned but hidden behaviours of Transformers such as hidden language dependencies and parts of speech (e.g., attentions linking root Verbs to their Direct Objects), coreference (e.g., "she" attending

to "mother"), entities (e.g., "Elon Musk" or "Iran"), and gender biases associated with particular roles (e.g., models predicting "he" as the necessary pronoun for "doctor"). Some of these learned patterns correlate to properties within the chemical domain. For example, coreference correlates to the learned atom-mapping behaviour discussed in this paper. We hope that others will be able to use RXNMapper-Vis to find meaningful patterns in the layers and heads of the molecular Transformer model and that these discoveries can enrich our knowledge and improve our tooling for the chemical domain.

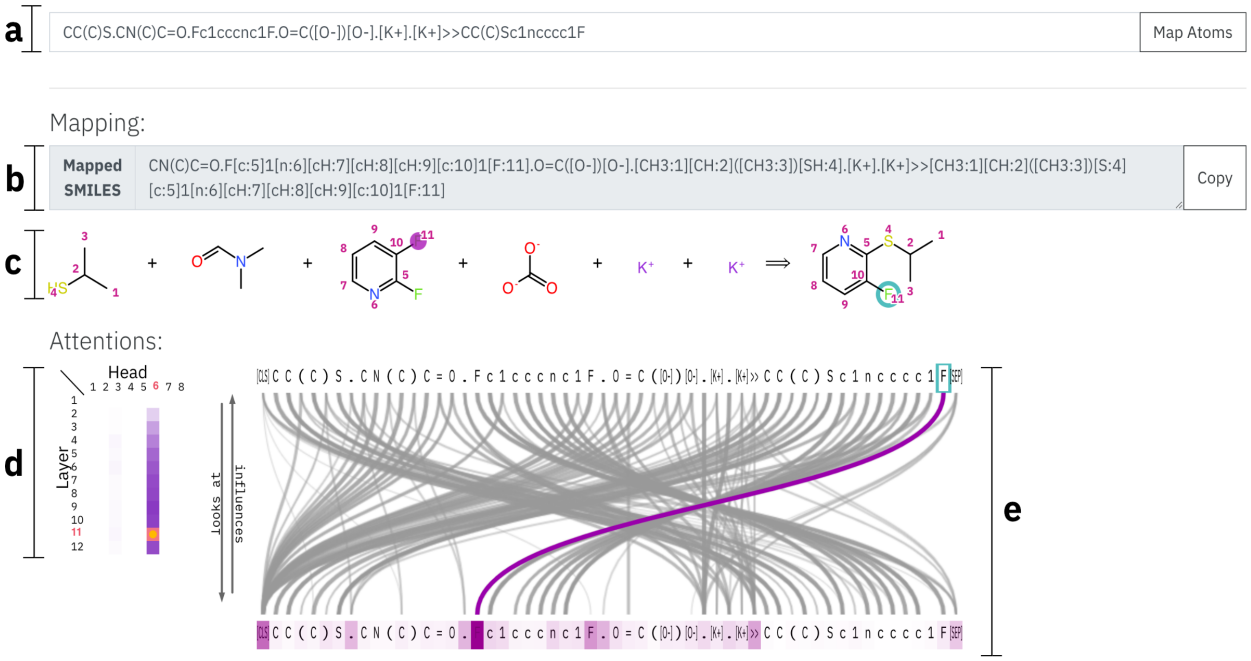


Figure S9: An overview of RXNMapper-Vis. Users can insert their reaction SMILES in (a), and the tool will display the atom-mapped string in (b). A 2D skeletal structure depiction of the SMILES is shown in (c). Hovering over any atom will show the attention weights out of that atom and onto all the other atoms. Clicking on an atom will freeze that particular attention view. The attentions of different heads and layers can be inspected in (d), where darker backgrounds of each cell indicate a higher performance at atom-mapping. Note that atom labels in (c) only show for the atom-mapping head. Changing the selected layer/head combination will update the attentions in (c) and (e). The attention graph in (e) shows the self-attention of the input as a connected graph, where darker and thicker curves indicate a higher attention weight out of tokens in the top row into each token in the bottom row. Hovering over any token highlights the connected attentions in the graph and the corresponding atoms in (c). Here, the Fluorine in the product is selected, and both the attention graph and the skeletal structure show the greatest attention to the correct reactant atom. The complete discrete probability distribution of the attentions is shown as a purple background over the input sequence.

References

(1) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic mapping of atoms across both simple and complex chemical reactions. *Nature communications* **2019**, *10*, 1–11.

(2) Marvin JS, ChemAxon. https://chemaxon.com, (Accessed Apr 02, 2020).

(3) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An efficient atom-mapping algorithm for chemical reactions. *Journal of chemical information and modeling* **2013**, *53*, 2812–2819.

(4) Indigo Toolkit. https://lifescience.opensource.epam.com/indigo/, (Accessed Apr 02, 2020).

(5) Wallis, S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* **2013**, *20*, 178–208.

(6) Zhang, Y.-D.; Ren, W.-W.; Lan, Y.; Xiao, Q.; Wang, K.; Xu, J.; Chen, J.-H.; Yang, Z. Stereoselective Construction of an Unprecedented 7- 8 Fused Ring System in Micrandilactone A by [3, 3]-Sigmatropic Rearrangement. *Organic letters* **2008**, *10*, 665–668.

(7) Sun, T.-W.; Ren, W.-W.; Xiao, Q.; Tang, Y.-F.; Zhang, Y.-D.; Li, Y.; Meng, F.-K.; Liu, Y.-F.; Zhao, M.-Z.; Xu, L.-M., et al. Diastereoselective Total Synthesis of (±)-Schindilactone A, Part 1: Construction of the ABC and FGH Ring Systems and Initial Attempts to Construct the CDEF Ring System. *Chemistry–An Asian Journal* **2012**, *7*, 2321–2333.

(8) Schweinitz, A.; Chtchemelinine, A.; Orellana, A. Synthesis of benzodiquinanes via tan-

dem palladium-catalyzed semipinacol rearrangement and direct arylation. *Organic letters* **2011**, *13*, 232–235.

(9) Acharyya, R. K.; Rej, R. K.; Nanda, S. Exploration of Ring Rearrangement Metathesis Reaction: A General and Flexible Approach for the Rapid Construction [5, n]-Fused Bicyclic Systems en Route to Linear Triquinanes. *The Journal of organic chemistry* **2018**, *83*, 2087–2103.

(10) Moni, L.; Banfi, L.; Basso, A.; Carcone, L.; Rasparini, M.; Riva, R. Ugi and Passerini reactions of biocatalytically derived chiral aldehydes: application to the synthesis of bicyclic pyrrolidines and of antiviral agent telaprevir. *The Journal of organic chemistry* **2015**, *80*, 3411–3428.

(11) Lowe, D. Chemical reactions from US patents (1976-Sep2016). 2017; https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

(12) 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*.

(13) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**,

(14) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations. 2020.

(15) Hoover, B.; Strobelt, H.; Gehrmann, S. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276* **2019**,

(16) Wiegreffe, S.; Pinter, Y. Attention is not not Explanation. **2019**,

(17) Vig, J. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714* **2019**,