Completion of partial reaction equations

Alain C. Vaucher
IBM Research Europe
Säumerstrasse 4, 8803 Rüschlikon, Switzerland
ava@zurich.ibm.com

Philippe Schwaller
IBM Research Europe
Säumerstrasse 4, 8803 Rüschlikon, Switzerland
phs@zurich.ibm.com

Teodoro Laino
IBM Research Europe
Säumerstrasse 4, 8803 Rüschlikon, Switzerland
teo@zurich.ibm.com

Abstract

We present a deep-learning model for inferring missing molecules in reaction equations. Such an algorithm features multiple interesting behaviors. First, it can infer the necessary reagents and solvents in chemical transformations specified only in terms of main compounds, as often resulting from retrosynthetic analyses. The completion with necessary reagents ensures that reaction equations are compatible with deep-learning models relying on a complete reaction specification. Second, it can cure existing datasets by detecting missing compounds, such as reagents that are essential for given classes of reactions. Finally, this model is a generalization of models for forward reaction prediction and retrosynthetic analysis, as both can be formulated in terms of incomplete reaction equations. We illustrate that a single trained model, based on the transformer architecture and acting on reaction SMILES strings, can address all three points.

1 Introduction

Deep-learning models applied to chemical reactions have received much attention in recent years: from the design of algorithms for forward reaction prediction [1–3] and retrosynthetic analysis [1, 4, 5] that help chemists plan the design and execution of chemical syntheses, to the generation of reaction fingerprints [6] and prediction of reaction classes [7, 6], yields [8], or activation energies [9].

Several of the latter predictive models were trained on fully-specified reactions — i.e., they rely on all the reagents being specified, including solvents and catalysts. Accordingly, when these models are applied to new reactions, a complete specification of the reagents is required. Unfortunately, both algorithms and chemists do not provide any guarantee of generating complete chemical reaction equations. It is therefore desirable to infer the missing molecules to provide higher quality data and to comply with a larger class of machine learning models.

In fact, an algorithm fulfilling this task can also be used for data curation. Many commonly-used reaction datasets [10, 11] were generated by automatically text-mining chemical knowledge from the unstructured data sources. The text-mining process is error-prone and often fails to recognize one or

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

several precursor molecules. An approach for the completion of partial reaction equations can either detect reactions that are potentially incomplete, or even automatically add the missing molecules, before the dataset is used for other downstream applications.

Furthermore, the task of inferring missing compounds in a reaction equation is a generalization of forward and single-step retrosynthetic prediction models. As a consequence, a properly tuned algorithm completing partial reaction equations also a forward or retrosynthetic prediction model.

In this work, we present a deep-learning model based on the transformer architecture that infers the molecules in partial reaction SMILES strings. This model does not contain any chemical knowledge except the one learned from the data during training. We illustrate its application for data curation, as well as its use for forward and retrosynthesis prediction. Figure 1 gives a few examples of partial reaction equations and how they can be completed.

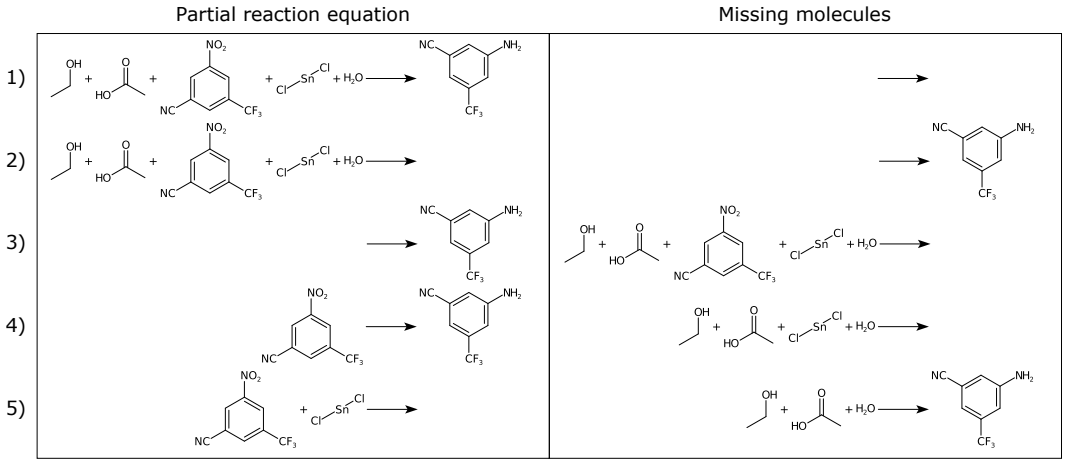

Figure 1: Partial reaction equations and associated missing molecules, corresponding to the input and output of the model presented in this work. Example 1) corresponds to a complete reaction equation, where no molecules are missing. Example 2) corresponds to a forward reaction prediction task, where the model must predict the reaction product. Example 3) corresponds to a retrosynthetic prediction, where the model infers potential precursors for a given molecule. Note that multiple outputs are possible in this case. Examples 4) and 5) illustrate other cases of completing partial reaction equations. The model introduced in this work, trained a single time, is able to tackle all these different cases.

2 Method

2.1 Model

We used a modified version of the Molecular Transformer [3]. Both input and output of the model are tokenized versions of reaction SMILES strings [12, 13]. Thereby, the input corresponds to the potentially incomplete reaction equation, and the output contains the missing precursors and/or products. When no precursor or product is missing, the reaction SMILES string will only contain the token separating precursors and products, ">>".

The transformer model is implemented with the `OpenNMT-py` library [14, 15]. The standard transformer implementations is applied with the following changes: the parameter `layers` is set to 4, `rnn_size` to 256, `word_vec_size` to 256, `max_generator_batches` to 32, `accum_count` to 4 and `label_smoothing` to 0. We trained the model for 1,000,000 steps.

2.2 Data

The model was trained and tested on data derived from the US patent reactions by Lowe [10], as post-processed by Pesciullesi et al. [16]. The data was obtained from the GitHub repository of Ref. [16] and the same data splits were used.

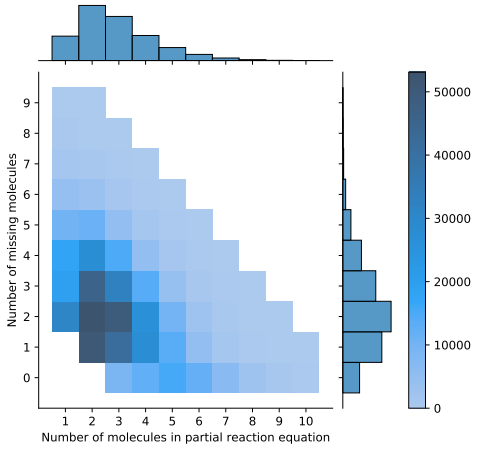

Figure 2: Frequencies of given number of molecules in the partial reaction equations and associated number of missing molecules.

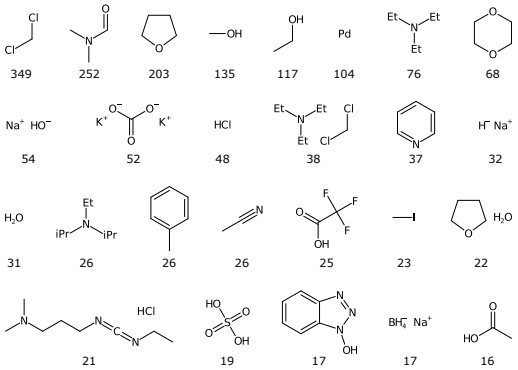

Figure 3: Most frequent molecules (or groups of molecules) inferred when applying the reaction completion model on the ground truth data. The number below each compound indicates the number of reactions for which that compound was predicted.

For training the reaction completion model, each reaction SMILES string in the original dataset was transformed into ten partial reaction SMILES strings by selecting randomly, from the precursors and products, the molecules to remove in the input. Thereby, this process generated reaction SMILES strings with a variable number of missing molecules. The only criterion was that the resulting partial SMILES string should contain enough tokens compared to the original reaction SMILES string, to avoid generating partial reaction SMILES strings containing common reagents only.

This resulted in training, validation, and test sets of sizes 10.9 M, 0.6 M, and 0.6 M, respectively. Figure 2 shows histograms for the number of molecules in the partial reaction SMILES strings, against the number of removed molecules.

3 Results and discussion

In the following, we address the different applications of the reaction completion model. All the calculations refer to the test split of the dataset. The "reference forward prediction model" refers to the pretrained model presented in Ref. [16]. In Figure 4, we illustrate the application of our model to a few reactions and compare the predictions with the expected value from the ground truth.

3.1 Partial reaction equations

The model achieves an accuracy of 30.4%. Note that the accuracy reflects exact matches only, where all the molecules in the resulting reaction equation are identical to the ground truth. An exhaustive evaluation of the reasons for not matching the 69.6% remaining cases lies outside the scope of this work; an initial analysis shows that often solvents are missing either in the prediction or in the ground truth, that some equivalent solvents or reagents are predicted instead of other ones, and that sometimes the partial reaction equation leaves multiple possibilities open as to what the reaction should be.

We evaluated the correctness of the resulting reaction equations (combining the partial reaction equation from the input and the model prediction) by assessing whether they are correct. To do so, we inspected whether the reference forward prediction model delivered a consistent product. We found out that this is the case for 77.6% of the reactions.

Applying the reaction completion model several times iteratively only leads to a marginal improvement of the accuracy to 30.5%.

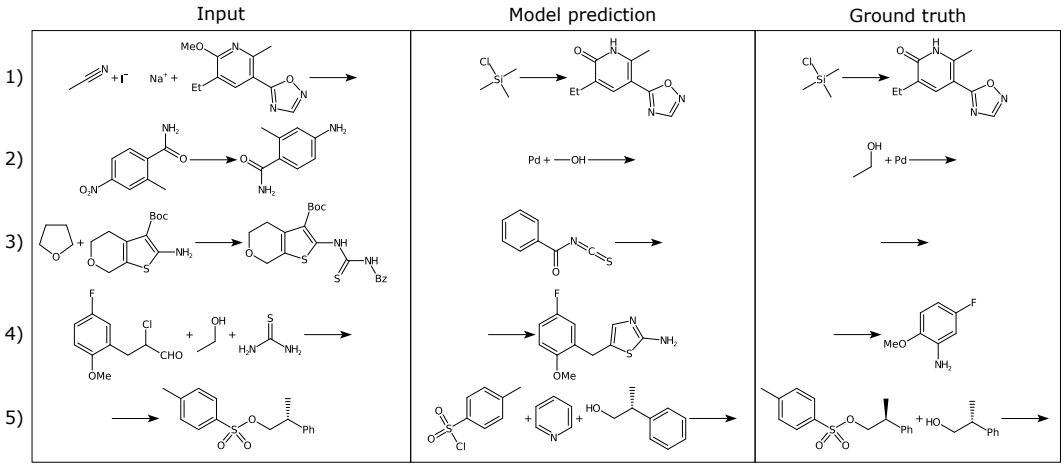

Figure 4: Examples of reaction completion. The first reaction is an example where the model finds correctly that a precursor and the product are missing. In the second example, the model predicts a different solvent (methanol) than the ground truth (ethanol). The third example illustrates a case where the application of the model to the original dataset reveals a missing precursor molecule. In the fourth example, we illustrate the application of the model as a forward prediction model. There, the dataset contained a mistaken product. The fifth example illustrates the application as a single-step retrosynthesis prediction model. The model chooses different precursors than present in the ground truth.

3.2 Application on ground truth data

By applying the reaction completion model to the the ground truth data, we found out that 2729 (out of 60548) reactions were considered to be incomplete. A total of 493 distinct molecules (or groups of molecules) were missing. The most frequently inferred molecules or groups of molecules are depicted in Figure 1. While many of those relate to solvents, upon analysis it is evident that many essential reagents are missing from the ground truth data.

3.3 Forward reaction prediction

By removing the products from the reaction equations, the reaction completion model can be applied for reaction prediction. We compare our predictions with the ground truth (products that were removed) and with the predictions of the reference forward prediction model. Our model achieves 68.1% accuracy, compared to 77.6% for the reference forward prediction model. Interestingly, for 9.4% of reactions, both models predict an identical product that differs from the ground truth. Upon manual inspection, many of these examples correspond to mistakes in the underlying ground truth.

We also note that for 7.0% of the forward predictions, our model predicted some precursors in addition to the products. Considering this, it may be that some of the predictions differing from the ground truth would be correct if one took the additional precursors into account.

3.4 Single-step retrosynthesis

For the single-step retrosynthesis task, partial reaction SMILES strings including only the reaction products were fed to the model. Accordingly, the predictions of the model contain only precursors. We assess the predictions by calculating the round-trip accuracy [5]. Taking, for the forward prediction, the model presented in this work, we obtain a round-trip accuracy of 81.5%. Using the reference forward prediction model, the round-trip accuracy is 82.6%. These values are slightly higher than the ones reported for transformer-based models for retrosynthesis [5]. A more extensive evaluation of the application of the model to retrosynthesis requires the inspection of other metrics that will be reported in a subsequent paper.

4 Conclusion

The model for reaction completion introduced in this work, while simple in its formulation, can tackle various tasks, including forward reaction prediction, single-step retrosynthesis and data curation. Our model requires only a single training, which automatically ensures compatibility among the tasks and may be beneficial for learning how molecules react. An interesting application of this model could also be guided retrosynthesis, where a chemist knows some of the precursors or reagents to obtain a given product and wishes to automatically complete the reaction equation. To improve the model and reach the accuracy of existing forward prediction and retrosynthesis model, we expect some effort to be needed in the generation of a more adequate dataset.

References

[1] Segler, M. H. S. & Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **23**, 5966–5971 (2017).

[2] Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).

[3] Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).

[4] Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).

[5] Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).

[6] Schwaller, P. *et al.* Mapping the Space of Chemical Reactions using Attention-Based Neural Networks. *ChemRxiv.9897365.v3* (2020).

[7] 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. *J. Chem. Inf. Model.* **55**, 39–53 (2015).

[8] Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of Chemical Reaction Yields using Deep Learning. *Chemrxiv.12758474.v1ChemRxiv.9897365.v3* (2020). Preprint at https://doi.org/10.26434/chemrxiv.12758474.v1.

[9] Jorner, K., Brinck, T., Norrby, P.-O. & Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *ChemRxiv.12758498.v1* (2020). Preprint at http://dx.doi.org/10.26434/chemrxiv.12758498.v1.

[10] Lowe, D. Chemical reactions from US patents (1976 - sep 2016) (2017).

[11] Nextmove Software Pistachio. (Accessed Oct 5, 2020).

[12] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).

[13] Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).

[14] Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, 67–72 (Association for Computational Linguistics, Vancouver, Canada, 2017).

[15] OpenNMT-py library, version 1.2.0. `https://github.com/OpenNMT/OpenNMT-py` (Accessed Nov 19, 2019).

[16] Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).